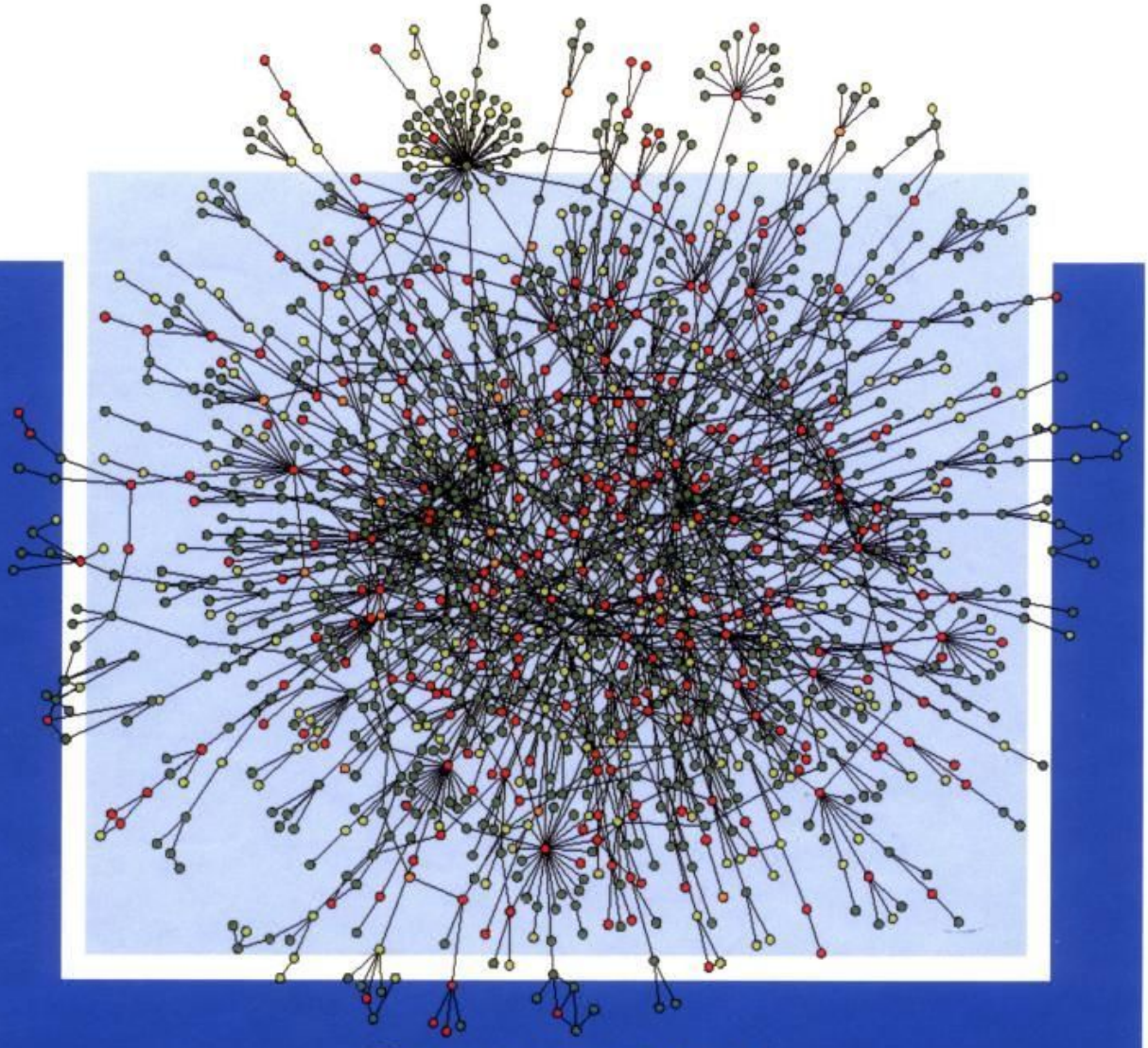
Stefan Bornholdt, Heinz Georg Schuster (Eds.)

Handbook of Graphs and Networks

From the Genome to the Internet



WILEY-VCH

This book defines the field of complex interacting networks in its infancy and presents the dynamics of networks and their structure as a key concept across disciplines. The contributions present common underlying principles of network dynamics and their theoretical description and are of interest to specialists as well as to the non-specialized reader looking for an introduction to this new exciting field.

Theoretical concepts include modeling networks as dynamical systems with numerical methods and new graph theoretical approaches, but also focus on networks that change their topology as in morphogenesis and self-organization. The authors offer concepts to model network structures and dynamics, focussing on approaches applicable across disciplines.

Contributors:
Lada A. Adamic, Uri Alon, Daniel ben-Avraham, Albert-Lásló Barabási, Béla Bollobás, Reuven Cohen, Sergei N. Dorogovtsev, Barbara Drossel, Shlomo Havlin, Bernardo A. Huberman, Sanjay Jain, Wolfgang Kinzel, Alan Kirman, Sandeep Krishna, Rajan M. Lukose, Sergei Maslov, Alan J. McKane, Jose F. F. Mendes, Kai Nagel, Mark E. J. Newman, Romualdo Pastor-Satorras, Oliver M. Riordan, Kim Sneppen, Ricard V. Solé, Sorin Solomon, Ralf J. Sommer, Alessandro Vespignani, Gérard Weisbuch

Stefan Bornholdt is Professor of Theoretical Physics and heads the Statistical Physics Group of the Interdisciplinary Center for Bioinformatics at the University of Leipzig, Germany. After studies at the University of Hamburg and UC Santa Barbara he received his doctorate in 1992. He held research positions at the Universities of Heidelberg and Kiel and in a biotech startup, and was visiting scientist at the Santa Fe Institute and the ITP Santa Barbara.

Heinz Georg Schuster is Professor of Theoretical Physics at the University of Kiel in Germany. In 1971 he attained his doctorate and in 1976 he was appointed Professor at the University of Frankfurt am Main in Germany. He was a visiting professor at the Weizmann-Institute of Science in Israel and at the California Institute of Technology in Pasadena, USA. He is author of several books, among others „Deterministic Chaos", which has been translated into five languages.

www.wiley-vch.de

WILEY-VCH

ISBN 3-527-40336-1

9 783527 403363

Stefan Bornholdt, Heinz Georg Schuster (Eds.)

Handbook of Graphs and Networks

From the Genome to the Internet

WILEY-VCH

Editors:

Stefan Bornholdt

University of Leipzig, Germany
e-mail: bornholdt@izbi.uni-leipzig.de

Hans Georg Schuster

University of Kiel, Germany
e-mail: schuster@theo-physik.uni-kiel.de

1st edition

Cover Picture

Graph representation of the network of known protein-protein interactions in yeast (with permission by A.-L. Barabási). After: H. Jeong, S. Mason, A.-L. Barabási, and Z. N. Oltvai, Centrality and lethality of protein networks, Nature 411 (2001) 41-42.

Library of Congress Card No.: applied for
British Library Cataloging-in-Publication Data:

A catalogue record for this book is available from the British Library

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

Printed in the Federal Republic of Germany

Printed on acid-free paper

Composition Uwe Krieg, Berlin

Printing Druckhaus Darmstadt GmbH, Darmstadt

Bookbinding Litges & Dopf Buchbinderei GmbH, Heppenheim

ISBN 3-527-40336-1

Preface

Understanding the complex world around us is a difficult task and simple principles that capture essential features of complex natural systems are always welcome. One such principle shared by a number of natural systems is their organization as networks of many interacting units: Interacting molecules in living cells, nerve cells in the brain, computers in a telecommunication network, and the social network of interacting people are just a few examples among many others. The concept of networks is an easy-to-use metaphor, but does it really help us in describing the real world?

Recent advances in the theory of complex networks indicate that this notion may be more than just a philosophical term. Triggered by recently available data on large real world networks (e.g. on the structure of the internet or on molecular networks in the living cell) combined with fast computer power on the scientist's desktop, an avalanche of quantitative research on network structure and dynamics currently stimulates diverse scientific fields. Results obtained so far span areas from the prevention of computer viruses to sexually transmitted diseases, and relate to the dynamics and stability of systems as diverse as the internet, regulatory circuits of the genome, and ecosystems.

This handbook aims at providing a snapshot of this active field in its early phase and gives a first-time compilation of the theory of complex network structure and dynamics and its application to different disciplines. All contributors are active researchers in the field and provide an up-to-date review of their respective research focus in self-contained chapters. The first five chapters focus on structure of networks, covering different theoretical aspects from graph theory to methods from theoretical physics, as well as models and applications. The second focus of this handbook is on biological themes, covered in the subsequent five chapters, spanning across scales from molecular networks to ecological systems. The remaining chapters cover even larger scales and concentrate on interdisciplinary applications as traffic, economics, social networks, internet and human language, and conclude with a chapter on network evolution models and a network perspective on the origin of life problem.

Our intention is not a fully comprehensive coverage of the field of complex networks, which for this young and dynamic field would be an impossible task. Perhaps it is even too early to predict where this field will finally settle. However, major themes of current research are covered in this handbook, as well as some topics that might be candidates to attract further attention in the near future. We hope that this handbook serves non-specialists and specialists alike as an inspiring introduction to this exciting new field. If readers take some inspiration from this handbook or even use it as a starting point for further research it would fully serve its purpose.

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

The idea for this handbook dates back to the "International Conference on Dynamical Networks in Complex Systems" held in Kiel in the summer of 2001 gathering a crowd of scientists with diverse backgrounds and the common interest of understanding dynamics and structure of complex networks from their respective disciplines. We hope that this handbook communicates some of the pioneering spirit of this meeting. We are grateful to our friends and colleagues who joined the endeavor of this handbook and provided review chapters on their fields of expertise. Finally we thank the Wiley-VCH staff for excellent professional support in producing this book.

Stefan Bornholdt and *Heinz Georg Schuster*

Leipzig and Kiel, Summer 2002

Contents

List of contributors

- *Lada A. Adamic*

  Hewlett Packard Labs
  Palo Alto
  CA 94304
  USA
  e-mail: ladamic@hpl.hp.com

- *Uri Alon*

  Department of Molecular Cell Biology and Department of Physics of Complex Systems
  Weizmann Institute of Science
  Rehovot 76100
  Israel
  e-mail: urialon@wisemail.weizmann.ac.il

- *Daniel ben-Avraham*

  Physics Department and Center for Statistical Physics (CISP)
  Clarkson University
  Potsdam
  NY 13699-5820
  USA

- *Albert-László Barabási*

  Department of Physics
  University of Notre Dame
  Notre Dame
  IN 46556
  USA
  e-mail: alb@nd.edu

- *Béla Bollobás*

  (1) Department of Mathematical Sciences
  University of Memphis
  Memphis TN 38152
  USA
  e-mail: bollobas@msci.memphis.edu

  (2) Trinity College
  Cambridge CB2 1TQ
  UK

- *Reuven Cohen*

  Minerva Center and Department of Physics
  Bar-Ilan University, Ramat-Gan
  Israel

- *Sergei N. Dorogovtsev*

  (1) Departamento de Física and Centro de Física do Porto
  Faculdade de Ciências
  Universidade do Porto
  Rua do Campo Alegre 687
  4169-007 Porto
  Portugal
  e-mail: sdorogov@fc.up.pt

  (2) A. F. Ioffe Physico-Technical Institute
  194021 St. Petersburg
  Russia

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

- *Barbara Drossel*

  Institut für Festkörperphysik
  TU Darmstadt
  Hochschulstr. 6
  64289 Darmstadt
  Germany
  e-mail: drossel@fkp.tu-darmstadt.de

- *Shlomo Havlin*

  Minerva Center and Department of Physics
  Bar-Ilan University
  Ramat-Gan
  Israel
  e-mail: havlin@ophir.ph.biu.ac.il

- *Bernardo A. Huberman*

  HP Labs
  Palo Alto
  CA 94304
  USA
  e-mail: huberman@hpl.hp.com

- *Sanjay Jain*

  (1) Department of Physics
  and Astrophysics
  University of Delhi
  Delhi 110007
  India
  e-mail: jain@physics.du.ac.in

  (2) Centre for Theoretical Studies
  Indian Institute of Science
  Bangalore 560 012
  India

  (3) Santa Fe Institute
  1399 Hyde Park Road
  Santa Fe, NM 87501
  USA

  (4) Jawaharlal Nehru Centre for
  Advanced Scientific Research
  Bangalore 560 064
  India

- *Wolfgang Kinzel*

  Institut für Theoretische Physik
  Universität Würzburg
  Am Hubland
  97074 Würzburg
  Germany
  e-mail: kinzel@physik.uni-wuerzburg.de

- *Alan Kirman*

  GREQAM
  2 Rue de la Charite
  13002 Marseille
  France
  e-mail: kirman@ehess.cnrs-mrs.fr

- *Sandeep Krishna*

  Centre for Theoretical Studies
  Indian Institute of Science
  Bangalore 560 012
  India
  e-mail: sandeep@physics.iisc.ernet.in

- *Rajan M. Lukose*

  Hewlett Packard Labs
  Palo Alto
  CA 94304
  USA
  e-mail: lukose@hpl.hp.com

- *Sergei Maslov*

  Department of Physics
  Brookhaven National Laboratory
  Upton, New York 11973
  USA
  e-mail: maslov@bnl.gov

- *Alan J. McKane*

  (1) Department of Theoretical Physics
  University of Manchester
  Manchester M13 9PL
  UK

(2) Departments of Physics and Biology
University of Virginia
Charlottesville, VA 22904
USA

- *Jose F. F. Mendes*

  Departamento de Física and Centro de Física do Porto
  Faculdade de Ciências
  Universidade do Porto
  Rua do Campo Alegre 687
  4169-007 Porto
  Portugal
  e-mail: jfmendes@fc.up.pt

- *Kai Nagel*

  Deptartment of Computer Science
  ETH Zürich
  CH-8092 Zürich
  Switzerland
  e-mail: nagel@inf.ethz.ch

- *Mark E. J. Newman*

  Santa Fe Institute
  1399 Hyde Park Road
  Santa Fe, NM 87501
  USA
  e-mail: mark@santafe.edu

- *Romualdo Pastor-Satorras*

  Departament de Física i Enginyeria Nuclear
  Universitat Politècnica de Catalunya
  Campus Nord B4
  08034 Barcelona
  Spain

- *Oliver M. Riordan*

  Trinity College
  Cambridge CB2 1TQ
  UK
  e-mail: O.M.Riordan@dpmms.cam.ac.uk

- *Kim Sneppen*

  Department of Physics
  Norwegian University of Science and Technology
  N-7491 Trondheim
  Norway
  e-mail: sneppen@nbi.dk

- *Ricard V. Solé*

  ICREA-Complex Systems Lab
  Universitat Pompeu Fabra-IMIM
  Dr Aiguader 80
  08003 Barcelona
  Spain
  e-mail: ricard.sole@cexs.upf.es

- *Sorin Solomon*

  Theoretical Physics Department
  Racah Institute of Physics
  Hebrew University of Jerusalem

- *Ralf J. Sommer*

  Max-Planck Institute for Developmental Biology
  Department for Evolutionary Biology
  Spemannstrasse 37/IV
  72076 Tübingen
  Germany
  e-mail: ralf.sommer@tuebingen.mpg.de

- *Alessandro Vespignani*

  The Abdus Salam International Centre for Theoretical Physics (ICTP)
  P. O. Box 586
  I-34100 Trieste
  Italy
  e-mail: alexv@ictp.trieste.it

- *Gérard Weisbuch*

  Laboratoire de Physique Statistique de l'Ecole Normale Supérieure
  24 rue Lhomond
  F-75231 Paris Cedex 5
  France
  e-mail: weisbuch@lps.ens.fr

1 Mathematical results on scale-free random graphs

Béla Bollobás and Oliver M. Riordan

1.1 Introduction

Recently there has been much interest in studying large-scale real-world networks and attempting to model their properties using random graphs. Although the study of real-world networks as graphs goes back some time, recent activity perhaps started with the paper of Watts and Strogatz [55] about the 'small-world phenomenon'. Since then the main focus of attention has shifted to the 'scale-free' nature of the networks concerned, evidenced by, for example, power-law degree distributions. It was quickly observed that the classical models of random graphs introduced by Erdős and Rényi [28] and Gilbert [33] are not appropriate for studying these networks, so many new models have been introduced. The work in this field falls very roughly into the following categories.

1. Direct studies of the real-world networks themselves, measuring various properties such as degree-distribution, diameter, clustering, etc.
2. Suggestions for new random graph models motivated by this study.
3. Computer simulations of the new models, measuring their properties.
4. Heuristic analysis of the new models to predict their properties.
5. Rigorous mathematical study of the new models, to prove theorems about their properties.

Although many hundreds of interesting papers have been written in this area (see, for example, the surveys [2, 27]), so far almost all of this work comes under 1-4; to date there has been very little rigorous mathematical work in the field. Our main aim in this article is to present some of this mathematical work, including several new results. Even an overview of the work in 1-4 lies outside our scope, so we shall present only those models which have been made mathematically precise and for which results have been proved, and mention only a few heuristic results for comparison with the theorems we present. For similar reasons, we cannot even survey the 'classical' theory of random graphs, referring the reader instead to [11] and [38]. However, we shall briefly describe the classical models, as well as some results relevant for comparison; much of the work on the new models has appeared in computer science and physics journals, and it may be that some of the authors are not aware of the related classical results.

The rest of this article is organized as follows. In the next section we briefly describe the classical models of random graphs. In section 1.3 we state some theorems about these

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

models chosen for comparison with recent results about the new models. Section 1.4 is a brief digression concerning the Watts-Strogatz 'small-world' model. The rest of the article concerns 'scale-free' models; a brief introduction is given in section 1.5. These models fall into two types. The first takes a power-law degree distribution as given, and then generates a graph with this distribution. Such models will not be considered here. The second type arises from attempts to *explain* the power law starting from basic assumptions about the growth of the graph. In section 1.6 we describe the Barabási-Albert (BA) model, noting that their definition does not make mathematical sense. A precisely defined model, the 'LCD model', along the lines of the BA model is described in section 1.7, followed by a generalization due to Buckley and Osthus [20] in the next section. In these and the following few sections we concentrate on the degree distribution, presenting results showing that the models are indeed scale-free. Sections 1.9 and 1.10 present such results for the 'copying' models of Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins and Upfal [40], and the very general models defined by Cooper and Frieze [24]. Section 1.11 describes a model for directed graphs with 'preferential attachment' using both in- and out-degrees, and gives the power laws for in- and out-degree distribution.

At this point we return to the LCD model, presenting results about properties other than degree sequence: the clustering coefficient is discussed in section 1.12, the diameter in section 1.13 and 'robustness' in section 1.14.

The last section concerns a special case of the BA model that had been studied considerably earlier; that of scale-free trees. In section 1.15, we present results for small subgraphs (useful for the LCD model) and distance distribution.

Finally, in section 1.16 we conclude with a few remarks.

1.2 Classical models of random graphs

The theory of random graphs was founded by Erdős and Rényi in a series of papers published in the late 1950s and early 1960s. Erdős and Rényi set out to investigate what a 'typical' graph with n labelled vertices and M edges looks like. They were not the first to study statistical properties of graphs; what set their work apart was the probabilistic point of view: they considered a probability space of graphs and viewed graph invariants as random variables. In this setting powerful tools of probability theory could be applied to what had previously been viewed as enumeration questions.

Throughout this section, and indeed the rest of this article, we consider models of *labelled* graphs. Although in the end one may choose to ignore the labels, the models are naturally defined as generating graphs on a certain set of distinguishable vertices, rather than isomorphism classes of graphs. For definiteness it is often convenient to assume that, when the graph has n vertices, the vertex set is $[n] = \{1, 2, \ldots, n\}$.

In modern notation Erdős and Rényi considered the space $\mathcal{G}_{n,M}$ of all $\binom{N}{M}$ graphs with vertex set $[n]$ having M edges, where $N = \binom{n}{2}$ is the number of all possible edges between vertices in $[n]$. The set $\mathcal{G}_{n,M}$ is made into a probability space by taking the elements of $\mathcal{G}_{n,M}$ equiprobable; $G_{n,M}$ will denote a random element of this space. We are interested in what happens as $n \to \infty$, with $M = M(n)$ a function of n. We say that $G_{n,M}$ has a certain

property $\mathcal{P}$ *with high probability* (**whp**) if

$$\Pr(G_{n,M} \text{ has } \mathcal{P}) \to 1$$

as $n \to \infty$. (Here and in what follows it is always understood that M is a function of n. The case when M is constant as $n \to \infty$ is rather uninteresting.) Following Erdős and Rényi, it is customary to say that a *typical* random graph $G_{n,M}$ has property $\mathcal{P}$ if $G_{n,M}$ has $\mathcal{P}$ **whp**.

One of the main discoveries of Erdős and Rényi was that as $M = M(n)$ increases, the structure of a typical $G_{n,M}$ tends to change suddenly. The following is a simple but fundamental result from [28] about connectedness.

Theorem 1. *Let $M_\omega = \frac{n}{2}(\log n + \omega)$, where $\omega = \omega(n)$ is a function of n. If $\omega \to -\infty$ then a typical G_{n,M_ω} is disconnected, while if $\omega \to \infty$, a typical G_{n,M_ω} is connected.*

In the 1950s, Austin, Fagen, Penney and Riordan [4], Gilbert [32, 33], and Riddell and Uhlenbeck [50] also studied statistical properties of graphs, but their approach was very different, using generating functions to obtain exact enumeration formulae and then approximating these. The results obtained this way were much weaker than those of Erdős and Rényi.

The model of random graphs introduced by Gilbert [33] (precisely at the time that Erdős and Rényi started their investigations of $\mathcal{G}_{n,M}$) is, perhaps, even more fundamental than $\mathcal{G}_{n,M}$, and is more convenient to use. To define Gilbert's model, $\mathcal{G}_{n,p}$, let $\{X_{ij} : 1 \le i < j \le n\}$ be an array of iid Bernoulli random variables, with $\Pr(X_{ij} = 1) = p$ and $\Pr(X_{ij} = 0) = 1 - p$, and let $G_{n,p}$ be the random graph on $[n]$ in which two vertices i and j are adjacent if $X_{ij} = 1$. Less formally, to construct a random $G_{n,p} \in \mathcal{G}_{n,p}$, put in edges with probability p, independently of each other. Again p is often a function of n, though the case p constant, $0 < p < 1$, makes perfect sense. For $M \sim pN$ the models $\mathcal{G}_{n,M}$ and $\mathcal{G}_{n,p}$ are almost interchangeable. (Note that, as usual, we commit a harmless abuse of notation, using $\mathcal{G}_{n,\cdot}$ for two different models. There is no danger of confusion, as $M \to \infty$ while $0 < p < 1$.)

Since the early 1960s several other 'classical' models of random graphs have been introduced. A *graph process* $\tilde{G}_n = (G_{n,t})_{t=0}^N$ on $[n]$ is a nested sequence of graphs, $G_{n,0} \subset G_{n,1} \subset \cdots \subset G_{n,N}$ such that $G_{n,t}$ has precisely t edges. The space $\tilde{\mathcal{G}}_n$ of *random graph processes* consists of all $N!$ graph processes on $[n]$, endowed with the uniform (normalized counting) measure. Note that this notation is consistent with that used earlier: the distribution of $G_{n,t}$, a random graph process stopped at time t, is precisely the distribution of $G_{n,t}$ as an element of $\mathcal{G}_{n,t}$. A random graph process has a natural interpretation as a dynamic Markov process; given $G_{n,0}, \ldots, G_{n,t}$, at the next step $G_{n,t+1}$ is obtained by adding one of the $N - t$ remaining possible edges to $G_{n,t}$ uniformly at random. In studying $\tilde{\mathcal{G}}_n$ one is mostly interested in the *hitting times* of certain properties (those preserved by adding edges), that is, the random variable given by the minimal t for which $G_{n,t}$ has the property. For example, Theorem 1 claims that **whp** the hitting time of connectedness is at least $\frac{n}{2}(\log n - \omega(n))$ and at most $\frac{n}{2}(\log n - \omega(n))$ whenever $\omega(n) \to \infty$. In fact, **whp**, the hitting time of connectedness is precisely the hitting time of having no isolated (degree 0) vertices.

To get a random element $G_{n,k\text{-out}}$ of the space $\mathcal{G}_{n,k\text{-out}}$, join each vertex i to k other vertices chosen at random and take the union of all these edges. Equivalently, let $\vec{G}_{n,k\text{-out}}$ be the random directed graph obtained by sending arcs from each vertex to a set of k other vertices chosen uniformly at random; the random graph $G_{n,k\text{-out}}$ is the underlying simple

graph of $\vec{G}_{n,k}$-out. Note that each $G_{n,k}$-out has at least $kn/2$ and at most kn edges; although kn is much smaller than $\frac{n}{2}\log n$, the threshold of connectedness given by Theorem 1, for all $k \geq 2$, **whp** $G_{n,k}$-out is connected.

The space $\mathcal{G}_{n,r}$-reg is simply the set of all r-regular graphs on $[n]$ with the uniform measure. Although this space is very easy to define, for larger values of r it is not so easy to study.

The study of random graphs really took off in the mid 1970s; since then several thousand papers have been written on the topic. Many of the results are presented in the monographs [11] and [38].

1.3 Results for classical random graphs

In this brief review it would be impossible to survey even the more important results about classical random graphs; all we shall do is present some results that are analogous to a number of results about scale-free random graphs we shall present later.

In addition to discovering the prevalence of 'phase transitions' for numerous properties of random graphs, Erdős and Rényi [29] proved that the component structure of a random graph process undergoes a sudden change around time $t \sim n/2$. This result about the emergence of the 'giant component' is the single most important theorem of Erdős and Rényi about random graphs. Here we state it only in a simplified form.

Theorem 2. *Let $c > 0$ be a constant, and set $p = c/n$. If $c < 1$ then* **whp** *every component of $G_{n,p}$ has order $O(\log n)$. If $c > 1$ then* **whp** *$G_{n,p}$ has a component with $(\alpha(c) + o(1))n$ vertices, where $\alpha(c) > 0$, and all other components have $O(\log n)$ vertices.*

Considerably more precise results have been proved by Bollobás [10], Łuczak [42], and Janson, Knuth, Łuczak and Pittel [37]. The component of order $\Theta(n)$ whose existence is guaranteed by Theorem 2 is usually called the *giant component*. If c is considerably larger than 1, then the giant component has a large robust (highly connected) subgraph.

For p constant, the degree sequence of $G_{n,p}$ is close to a sequence of n iid Binomial random variables with probability p and mean np. (A very strong precise result along these lines is given in [46].) For $p = c/n$, where c is constant, the degree sequence is well approximated by a sequence of n iid Poisson random variables with mean c. In particular, one has the following very weak result.

Theorem 3. *Let X_k be the number of vertices of degree k in $G_{n,p}$ where $p = c/n$, with $c > 0$ constant. Then for $k = 0, 1, \ldots$*

$$\Pr\left((1-\epsilon)\frac{c^k e^{-c}}{k!} \leq \frac{X_k}{n} \leq (1+\epsilon)\frac{c^k e^{-c}}{k!}\right) \to 1$$

as $n \to \infty$.

In a graph G, the *distance* $d(u, v)$ between two vertices u and v is the length (number of edges) of the shortest path between them. The *diameter* $\mathrm{diam}(G)$ of a connected graph G is the maximum distance between two vertices; a disconnected graph is taken to have infinite diameter. The diameter of a random graph has been studied by a great many people, including

Burtin [21, 22], Bollobás [9] and Bollobás and de la Vega [14]. If $pn/\log n \to \infty$ and $\log n/\log(pn) \to \infty$ then **whp** the diameter of $G_{n,p}$ is asymptotic to $\log n/\log(pn)$. In the range we are interested in here, corresponding to the $\Theta(n)$ edges in scale-free random graphs, $G_{n,p}$ is disconnected, so the the diameter of $G_{n,k}$-out or $G_{n,r}$-reg is more relevant. Let us state a weak form of a result from [14].

Theorem 4. *Let $r \geq 3$ and $\epsilon > 0$ be fixed. Then*

$$\Pr\left((1-\epsilon)\frac{\log n}{\log(r-1)} \leq \mathrm{diam}(G_{n,r}\text{-}\textit{reg}) \leq (1+\epsilon)\frac{\log n}{\log(r-1)}\right) \to 1$$

as $n \to \infty$.

As we shall see, results vaguely resembling Theorem 4 hold for scale-free random graphs. More or less by definition, the results corresponding to Theorem 3 are rather different.

1.4 The Watts-Strogatz 'small-world' model

In 1998, Watts and Strogatz [55] raised the possibility of constructing random graphs that have some of the important properties of 'real-world' networks. The real-world networks they considered included neural networks, the power grid of the western United States and the collaboration graph of film actors. Watts and Strogatz noticed that these networks were 'small-world' networks: their diameters were considerably smaller than those of regularly constructed graphs (such as lattices, or grid graphs) with the same number of vertices and edges. More precisely, Watts and Strogatz found that real-world networks tend to be highly clustered, like lattices, but have small diameters, like random graphs. That large social networks have rather small diameters had been noticed considerably earlier, in the 1960s, by Milgram [47] and others, and was greatly popularized by Guare's popular play 'six degrees of separation' in 1990.

The importance of the Watts and Strogatz paper is due to the fact that it started the active and important field of modelling large-scale networks by random graphs defined by simple rules. As it happens, from a mathematical point of view, the experimental results in [55] were far from surprising.

Instead of the usual diameter $\mathrm{diam}(G)$ of a graph G, Watts and Strogatz considered the average distance

$$L(G) = \sum_{\{u,v\}\subset V, u\neq v} d(u,v)/\binom{n}{2},$$

where V is the vertex set of G and n is the number of vertices. Clearly $L(G) \leq \mathrm{diam}(G)$, but in 'most' cases $L(G)$ is not much smaller than $\mathrm{diam}(G)$. (For example, for $G_{n,r}$-reg, $r \geq 3$, **whp** these quantities are asymptotically equal.)

To measure the 'cliquishness' of a graph, for a graph G and vertex v, let $C_v(G)$ be the proportion of pairs of neighbours of v that are themselves neighbours, and let $C_1(G)$ be the average of $C_v(G)$ as v runs over the vertices. In section 1.12 we shall give a more formal definition of this *clustering coefficient* $C_1(G)$, together with a variant of it.

For a random r-regular graph, $C_1(G_{n,r}\text{-reg}) \sim \frac{r-1}{n}$, while

$$\operatorname{diam}(G_{n,r}\text{-reg}) \sim \log n / \log(r-1) :$$

the clustering coefficient is small, and so is the diameter. On the other hand, as pointed out by Watts and Strogatz, many real-world networks tend to have a largish clustering coefficient *and* small diameter. To construct graphs with these properties, Watts and Strogatz suggested starting with a fixed graph with large clustering coefficient and 'rewiring' some of the edges.

To be precise, let G be the graph C_n^r, the r^{th} power of an n-cycle, where $n > 2r$. Thus G is a $2r$-regular graph of order n; two vertices are joined in G if their distance in the n-cycle C_n is at most r. For $n = 2rs$, $s \geq 2$, say, we have $\operatorname{diam}(G) = s$, and $L(G) \sim s/2$ as $s \to \infty$, while $C_1(G) = \frac{3(r-1)}{2(2r-1)}$. Let $G(p)$ be the random graph obtained from G by deleting each edge at random with probability p, independently of the other edges, and then adding the same number of edges back at random. Almost equivalently, $G(p)$ is obtained from G by 'rewiring' a proportion p of the edges. What Watts and Strogatz found was that, even for a small value of p, $L(G(p))$ drops down to $O(\log n)$, while $C_1(G(p))$ stays close to $3/4$; the introduction of a small number of random edges reduces the diameter to $O(\log n)$.

Following this observation, much research was devoted to the 'surprising' phenomenon that the introduction of a little randomness makes the diameter small (while various other graph invariants remain essentially unchanged). In fact, it is far from surprising that a few random edges superimposed on a connected ground graph give a graph of small diameter. For example, Bollobás and Chung [13] proved that a random matching added to a cycle gives a graph whose diameter is about that of a random cubic graph. Similarly, for $c > 0$, adding cn random edges to a tree of order n results in a graph of diameter $O(\log n)$. These results (though not the precise constant given in [13]) are particular instances of a general phenomenon which has been known much longer; they follow from the fact that the diameter of $G_{n,r}$-reg (or of the giant component of $G_{n,p}$, $p = c/n$) is $O(\log n)$.

The graphs obtained by rewiring some of the edges of a power of a cycle do not resemble large-scale real-world networks, although they share some of their characteristics. To model these networks, it is desirable to define new families of random graphs rather different from from the classical models. This is the topic of the next several sections.

1.5 Scale-free models

In 1999, Faloutsos, Faloutsos and Faloutsos [30] suggested certain 'scale-free' power laws for the graph of the Internet, and showed that these power laws fit the real data very well. In particular, they suggested that the degree distribution follows a power law, in contrast to the Poisson distribution for classical random graphs given in Theorem 3. This was soon followed by work on rather vaguely described random graph models aiming to explain these power laws, and others seen in features of many real-world networks.

In fact, power-law distributions had been observed considerably earlier; in particular, in 1926 Lotka [41] claimed that citations in academic literature follow a power law, and in 1997 Gilbert [34] suggested a probabilistic model supporting 'Lotka's law'. Other early investigations into power-law distributions are due to Simon [51] and Zipf [56].

The degree distribution of the graph of telephone calls seems to follow a power law as well; motivated by this, Aiello, Chung and Lu [1] proposed a model for 'massive graphs'. This model ensures that the degree distribution follows a power law by *fixing* a degree sequence in advance to fit the required power law, and then taking the space of random graphs with this degree sequence. Thus their approach is very different from the models we are interested in, where the aim is to understand how power laws might arise, by finding simple rules that generate random graphs satisfying such laws.

In the next sections we present several of these models, concentrating for the moment on the degree sequence. Later in the article we return to one particular model, the LCD model, presenting results about several other properties.

1.6 The Barabási-Albert model

Perhaps the most basic and important of the 'scale-free' random graph models, i.e., models producing power-law or 'scale-free' behaviour from simple rules, is the 'BA model'. This was introduced by Barabási and Albert [5] in 1999:

> ... starting with a small number (m_0) of vertices, at every time step we add a new vertex with $m(\leq m_0)$ edges that link the new vertex to m different vertices already present in the system. To incorporate preferential attachment, we assume that the probability Π that a new vertex will be connected to a vertex i depends on the connectivity k_i of that vertex, so that $\Pi(k_i) = k_i / \sum_j k_j$. After t steps the model leads to a random network with $t + m_0$ vertices and mt edges.

The basic motivation is to provide a highly simplified model of the growth of, for example, the world-wide web. New sites (or pages) are added one at a time, and link to earlier sites chosen with probabilities depending on their current 'popularity'; this is the principle that 'popularity is attractive'; this principle presumably plays a role in the growth of real networks in a wide range of contexts. It is customary to call this the 'preferential attachment' rule. Barabási and Albert themselves, and many other people, gave experimental and heuristic results about the BA model; we will return to a few of these later. From a mathematical point of view, however, the description above, repeated in many papers, does not make sense.

The first problem is getting started: how do we take probabilities proportional to the degrees when these are all zero? Perhaps it makes sense to ignore the explicit start from no edges given by Barabási and Albert, and start instead from a small graph G_0 with no isolated vertices, hoping that the choice of G_0 makes little difference. While for many properties G_0 turns out not to matter, for others it matters very much. For example, in the case $m = 1$ the BA model describes the growth of a tree, *provided G_0 is a tree*. If G_0 is disconnected, say, then at all later stages the graph produced will also be disconnected. For general m the initial graph G_0 also has significant lasting effects, for example on the expected maximum degree, which can change by a constant factor when G_0 is changed.

The second problem is with the preferential attachment rule itself, and arises only for $m \geq 2$; when we add a new vertex, say the $t + 1^{\text{st}}$, we must join it to a random set N_{t+1} of m earlier vertices. In our notation, working always with graphs on $\{1, 2, \ldots\}$, the BA model

says only that, for $1 \le i \le t$,

$$\Pr(i \in N_{t+1}) = m d_t(i) / \sum_{j=1}^{t} d_t(j), \tag{1.1}$$

where $d_t(i)$ is the degree of vertex i in the growing graph at time t. (Actually, as can be seen from the quotation above, Barabási and Albert give this formula without the factor of m. If we assume their formula is intended to hold separately for each edge added, then (1.1) follows. However, their description does not allow us to add edges one by one independently, as it is specified that the edges go to different vertices.) To fully describe the model, we must specify the distribution of N_{t+1}, i.e., the probability that $N_{t+1} = S$ for each of the $\binom{t}{m}$ possible sets S of earlier vertices. This distribution is not uniquely specified by giving the marginal probabilities that $i \in N_{t+1}$ for each earlier vertex i. To see this note, for example, that the distribution of N_{t+1} has $\binom{t}{m} - 1$ degrees of freedom (the $\binom{t}{m}$ probabilities must add up to 1) while there are only t marginal probabilities specified by the BA description. Again one might hope that the exact choice does not make much difference, and again this turns out to be false. As shown by the following result, there is a range of models fitting the BA description with very different properties.

Theorem 5. *Let $f(n)$, $n \ge 2$, be any integer valued function with $f(2) = 0$ and $f(n) \le f(n+1) \le f(n) + 1$ for every $n \ge 2$, such that $f(n) \to \infty$ as $n \to \infty$. Then there is a random graph process $T^{(n)}$ satisfying* (1.1) *with $m = 2$ such that, with probability 1, $T^{(n)}$ has exactly $f(n)$ triangles for all sufficiently large n.*

In less formal language, Theorem 5 says, for example, that if you want $\log n$ triangles when the graph has n vertices, there is a precise model satisfying the BA description (except for the start, which cannot be satisfied) which achieves this. Similarly, if you want n^α triangles for any $0 < \alpha \le 1$, or any other plausible function. Thus the clustering coefficient (see section 1.12) may also be tuned. The only tiny caveat is that you may be forced to create a finite number of triangles at the start. Note that this is different from the result in [36], which considers a model outside the Barabási-Albert definition (triangles are created by adding edges between existing vertices).

Proof. We give only an outline of the proof. We will work entirely with simple graphs, with no loops or multiple edges, starting with $T^{(2)}$ a single edge. When adding a new vertex v to a simple graph and joining it to two distinct existing vertices, x and y, the number of triangles either remains the same, or goes up by one. It goes up by one if and only if xy is an edge. Restating the theorem, we must show that given $T^{(n)}$ we have two ways choosing x and y to define $T^{(n+1)}$, each satisfying the Barabási-Albert preferential attachment rule (1.1): one where xy is always an edge of $T^{(n)}$, and one where, except perhaps for finitely many steps near the start, it never is.

The first case is easy: to guarantee a new triangle, take xy to be a random edge of $T^{(n)}$. By definition of degree, the probability that a particular vertex w is chosen as one of x and y is just the degree $d(w)$ of w in $T^{(n)}$ over the total number $(2n-3)$ of edges of $T^{(n)}$, so (1.1) is satisfied.

For the second case we must assign non-negative weights $p_{xy} = p_{yx}$ to pairs $\{x, y\} \subset V(T^{(n)})$ with p_{xy} zero for every edge, such that $\sum_{y \ne x} p_{xy} = d(x)/(2n-3)$.

Then $\sum_{\{x,y\}} p_{xy} = 1$, so we may take p_{xy} as the probability of joining the new vertex to x and y. Such an assignment is possible under very mild conditions; for example, the maximum degree of $T^{(n)}$ being at most $n/3$ is more than sufficient. It is easy to check that in any process satisfying (1.1), the maximum degree is at most $O(n^{1/2})$ **whp**, so the result follows. □

An extreme case of the process above, in which a triangle is added at every step, was actually considered by Dorogovtsev and Mendes [27] (section IX C), without noting that it satisfies the Barabási-Albert description. In fact, it is introduced there as a simpler alternative model for easier analysis.

As seen from the example above, in order to prove results about the BA model, one must first decide on the details of the model itself. In the next section we present one particular choice for how to do this which has several advantages.

1.7 The LCD model and $G_m^{(n)}$

In this section we define precisely a random graph model introduced in [16] satisfying the vague description given by Barabási and Albert. It turns out to be convenient to allow multiple edges and loops; there will not be very many of these, and in any case there seems no reason to exclude them from the point of view of the interpretation: one web site may contain several links to another, for example, or links to itself.

Consider a fixed sequence of vertices $v_1, v_2, \ldots$. (Later we shall take $v_i = i$; the general case simplifies the notation when we merge vertices.) Let us write $d_G(v)$ for the degree of the vertex v in the graph G. We define inductively a random graph process $(G_1^{(t)})_{t\geq 0}$ so that $G_1^{(t)}$ is a graph on $\{v_i : 1 \leq i \leq t\}$, as follows: start with $G_1^{(0)}$ the empty 'graph' with no vertices, or with $G_1^{(1)}$ the graph with one vertex and one loop. Given $G_1^{(t-1)}$, form $G_1^{(t)}$ by adding the vertex v_t together with a single edge between v_t and v_i, where i is chosen randomly with

$$\Pr(i = s) = \begin{cases} d_{G_1^{(t-1)}}(v_s)/(2t-1) & 1 \leq s \leq t-1, \\ 1/(2t-1) & s = t. \end{cases} \tag{1.2}$$

In other words, send an edge e from v_t to a random vertex v_i, where the probability that a vertex is chosen as v_i is proportional to its degree at the time, counting e as already contributing one to the degree of v_t. (We shall see why this is convenient later.) For $m > 1$, add m edges from v_t one at a time, counting the previous edges as well as the 'outward half' of the edge being added as already contributing to the degrees. We choose this precise rule because it leads to the following equivalent definition: define the process $(G_m^{(t)})_{t\geq 0}$ by running the process $(G_1^{(t)})$ on a sequence $v_1', v_2', \ldots$, and forming the graph $G_m^{(t)}$ from $G_1^{(mt)}$ by identifying the vertices $v_1', v_2', \ldots, v_m'$ to form v_1, identifying $v_{m+1}', v_{m+2}', \ldots, v_{2m}'$ to form v_2, and so on.

For the rest of the article we shall take $v_i = i$, so $G_m^{(t)}$ is a graph on $[t] = \{1, 2, \ldots, t\}$. Note that the edges of $G_m^{(t)}$ have a natural orientation, from later vertices to earlier vertices, so ij is oriented from i to j if $i > j$. However, as for studies of the BA model, we shall generally treat the graph as unoriented. For these models the orientation is not very interesting (indeed it may be reconstructed from the graph even if the vertex labels are not given).

In addition to satisfying the basic mathematical criterion of being precisely specified, the process $G_m^{(t)}$ has several useful properties. One is that $G_m^{(t)}$ can be defined in terms of $G_1^{(mt)}$, a much simpler object, so questions about $G_m^{(t)}$ can be re-written in terms of $G_1^{(mt)}$, and results can be proved this way. Another very important property is the following: while the process $G_1^{(t)}$ is dynamic, the distribution of the graph $G_1^{(n)}$ obtained at a particular time $t = n$ has a simple *static* description, the *linearized chord diagram* or LCD description, given in [16]:

> An *n-pairing* is a partition of the set $\{1, 2, \ldots, 2n\}$ into pairs, so there are $(2n)!/(n!2^n)$ n-pairings. These objects are sometimes thought of as *linearized chord diagrams* (or *LCDs*) [15, 52], where an LCD with n chords consists of $2n$ distinct points on the x-axis paired off by semi-circular chords in the upper half plane. Two LCDs are considered to be the same when one can be turned into the other by moving the points on the x-axis without changing their order. Thinking of pairings as LCDs, we shall talk of chords and their left and right endpoints. We form a graph $\phi(L)$ from an LCD L as follows: starting from the left, identify all endpoints up to and including the first right endpoint reached to form vertex 1. Then identify all further endpoints up to the next right endpoint to form vertex 2, and so on. For the edges, replace each chord by an edge joining the vertex corresponding to its right endpoint to that corresponding to its left endpoint.

As stated in [16], if L is chosen uniformly at random from all $(2n)!/(n!2^n)$ LCDs with n chords (i.e., n-pairings), then $\phi(L)$ has the same distribution as a random $G_1^{(n)}$ defined via the process $G_1^{(t)}$ given earlier.

> To see this note that L can be obtained by taking a random LCD L' with $n-1$ chords and adding a new chord whose right endpoint is to the right of all $n-1$ chords, and whose left endpoint lies in one of the $2n-1$ possible places, each chosen with equal probability. This corresponds to adding a new vertex to $\phi(L')$ and joining it to another vertex with probabilities according to the degrees, exactly as in the description of $(G_1^{(n)})$.

A simple result proved in [19] using the LCD description, which can also be proved in other ways, concerns the degree sequence. We write $\#_m^n(d)$ for the number of vertices of $G_m^{(n)}$ with *in-degree* equal to d, i.e., with (total) degree $m + d$.

Theorem 6. *Let $m \geq 1$ and $\epsilon > 0$ be fixed, and set*

$$\alpha_{m,d} = \frac{2m(m+1)}{(d+m)(d+m+1)(d+m+2)}.$$

Then with probability tending to 1 as $n \to \infty$ we have

$$(1-\epsilon)\alpha_{m,d} \leq \frac{\#_m^n(d)}{n} \leq (1+\epsilon)\alpha_{m,d}$$

for every d in the range $0 \leq d \leq n^{1/15}$.

This result gives a rigorous justification of the power-law dependence of the degrees described in [6].

Let us remark that in the case $m = 1$, essentially this result had been proved much earlier by Szymański [54] in a slightly different context (see section 1.15).

In the next few sections we describe other scale-free models for which power-law degree distribution has been proved.

1.8 The Buckley-Osthus model

Two groups, Dorogovtsev, Mendes and Samukhin [26] and Drinea, Enachescu and Mitzenmacher [25], introduced a variation on the BA model in which vertices have an 'initial attractiveness': the probability that an old vertex is chosen to be a neighbour of the new vertex is proportional to its in-degree plus a constant 'initial attractiveness', which we shall write as am. The case $a = 1$ is just the BA model, since there total degree is used, and each out-degree is m. Buckley and Osthus [20] made this more general model precise along the lines of the LCD model; for a fixed positive integer a, they define a process $H_{a,1}^{(t)}$ exactly as $G_1^{(t)}$ is defined above, but replacing (1.2) with

$$\Pr(i = s) = \begin{cases} \frac{d^{in}_{H_{a,1}^{(t-1)}}(v_s)+a}{(a+1)t-1} & 1 \le s \le t-1, \\ \frac{a}{(a+1)t-1} & s = t. \end{cases}$$

Note that when $a = 1$ the definition of $H_{a,1}^{(t)}$ reduces exactly to that of $G_1^{(t)}$. As for $G_m^{(t)}$, a process $H_{a,m}^{(t)}$ is defined in [20] by identifying vertices in groups of m. Buckley and Osthus established that the degree distribution of this model also obeys a power law. Let us write $\#_{a,m}^n(d)$ for the number of vertices of $H_{a,m}^{(n)}$ with in-degree d.

Theorem 7. *Let $m \ge 1$ and $a \ge 1$ be fixed integers, and set*

$$\alpha_{a,m,d} = (a+1)(am+a)! \binom{d+am-1}{am-1} \frac{d!}{(d+am+a+1)!}.$$

Let $\epsilon > 0$ be fixed. Then **whp** *we have*

$$(1-\epsilon)\alpha_{a,m,d} \le \frac{\#_{a,m}^n(d)}{n} \le (1+\epsilon)\alpha_{a,m,d}$$

for all d in the range $0 \le d \le n^{1/100(a+1)}$. In particular, **whp** *for all d in this range we have*

$$\frac{\#_{a,m}^n(d)}{n} = \Theta(d^{-2-a}).$$

The proof is rather difficult, as the equivalent for $H_{a,1}^{(t)}$ of the LCD model for $G_1^{(t)}$ is much more complicated. Dorogovtsev, Mendes and Samukhin [26] gave a non-rigorous argument for a weaker form of this result, where the range of d considered is bounded.

1.9 The copying model

Around the same time as the BA model, Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins and Upfal [40] gave rather different models to explain the observed power laws in the web graph. The basic idea is that a new web page is often made by copying an old one, and then changing some of the links. Let us define one of these models by quoting almost verbatim from [40]:

> The linear growth copying model is parametrized by a *copy factor* $\alpha \in (0,1)$ and a constant out-degree $d \geq 1$. At each time step, one vertex u is added, and u is then given d out-links for some constant d. To generate the out-links, we begin by choosing a 'prototype' vertex p uniformly at random from V_t (the old vertices). The i^{th} out-link of u is then chosen as follows. With probability α, the destination is chosen uniformly at random from V_t, and with the remaining probability the out-link is taken to be the i^{th} out-link of p. Thus, the prototype is chosen once in advance. The d out-links are chosen by α-biased independent coin flips, either randomly from V_t, or by copying the corresponding out-link of the prototype.
>
> The intuition behind this model is the following. When an author decides to create a new web page, the author is likely to have some topic in mind. The choice of prototype represents the choice of topic—larger topics are more likely to be chosen. The Bernoulli copying events reflect the following intuition: a new viewpoint about the topic will probably link to many pages 'within' the topic (i.e., pages already linked to by existing resource lists about the topic), but will also probably introduce a new spin on the topic, linking to some new pages whose connection to the topic was previously unrecognized.

As for the BA model, it turns out that the degree distribution does follow a power law. Let $N_{t,r}$ be the expected number of vertices of degree r in the graph formed by the model outlined above at time t (when the graph has t vertices). Among other results, the following was proved in [40].

Theorem 8. *For $r > 0$, the limit $P_r = \lim_{t\to\infty} N_{t,r}/t$ exists, and satisfies*

$$P_r = P_0 \prod_{i=1}^{r} \frac{1 + \alpha/(i(1-\alpha))}{1 + 2/(i(1-\alpha))}$$

and

$$P_r = \Theta\left(r^{-\frac{2-\alpha}{1-\alpha}}\right).$$

When one looks only at the degree sequence, this copying model behaves very similarly to models with preferential attachment; we shall return to this in the next section. In other ways, however, the model is essentially different. An obvious example is that copying will give rise to many more dense bipartite subgraphs; part of the original motivation was to explain the appearance of these in the web graph.

1.10 The Cooper-Frieze model

Recently, Cooper and Frieze [24] introduced a model with many parameters which includes the models of the last three sections as special cases, and proved a very general result about the power-law distribution of degrees. In the undirected case, the model describes a (multi-)graph process $G(t)$, starting from $G(0)$ a single vertex with no edges. Their attachment rule is a mixture of preferential (by degree) and uniform (uniformly at random, or 'u.a.r'). Quoting from [24]:

> Initially, at step $t = 0$, there is a single vertex v_0. At any step $t = 1, 2, \ldots, T, \ldots$ there is a birth process in which either new vertices or new edges are added. Specifically, either a procedure NEW is followed with probability $1 - \alpha$, or a procedure OLD is followed with probability α. In procedure NEW, a new vertex v is added to $G(t-1)$ with one or more edges added between v and $G(t-1)$. In procedure OLD, an existing vertex v is selected and extra edges are added at v.
>
> The recipe for adding edges typically permits the choice of initial vertex v (in the case of OLD) and the terminal vertices (in both cases) to be made either u.a.r or according to vertex degree, or a mixture of these two based on further sampling. The number of edges added to vertex v at step t by the procedures (NEW, OLD) is given by distributions specific to the procedure. The details of these choices are given below.

The parameters fixed in advance are integers $j_0, j_1 \geq 1$, and real numbers α, β, γ, δ, $p_1, \ldots, p_{j_0}$, $q_1, \ldots q_{j_1}$ between 0 and 1, with $\alpha < 1$ and $\sum p_i = \sum q_i = 1$. The procedure for defining $G(t)$ from $G(t-1)$ is as follows (from [24]):

> Choice of procedure at step t.
> α: Probability that an OLD node generates edges.
> $1 - \alpha$: Probability that a NEW node is created.
> Procedure NEW
> $\boldsymbol{p} = (p_i : i \geq 1)$: Probability that new node generates i new edges.
> β: Probability that choices of terminal vertices are made uniformly.
> $1 - \beta$: Probability that choices of terminal vertices are made according to degree.
> Procedure OLD
> $\boldsymbol{q} = (q_i : i \geq 1)$: Probability that old node generates i new edges.
> δ: Probability that the initial node is selected uniformly.
> $1 - \delta$: Probability that the initial node is selected according to degree.
> γ: Probability that choices of terminal vertices are made uniformly.
> $1 - \gamma$: Probability that choices of terminal vertices are made according to degree.

In words:

> The model creates edges in the following way: An initial vertex v is selected. If the terminal vertex w is chosen u.a.r, we say v is *assigned* to w. If the terminal vertex w is chosen according to its vertex degree, we say v is *copied* to w. In either

case the edge has an intrinsic direction (v, w), which we may choose to ignore. We note that sampling according to vertex degree is equivalent to selecting an edge u.a.r and then selecting an endpoint u.a.r.

Note that although this 'copying' mechanism is not at all that of the 'copying model' described in the previous section, as pointed out by Cooper and Frieze, as far as the evolution of the degree sequence is concerned, the two are more or less interchangeable. Note also that the mixture of uniform and preferential attachment is easily seen to be equivalent to the preferential attachment with constant initial attractiveness considered in section 1.8.

Given the generality of the Cooper-Frieze model, it is not surprising that their result is rather difficult to state. Quoting again from [24], we must first start with some notation.

Notation

Let $\mu_p = \sum_{j=0}^{j_0} jp_j$, $\mu_q = \sum_{j=0}^{j_1} iq_j$ and let $\theta = 2((1-\alpha)\mu_p + \alpha\mu_q)$. To simplify subsequent notation, we transform the parameters as follows:

$$\begin{aligned}
a &= 1 + \beta\mu_p + \frac{\alpha\gamma\mu_q}{1-\alpha} + \frac{\alpha\delta}{1-\alpha},\\
b &= \frac{(1-\alpha)(1-\beta)\mu_p}{\theta} + \frac{\alpha(1-\gamma)\mu_q}{\theta} + \frac{\alpha(1-\delta)}{\theta},\\
c &= \beta\mu_p + \frac{\alpha\gamma\mu_q}{1-\alpha},\\
d &= \frac{(1-\alpha)(1-\beta)\mu_p}{\theta} + \frac{\alpha(1-\gamma)\mu_q}{\theta},\\
e &= \frac{\alpha\delta}{1-\alpha}, \qquad\qquad f = \frac{\alpha(1-\delta)}{\theta}.
\end{aligned}$$

We note that

$$c + e = a - 1 \text{ and } b = d + f. \tag{1.3}$$

Now define the sequence $(d_0, d_1, \ldots, d_k, \ldots)$ by $d_0 = 0$ and for $k \geq 1$

$$d_k(a+bk) = (1-\alpha)p_k + (c+d(k-1))d_{k-1} + \sum_{j=1}^{k-1}(e+f(k-j))q_j d_{k-j}. \tag{1.4}$$

Since $a \geq 1$, this system of equations has a unique solution.

Statement of results

The main quantity we study is the random variable $D_k(t)$, the number of vertices of degree k at step t. We let $\overline{D}_k(t) = \mathbb{E}(D_k(t))$. We prove that for small k, $\overline{D}_k(t) \approx d_k t$ as $t \to \infty$.

Theorem 9. *There exists a constant $M > 0$ such that for $t, k = 1, 2, \ldots,$*

$$|\overline{D}_k(t) - td_k| \leq Mt^{1/2}\log t.$$

The number of vertices $\nu(t)$ at step t is **whp** asymptotic to $(1-\alpha)t$. It follows that

$$\bar{d}_k = \frac{d_k}{1-\alpha}.$$

The next theorem summarizes what we know about the d_k:

Theorem 10.

(i) *$Ak^{-\zeta} \le d_k \le B\min\{k^{-1}, k^{-\zeta/j_1}\}$ where $\zeta = (1+d+f\mu_q)/(d+f)$.*

(ii) *If $j_1 = 1$ then $d_k \sim Ck^{-(1+1/(d+f))}$.*

(iii) *If $f = 0$ then $d_k \sim Ck^{-(1+1/d)}$.*

(iv) *If the* SOLUTION CONDITIONS *hold then*

$$d_k = C\left(1 + O\left(\frac{1}{k}\right)\right)k^{-x},$$

where C is constant and

$$x = 1 + \frac{1}{d+f\mu_q}. \tag{1.5}$$

We say that $\{q_j : j = 1,\dots,j_1\}$ is *periodic* if there exists $m > 1$ such that $q_j = 0$ unless $j \in \{m, 2m, 3m, \dots\}$.

Let

$$\phi_1(y) = y^{j_1} - \left(\frac{d+q_1 f}{b}y^{j_1-1} + \frac{q_2 f}{b}y^{j_1-2} + \cdots + \frac{q_{j_1} f}{b}\right).$$

Our SOLUTION CONDITIONS are:

S(i) $f > 0$ and either (a) $d + q_1 f > 0$ or (b) $\{q_j\}$ is not periodic.

S(ii) The polynomial $\phi_1(y)$ has no repeated roots.

In summary, these results say that the 'expected degree sequence' converges in a strong sense to the solution of a certain recurrence relation, and that under rather weak conditions, this solution follows a power law with an explicitly determined exponent and a bound on the error term. Cooper and Frieze also prove a simple concentration result, which we will not state, showing the number of vertices of a certain degree is close to its expectation.

In addition to the undirected case, Cooper and Frieze consider what one might call 'semi-directed' models, where one uses in- or out-degree for the preferential attachment rule, but not both. In the next section we describe a simple model which uses both.

1.11 Directed scale-free graphs

Although sometimes described in terms of directed graphs, with the exception noted above all the models presented so far are to all intents and purposes undirected, in the sense that the edge orientations are not used in an essential way in defining the model. As the real-world

networks motivating scale-free random graphs are often directed, it makes sense to consider directed models, and it is natural to consider 'preferential attachment' which depends on in- and out-degrees. Such a model was introduced by Bollobás, Borgs, Chayes and Riordan in [12]:

> We consider a graph which grows by adding single edges at discrete time steps. At each such step a vertex may or may not also be added. For simplicity we allow multiple edges and loops. More precisely, let α, β, γ, δ_{in} and δ_{out} be non-negative real numbers, with $\alpha + \beta + \gamma = 1$. Let G_0 be any fixed initial graph, for example a single vertex without edges, and let t_0 be the number of edges of G_0. (Depending on the parameters, we may have to assume $t_0 \geq 1$ for the first few steps of our process to make sense.) We set $G(t_0) = G_0$, so at time t the graph $G(t)$ has exactly t edges, and a random number $n(t)$ of vertices. In what follows, *to choose* a vertex v of $G(t)$ *according to* $d_{out} + \delta_{out}$ means to choose v so that $\Pr(v = v_i)$ is proportional to $d_{out}(v_i) + \delta_{out}$, i.e., so that $\Pr(v = v_i) = (d_{out}(v_i) + \delta_{out})/(t + \delta_{out}n(t))$. To *choose* v *according to* $d_{in} + \delta_{in}$ means to choose v so that $\Pr(v = v_i) = (d_{in}(v_i) + \delta_{in})/(t + \delta_{in}n(t))$, where all degrees are measured in $G(t)$.
>
> For $t \geq t_0$ we form $G(t+1)$ from $G(t)$ according the the following rules:
>
> (A) With probability α, add a new vertex v together with an edge from v to an existing vertex w, where w is chosen according to $d_{in} + \delta_{in}$.
>
> (B) With probability β, add an edge from an existing vertex v to an existing vertex w, where v and w are chosen independently, v according to $d_{out} + \delta_{out}$, and w according to $d_{in} + \delta_{in}$.
>
> (C) With probability γ, add a new vertex w and an edge from an existing vertex v to w, where v is chosen according to $d_{out} + \delta_{out}$.
>
> The probabilities α, β and γ clearly should add up to one. To avoid trivialities, we will also assume that $\alpha + \gamma > 0$. When considering the web graph we take $\delta_{out} = 0$; the motivation is that vertices added under step (C) correspond to web pages which purely provide content - such pages never change, are born without out-links and remain without out-links. Vertices added under step (A) correspond to usual pages, to which links may be later added. While mathematically it seems natural to take $\delta_{in} = 0$ in addition to $\delta_{out} = 0$, this gives a model in which every page not in G_0 has either no in-links or no out-links, which is rather unrealistic and uninteresting! A non-zero value of δ_{in} corresponds to insisting that a page is not considered part of the web until something points to it, typically one of the big search engines. It is natural to consider these edges from search engines separately from the rest of the graph, as they are of a rather different nature; for the same reason it is natural not to insist that δ_{in} be an integer. We include the parameter δ_{out} to make the model symmetric with respect to reversing the directions of edges (swapping α with γ and δ_{in} with δ_{out}), and because we expect the model to be applicable in contexts other than that of the web graph.
>
> Our model allows loops and multiple edges; there seems no reason to exclude them. However, there will not be very many, so excluding them would not significantly affect our conclusions.

Note also that our model includes (a precise version of) the $m = 1$ case of the original model of Barabási and Albert as a special case, taking $\beta = \gamma = \delta_{out} = 0$ and $\alpha = \delta_{in} = 1$. We could introduce more parameters, adding m edges for each new vertex, or (as in [24]) a random number with a certain distribution, but one of our aims is to keep the model simple, and the main effect, of varying the overall average degree, can be achieved by varying β.

As for the other models, power law degree distribution is proved, this time for in- and out-degrees separately. Setting

$$c_1 = \frac{\alpha + \beta}{1 + \delta_{in}(\alpha + \gamma)} \quad \text{and} \quad c_2 = \frac{\beta + \gamma}{1 + \delta_{out}(\alpha + \gamma)},$$

and writing $x_i(t)$ for the number of vertices of $G(t)$ with in-degree i, and $y_i(t)$ for the number with out-degree i, the following result is proved in [12].

Theorem 11. *Let $i \geq 0$ be fixed. There are constants p_i and q_i such that $x_i(t) = p_i t + o(t)$ and $y_i(t) = q_i t + o(t)$ hold with probability 1. Furthermore, if $\alpha\delta_{in} + \gamma > 0$ and $\gamma < 1$, then as $i \to \infty$ we have*

$$p_i \sim C_{IN} i^{-X_{IN}},$$

where $X_{IN} = 1 + 1/c_1$ and C_{IN} is a positive constant. If $\gamma\delta_{out} + \alpha > 0$ and $\alpha < 1$, then as $i \to \infty$ we have

$$q_i \sim C_{OUT} i^{-X_{OUT}},$$

with $X_{OUT} = 1 + 1/c_2$ and C_{OUT} is a positive constant.

In the statement above, the $o(t)$ notation refers to $t \to \infty$ with i fixed, while $a(i) \sim b(i)$ means $a(i)/b(i) \to 1$ as $i \to \infty$.

In addition, the joint distribution of in- and out-degrees is studied; formulae are given for the limiting fraction $r_{i,j}$ of vertices with in-degree i and out-degree j. As these are not very simple, we shall not reproduce them here.

For the rest of the article we return to the LCD model, turning our attention to properties other than the degree sequence.

1.12 Clustering coefficient and small subgraphs

Following Watts and Strogatz, one of the basic properties of the scale-free random graphs considered in many papers is the *clustering coefficient* C. As we have seen in section 1.4, this coefficient describes 'what proportion of the acquaintances of a vertex know each other'. Formally, given a simple graph G (without loops and multiple edges), and a vertex v (with at least two neighbours, say), the *local clustering coefficient* at v is given by

$$C_v(G) = \frac{\text{number of edges between neighbours of } v}{\binom{d_G(v)}{2}}, \tag{1.6}$$

where $d_G(v)$ is the degree of v in G, so the denominator is the maximum possible number of edges between neighbours of v, and $0 \le C_v(G) \le 1$. There are then two possible definitions for the clustering coefficient $C = C(G)$ of the whole graph. Perhaps the most often stated, is '$C(G)$ is the average of $C_v(G)$', i.e., taking the vertex set to be $[n]$,

$$C(G) = C_1(G) = \sum_{v=1}^{n} C_v(G)/n. \tag{1.7}$$

(Again we commit a slight abuse of notation, as 1 is also a vertex of the graph.) This kind of 'average of an average' is often not very informative; the more natural alternative is to weight by the denominator of (1.6), giving

$$C(G) = C_2(G) = \left(\sum_{v=1}^{n} \binom{d_G(v)}{2} C_v(G) \right) / \sum_{v=1}^{n} \binom{d_G(v)}{2}. \tag{1.8}$$

This second definition is easily seen to have the following natural equivalent form:

$$C_2(G) = \frac{\text{no. of pairs } ab, ac \text{ of adjacent edges for which } bc \text{ is an edge}}{\text{no. of pairs } ab, ac \text{ of adjacent edges}},$$

which has the advantage that it makes sense when some degrees are less than 2. In turn we can re-write the equation above as

$$C_2(G) = \frac{3 \times \text{number of triangles}}{\text{number of pairs of adjacent edges}}. \tag{1.9}$$

In this form there is no problem applying the definition to multigraphs.

In some papers it is not clear which of the two definitions above is intended; when it is clear, sometimes C_1 is used, and sometimes C_2. It is not often pointed out that these definitions are different: for an extreme example, take G to be a double star, where vertices 1 and 2 are joined to each other and to all other vertices, and there are no other edges. Then $C_v(G)$ is 1 for $v \ge 3$ and $2/(n-1)$ for $v = 1, 2$. It follows that $C_1(G) = 1 - o(1)$, while $C_2(G) \sim 2/n$. In more balanced graphs the definitions will give more similar values, but they will still differ by at least a constant factor much of the time.

For this section we shall use $C_2(G)$ as the definition of the clustering coefficient $C(G)$, and we shall prove the following result.

Theorem 12. *Let $m \ge 1$ be fixed. The expected value of the clustering coefficient $C(G_m^{(n)}) = C_2(G_m^{(n)})$ satisfies*

$$\mathbb{E}(C(G_m^{(n)})) \sim \frac{m-1}{8} \frac{(\log n)^2}{n}$$

as $n \to \infty$.

To prove Theorem 12 we will count the number of triangles in $G_m^{(n)}$. More generally, we describe a method for counting subgraphs isomorphic to any small fixed graph.

When $m = 1$, the Barabási-Albert or LCD model is very simple, giving either a tree or a forest with loops according to the precise definition chosen. Although this model is less interesting than the general case, it has the advantage that its small subgraphs can be analyzed precisely (see section 1.15). Because of the exact choice made in the definition of $G_m^{(n)}$, these results then carry over to this full model.

Let S be a graph on $\{1, 2, \ldots, n\}$ with loops allowed. Orient each edge ij of S with $i \le j$ from j to i. Let us write $V^+(S)$ for the set of vertices of S from which edges leave, and $V^-(S)$ for those vertices at which edges arrive. (These sets are, of course, not in general disjoint.) For $i \in V(S)$ let $d_S^{in}(i)$ be the in-degree of i in S and let $d_S^{out}(i)$ be the out-degree. (A loop at i contributes 1 to each of $d_S^{in}(i)$ and $d_S^{out}(i)$.) Finally, let $C_S(t)$ be the number of edges of S 'crossing' t, i.e., the number of edges ij of S with $i \le t$ and $j \ge t$.

Note that S is a *fixed* graph, not an isomorphism class of graphs; there are $\binom{n}{3}$ different graphs S which are triangles, for example. When we say S is a subgraph of $G_1^{(n)}$, or write $S \subset G_1^{(n)}$, we shall mean that exactly the edges of S occur in $G_1^{(n)}$, not that $G_1^{(n)}$ has a subgraph isomorphic to S. Suppose that $d_S^{out}(i) \le 1$ for every i, so S is a possible subgraph of $G_1^{(n)}$.

Theorem 13. *Let S be a possible subgraph of $G_1^{(n)}$. With the notation above, the probability p_S that $S \subset G_1^{(n)}$ satisfies*

$$p_S = \prod_{i \in V^-(S)} d_S^{in}(i)! \prod_{i \in V^+(S)} \frac{1}{2i-1} \prod_{t \notin V^+(S)} \left(1 + \frac{C_S(t)}{2t-1}\right). \tag{1.10}$$

Furthermore,

$$p_S = \prod_{i \in V^-(S)} d_S^{in}(i)! \prod_{ij \in E(S)} \frac{1}{2\sqrt{ij}} \exp\left(O\left(\sum_{i \in V(S)} C_S(i)^2/i\right)\right). \tag{1.11}$$

The result is simple to prove once one finds the correct quantity to calculate inductively; the details are given for a closely related model in section 1.15. Note that the second product in (1.11) gives essentially what one would expect if edges were present in $G_1^{(n)}$ independently of one another. The first product (and the final factor) show that they are not.

We now pass to $G_m^{(n)}$. Rather than writing down a cumbersome general formula, let us consider the case of triangles.

Theorem 14. *Let $m \ge 1$ be fixed. The expected number of triangles in $G_m^{(n)}$ is given by*

$$(1 + o(1))\frac{m(m-1)(m+1)}{48}(\log n)^3$$

as $n \to \infty$.

Proof. Recall that that $G_m^{(n)}$ is obtained from $G_1^{(mn)}$ by identifying the vertices in groups of m. Let a, b, c with $1 \le a < b < c \le n$ be given. Then abc is a triangle in $G_m^{(n)}$ if and only if there are integers $m(a-1) < i, i' \le ma$, $m(b-1) < j, j' \le mb$, $m(c-1) < k, k' \le mc$

such that the graph S with edges ij', jk' and $i'k$ is a subgraph of $G_1^{(mn)}$. Now for this S, provided $d_S^{out}(v) \le 1$ for all v, we have from (1.11) that

$$p_S = \eta_1 \prod_{x \in V^-(S)} d_S^{in}(x)! \prod_{xy \in E(S)} \frac{1}{2\sqrt{xy}} = \eta_2 \prod_{x \in V^-(S)} d_S^{in}(x)! \frac{1}{8m^3abc},$$

where the 'correction factors' η_1, η_2 are bounded, and tend to 1 if $a \to \infty$. Given $1 \le a < b < c \le n$, what are the possible choices for i, i', j, j', k, k'? Note first that k, k' must be distinct, giving $m(m-1)$ choices, as if $k = k'$ then $d_S^{out}(k) = 2$. There are m^2 choices for j, j'. Finally we may have $i = i'$, in which case $d_S^{in}(i) = 2$ and $d_S^{in}(v) \le 1$ for every other v, or $i \ne i'$ in which case $d_S^{in}(v) \le 1$ for all v. There are $m(m-1)m^2m = m^4(m-1)$ choices with $i = i'$ and $m(m-1)m^2m(m-1) = m^4(m-1)^2$ choices with $i \ne i'$. Hence the expected number of triangles with vertices a, b, c in $G_m^{(n)}$ (recall that $G_m^{(n)}$ is a multigraph, so may contain several triangles with the same vertex set) is given by

$$\eta_3 \left(m^4(m-1)2\frac{1}{8m^3abc} + m^4(m-1)^2 1 \frac{1}{8m^3abc} \right) = \eta_3 \frac{m(m-1)(m+1)}{8abc},$$

where η_3 is bounded and tends to 1 as $a \to \infty$. Summing over a, b and c with $1 \le a < b < c \le n$ we see that the main contribution is from terms with $a \to \infty$, and the expected number of triangles in $G_m^{(n)}$ is given by

$$(1+o(1)) \sum_{1 \le a < b < c \le n} \frac{m(m-1)(m+1)}{8abc} \sim \frac{m(m-1)(m+1)}{48} (\log n)^3,$$

proving Theorem 14. □

One can use the same method to prove the following more general result.

Theorem 15. *Let $l \ge 3$ be fixed. Then the expected number of l-cycles in $G_m^{(n)}$ is of the form*

$$(1+o(1)) C_{m,l} (\log n)^l$$

as $n \to \infty$ with $m \ge 2$ fixed, where $C_{m,l}$ is a positive constant. Furthermore, as $m \to \infty$ we have $C_{m,l} = \Theta(m^l)$.

Finding the exact constants in the result above becomes harder as l increases; for example, given $1 \le a < b < c < d < e \le n$ there are 12 ways to arrange a 5-cycle with these vertices. In 8 of these arrangements there are two vertices with two edges coming in from the right. In the other four there is only one such vertex. When passing to $G_1^{(mn)}$ there may thus be 0,1 or 2 vertices with in-degree 2.

Note that Theorems 14 and 15 contradict the heuristic value of $n^{l/4}$ for the number of l-cycles given by Farkas, Derényi, Barabási and Vicsek [31] on the basis of eigenvalue distribution.

Let us finish this section by returning to the clustering coefficient, calculated according to (1.9). Having estimated the number of triangles, we only need to know the number of pairs of adjacent edges. Let us write $P_2 = P_2(G_m^{(n)})$ for the number of pairs of adjacent edges ab, ac in $G_m^{(n)}$.

Theorem 16. *Let $m \geq 1$ and $\epsilon > 0$ be fixed. Then*

$$(1-\epsilon)\frac{m(m+1)}{2} n \log n \leq P_2(G_m^{(n)}) \leq (1+\epsilon)\frac{m(m+1)}{2} n \log n$$

holds **whp** *as $n \to \infty$.*

Proof. The result is easy to prove using the methods above, so we give only a brief sketch.

There are three types of contribution to P_2: we may have $b, c \leq a$, $b \leq a < c$ (equivalent to $c \leq a < b$) or $a < b, c$. Since all out-degrees in $G_m^{(n)}$ are at most m, there can only be $O(n)$ pairs of the first two types. Hence, using the methods above (skipping the details),

$$\mathbb{E}(P_2(G_m^{(n)})) = O(n) + (1+o(1)) \sum_{1 \leq a < b < c} \frac{2m^3 + m^3(m-1)}{4m^2 a\sqrt{bc}}$$
$$\sim \frac{m(m+1)}{2} n \log n.$$

Standard martingale methods can be used to show that P_2 is concentrated within $O(n)$ of its mean, completing the proof. □

Note that both the number of triangles and the number of pairs of adjacent edges are not quite what one might expect just by looking at the individual edge probabilities (or the degrees). One difference is an extra factor of $(m+1)/m$ appearing in both, from the correlation between the presence of edges ab and ac when $a < b < c$. The other is a factor $(m-1)/m$ only in the number of triangles, from the fixed out-degrees. These factors are often ignored, for example in [39].

Combined with Theorem 14 and the definition (1.9), the result above shows that the expected clustering coefficient C of $G_m^{(n)}$ is asymptotically

$$\frac{m-1}{8}\frac{(\log n)^2}{n},$$

proving Theorem 12. Note that the clustering coefficient is very different from the experimental value $C \sim n^{-0.75}$ given for $m = 2$ by Barabási and Albert [2], or the heuristic $C \sim n^{-0.25}$ that would follow from the claims of Farkas, Derényi, Barabási and Vicsek [31]. Klemm and Eguiluz [39] give an 'analytic' value of

$$C_1(G) \sim \frac{m}{8}\frac{(\log n)^2}{n}$$

for $C_1(G)$; this is off by a constant factor for two reasons. One is that the heuristic used ignores the factor $(m-1)/m$ mentioned above. The other is that, while the aim is clearly to calculate $C_1(G)$, a heuristic used is to replace the top and bottom of (1.6) by their expectations. This introduces an error which, for $G_m^{(n)}$, one can check is a roughly constant factor; it turns out that this error is roughly the ratio between $C_2(G)$ and $C_1(G)$, explaining the similarity between the formula above and the true form of $C_2(G)$ given in Theorem 12.

1.13 Pairings on $[0,1]$ and the diameter of the LCD model

So far the results concerning the LCD model have been obtained directly either from (1.2) or from the discrete combinatorial interpretation in terms of pairings. In [16] another way of generating $G_m^{(n)}$ was introduced; this formulation, in terms of pairings of random real numbers, is useful for proving more complicated results, as it allows the re-introduction of independence to a significant extent.

Let $N = mn$. The idea is that to obtain our LCD with N chords, instead of pairing off fixed points $1, 2, \dots 2N$, we shall pair off random points in the interval $[0,1]$. In fact, taking iid uniformly random points $x_1, \dots, x_{2N}$, we may as well pair x_{2i-1} with x_{2i} for all i; the randomness of the order in which the x_i appear when moving from 0 to 1 guarantees that the LCD obtained is the uniformly random LCD we require.

We now consider generating the pairing starting with the right endpoints. As in [16], we call a random variable with density function $2x$, $0 < x < 1$, an $M_2(0,1)$ *random variable*. Let us write l_i, r_i for the left and right endpoints of the chord $x_{2i-1}x_{2i}$, so $\{l_i, r_i\} = \{x_{2i-1}, x_{2i}\}$ with $l_i < r_i$. Then $\Pr(r_i \le t) = \Pr(x_{2i-1}, x_{2i} \le t) = t^2$, so the r_i are iid $M_2(0,1)$ random variables. Also, given $r_1, \dots, r_N$, the random variables $l_1, \dots, l_N$ are independent with l_i uniformly distributed on $[0, r_i]$.

To express the LCD we have defined as a pairing on $\{1, 2, \dots, 2N\}$, we must sort all the x_i together. We do this by first sorting the r_i, and then considering between which r_i each l_i lies.

The construction for $G_1^{(mn)}$ is as follows: we start with $N = mn$ iid $M_2(0,1)$ random variables, $r_1, \dots, r_N$. Sort these into increasing order, to obtain $R_1, \dots, R_N$, setting $R_0 = 0$. Let $L'_1, \dots, L'_N$ be independent, with L'_i uniform on $[0, R_i]$. Then our LCD $\mathcal{L}$ is given by pairing L'_i and R_i. As the right endpoints $R_1, \dots, R_N$ are already in order, if $R_{j-1} < L'_i < R_j$ then in the graph $G_1^{(mn)}$ obtained as $\phi(\mathcal{L})$ (see section 1.7), vertex i sends its outgoing edge to vertex j. (Throughout we of course ignore the probability zero event that two endpoints are the same.)

For $G_m^{(n)}$ we must merge vertices in groups of m, so what will really matter is where the m^{th}, $2m^{\text{th}}$ etc. right endpoints lie. Simplifying very slightly, the construction is as follows: let the R_i be defined as above. For $1 \le i \le n$ set $W_i = R_{mi}$, taking $W_0 = 0$. To obtain exactly $G_m^{(n)}$ we should consider N independent random variables which we denote $L_{i,r}$, $1 \le i \le n$, $1 \le r \le m$, with $L_{i,r}$ uniform on $[0, R_{(m-1)i+r}]$. In fact it is often good enough to work only with the W_i, taking $L_{i,r}$ uniform on $[0, W_i] = [0, R_{mi}]$. The graph $G_m^{(n)}$ is obtained by taking edges from i to m (not necessarily distinct) vertices $t_{i,1}, \dots, t_{i,m}$ obtained as follows: $t_{i,r}$ is the integer t for which $W_{t-1} < L_{i,r} < W_t$.

In summary, the following is an almost exact alternative description of $G_m^{(n)}$. (The modifications to make it exact are implicit in the paragraph above.)

Let random variables W_i be defined as above, and set $w_i = W_i - W_{i-1}$. Given the W_i, define independent random variables $t_{i,r}$, $1 \le i \le n$, $1 \le r \le m$, with

$$\Pr(t_{i,r} = j) = \begin{cases} w_j/W_i & j \le i, \\ 0 & j > i. \end{cases} \tag{1.12}$$

Then the graph formed by taking edges from i to $t_{i,r}$ has (essentially) the same distribution

as $G_m^{(n)}$. The power of this approach is that we may condition on the W_i, assuming they have 'typical' properties. Then the $t_{i,r}$ are conditionally independent.

As a simple application of this approach, we observe that the maximum degree of $G_m^{(n)}$ has the following rather unpleasant description. Let $X_1, X_2, \dots$ be the points of a Poisson process on $[0,\infty]$ with rate m, so, setting $X_0 = 0$, the variables $X_i - X_{i-1}$ are iid exponentials with mean $1/m$. Let $Y_i = \sqrt{X_{mi}}$, and let $D_m = \max\{Y_i - Y_{i-1}, 1 \le i < \infty\}$, noting that this maximum exists with probability one. Note that the distribution of D_m depends on m only. Let $\Delta(G_m^{(n)})$ denote the maximum degree of $G_m^{(n)}$.

Theorem 17. *Let $m \ge 1$ be fixed. Then $\Delta(G_m^{(n)})/(2m\sqrt{n})$ converges in distribution to D_m as $n \to \infty$.*

Proof. As before, here we can only give a sketch. Note that if U is a random variable which is uniform on $[0,1]$, then $\sqrt{U}$ has a $M_2(0,1)$ distribution, since $\Pr(\sqrt{U} \le t) = \Pr(U \le t^2) = t^2$. Loosely speaking, it follows that for large n, the distribution of the squares of the first few R_i is given by a Poisson process of rate mn on $[0,\epsilon]$, for ϵ sufficiently small, so we may take $R_i^2 = X_i/n$, or $R_i = \sqrt{X_i/n}$. Then the W_i are given by $Y_i/\sqrt{n}$, so $\max\{w_i\}$ is given by $D_m/\sqrt{n}$. Finally, *given the* W_i, the degree in $G_m^{(n)}$ of a particular early vertex j is concentrated about its expectation of $(2+o(1))mnw_j$. □

A much more substantial result proved using this description of $G_m^{(n)}$ is the diameter formula in [16]. Before stating this result, let us pause for a moment to consider what we might expect the diameter to be. Computer experiments presented by Barabási, Albert and Jeong [3, 7] and heuristic arguments given by Newman, Strogatz and Watts [48] suggest that $G_m^{(n)}$ should have diameter of the form $A + B\log n$. At first sight, such a small diameter might seem surprising, but it is in line with the Watts-Strogatz small-world phenomenon described in section 1.4. What would we expect from the point of view of random graphs? Certainly *at most* $\Theta(\log n)$: as described in section 1.4, even a small amount of global randomness gives logarithmic diameter. In fact one might expect the diameter to be even smaller: the unbalanced degree distribution pushes up the number of small paths, and thus, perhaps, pushes the diameter down. As shown in [16], this is indeed the case, though it is not very easy to prove.

Theorem 18. *Fix an integer $m \ge 2$ and a positive real number ϵ. Then* **whp** *$G_m^{(n)}$ is connected and has diameter $\mathrm{diam}(G_m^{(n)})$ satisfying*

$$(1-\epsilon)\log n/\log\log n \le \mathrm{diam}(G_m^{(n)}) \le (1+\epsilon)\log n/\log\log n.$$

The lower bound is relatively straightforward, based on counting the expected number of paths between two fixed vertices using techniques similar to those in section 1.12. The upper bound, proved via a neighbourhood expansion argument, is much more complicated. Essential use is made of the independence introduced by conditioning on the W_i, but even with this there are many complications.

As pointed out in [16], and independently by Cohen and Havlin [23], there is a heuristic argument giving the correct diameter of $\log n/\log\log n$. (This is just the standard neighbourhood expansion argument without the details; it is important to take the whole degree sequence and not apply some form of cutoff.) However one must be careful with such heuristics. For

example, they apply also to the case $m = 1$, where, as shown by Pittel [49] in the context of scale-free trees, the diameter is $\Theta(\log n)$.

1.14 Robustness and vulnerability

Another property of scale-free graphs and the real-world networks inspiring them which has received much attention is their 'robustness'.

Suppose we delete vertices independently at random from $G_m^{(n)}$, each with probability q. What is the structure of the remaining graph? Is it connected? Does it have a giant component? A precise form of the question is: fix $0 < q < 1$. Suppose vertices of $G_m^{(n)}$ are deleted independently at random with probability $q = 1 - p$. Let the graph resulting be denoted G_p. For which p is there a constant $c = c(p) > 0$ independent of n such that with high probability G_p has a component with at least cn vertices? What is the critical value of p below which no such constant exists?

As noted in [16], it is easy to see from the neighbourhood expansion argument used there that there is no critical p. Once the neighbourhoods of a given vertex reach a certain size (which happens with some positive probability), they continue expanding, and the vertex is almost certainly joined to the vertex surviving in G_p with lowest index. However, it turns out that this 'giant' component becomes very small as p approaches zero. To estimate its size we can use the pairing model to relate the structure of $G_m^{(n)}$ to a certain scale-free branching process; here we shall only give an outline of the argument, referring the reader to [18] for the rather technical details.

Theorem 19. *Let $m \geq 2$ and $0 < p < 1$ be fixed, and let G_p be obtained from $G_m^{(n)}$ by deleting vertices independently with probability $1 - p$. Then as $n \to \infty$,* **whp** *the largest component of G_p has order $(c(p,m) + o(1))n$ for some constant $c(p,m) > 0$. Furthermore, as $p \to 0$ with m fixed, $c(p,m) = \exp(1/O(p))$.*

Proof. Recall from the pairing model that each vertex i sends out m edges, to targets $t_{i,1}, \ldots, t_{i,m}$, where the $t_{i,r}$ are independent and for $j \leq i$ we have, essentially,

$$\Pr(t_{i,r} = j) = w_j / W_i$$

for random quantities w_j and $W_i = \sum_{j=0}^{i-1} w_i$ defined earlier. To a good approximation (in a sense we shall not make precise here), as long as $i \to \infty$ we can replace W_i by the value $\sqrt{i/n}$ around which it is concentrated. Also, w_i is the 'waiting time' from W_{i-1} for m samples from a distribution with density $2mnx$, $0 < x < 1$. So to a good approximation the w_i are given by

$$w_i = \frac{Z_i}{2mnW_i} = \frac{Z_i}{2m\sqrt{in}},$$

where the Z_i are iid, each with a distribution Z the sum of m independent exponential random variables with parameter 1, so Z has density

$$f_Z(x) = \frac{x^{m-1}e^{-x}}{(m-1)!}. \tag{1.13}$$

To this degree of approximation we have

$$\Pr(t_{i,r} = j) = \frac{w_j}{W_i} \sim \frac{Z_j}{2m\sqrt{jn}}\sqrt{\frac{n}{i}} = \frac{Z_j}{2m\sqrt{ij}}. \tag{1.14}$$

Re-scaling, let us write $i = \alpha n$, and consider the probability that $t_{i,r} = j$ for some j with $j \in [\beta n, (\beta+\mathrm{d}\beta)n]$ and $Z_j \in [y, y+\mathrm{d}y]$. Since there are $n\mathrm{d}\beta$ vertices j in this range, of which a fraction $f_Z(y)\mathrm{d}y$ will have Z_j in the required interval, this probability is approximately

$$n\mathrm{d}\beta f_Z(y)\mathrm{d}y\frac{y}{2m\sqrt{\alpha n\beta n}} = \frac{yf_Z(y)}{2m\sqrt{\alpha\beta}}\mathrm{d}y\mathrm{d}\beta, \tag{1.15}$$

provided $\beta < \alpha$. Similarly, let us fix a vertex i with $i = \alpha n$ and $Z_i = x$ and consider the probability that there is a vertex $j > i$ with $j \in [\beta n, (\beta + \mathrm{d}\beta)n]$ and $Z_j \in [y, y + \mathrm{d}y]$ sending an edge to i. Again there are $n\mathrm{d}\beta$ vertices j in the right range, a fraction $f_Z(y)\mathrm{d}y$ of which have Z_j in the right range, so using (1.14) again this probability is approximately

$$mn\mathrm{d}\beta f_Z(y)\mathrm{d}y\frac{x}{2m\sqrt{\alpha n\beta n}} = \frac{xf_Z(y)}{2\sqrt{\alpha\beta}}\mathrm{d}y\mathrm{d}\beta. \tag{1.16}$$

(Here the initial factor of m comes from each vertex sending out m edges independently).

Motivated by the above let us define a birth process as follows: in generation $t \geq 0$ there will be a finite number $N(t)$ of 'vertices'. Each vertex v has three numbers associated with it: $\alpha(v) \in (0, 1)$, corresponding to the α in $i = \alpha n$ above, $x(v)$, corresponding to Z_i above, and an integer $l(v)$ which will be either m or $m - 1$. This tells us the number of 'left-children' of v: as we work outwards from an initial vertex of $G_m^{(n)}$ finding all vertices at distance 1, then distance 2, etc., there are two ways we can reach a new vertex w from an old vertex w'; from the right (so $w < w'$ and $t_{w',r} = w$ for some r) or from the left. In the next step there will be m or $m - 1$ new left-children of w (vertices $t_{w,s}$, $1 \leq s \leq m$) respectively.

A vertex v in generation t with $\alpha(v) = \alpha$ and $x(v) = x$ gives rise to provisional offspring in generation $t + 1$ as follows: v gives rise independently to exactly $l(v)$ provisional left-children $w_1, \ldots, w_{l(v)}$. For each i we have $l(w_i) = m$, and the values $\beta_i = \alpha(w_i)$ and $y_i = x(w_i)$ are chosen according to the density (1.15), with $0 < \beta_i \leq \alpha$. Also, v gives rise to a Poisson number of provisional right-children w, each with $l(w) = m-1$, with the chance of v giving rise to a provisional right-child w having $\alpha(w) \in [\beta, \beta+\mathrm{d}\beta]$ and $x(w) \in [x, x+\mathrm{d}x]$ given by (1.16), for $\alpha \leq \beta < 1$.

To obtain the next generation, we take all the provisional children of the current generation, and keep each with probability p, independently of the others.

Let us write $\mathcal{N} = \mathcal{N}_p^{l,\alpha,x}$ for the process defined above, starting with a single vertex v having $l(v) = l$, $\alpha(v) = \alpha$ and $x(v) = x$. Note that the definition of $\mathcal{N}$ does not involve n. Let us write $c(v)$ for the total number of descendants of v in all generations, which may be infinite. There is a certain 'survival probability' $s(p, l, \alpha, x) = \Pr(c(v) = \infty)$. By elementary probability theory, we have

$$s(p, l, \alpha, x) = \lim_{t\to\infty} \Pr(c(v) \geq t),$$

and hence

$$s(p, l, \alpha, x) = \Pr(c(v) \geq (\log n)^{10}) + o(1), \tag{1.17}$$

say, as $n \to \infty$. From the remarks of the last few paragraphs, when n is large the process $\mathcal{N}_1^{m,\alpha,x}$ gives a good approximation to the initial growth (up to $(\log n)^{10}$ vertices, say) of the neighbourhoods of a vertex i in $G_m^{(n)}$ with $i = \alpha n$ and $Z_i = x$. Once the neighbourhoods reach size $(\log n)^{10}$, with high probability i is in the giant component. Using (1.17), it follows that for α bounded away from 0 and 1 and for any fixed x, the probability that such a vertex lies in the giant component of G_p is given by $s(p, m, \alpha, x) + o(1)$.

Fix $0 < p < 1$; from now on we suppress dependence on p. Let us write $L(\alpha)$ for the chance that a particular potential left-child w of a vertex v with $\alpha(v) = \alpha$ and $x(v) = x$ itself survives, and has $c(w) = \infty$. Note that this probability does not depend on x. Also, let us write $r(\alpha, x)$ for the chance that *some* potential right-child w of v survives and has $c(w) = \infty$. Since $c(v)$ is finite if and only if $c(w)$ is finite for all surviving children w of v, we have

$$s(p, l, \alpha, x) = 1 - (1 - L(\alpha))^l (1 - r(\alpha, x)). \tag{1.18}$$

Since the density (1.16) is proportional to x, it is easy to see that $r(\alpha, x)$ has the form

$$r(\alpha, x) = 1 - e^{-xR(\alpha)}$$

for some function R depending on α only. (R is the negative of the log of the chance that all right-children of v die out, if $\alpha(v) = \alpha$ and $x(v) = 1$.) Using again that $c(v) = \infty$ if and only if at least one child w of v survives and has $c(w) = \infty$, it is easy to see that

$$L(\alpha) = p \int_{\beta=0}^{\alpha} \int_{y=0}^{\infty} \left(1 - (1 - L(\beta)^m) e^{-yR(\beta)}\right) \frac{y f_Z(y)}{2m\sqrt{\alpha\beta}} \mathrm{d}y \mathrm{d}\beta.$$

Indeed, given v, the last factor gives (from (1.15)) the chance that the particular potential left-child w of v we are considering has $\alpha(w) = \beta$ and $x(w) = y$. From (1.18) the factor in brackets is the chance that $c(w) = \infty$. The first factor p is the chance that w itself is actually born in the first place. From the form (1.13) of $f_Z(y)$ it turns out that the y integral is easy to do, giving

$$L(\alpha) = \frac{p}{2\sqrt{\alpha}} \int_{\beta=0}^{\alpha} \frac{1}{\sqrt{\beta}} \left(1 - \frac{(1 - L(\beta)^m)}{(1 + R(\beta))^{m+1}}\right) \mathrm{d}\beta. \tag{1.19}$$

Similarly, one can check that

$$R(\alpha) = \frac{p}{2\sqrt{\alpha}} \int_{\beta=\alpha}^{1} \frac{1}{\sqrt{\beta}} \left(1 - \frac{(1 - L(\beta)^{m-1})}{(1 + R(\beta))^m}\right) \mathrm{d}\beta. \tag{1.20}$$

The pair of equations above may have more than one solution (in particular, both L and R identically zero is a solution); standard probability theory tells us that the functions L and R are the (unique) maximal solution to (1.19),(1.20).

From the equations above it is easy to deduce Theorem 19 (as well as more detailed results). Using monotonicity of the functionals on the right hand side one can find upper and lower bounds on $L(\alpha)$ and $R(\alpha)$ and hence on the expected size of the giant component of G_p. Concentration follows immediately since the proof shows that all vertices are in very small components (order $O((\log n)^{10})$, say) or the giant component, and the number of vertices in very small components can be shown to be concentrated by standard martingale methods. □

The situation when the graph $G_m^{(n)}$ is deliberately attacked is rather different. Given the actual (random) graph, determining the 'best' attack is not an easy problem, and is in any case not realistic. Here we consider the natural, simple approach of deleting the earliest vertices, up to some cutoff cn. This time there is a value $c < 1$ beyond which there is no giant component in what remains.

Theorem 20. *Let G_c be obtained from $G_m^{(n)}$ by deleting all vertices with index less than cn, where $0 < c < 1$ is constant, and let $c_m = (m-1)/(m+1)$. If $c < c_m$ then* **whp** *G_c has a component with $\Theta(n)$ vertices. If $c > c_m$ then* **whp** *G_c has no such component.*

Proof. We just give an outline, using the methods above. Considering the quantities analogous to $L(\alpha)$ and $R(\alpha)$ but defined for G_c, we obtain equations identical to (1.19) and (1.20) except that now $p = 1$, and the lower limit in the integral in (1.19) is c rather than 0. Near the critical probability, the functions L and R will be small, and hence close to the solution of the linearized form of the equations. It is easy to solve these linearized equations; a non-zero solution exists if and only if $c = c_m$, and one can deduce the result. □

1.15 The case $[0, 1]$: plane-oriented recursive trees

A simple special case of the BA model that has been considered in several papers is the $m = 1$ case, where each vertex sends a single edge to an earlier vertex, giving rise to a tree. In this context the LCD model is not the most natural interpretation of the BA description, as it gives rise to a forest with loops. It turns out that essentially the $m = 1$ case of the BA model had been considered more than a decade before [5].

For $m = 1$, the generally accepted interpretation of the imprecise description in [5] is to start with one vertex, the *root* which has an extra 'virtual edge' coming in to it from nowhere, so the degree of the root at the start counts as 1. Thus at time t, when there are t vertices, although there are $t - 1$ edges in the tree, the effective sum of the degrees is $2(t-1)+1$. As has occasionally been pointed out, this precise version of the $m = 1$ model is not at all new; it is exactly the standard model for random plane-oriented recursive trees. A tree on a labelled vertex set $V = \{1, 2, \ldots, t\}$ is *recursive* if each vertex other than 1 is joined to exactly one earlier vertex. In other words, the tree can be grown by adding the vertices in numerical order, joining each new vertex to some old vertex. Uniform random recursive trees, grown one vertex at a time by joining the new vertex to an old vertex chosen uniformly at random, have been studied for some time; see, for example, the survey [44].

A *plane-oriented* tree is one with a cyclic order at each vertex, induced, for example, by drawing the tree in the plane. When a new vertex v is added to a plane-oriented recursive tree T and joined to an existing vertex w, the number of different plane-oriented recursive trees that may result is given by the degree d of w in T, as there d different ways in which the new edge can meet the vertex w. In fact, as in the BA model, the standard definition treats the first vertex, the root, differently, effectively imagining an edge from the root going off to infinity. In this way branches of plane-oriented recursive trees are again plane-oriented recursive trees. Plane oriented recursive trees were introduced by Szymański [54] in 1987 (although with a slightly different treatment of the root) and have been studied since in several papers, including [8, 43, 45, 54].

For this section by T_n we shall mean the random plane-oriented recursive tree with n vertices, with vertex set $\{1, 2, \dots n\}$, or, equivalently, the Barabási-Albert scale-free random tree given by (a precise version of) the $m = 1$ case of the model introduced in [5]. Formally, T_1 consists of a single vertex 1 with no edges. For $n \geq 2$, given T_{n-1}, the tree T_n is constructed by adding a new vertex n and joining it to and old vertex v, $1 \leq v \leq n-1$, with

$$\Pr(v = j) = \frac{d_{n-1}(j)}{2n-3}$$

for $j \geq 2$ and

$$\Pr(v = 1) = \frac{d_{n-1}(1)+1}{2n-3},$$

where $d_{n-1}(j)$ is the degree of the vertex j in the tree T_{n-1}. As the tree grows out from the root we shall say that the new vertex added as above is a *child* of v, and that v is its *parent*.

Note that definition of T_n is very similar to that of $G_1^{(n)}$ as given in section 1.7.

Since T_n has been around for some time, it is not surprising that various properties are known already. For example, the 'load scaling' considered in [53] and [35] was already determined in a more precise form in [45] (see also [17]). In contrast, while [53] claims that the distribution of shortest path lengths has also been established, this is not the case as far as we know; it is certainly not in the references cited.

Here we shall show how to calculate exactly the probability that a certain subgraph is present in T_n. A form of this result for $G_1^{(n)}$ was used in section 1.12 to study clustering and small subgraphs of $G_m^{(n)}$. From this result one can also obtain the distribution of shortest paths.

Based on the work in [16] it is possible to write down a formula for the probability that a particular subgraph is present in T_n. The key is to consider a certain sequence of expectations that can be calculated inductively. As the formulae are a little complicated, we start with some motivation. Note that throughout this section, as in section 1.12, we are asking whether a particular subgraph S is present in T_n, not whether some subgraph *isomorphic* to S is present. Thus for example S may consist of the edges $\{2, 3\}$, $\{3, 7\}$ and $\{2, 8\}$. Then $\Pr(S \subset T_n)$ is taken to mean the probability that in T_n we have exactly these edges (so the parent of 3 is 2, that of 7 is 3 and that of 8 is 2), not the probability that T_n contains a path of 3 edges.

Let us write f_t for the parent of vertex t. Then, given T_t, the probability that f_{t+1} is i is proportional to the degree of i in T_t; more precisely, from the definition of T_n we have $\Pr(f_{t+1} = i \mid T_t) = \frac{d_t(i)}{2t-1}$, where $d_t(i)$ is the degree of i in T_t. Taking expectations, we see that $\Pr(f_{t+1} = i) = \mathbb{E}(d_t(i))/(2t-1)$, so we would like to know the expectations $\mathbb{E}(d_t(i))$. These are easy to calculate: $\mathbb{E}(d_i(i)) = 1$ for all $i \geq 1$, and in going from T_t to T_{t+1} the degree of i increases by one if $f_{t+1} = i$, which happens with probability $d_t(i)/(2t-1)$. Thus

$$\mathbb{E}(d_{t+1}(i)) = \mathbb{E}(d_t(i)) + \frac{\mathbb{E}(d_t(i))}{2t-1} = \frac{2t}{2t-1}\,\mathbb{E}(d_t(i)),$$

giving $\mathbb{E}(d_t(i)) = \prod_{s=i}^{t-1} \frac{2s}{2s-1}$.

More generally, suppose that S is a fixed graph consisting of a set S' of edges with both ends numbered less than j, say, and one more edge from k to i, with $i \leq j < k$. Given that

$S' \subset T_j$ we will have $S \subset T_k$ if and only if $f_k = i$, and the probability of the latter event depends on the degree of i, as before. Thus we would like to calculate $\mathbb{E}(d_t(i) \mid S' \subset T_t)$. Once we have this expectation for $t = j$, its values for $t = j+1, \ldots, k-1$ can be calculated inductively as before. The problem occurs when two edges end at the same vertex: if $S' = S'' \cup \{ij\}$, say, then while we have $\Pr(S' \subset T_j) = \mathbb{E}(d_{j-1}(i) \mid S'' \subset T_{j-1})/(2j-3)$, the event that $f_j = i$ is more likely when $d_{j-1}(i)$ is large, so $\mathbb{E}(d_j(i) \mid S' \subset T_j)$ is not related in a simple way to quantities we have already calculated.

The key turns out to be to consider rising factorials: $[d]_r = d(d+1)\cdots(d+r)$. We keep track of an expectation involving $[d_t(i)]_r$ for each vertex i of the subgraph S which will have r edges coming into it in the future (times later than t).

From now on, fix a graph S which is possible as a subgraph of T_n for large n. (So S has no loops, and for each vertex i, there is at most one edge in S from i to an earlier vertex.) For $t \geq i$ let $R_t(i)$ be the number of $j > t$ such that ij is an edge of S, i.e., the number of edges in S coming in to i after time t. Let S_t consist of those edges ij of S for which $i, j \leq t$, let

$$X_t = \prod_{ij \in E(S_t)} \mathbb{I}_{ij \in E(T_t)} \prod_{i \in V(S), i \leq t} [d_t(i)]_{R_t(i)},$$

and set $\lambda_t = \mathbb{E}(X_t)$. Here $\mathbb{I}_{\mathcal{A}}$ is the indicator function of the event $\mathcal{A}$. Note that $\lambda_0 = 1$, while for t large (at least the largest vertex in S) we have $X_t = \mathbb{I}_{S \subset T_t}$, so $\lambda_t = \Pr(S \subset T_t)$, the quantity we wish to calculate.

Lemma 21. *For $t \geq 0$ we have*

$$\lambda_{t+1} = R_{t+1}(t+1)! \frac{1}{2t-1} \lambda_t \tag{1.21}$$

if there is an edge $\{k, t+1\}$ in S with $k \leq t$, and

$$\lambda_{t+1} = R_{t+1}(t+1)! \left(1 + \frac{C_S(t+1)}{2t-1}\right) \lambda_t \tag{1.22}$$

otherwise, where $C_S(t+1)$ is the number of edges $ij \in E(S)$ with $i \leq t$ and $j > t$.

Proof. Since in T_{t+1} the degree of $t+1$ is always exactly 1, we can write X_{t+1} as

$$X_{t+1} = R_{t+1}(t+1)!Y, \tag{1.23}$$

where

$$Y = \prod_{ij \in E(S_{t+1})} \mathbb{I}_{ij \in E(T_{t+1})} \prod_{i \in V(S), i \leq t} [d_{t+1}(i)]_{R_{t+1}(i)}.$$

Suppose first that S does not contain an edge $\{k, t+1\}$ with $k \leq t$. Then $S_{t+1} = S_t$, and for each $i \leq t$ we have $R_{t+1}(i) = R_t(i)$. Also, as each edge ij of $S_{t+1} = S_t$ has $i, j \leq t$, for each such edge we have $\mathbb{I}_{ij \in E(T_{t+1})} = \mathbb{I}_{ij \in E(T_t)}$. Thus, in this case,

$$Y = \prod_{ij \in E(S_t)} \mathbb{I}_{ij \in E(T_t)} \prod_{i \in V(S), i \leq t} [d_{t+1}(i)]_{R_t(i)}.$$

Note that this is exactly the formula for X_t except that $d_t(i)$ has been replaced by $d_{t+1}(i)$. Now let us fix T_t, and hence X_t, and consider the random choice of f_{t+1}, the vertex that the next new vertex joins to. We will have $Y = X_t$ unless f_{t+1} is a vertex of S, as all relevant degrees will be the same in T_{t+1} as in T_t. What happens if $f_{t+1} = j$ for some $j \in V(S)$? Then $d_{t+1}(j) = d_t(j) + 1$, so

$$[d_{t+1}(j)]_{R_t(j)} = [d_t(j) + 1]_{R_t(j)} = \frac{d_t(j) + R_t(j)}{d_t(j)} [d_t(j)]_{R_t(j)},$$

and, as all other degrees stay the same, $Y - X_t = X_t R_t(j)/d_t(j)$. Now for each $j \in V(S)$, $j \leq t$, the probability that $f_{t+1} = j$ is just $d_t(j)/(2t-1)$. Thus the expected difference $Y - X_t$ is given by

$$\mathbb{E}(Y - X_t) = \sum_{j \in V(S), j \leq t} \frac{d_t(j)}{2t-1} X_t R_t(j)/d_t(j) = X_t \frac{C_S(t+1)}{2t-1}.$$

Taking expectations of both sides gives $\mathbb{E}(Y) = (1 + C_S(t+1)/(2t-1))\,\mathbb{E}(X_t)$, which together with (1.23) proves (1.22).

We now turn to (1.21). Suppose that $\{k, t+1\}$ is an edge of S with $k \leq t$. Given T_t, we have $Y = 0$ unless $f_{t+1} = k$, which happens with probability $d_t(k)/(2t-1)$. Supposing that f_{t+1} is equal to k, what is Y? In this case $S_{t+1} = S_t \cup \{k, t+1\}$, and, since we have $\{k, t+1\} \in E(T_{t+1})$, we have

$$\prod_{ij \in E(S_{t+1})} \mathbb{I}_{ij \in E(T_{t+1})} = \prod_{ij \in E(S_t)} \mathbb{I}_{ij \in E(T_t)}.$$

For $i \leq t$, $i \neq k$ we have $d_{t+1}(i) = d_t(i)$, while $d_{t+1}(k) = d_t(k) + 1$. Also, $R_{t+1}(i) = R_t(i)$ for $i \neq k$, while $R_{t+1}(k) = R_t(k) - 1$. Thus

$$\begin{aligned} \prod_{i \in V(S), i \leq t} [d_{t+1}(i)]_{R_{t+1}(i)} &= [d_t(k) + 1]_{R_t(k)-1} \prod_{i \in V(S), i \leq t, i \neq k} [d_t(i)]_{R_t(i)} \\ &= \frac{1}{d_t(k)} \prod_{i \in V(S), i \leq t} [d_t(i)]_{R_t(i)}. \end{aligned}$$

Thus, in this case, if $f_{t+1} = k$ then we have $Y = X_t/d_t(k)$. Since this event has probability $d_t(k)/(2t-1)$, we have $\mathbb{E}(Y \mid T_t) = X_t/(2t-1)$. Taking the expectation of both sides gives $\mathbb{E}(Y) = \mathbb{E}(X_t)/(2t-1) = \lambda_t/(2t-1)$. Together with (1.23) this proves (1.21). □

Lemma 21 has the following immediate consequence. For a possible subgraph S of T_n, orient each edge $ij \in E(S)$ with $i < j$ from j to i. As in section 1.12, we write $V^+(S)$ for the set vertices of S from which edges leave, and $V^-(S)$ for those vertices at which edges arrive. (These sets are not in general disjoint.) For $i \in V^-(S)$ let $d_S^{in}(i)$ be the in-degree of i in S (so $d_S^{in}(i) = R_i(i)$), let $d_S^{out}(i)$ be the out-degree, and $d_S(i)$ the total degree of i in S.

Corollary 22. *Let S be a possible subgraph of T_n. With the notation above, the probability p_S that $S \subset T_n$ satisfies*

$$p_S = \prod_{i \in V^-(S)} d_S^{in}(i)! \prod_{i \in V^+(S)} \frac{1}{2i-3} \prod_{t \notin V^+(S)} \left(1 + \frac{C_S(t)}{2t-3}\right). \tag{1.24}$$

Furthermore,

$$p_S = \prod_{i \in V^-(S)} d_S^{in}(i)! \prod_{ij \in E(S)} \frac{1}{2\sqrt{ij}} \exp\left(O\left(\sum_{i \in V(S)} C_S(i)^2/i \right) \right). \tag{1.25}$$

Proof. The first statement follows immediately from Lemma 21; replace t by $t-1$ in (1.21) and (1.22), and write $p_S = \lambda_n = \lambda_n/\lambda_0$ as the product of the factors appearing in these equations.

The second statement follows by simple approximations: for all $x \geq 0$ we have $\log(1+x) = x + O(x^2)$. Thus for $c \geq 0$,

$$\log\left(\prod_{t=i+1}^{j-1} \left(1 + \frac{c}{2t-3}\right) \right) = \sum_{t=i+1}^{j-1} \left(\frac{c}{2t} + O(c^2/t^2) \right) = \frac{c}{2} \log\left(\frac{j}{i}\right) + O(c^2/i).$$

Writing the vertices of S in order as $v_1, \ldots, v_l$, let $c_k = C_S(v_k + 1)$ be the number of edges in S from $\{v_{k+1}, \ldots, v_l\}$ to $\{v_1, \ldots, v_k\}$. Then for $v_k < t < v_{k+1}$ we have $C_S(t) = c_k$, so

$$p_S = \prod_{i \in V^-(S)} d_S^{in}(i)! \prod_{i \in V^+(S)} \frac{1}{2i} \prod_{k=1}^{l-1} (v_{k+1}/v_k)^{c_k/2} \exp\left(O\left(\sum_{i \in V(S)} C_S(i)^2/i \right) \right).$$

(One can easily check that the error from replacing $2i-3$ by $2i$ in the second product is absorbed by the final error term.) Now the final exponent of v_k is just

$$-d_S^{out}(v_k) + c_{k-1}/2 - c_k/2.$$

Since $c_k = c_{k-1} - d_S^{out}(v_k) + d_S^{in}(v_k)$, this is just $-d_S(v_k)/2$. Finally, there is one factor of two in the denominator for each edge, so (1.25) follows. □

Essentially this result, but stated for the related model $G_1^{(n)}$, was used in section 1.12 to find the clustering coefficient of $G_m^{(n)}$. We finish this section with a direct application of Corollary 22 to T_n from [17]. We write $E_k = E_k(n)$ for the expected number of (shortest) paths in T_n of length k, so $\sum_{k=1}^{\infty} E_k(n) = \binom{n}{2}$.

Theorem 23. *Suppose that $k = k(n)$ satisfies $k/\log n \to \alpha$, where $0 < \alpha < e$. Then*

$$E_k = \Theta(n^{1+\alpha \log(e/\alpha)}/\sqrt{\log n}), \tag{1.26}$$

as $n \to \infty$. Furthermore, if $k = \log n + x\sqrt{\log n}$ where $x = x(n) = o(\log n)$, then

$$E_k \sim \frac{n^2}{2} \frac{1}{\sqrt{2\pi \log n}} e^{-x^2/2} \tag{1.27}$$

as $n \to \infty$.

Note that the second statement says that the distribution of path lengths is asymptotically normal with mean and variance $\log n$.

1.16 Conclusion

Most but not all of the rigorous results concerning models of large-scale real-world networks we have reviewed confirm the computer experiments and heuristic calculations performed on these models. In many cases, the results are surprisingly difficult to prove, and need techniques not used in the theory of classical random graphs. However, much remains to be done.

There is a great need for models of real-life networks that incorporate many of the important features of these systems, but can still be analyzed rigorously. The models defined and analyzed in this article are too simple for many applications, but there is also a danger in constructing models which take into account too many features of the real-life networks. Beyond a certain point, a complicated model hardly does more than describe the particular network that inspired it.

If the right balance can be struck, well constructed models and their careful analysis should give a sound understanding of growing networks that can be used to answer practical questions about their current state, as well as to predict their future development.

Acknowledgements

This research is supported by NSF grant DSM 9971788 and DARPA grant F33615-01-C-1900.

References

[1] W. Aiello, F. Chung and L. Lu, A random graph model for power law graphs, *Experiment. Math.* **10** (2001), 53–66.

[2] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74** (2002), 47–97.

[3] R. Albert, H. Jeong and A.-L. Barabási, Diameter of the world-wide web, *Nature* **401** (1999), 130–131.

[4] Austin, T.L., Fagen, R.E., Penney, W.F., and Riordan, J., The number of components of random linear graphs, *Ann. Math. Statist.* **30** (1959), 747–754.

[5] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, *Science* **286** (1999), 509–512.

[6] A.-L. Barabási, R. Albert and H. Jeong, Mean-field theory for scale-free random networks, *Physica A* **272** (1999), 173-187.

[7] A.-L. Barabási, R. Albert and H. Jeong, Scale-free characteristics of random networks: the topology of the world-wide web, *Physica A* **281** (2000), 69–77.

[8] F. Bergeron, P. Flajolet and B. Salvy, Varieties of increasing trees, in Proc. 17th Colloq. on Trees in Algebra and Programming held in Rennes, *Lecture Notes in Computer Science* **581**, Springer Verlag, Berlin, 1992, pp 24–48.

[9] Bollobás, B., The diameter of random graphs, *Trans. Amer. Math. Soc.* **267** (1981), 41–52.

[10] Bollobás, B., The evolution of random graphs. *Trans. Amer. Math. Soc.* **286** (1984), 257–274.

[11] B. Bollobás, *Random Graphs, Second Edition,* Cambridge studies in advanced mathematics, vol. 73, Cambridge University Press, Cambridge, 2001, xviii + 498 pp.

[12] Bollobás, B., Borgs, C., Chayes, T., and Riordan, O.M., Directed scale-free graphs, preprint.

[13] B. Bollobás and F.R.K. Chung, The diameter of a cycle plus a random matching, *SIAM J. Discrete Math.* **1** (1988), 328–333.

[14] Bollobás, B. and Fernandez de la Vega, W., The diameter of random regular graphs. *Combinatorica* **2** (1982), 125–134.

[15] B. Bollobás and O.M. Riordan, Linearized chord diagrams and an upper bound for Vassiliev invariants, *J. Knot Theory Ramifications* **9** (2000), 847–853.

[16] Bollobás, B. and Riordan, O.M., The diameter of a scale-free random graph, to appear in *Combinatorica*. (Preprint available from http://www.dpmms.cam.ac.uk/∼omr10/.)

[17] Bollobás, B. and Riordan, O.M., On shortest paths and load-scaling in scale-free trees, preprint.

[18] Bollobás, B. and Riordan, O.M., Robustness and vulnerability of scale-free random graph, in preparation.

[19] B. Bollobás, O. Riordan, J. Spencer and G. Tusnády, The degree sequence of a scale-free random graph process, *Random Structures and Algorithms* **18** (2001), 279–290.

[20] Buckley, P.G. and Osthus, D., Popularity based random graph models leading to a scale-free degree sequence, preprint available from http://www.informatik.hu-berlin.de/∼osthus/.

[21] Burtin, Ju.D., Asymptotic estimates of the diameter and the independence and domination numbers of a random graph, *Dokl. Akad. Nauk SSSR* **209** (1973), 765–768, translated in *Soviet Math. Dokl.* **14** (1973), 497–501.

[22] Burtin, Ju.D., Extremal metric characteristics of a random graph. I, *Teor. Verojatnost. i Primenen.* **19** (1974), 740–754.

[23] Cohen, R., and Havlin, S., Ultra small world in scale-free networks, http://www.arxiv.org/abs/cond-mat.

[24] C. Cooper and A. Frieze, A general model of web graphs, preprint.

[25] E. Drinea, M. Enachescu and M. Mitzenmacher, Variations on random graph models for the web, technical report, Harvard University, Department of Computer Science, 2001.

[26] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Structure of growing networks with preferential linking *Phys. Rev. Lett.* **85** (2000), 4633.

[27] S.N. Dorogovtsev and J.F.F. Mendes, Evolution of networks, *Adv. Phys.* **51** (2002), 1079.

[28] Erdős, P., and Rényi, A., On random graphs I., *Publicationes Mathematicae Debrecen* **5** (1959), 290–297.

[29] Erdős, P., and Rényi, A., On the evolution of random graphs, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** (1960), 17–61.

[30] M. Faloutsos, P. Faloutsos and C. Faloutsos, On power-law relationships of the internet topology, SIGCOMM 1999, *Comput. Commun. Rev.* **29** (1999), 251.

[31] I.J. Farkas, I. Derényi, A.-L. Barabási and T. Vicsek, Spectra of "real-world" graphs: Beyond the semicircle law, *Phys. Rev. E* **64** (2001), 026704.

[32] Gilbert, E.N., Enumeration of labelled graphs, *Canad. J. Math.* **8** (1956), 405–411.
[33] Gilbert, E.N., Random graphs, *Annals of Mathematical Statistics* **30** (1959), 1141–1144.
[34] Gilbert, N., A simulation of the structure of academic science, *Sociological Research Online* **2** (1997).
[35] K.-I. Goh, B. Kahng and D. Kim, Universal behavior of load distribution in scale-free networks, *Physical Review Letters* **87** (2001), 278701.
[36] P. Holme and B.J. Kim, Growing scale-free networks with tunable clustering, *Phys. Rev. E* **65** (2002), 026107.
[37] Janson, S., Knuth, D., E., Łuczak, T., and Pittel, B., The birth of the giant component, *Random Structures and Algorithms* **3** (1993), 233–358.
[38] Janson, S., Łuczak, T., and Ruciński, A. (2000). *Random Graphs*. John Wiley and Sons, New York, xi+ 333pp.
[39] K. Klemm and V.M. Eguiluz, *Phys. Rev. E* **65** (2002), 057102.
[40] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal, Stochastic models for the web graph, FOCS 2000.
[41] Lotka, A.J., The frequency distribution of scientific productivity, *J. of the Washington Acad. of Sci.* **16** (1926), 317.
[42] Łuczak, T., Component behavior near the critical point of the random graph process, *Random Structures and Algorithms* **1** (1990), 287–310.
[43] H.M. Mahmoud, Distances in random plane-oriented recursive trees, *J. Comput. and Applied Math.* **41** (1992), 237–245.
[44] H.M. Mahmoud and R.T. Smythe, A survey of recursive trees, *Th. of Probability and Math. Statistics* **51** (1995), 1–27.
[45] H.M. Mahmoud, R.T. Smythe and J. Szymański, On the structure of random plane-oriented recursive trees and their branches, *Random Struct. Alg.* **4** (1993), 151–176.
[46] McKay, B.D., and Wormald, N.C., The degree sequence of a random graph. I. The models, *Random Structures and Algorithms* **11** (1997), 97–117.
[47] Milgram, S., The small world phenomenon, *Psychol. Today* **2** (1967), 60–67.
[48] M.E.J. Newman, S.H. Strogatz and D.J. Watts, Random graphs with arbitrary degree distribution and their applications, *Physical Review E* **64** (2001), 026118.
[49] B. Pittel, Note on the heights of random recursive trees and random m-ary search trees, *Random Struct. Alg.* **5** (1994), 337–347.
[50] Riddell, R.J.Jr., and Uhlenbeck, G.E., On the theory of virial development of the equation of state of monoatomic gases, *J. Chem. Phys.* **21** (1953), 2056–2064.
[51] Simon, H., On a class of skew distribution functions, *Biometrika* **42** (1955), 425–440.
[52] A. Stoimenow, Enumeration of chord diagrams and an upper bound for Vassiliev invariants, *J. Knot Theory Ramifications* **7** (1998), 93–114.
[53] G. Szabó, M. Alava and J. Kertész, Shortest paths and load scaling in scale-free trees, *Phys. Rev. E* **66** (2002), 026101.
[54] J. Szymański, On a nonuniform random recursive tree, *A.. Discrete Math.* **33** (1987), 297–306.
[55] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* **393** (1998), 440–442.
[56] Zipf, G.K., Human behavior and the principle of least effort, New York: Hafner, 1949.

2 Random graphs as models of networks

Mark E. J. Newman

Abstract

The random graph of Erdős and Rényi is one of the oldest and best studied models of a network, and possesses the considerable advantage of being exactly solvable for many of its average properties. However, as a model of real-world networks such as the Internet, social networks or biological networks it leaves a lot to be desired. In particular, it differs from real networks in two crucial ways: it lacks network clustering or transitivity, and it has an unrealistic Poissonian degree distribution. In this paper we review some recent work on generalizations of the random graph aimed at correcting these shortcomings. We describe generalized random graph models of both directed and undirected networks that incorporate arbitrary non-Poisson degree distributions, and extensions of these models that incorporate clustering too. We also describe two recent applications of random graph models to the problems of network robustness and of epidemics spreading on contact networks.

2.1 Introduction

In a series of seminal papers in the 1950s and 1960s, Paul Erdős and Alfréd Rényi proposed and studied one of the earliest theoretical models of a network, the **random graph** (Erdős and Rényi, 1959, 1960, 1961). This minimal model consists of n nodes or **vertices**, joined by links or **edges** which are placed between pairs of vertices chosen uniformly at random. Erdős and Rényi gave a number of versions of their model. The most commonly studied is the one denoted $G_{n,p}$, in which each possible edge between two vertices is present with independent probability p, and absent with probability $1 - p$. Technically, in fact, $G_{n,p}$ is the *ensemble* of graphs of n vertices in which each graph appears with the probability appropriate to its number of edges.[1]

Often one wishes to express properties of $G_{n,p}$ not in terms of p but in terms of the average degree z of a vertex. (The **degree** of a vertex is the number of edges connected to that vertex.) The average number of edges on the graph as a whole is $\frac{1}{2}n(n-1)p$, and the average number of *ends* of edges is twice this, since each edge has two ends. So the average degree of a vertex is

$$z = \frac{n(n-1)p}{n} = (n-1)p \simeq np, \tag{2.1}$$

[1] For a graph with n vertices and m edges this probability is $p^m(1-p)^{M-m}$, where $M = \frac{1}{2}n(n-1)$.

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

where the last approximate equality is good for large n. Thus, once we know n, any property that can be expressed in terms of p can also be expressed in terms of z.

The Erdős–Rényi random graph has a number of desirable properties as a model of a network. In particular it is found that many of its ensemble average properties can be calculated exactly in the limit of large n (Bollobás, 1985; Janson *et al.*, 1999). For example, one interesting feature, which was demonstrated in the original papers by Erdős and Rényi, is that the model shows a phase transition[2] with increasing z at which a **giant component** forms. A **component** is a subset of vertices in the graph each of which is reachable from the others by some path through the network. For small values of z, when there are few edges in the graph, it is not surprising to find that most vertices are disconnected from one another, and components are small, having an average size that remains constant as the graph becomes large. However, there is a critical value of z above which the one largest component in the graph contains a finite fraction S of the total number of vertices, i.e., its size nS scales linearly with the size of the whole graph. This largest component is the giant component. In general there will be other components in addition to the giant component, but these are still small, having an average size that remains constant as the graph grows larger. The phase transition at which the giant component forms occurs precisely at $z = 1$. If we regard the fraction S of the graph occupied by the largest component as an order parameter, then the transition falls in the same universality class as the mean-field percolation transition (Stauffer and Aharony, 1992).

The formation of a giant component in the random graph is reminiscent of the behaviour of many real-world networks. One can imagine loose-knit networks for which there are so few edges that, presumably, the network has no giant component, and all vertices are connected to only a few others. The social network in which pairs of people are connected if they have had a conversation within the last 60 seconds, for example, is probably so sparse that it has no giant component. The network in which people are connected if they have *ever* had a conversation, on the other hand, is very densely connected and certainly has a giant component.

However, the random graph differs from real-world networks in some fundamental ways also. Two differences in particular have been noted in the recent literature (Strogatz, 2001; Albert and Barabási, 2002). First, as pointed out by Watts and Strogatz (1998; Watts 1999) real-world networks show strong **clustering** or **network transitivity**, where Erdős and Rényi's model does not. A network is said to show clustering if the probability of two vertices being connected by an edge is higher when the vertices in question have a common neighbour. That is, there is another vertex in the network to which they are both attached. Watts and Strogatz measured this clustering by defining a **clustering coefficient** C, which is the average probability that two neighbours of a given vertex are also neighbours of one another. In many real-world networks the clustering coefficient is found to have a high value, anywhere from a few percent to 50 percent or even more. In the random graph of Erdős and Rényi on the other hand, the probabilities of vertex pairs being connected by edges are by definition independent, so that there is no greater probability of two vertices being connected if they have a mutual neighbour than if they do not. This means that the clustering coefficient for a random graph is simply $C = p$, or equivalently $C \simeq z/n$. In Table 2.1 we compare clustering coefficients for a number of real-world networks with their values on a random graph with the same number of vertices and edges. The graphs listed in the table are:

[2] Erdős and Rényi didn't call it that, but that's what it is.

network	n	z	clustering coefficient C measured	clustering coefficient C random graph
Internet (autonomous systems)[a]	6 374	3.8	0.24	0.00060
World-Wide Web (sites)[b]	153 127	35.2	0.11	0.00023
power grid[c]	4 941	2.7	0.080	0.00054
biology collaborations[d]	1 520 251	15.5	0.081	0.000010
mathematics collaborations[e]	253 339	3.9	0.15	0.000015
film actor collaborations[f]	449 913	113.4	0.20	0.00025
company directors[f]	7 673	14.4	0.59	0.0019
word co-occurrence[g]	460 902	70.1	0.44	0.00015
neural network[c]	282	14.0	0.28	0.049
metabolic network[h]	315	28.3	0.59	0.090
food web[i]	134	8.7	0.22	0.065

Table 2.1: Number of vertices n, mean degree z, and clustering coefficient C for a number of different networks. Numbers are taken from [a]Pastor-Satorras *et al.* (2001), [b]Adamic (1999), [c]Watts and Strogatz (1998), [d]Newman (2001b), [e]Newman (2001d), [f]Newman *et al.* (2001), [g]i Cancho and Solé (2001), [h]Montoya and Solé (2001), [i]Fell and Wagner (2000).

- Internet: a graph of the fibre optic connections that comprise the Internet, at the level of so-called "autonomous systems." An autonomous system is a group of computers within which data flow is handled autonomously, while data flow between groups is conveyed over the public Internet. Examples of autonomous systems might be the computers at a company, a university, or an Internet service provider.
- World-Wide Web: a graph of sites on the World-Wide Web in which edges represent "hyperlinks" connecting one site to another. A site in this case means a collection of pages residing on a server with a given name. Although hyperlinks are directional, their direction has been ignored in this calculation of the clustering coefficient.
- Power grid: a graph of the Western States electricity transmission grid in the United States. Vertices represent stations and substations; edges represent transmission lines.
- Biology collaborations: a graph of collaborations between researchers working in biology and medicine. A collaboration between two scientists is defined in this case as coauthorship of a paper that was catalogued in the Medline bibliographic database between 1995 and 1999 inclusive.
- Mathematics collaborations: a similar collaboration graph for mathematicians, derived from the archives of *Mathematical Reviews*.
- Film actor collaborations: a graph of collaborations between film actors, where a collaboration means that the two actors in question have appeared in a film together. The data are from the Internet Movie Database.
- Company directors: a collaboration graph of the directors of companies in the Fortune 1000 for 1999. (The Fortune 1000 is the 1000 US companies with the highest revenues during the year in question.) Collaboration in this case means that two directors served on the board of a Fortune 1000 company together.

- Word co-occurrences: a graph in which the vertices represent words in the English language, and an edge signifies that the vertices it connects frequently occur in adjacent positions in sentences.
- Neural network: a graph of the neural network of the worm *C. Elegans.*
- Metabolic network: a graph of interactions forming a part of the energy generation and small building block synthesis metabolism of the bacterium *E. Coli.* Vertices represent substrates and products, and edges represent interactions.
- Food web: the food web of predator–prey interactions between species in Ythan Estuary, a marine estuary near Aberdeen, Scotland. Like the links in the World-Wide Web graph, the directed nature of the interactions in this food web have been neglected for the purposes of calculating the clustering coefficient.

As the table shows, the agreement between the clustering coefficients in the real networks and in the corresponding random graphs is not good. The real and theoretical figures differ by as much as four orders of magnitude in some cases. Clearly, the random graph does a poor job of capturing this particular property of networks.

A second way in which random graphs differ from their real-world counterparts is in their degree distributions, a point which has been emphasized particularly in the work of Albert, Barabási, and collaborators (Albert *et al.*, 1999; Barabási and Albert, 1999). The probability p_k that a vertex in an Erdős–Rényi random graph has degree exactly k is given by the binomial distribution:

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}. \tag{2.2}$$

In the limit where $n \gg kz$, this becomes

$$p_k = \frac{z^k \epsilon^{-z}}{k!}, \tag{2.3}$$

which is the well-known Poisson distribution. Both binomial and Poisson distributions are strongly peaked about the mean z, and have a large-k tail that decays rapidly as $1/k!$. We can compare these predictions to the degree distributions of real networks by constructing histograms of the degrees of vertices in the real networks. We show some examples, taken from the networks described above, in Fig. 2.1. As the figure shows, in most cases the degree distribution of the real network is very different from the Poisson distribution. Many of the networks, including Internet and World-Wide Web graphs, appear to have power-law degree distributions (Albert *et al.*, 1999; Faloutsos *et al.*, 1999; Broder *et al.*, 2000), which means that a small but non-negligible fraction of the vertices in these networks have very large degree. This behaviour is quite unlike the rapidly decaying Poisson degree distribution, and can have profound effects on the behaviour of the network, as we will see later in this chapter. Other networks, particularly the collaboration graphs, appear to have power-law degree distributions with an exponential cutoff at high degree (Amaral *et al.*, 2000; Newman, 2001a,b), while others still, such as the graph of company directors, seem to have degree distributions with a purely exponential tail (Newman *et al.*, 2001). The power grid of Table 2.1 is another example of a network that has an exponential degree distribution (Amaral *et al.*, 2000).

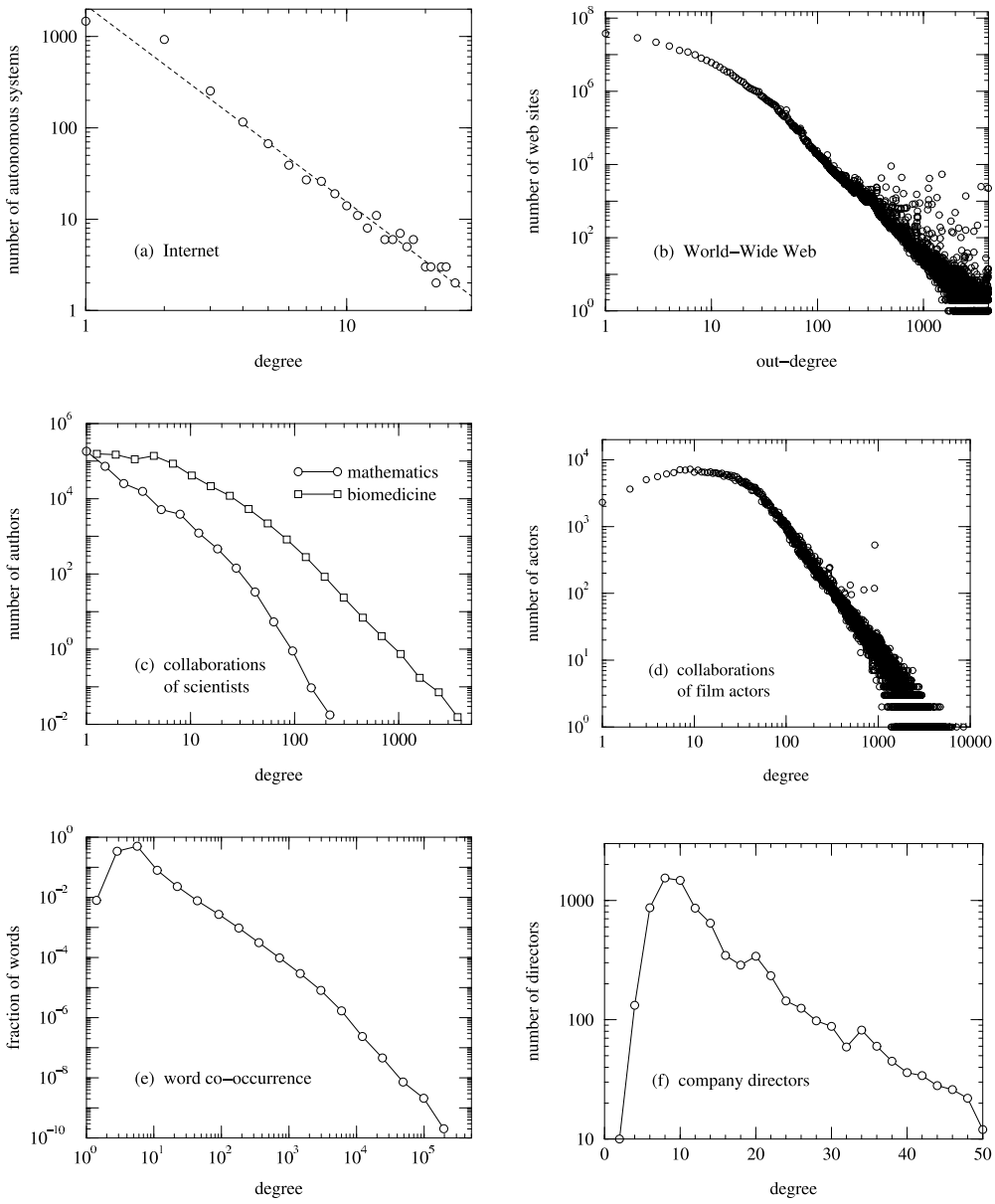


Figure 2.1: Measured degree distributions for a number of different networks. (a) Physical connections between autonomous systems on the Internet, *circa* 1997 (Faloutsos *et al.*, 1999). (b) A 200 million page subset of the World-Wide Web, *circa* 1999 (Broder *et al.*, 2000). The figure shows the out-degree of pages, i.e., numbers of links pointing from those pages to other pages. (c) Collaborations between biomedical scientists and between mathematicians (Newman, 2001b,d). (d) Collaborations of film actors (Amaral *et al.*, 2000). (e) Co-occurrence of words in the English language (i Cancho and Solé, 2001). (f) Board membership of directors of Fortune 1000 companies for year 1999 (Newman *et al.*, 2001).

In this chapter we show how to generalize the Erdős–Rényi random graph to mimic the clustering and degree properties of real-world networks. In fact, most of the chapter is devoted to extensions that correct the degree distribution, for which an elegant body of theory has been developed in the last few years. However, towards the end of the chapter we also consider ways in which clustering can be introduced into random graphs. Work on this latter problem is significantly less far advanced than work on degree distributions, and we have at present only a few preliminary results. Whether these results can be extended, and how, are open questions.

2.2 Random graphs with specified degree distributions

It is relatively straightforward to generate random graphs that have non-Poisson degree distributions. The method for doing this has been discussed a number of times in the literature, but appears to have been put forward first by Bender and Canfield (1978). The trick is to restrict oneself to a specific degree *sequence*, i.e., to a specified set $\{k_i\}$ of the degrees of the vertices $i = 1 \ldots n$. Typically this set will be chosen in such a way that the fraction of vertices having degree k will tend to the desired degree distribution p_k as n becomes large. For practical purposes however, such as numerical simulation, it is almost always adequate simply to draw a degree sequence $\{k_i\}$ from the distribution p_k directly.

Once one has one's degree sequence, the method for generating the graph is as follows: one gives each vertex i a number k_i of "stubs"—ends of edges emerging from the vertex—and then one chooses pairs of these stubs uniformly at random and joins them together to make complete edges. When all stubs have been used up, the resulting graph is a random member of the ensemble of graphs with the desired degree sequence.[3] Note that, because of the $k_i!$ possible permutations of the stubs emerging from the ith vertex, there are $\prod_i k_i!$ different ways of generating each graph in the ensemble. However, this factor is constant so long as the degree sequence $\{k_i\}$ is held fixed, so it does not prevent the method from sampling the ensemble correctly. This is the reason why we restrict ourselves to a fixed degree sequence—merely fixing the degree distribution is not adequate to ensure that the method described here generates graphs uniformly at random from the desired ensemble.

The method of Bender and Canfield does not allow us to specify a clustering coefficient for our graph. (The clustering coefficient had not been invented yet when Bender and Canfield were writing in 1978.) Indeed the fact that the clustering coefficient is not specified is one of the crucial properties of these graphs that makes it possible, as we will show, to solve exactly for many of their properties in the limit of large graph size. As an example of why this is important, consider the following simple calculation. The mean number of neighbours of a randomly chosen vertex A in a graph with degree distribution p_k is $z = \langle k \rangle = \sum_k k p_k$. Suppose however that we want to know the mean number of second neighbours of vertex A, i.e., the mean number of vertices two steps away from A in the graph. In a network with clustering, many of the second neighbours of a vertex are also first neighbours—the friend of my friend is also my friend—and we would have to allow for this effect to order avoid

[3] The only small catch to this algorithm is that the total number of stubs must be even if we are not to have one stub left over at the end of the pairing process. Thus we should restrict ourselves to degree sequences for which $\sum_i k_i$ is even.

overcounting the number of second neighbours. In our random graphs however, no allowances need be made. The probability that one of the second neighbours of A is also a first neighbour goes as n^{-1} in the random graph, regardless of degree distribution, and hence can be ignored in the limit of large n.

There is another effect, however, that we certainly must take into account if we wish to compute correctly the number of second neighbours: the degree distribution of the first neighbour of a vertex is not the same as the degree distribution of vertices on the graph as a whole. Because a high-degree vertex has more edges connected to it, there is a higher chance that any given edge on the graph will be connected to it, in precise proportion to the vertex's degree. Thus the probability distribution of the degree of the vertex to which an edge leads is proportional to kp_k and not just p_k (Feld, 1991; Molloy and Reed, 1995; Newman, 2001d). This distinction is absolutely crucial to all the further developments of this paper, and the reader will find it worthwhile to make sure that he or she is comfortable with it before continuing.

In fact, we are interested here not in the complete degree of the vertex reached by following an edge from A, but in the number of edges emerging from such a vertex other than the one we arrived along, since the latter edge only leads back to vertex A and so does not contribute to the number of second neighbours of A. This number is one less than the total degree of the vertex and its correctly normalized distribution is therefore $q_{k-1} = kp_k/\sum_j jp_j$, or equivalently

$$q_k = \frac{(k+1)p_{k+1}}{\sum_j jp_j}. \tag{2.4}$$

The average degree of such a vertex is then

$$\sum_{k=0}^{\infty} kq_k = \frac{\sum_{k=0}^{\infty} k(k+1)p_{k+1}}{\sum_j jp_j} = \frac{\sum_{k=0}^{\infty}(k-1)kp_k}{\sum_j jp_j} = \frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}. \tag{2.5}$$

This is the average number of vertices two steps away from our vertex A via a particular one of its neighbours. Multiplying this by the mean degree of A, which is just $z = \langle k\rangle$, we thus find that the mean number of second neighbours of a vertex is

$$z_2 = \langle k^2\rangle - \langle k\rangle. \tag{2.6}$$

If we evaluate this expression using the Poisson degree distribution, Eq. (2.3), then we get $z_2 = \langle k\rangle^2$—the mean number of second neighbours of a vertex in an Erdős–Rényi random graph is just the square of the mean number of first neighbours. This is a special case however. For most degree distributions Eq. (2.6) will be dominated by the term $\langle k^2\rangle$, so the number of second neighbours is roughly the mean square degree, rather than the square of the mean. For broad distributions such as those seen in Fig. 2.1, these two quantities can be very different (Newman, 2001d).

We can extend this calculation to further neighbours also. The average number of edges leading from each second neighbour, other than the one we arrived along, is also given by (2.5), and indeed this is true at any distance m away from vertex A. Thus the average number of neighbours at distance m is

$$z_m = \frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle} z_{m-1} = \frac{z_2}{z_1} z_{m-1}, \tag{2.7}$$

where $z_1 \equiv z = \langle k \rangle$ and z_2 is given by Eq. (2.6). Iterating this equation we then determine that

$$z_m = \left[\frac{z_2}{z_1}\right]^{m-1} z_1. \tag{2.8}$$

Depending on whether z_2 is greater than z_1 or not, this expression will either diverge or converge exponentially as m becomes large, so that the average total number of neighbours of vertex A at all distances is finite if $z_2 < z_1$ or infinite if $z_2 > z_1$ (in the limit of infinite n).[4] If this number is finite, then clearly there can be no giant component in the graph. Conversely, if it is infinite, then there must be a giant component. Thus the graph shows a phase transition similar to that of the Erdős–Rényi graph precisely at the point where $z_2 = z_1$. Making use of Eq. (2.6) and rearranging, we find that this condition is also equivalent to $\langle k^2 \rangle - 2\langle k \rangle = 0$, or, as it is more commonly written,

$$\sum_{k=0}^{\infty} k(k-2)p_k = 0. \tag{2.9}$$

This condition for the position of the phase transition in a random graph with arbitrary degree sequence was first given by Molloy and Reed (1995).

An interesting feature of Eq. (2.9) is that, because of the factor $k(k-2)$, vertices of degree zero and degree two contribute nothing to the sum, and therefore the number of such vertices does not affect the position of the phase transition or the existence of the giant component. It is easy to see why this should be the case for vertices of degree zero; obviously one can remove (or add) degree-zero vertices without changing the fact of whether a giant component does or does not exist in a graph. But why vertices of degree two? This has a simple explanation also: removing vertices of degree two does not change the topological structure of a graph because all such vertices fall in the middle of edges between other pairs of vertices. We can therefore remove (or add) any number of such vertices without affecting the existence of the giant component.

Another quantity of interest in many networks is the typical distance through the network between pairs of vertices (Milgram, 1967; Travers and Milgram, 1969; Pool and Kochen, 1978; Watts and Strogatz, 1998; Amaral *et al.*, 2000). We can use Eq. (2.8) to make a calculation of this quantity for our random graph as follows. If we are "below" the phase transition of Eq. (2.9), in the regime where there is no giant component, then most pairs of vertices will not be connected to one another at all, so vertex–vertex distance has little meaning. Well above the transition on the other hand, where there is a giant component, all vertices in this giant component are connected by some path to all others. Eq. (2.8) tells us the average number of vertices a distance m away from a given vertex A in the giant component. When the total number of vertices within distance m is equal to the size n of the whole graph, m is equal to the so-called "radius" r of the network around vertex A. Indeed, since $z_2/z_1 \gg 1$ well above the transition, the number of vertices at distance m grows quickly with m in this

[4] The case of $z_1 = z_2$ is deliberately missed out here, since it is non-trivial to show how the graph behaves exactly at this transition point (Bollobás, 1985). For our current practical purposes however, this matters little, since the chances of any real graph being precisely at the transition point are negligible.

regime (see Eq. (2.8) again), which means that most of the vertices in the network will be far from A, around distance r, and r is thus also approximately equal to the average vertex–vertex distance ℓ. Well above the transition therefore, ℓ is given approximately by $z_\ell \simeq n$, or

$$\ell = \frac{\log(n/z_1)}{\log(z_2/z_1)} + 1. \tag{2.10}$$

For the special case of the Erdős–Rényi random graph, for which $z_1 = z$ and $z_2 = z^2$ as noted above, this expression reduces to the well-known standard formula for this case: $\ell = \log n / \log z$ (Bollobás, 1985).

The important point to notice about Eq. (2.10) is that the vertex–vertex distance increases logarithmically with the graph size n, i.e., it grows rather slowly.[5] Even for very large networks we expect the typical distance through the network from one vertex to another to be quite small. In social networks this effect is known as the **small-world effect**,[6] and was famously observed by the experimental psychologist Stanley Milgram in the letter-passing experiments he conducted in the 1960s (Milgram, 1967; Travers and Milgram, 1969; Kleinfeld, 2000). More recently it has been observed also in many other networks including non-social networks (Watts and Strogatz, 1998; Amaral *et al.*, 2000). This should come as no great surprise to us however. On the contrary, it would be surprising if most networks did *not* show the small-world effect. If we define the **diameter** d of a graph to be the *maximum* distance between any two connected vertices in the graph, then it can be proven rigorously that the fraction of all possible graphs for which $d > c \log n$ for some constant c tends to zero as n becomes large (Bollobás, 1985). And clearly if the diameter increases as $\log n$ or slower, then so also must the average vertex–vertex distance. Thus our chances of finding a network that does not show the small-world effect are very small for large n.

As a test of Eq. (2.10), Fig. 2.2 compares our predictions of average distance ℓ with direct measurements for fourteen different scientific collaboration networks, including the biology and mathematics networks of Table 2.1. In this figure, each network is represented by a single point, whose position along the horizontal axis corresponds to the theoretically predicted value of ℓ and along the vertical axis the measured value. If Eq. (2.10) were exactly correct, all the points in the figure would fall on the dotted diagonal line. Since we know that the equation is only approximate, it comes as no surprise that the points do not fall perfectly along this line, but the results are encouraging nonetheless; in most cases the theoretical prediction is close to the correct result and the overall scaling of ℓ with $\log n$ is clear. If the theory were equally successful for networks of other types, it would provide a useful way of estimating average vertex–vertex separation. Since z_1 and z_2 are local quantities that can be calculated at least approximately from measurements on only a small portion of the network, it would in many cases be considerably simpler and more practical to apply Eq. (2.10) than to measure ℓ directly.

[5] Krzywicki (2001) points out that this is true only for components such as the giant component that contain loops. For tree-like components that contain no loops the mean vertex–vertex distance typically scales as a power of n. Since the giant components of neither our models nor our real-world networks are tree-like, however, this is not a problem.

[6] Some authors, notably Watts and Strogatz (1998), have used the expression "small-world network" to refer to a network that simultaneously shows both the small-world effect and high clustering. To prevent confusion however we will avoid this usage here.

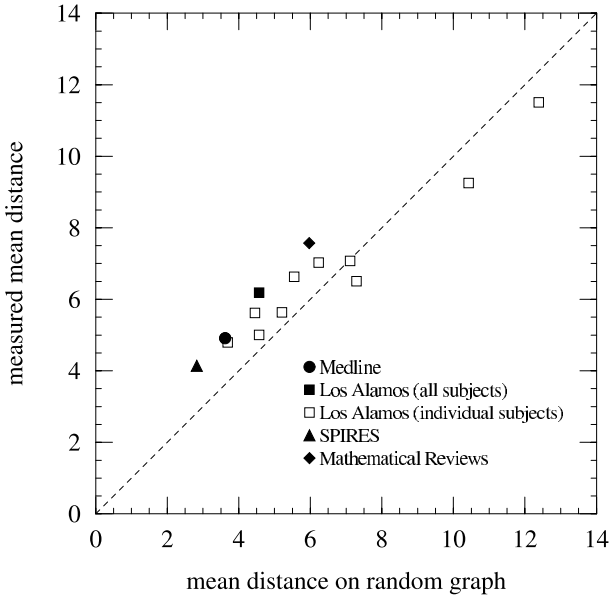


Figure 2.2: Comparison of mean vertex–vertex distance measured in fourteen collaboration networks against our theoretical predictions of the same quantities from Eq. (2.10). The networks are constructed using bibliographic data for papers in biology and medicine (Medline), physics (Los Alamos E-print Archive), high-energy physics (SPIRES), and mathematics (Mathematical Reviews). If empirical results and theory agreed perfectly, the points would fall on the dotted diagonal line. After Newman (2001c).

Although our random graph model does not allow us to fix the level of clustering in the network, we can still calculate an average clustering coefficient for the Bender–Canfield ensemble easily enough. Consider a particular vertex A again. The ith neighbour of A has k_i edges emerging from it other than the edge attached to A, and k_i is distributed according to the distribution q_k, Eq. (2.4). The probability that this vertex is connected to another neighbour j is $k_i k_j/(nz)$, where k_j is also distributed according to q_k, and average of this probability is precisely the clustering coefficient:

$$C = \frac{\langle k_i k_j \rangle}{nz} = \frac{1}{nz}\Bigl[\sum_k k q_k\Bigr]^2 = \frac{z}{n}\left[\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle^2}\right]^2 = \frac{z}{n}\left[c_v^2 + \frac{z-1}{z}\right]^2. \tag{2.11}$$

The quantity c_v is the so-called coefficient of variation of the degree distribution—the ratio of the standard deviation to the mean. Thus the clustering coefficient for the random graph with a non-Poisson degree distribution is equal to its value z/n for the Poisson-distributed case, times a function whose leading term goes as the fourth power of the coefficient of variation of the degree distribution. So the clustering coefficient still vanishes with increasing graph size, but may have a much larger leading coefficient, since c_v can be quite large, especially for degree distributions with long tails, such as those seen in Fig. 2.1.

Take for example the World-Wide Web. If one ignores the directed nature of links on the Web, then the resulting graph is measured to have quite a high clustering coefficient of

0.11 (Adamic, 1999), as shown in Table 2.1. The Erdős–Rényi random graph with the same n and z, by contrast, has a clustering coefficient of only 0.00023. However, if we use the degree distribution shown in Fig. 2.1a to calculate a mean degree and coefficient of variation for the Web, we get $z = 10.23$ and $c_v = 3.685$, which means that $(c_v^2 + (z-1)/z)^2 = 209.7$. Eq. (2.11) then tells us that the random graph with the correct degree distribution would actually have a clustering coefficient of $C = 0.00023 \times 209.7 = 0.048$. This is still about a factor of two away from the correct answer, but a lot closer to the mark than the original estimate, which was off by a factor of more than 400. Furthermore, the degree distribution used in this calculation was truncated at $k = 4096$. (The data were supplied to author in this form.) Without this truncation, the coefficient of variation would presumably be larger still. It seems possible therefore, that most, if not all, of the clustering seen in the Web can be accounted for merely as a result of the long-tailed degree distribution. Thus the fact that our random graph models do not explicitly include clustering is not necessarily a problem.

On the other hand, some of the other networks of Table 2.1 do show significantly higher clustering than would be predicted by Eq. (2.11). For these, our random graphs will be an imperfect model, although as we will see they still have much to contribute. Extension of our models to include clustering explicitly is discussed in Section 2.6.

It would be possible to continue the analysis of our random graph models using the simple methods of this section. However, this leads to a lot of tedious algebra which can be avoided by introducing an elegant tool, the probability generating function.

2.3 Probability generating functions

In this section we describe the use of probability generating functions to calculate the properties of random graphs. Our presentation closely follows that of Newman *et al.* (2001).

A **probability generating function** is an alternative representation of a probability distribution. Take the probability distribution p_k introduced in the previous section, for instance, which is the distribution of vertex degrees in a graph. The corresponding generating function is

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k. \tag{2.12}$$

It is clear that this function captures all of the information present in the original distribution p_k, since we can recover p_k from $G_0(x)$ by simple differentiation:

$$p_k = \frac{1}{k!} \frac{\delta^k G_0}{\delta x^k}\bigg|_{x=0}. \tag{2.13}$$

We say that the function G_0 "generates" the probability distribution p_k.

We can also define a generating function for the distribution q_k, Eq. (2.4), of other edges leaving the vertex we reach by following an edge in the graph:

$$G_1(x) = \sum_{k=0}^{\infty} q_k x^k = \frac{\sum_{k=0}^{\infty}(k+1)p_{k+1}x^k}{\sum_j j p_j} = \frac{\sum_{k=0}^{\infty} k p_k x^{k-1}}{\sum_j j p_j} = \frac{G_0'(x)}{z}, \tag{2.14}$$

where $G_0'(x)$ denotes the first derivative of $G_0(x)$ with respect to its argument. This generating function will be useful to us in following developments.

2.3.1 Properties of generating functions

Generating functions have some properties that will be of use in this chapter. First, if the distribution they generate is properly normalized then

$$G_0(1) = \sum_k p_k = 1. \tag{2.15}$$

Second, the mean of the distribution can be calculated directly by differentiation:

$$G_0'(1) = \sum_k k p_k = \langle k \rangle. \tag{2.16}$$

Indeed we can calculate any moment of the distribution by taking a suitable derivative. In general,

$$\langle k^n \rangle = \sum_k k^n p_k = \left[\left(x \frac{\delta}{\delta x} \right)^n G_0(x) \right]_{x=1}. \tag{2.17}$$

Third, and most important, if a generating function generates the probability distribution of some property k of an object, such as the degree of a vertex, then the sum of that property over n independent such objects is distributed according to the nth power of the generating function. Thus the sum of the degrees of n randomly chosen vertices on our graph has a distribution which is generated by the function $[G_0(x)]^n$. To see this, note that the coefficient of x^m in $[G_0(x)]^n$ has one term of the form $p_{k_1} p_{k_2} \dots p_{k_n}$ for every set $\{k_i\}$ of the degrees of the n vertices such that $\sum_i k_i = m$. But these terms are precisely the probabilities that the degrees sum to m in every possible way, and hence $[G_0(x)]^n$ is the correct generating function. This property is the reason why generating functions are useful in the study of random graphs. Most of the results of this chapter rely on it.

2.3.2 Examples

To make these ideas more concrete, let us consider some specific examples of generating functions. Suppose for instance that we are interested in the standard Erdős–Rényi random graph, with its Poisson degree distribution. Substituting Eq. (2.3) into (2.12), we get

$$G_0(x) = e^{-z} \sum_{k=0}^{\infty} \frac{z^k}{k!} x^k = e^{z(x-1)}. \tag{2.18}$$

This is the generating function for the Poisson distribution. The generating function $G_1(x)$ for vertices reached by following an edge is also easily found, from Eq. (2.14):

$$G_1(x) = \frac{G_0'(x)}{z} = e^{z(x-1)}. \tag{2.19}$$

Thus, for the case of the Poisson distribution we have $G_1(x) = G_0(x)$. This identity is the reason why the properties of the Erdős–Rényi random graph are particularly simple to solve analytically.[7]

As a second example, consider a graph with an exponential degree distribution:

$$p_k = (1 - \epsilon^{-1/\kappa})\epsilon^{-k/\kappa}, \tag{2.20}$$

where κ is a constant. The generating function for this distribution is

$$G_0(x) = (1 - \epsilon^{-1/\kappa}) \sum_{k=0}^{\infty} \epsilon^{-k/\kappa} x^k = \frac{1 - \epsilon^{-1/\kappa}}{1 - x\epsilon^{-1/\kappa}}, \tag{2.21}$$

and

$$G_1(x) = \left[\frac{1 - \epsilon^{-1/\kappa}}{1 - x\epsilon^{-1/\kappa}} \right]^2. \tag{2.22}$$

As a third example, consider a graph in which all vertices have degree 0, 1, 2, or 3 with probabilities $p_0 \ldots p_3$. Then the generating functions take the form of simple polynomials

$$G_0(x) = p_3 x^3 + p_2 x^2 + p_1 x + p_0, \tag{2.23}$$

$$G_1(x) = q_2 x^2 + q_1 x + q_0 = \frac{3p_3 x^2 + 2p_2 x + p_1}{3p_3 + 2p_2 + p_1}. \tag{2.24}$$

2.4 Properties of undirected graphs

We now apply our generating functions to the calculation of a variety of properties of undirected graphs. In Section 2.5 we extend the method to directed graphs as well.

2.4.1 Distribution of component sizes

The most basic property we will consider is the distribution of the sizes of connected components of vertices in the graph. Let us suppose for the moment that we are below the phase transition, in the regime in which there is no giant component. (We will consider the regime above the phase transition in a moment.) As discussed in Section 2.2, the calculations will depend crucially on the fact that our graphs do not have significant clustering. Instead, the clustering coefficient—the probability that two of your friends are also friends of one another—is given by Eq. (2.11), which tends to zero as $n \to \infty$. The probability of any two randomly chosen vertices i and j with degrees k_i and k_j being connected is the same regardless of where the vertices are. It is always equal to $k_i k_j/(nz)$, and hence also tends to zero as $n \to \infty$. This means that *any finite component of connected vertices has no closed loops in it,* and this is the crucial property that makes exact solutions possible. In physics jargon, all finite components are **tree-like**.

[7] This result is also closely connected to our earlier result that the mean number of second neighbours of a vertex on an Erdős–Rényi graph is simply the square of the mean number of first neighbours.

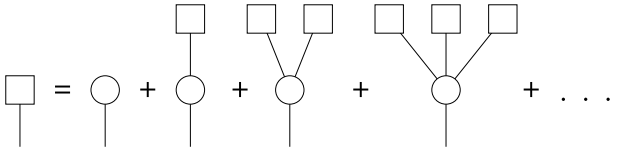


Figure 2.3: Schematic representation of the possible forms for the connected component of vertices reached by following a randomly chosen edge. The total probability of all possible forms (left-hand side) can be represented self-consistently as the sum of the probabilities (right-hand side) of having only a single vertex (the circle), having a single vertex connected to one other component, or two other components, and so forth. The entire sum can be expressed in closed form as Eq. (2.25).

Given this, we can calculate the distribution of component sizes below the transition as follows. Consider a randomly chosen edge somewhere in our graph and imagine following that edge to one of its ends and then to every other vertex reachable from that end. This set of vertices we refer to as the cluster at the end of a randomly chosen edge. Let $H_1(x)$ be the generating function that generates the distribution of sizes of such clusters, in terms of numbers of vertices. Each cluster can take many different forms, as shown in Fig. 2.3. We can follow our randomly chosen edge and find only a single vertex at its end, with no further edges emanating from it. Or we can find a vertex with one or more edges emanating from it. Each edge then leads to another complete cluster whose size is also distributed according to $H_1(x)$.

The number of edges k emanating from our vertex, other than the one along which we arrived, is distributed according to the distribution q_k of Eq. (2.4), and, using the multiplication property of generating functions from Section 2.3.1, the distribution of the sum of the sizes of the k clusters that they lead to is generated by $[H_1(x)]^k$. Thus the total number of vertices reachable by following our randomly chosen edge is generated by

$$H_1(x) = x \sum_{k=0}^{\infty} q_k [H_1(x)]^k = x G_1(H_1(x)), \tag{2.25}$$

where the leading factor of x accounts for the one vertex at the end of our edge, and we have made use of Eq. (2.14).

The quantity we actually want to know is the distribution of the sizes of the clusters to which a randomly chosen vertex belongs. The number of edges emanating from such a vertex is distributed according to the degree distribution p_k, and each such edge leads to a cluster whose size in vertices is drawn from the distribution generated by the function $H_1(x)$ above. Thus the size of the complete component to which a randomly vertex belongs is generated by

$$H_0(x) = x \sum_{k=0}^{\infty} p_k [H_1(x)]^k = x G_0(H_1(x)). \tag{2.26}$$

Now we can calculate the complete distribution of component sizes by solving (2.25) self-consistently for $H_1(x)$ and then substituting the result into (2.26).

Consider for instance the third example from Section 2.3.2, of a graph in which all vertices have degree three or less. Then Eq. (2.25) implies that $u = H_1(x)$ is a solution of the quadratic

equation

$$q_2 u^2 + \left(q_1 - \frac{1}{x}\right) u + q_0 = 0, \tag{2.27}$$

or

$$H_1(x) = \frac{\frac{1}{x} - q_1 \pm \sqrt{\left(q_1 - \frac{1}{x}\right)^2 - 4q_0 q_2}}{2q_2}. \tag{2.28}$$

Substituting this into Eq. (2.26) and differentiating m times then gives the probability that a randomly chosen vertex belongs to a component of exactly m vertices total.

Unfortunately, cases such as this in which we can solve exactly for $H_0(x)$ and $H_1(x)$ are rare. More often no closed-form solution is possible. (For the simple Poissonian case of the Erdős–Rényi random graph, for instance, Eq. (2.25) is transcendental and has no closed-form solution.) We can still find closed-form expressions for the generating functions up to any finite order in x however, by iteration of (2.25). To see this, suppose that we have an approximate expression for $H_1(x)$ that is correct up to some finite order x^m, but possibly incorrect at order x^{m+1} and higher. If we substitute this approximate expression into the right-hand side of Eq. (2.25), we get a new expression for $H_1(x)$ and, because of the leading factor of x, the only contributions to the coefficient of x^{m+1} in this expression come from the coefficients of x^m and lower in the old expression. Since these lower coefficients were exactly correct, it immediately follows that the coefficient of x^{m+1} in the new expression is correct also. Thus, if we start with the expression $H_1(x) = q_0 x$, which is correct to order x^1, substitute it into (2.25), and iterate, then on each iteration we will generate an expression for $H_1(x)$ that is accurate to one order higher. After m iterations, we will have an expression in which the coefficients for all orders up to and including x^{m+1} are exactly correct.

Take for example the Erdős–Rényi random graph with its Poisson degree distribution, for which $G_0(x) = G_1(x) = \epsilon^{z(x-1)}$, as shown in Section 2.3.2. Then, noting that $q_0 = \epsilon^{-z}$ for this case, we find that the first few iterations of Eq. (2.25) give

$$zH_1^{(1)}(x) = xz\epsilon^{-z} + \mathcal{O}(x^2), \tag{2.29a}$$

$$zH_1^{(2)}(x) = xz\epsilon^{-z} + (xz\epsilon^{-z})^2 + \mathcal{O}(x^3), \tag{2.29b}$$

$$\vdots$$

$$zH_1^{(5)}(x) = xz\epsilon^{-z} + (xz\epsilon^{-z})^2 + \tfrac{3}{2}(xz\epsilon^{-z})^3 + \tfrac{5}{3}(xz\epsilon^{-z})^4 + \tfrac{8}{3}(xz\epsilon^{-z})^5 + \mathcal{O}(x^6), \tag{2.29c}$$

and so forth, from which we conclude that the probabilities P_s of a randomly chosen site belonging to components of size $s = 1, 2, 3 \ldots$ are

$$P_1 = \epsilon^{-z}, \quad P_2 = z\epsilon^{-2z}, \quad P_3 = \tfrac{3}{2}z^2\epsilon^{-3z}, \quad P_4 = \tfrac{5}{3}z^3\epsilon^{-4z}, \quad P_5 = \tfrac{8}{3}z^4\epsilon^{-5z}. \tag{2.30}$$

With a good symbolic manipulation program it is straightforward to calculate such probabilities to order 100 or so. If we require probabilities to higher order it is still possible to use Eqs. (2.25) and (2.26) to get answers, by iterating (2.25) numerically from a starting value of

$H_1(x) = q_0 x$. Doing this for a variety of different values of x close to $x = 0$, we can use the results to calculate the derivatives of $H_0(x)$ and so evaluate the P_s. Unfortunately, this technique is only usable for the first few P_s, because, as is usually the case with numerical derivatives, limits on the precision of floating-point numbers result in large errors at higher orders. To circumvent this problem we can employ a technique suggested by Moore and Newman (2000), and evaluate the derivatives instead by numerically integrating the Cauchy formula

$$P_s = \frac{1}{s!}\frac{\partial^s H_0}{\partial x^s}\bigg|_{x=0} = \frac{1}{2\pi \mathrm{i}} \oint \frac{H_0(\zeta)\,\delta\zeta}{\zeta^s}, \tag{2.31}$$

where the integral is performed around any contour surrounding the origin but inside the first pole in $H_0(\zeta)$. For the best precision, Moore and Newman suggest using the largest such contour possible. In the present case, where P_s is a properly normalized probability distribution, it is straightforward to show that $H_0(\zeta)$ must always converge within the unit circle and hence we recommend using this circle as the contour. Doing so appears to give excellent results in practice (Newman *et al.*, 2001), with a thousand or more derivatives easily calculable in reasonable time.

2.4.2 Mean component size

Although, as we have seen, it is not usually possible to calculate the probability distribution of component sizes P_s to all orders in closed form, we can calculate moments of the distribution, which in many cases is more useful anyway. The simplest case is the first moment, the mean component size. As we saw in Section 2.3.1, the mean of the distribution generated by a generating function is given by the derivative of the generating function evaluated at unity (Eq. (2.16)). Below the phase transition, the component size distribution is generated by $H_0(x)$, Eq. (2.26), and hence the mean component size below the transition is

$$\langle s\rangle = H_0'(1) = \bigl[G_0(H_1(x)) + xG_0'(H_1(x))H_1'(x)\bigr]_{x=1} = 1 + G_0'(1)H_1'(1), \tag{2.32}$$

where we have made use of the fact, Eq. (2.15), that properly normalized generating functions are equal to 1 at $x = 1$, so that $G_0(1) = H_1(1) = 1$. The value of $H_1'(1)$ we can calculate from Eq. (2.25) by differentiating and rearranging to give

$$H_1'(1) = \frac{1}{1 - G_1'(1)}, \tag{2.33}$$

and substituting into (2.32) we find

$$\langle s\rangle = 1 + \frac{G_0'(1)}{1 - G_1'(1)}. \tag{2.34}$$

This expression can also be written in a number of other forms. For example, we note that

$$G_0'(1) = \sum_k k p_k = \langle k\rangle = z_1, \tag{2.35}$$

$$G_1'(1) = \frac{\sum_k k(k-1)p_k}{\sum_k k p_k} = \frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle} = \frac{z_2}{z_1}, \tag{2.36}$$

where we have made use of Eq. (2.6). Substituting into (2.34) then gives the average component size below the transition as

$$\langle s \rangle = 1 + \frac{z_1^2}{z_1 - z_2}. \tag{2.37}$$

This expression has a divergence at $z_1 = z_2$, which signifies the formation of the giant component and gives an alternative and more rigorous derivation of the position of the critical point to that given in Section 2.2. Using Eq. (2.34), we could also write the condition for the phase transition as $G_1'(1) = 1$.

2.4.3 Above the phase transition

The calculations of the previous sections concerned the behaviour of the graph below the phase transition where there is no giant component in the graph. Almost all graphs studied empirically seem to be in the regime above the transition and do have a giant component. (This may be a tautologous statement, since it probably rarely occurs to researchers to consider a network representation of a set of objects or people so loosely linked that there is no connection between most pairs.) Can our generating function techniques be extended to this regime? As we now show, they can, although we will have to use some tricks to make things work. The problem is that the giant component is not a component like those we have considered so far. Those components had a finite average size, which meant that in the limit of large graph size they were all tree-like, containing no closed loops, as discussed in Section 2.4.1. The giant component, on the other hand, scales, by definition, as the size of the graph as a whole, and therefore becomes infinite as $n \to \infty$. This means that there will in general be loops in the giant component, which makes all the arguments of the previous sections break down. This problem can be fixed however by the following simple ploy. Above the transition, we define $H_0(x)$ and $H_1(x)$ to be the generating functions for the distributions of component sizes *excluding* the giant component. The non-giant components are still tree-like even above the transition, so Eqs. (2.25) and (2.26) are correct for this definition. The only difference is that now $H_0(1)$ is no longer equal to 1 (and neither is $H_1(1)$). Instead,

$$H_0(1) = \sum_s P_s = \text{fraction of vertices not in giant component}, \tag{2.38}$$

which follows because the sum over s is now over only the non-giant components, so the probabilities P_s no longer add up to 1. This result is very useful; it allows us to calculate the size S of the giant component above the transition as a fraction of the total graph size, since $S = 1 - H_0(1)$. From Eqs. (2.25) and (2.26), we can see that S must be the solution of the equations

$$S = 1 - G_0(v), \qquad v = G_1(v), \tag{2.39}$$

where $v \equiv H_1(1)$. As with the calculation of the component size distribution in Section 2.4.1, these equations are not normally solvable in closed form, but a solution can be found to arbitrary numerical accuracy by iteration starting from a suitable initial value of v, such as $v = 0$.

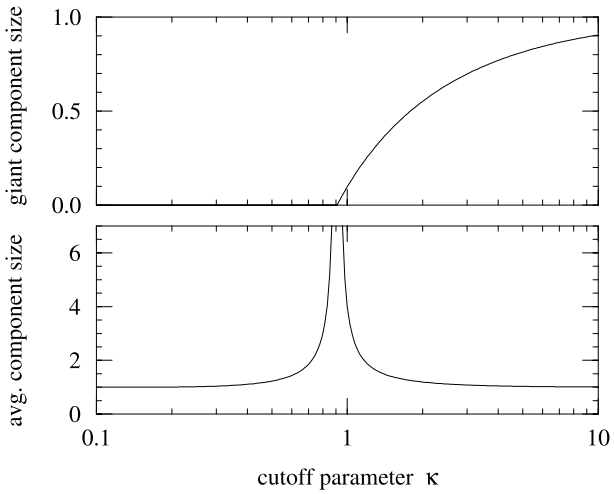


Figure 2.4: Behaviour of a random graph with an exponential degree distribution of the form of Eq. (2.20). Top: fraction of the graph occupied by the giant component. Bottom: average component size. Note that the horizontal axis is logarithmic.

We can also calculate the average sizes of the non-giant components in the standard way by differentiating Eq. (2.26). We must be careful however, for a couple of reasons. First, we can no longer assume that $H_0(1) = H_1(1) = 1$ as is the case below the transition. Second, since the distribution P_s is not normalized to 1, we have to perform the normalization ourselves. The correct expression for the average component size is

$$\begin{aligned}\langle s\rangle &= \frac{H_0'(1)}{H_0(1)} = \frac{1}{H_0(1)}\biggl[G_0(H_1(1)) + \frac{G_0'(H_1(1))G_1(H_1(1))}{1-G_1'(H_1(1))}\biggr] \\ &= 1 + \frac{zv^2}{[1-S][1-G_1'(v)]},\end{aligned} \tag{2.40}$$

where v and S are found from Eq. (2.39). It is straightforward to verify that this becomes equal to Eq. (2.34) when we are below the transition and $S = 0$, $v = 1$.

As an example of these results, we show in Fig. 2.4 the size of the giant component and the average (non-giant) component size for graphs with an exponential degree distribution of the form of Eq. (2.20), as a function of the exponential constant κ. As the figure shows, there is a divergence in the average component size at the phase transition, with the giant component becoming non-zero smoothly above the transition. Those accustomed to the physics of continuous phase transitions will find this behaviour familiar; the size of the giant component acts as an order parameter here, as it did in the Erdős–Rényi random graph in the introduction to this chapter, and the average component size behaves like a susceptibility. Indeed one can define and calculate critical exponents for the transition using this analogy, and as with the Erdős–Rényi model, their values put us in the same universality class as the mean-field (i.e., infinite dimension) percolation transition (Newman *et al.*, 2001). The phase transition in Fig. 2.4 takes place just a little below $\kappa = 1$ when $G_1'(1) = 1$, which gives a critical value of $\kappa_c = (\log 3)^{-1} = 0.910\ldots$

2.5 Properties of directed graphs

Some of the graphs discussed in the introduction to this chapter are directed graphs. That is, the edges in the network have a direction to them. Examples are the World-Wide Web, in which hyperlinks from one page to another point in only one direction, and food webs, in which predator–prey interactions are asymmetric and can be thought of as pointing from predator to prey. Other recently studied examples of directed networks include telephone call graphs (Abello *et al.*, 1998; Hayes, 2000; Aiello *et al.*, 2000), citation networks (Redner, 1998; Vazquez, 2001), and email networks (Ebel *et al.*, 2002).

Directed networks are more complex than their undirected counterparts. For a start, each vertex in an directed network has two degrees, an **in-degree**, which is the number of edges that point into the vertex, and an **out-degree**, which is the number pointing out. There are also, correspondingly, two degree distributions. In fact, to be completely general, we must allow for a **joint degree distribution** of in- and out-degree: we define p_{jk} to be the probability that a randomly chosen vertex simultaneously has in-degree j and out-degree k. Defining a joint distribution like this allows for the possibility that the in- and out-degrees may be correlated. For example in a graph where every vertex had precisely the same in- and out-degree, p_{jk} would be non-zero if and only if $j = k$.

The component structure of a directed graph is more complex than that of an undirected graph also, because a directed path may exist through the network from vertex A to vertex B, but that does not guarantee that one exists from B to A. As a result, any vertex A belongs to components of four different types:

1. The **in-component** is the set of vertices from which A can be reached.

2. The **out-component** is the set of vertices which can be reached from A.

3. The **strongly connected component** is the set of vertices from which vertex A can be reached *and* which can be reached from A.

4. The **weakly connected component** is the set of vertices that can be reached from A ignoring the directed nature of the edges altogether.

The weakly connected component is just the normal component to which A belongs if one treats the graph as undirected. Clearly the details of weakly connected components can be worked out using the formalism of Section 2.4, so we will ignore this case. For vertex A to belong to a strongly connected component of size greater than one, there must be at least one other vertex that can both be reached from A and from which A can be reached. This however implies that there is a closed loop of directed edges in the graph, something which, as we saw in Section 2.4.1, does not happen in the limit of large graph size. So we ignore this case also. The two remaining cases, the in- and out-components, we consider in more detail in the following sections.

2.5.1 Generating functions

Because the degree distribution p_{jk} for a directed graph is a function of two variables, the corresponding generating function is also:

$$\mathcal{G}(x,y) = \sum_{j,k=0}^{\infty} p_{jk} x^j y^k. \tag{2.41}$$

This function satisfies the normalization condition $\mathcal{G}(1,1) = 1$, and the means of the in- and out-degree distributions are given by its first derivatives with respect to x and y. However, there is *only one* mean degree z for a directed graph, since every edge must start and end at a site. This means that the total and hence also the average numbers of in-going and out-going edges are the same. This gives rise to a constraint on the generating function of the form

$$\left.\frac{\partial \mathcal{G}}{\partial x}\right|_{x,y=1} = z = \left.\frac{\partial \mathcal{G}}{\partial y}\right|_{x,y=1}, \tag{2.42}$$

and there is a corresponding constraint on the probability distribution p_{jk} itself, which can be written

$$\sum_{jk} (j-k) p_{jk} = 0. \tag{2.43}$$

From $\mathcal{G}(x,y)$, we can now define single-argument generating functions G_0 and G_1 for the number of out-going edges leaving a randomly chosen vertex, and the number leaving the vertex reached by following a randomly chosen edge. These play a similar role to the functions of the same name in Section 2.4. We can also define generating functions F_0 and F_1 for the number of edges arriving at a vertex. These functions are given by

$$F_0(x) = \mathcal{G}(x,1), \qquad F_1(x) = \frac{1}{z}\left.\frac{\partial \mathcal{G}}{\partial y}\right|_{y=1}, \tag{2.44}$$

$$G_0(y) = \mathcal{G}(1,y), \qquad G_1(y) = \frac{1}{z}\left.\frac{\partial \mathcal{G}}{\partial x}\right|_{x=1}. \tag{2.45}$$

Once we have these functions, many results follow as before.

2.5.2 Results

The probability distribution of the numbers of vertices reachable from a randomly chosen vertex in a directed graph—i.e., of the sizes of the out-components—is generated by the function $H_0(y) = yG_0(H_1(y))$, where $H_1(y)$ is a solution of $H_1(y) = yG_1(H_1(y))$, just as before. (A similar and obvious pair of equations governs the sizes of the in-components.) The average out-component size for the case where there is no giant component is then given by Eq. (2.34), and thus the point at which a giant component first appears is given once more by $G_1'(1) = 1$. Substituting Eq. (2.45) into this expression gives the explicit condition

$$\sum_{jk} (2jk - j - k) p_{jk} = 0 \tag{2.46}$$

for the first appearance of the giant component. This expression is the equivalent for the directed graph of Eq. (2.9). It is also possible, and equally valid, to define the position at which the giant component appears by $F_1'(1) = 1$, which provides an alternative derivation for Eq. (2.46).

But this raises an interesting issue. Which giant component are we talking about? Just as with the small components, there are four types of giant component, the giant in- and out-components, and the giant weakly and strongly connected components. Furthermore, while the giant weakly connected component is as before trivial, the giant strongly connected component does not normally vanish as the other strongly connected components do. There is no reason why a giant component should contain no loops, and therefore no reason why we should not have a non-zero giant strongly connected component.

The condition for the position of the phase transition given above is derived from the point at which the mean size of the out-component reachable from a vertex diverges, and thus this is the position at which the giant *in*-component forms (since above this point an extensive number of vertices can be reached starting from one vertex, and hence that vertex must belong to the giant in-component). Furthermore, as we have seen, we get the same condition if we ask where the mean in-component size diverges, i.e., where the giant *out*-component forms, and so we conclude that both giant in- and out-components appear at the same time, at the point given by Eq. (2.46).

The sizes of these two giant components can also be calculated with only a little extra effort. As before, we can generalize the functions $H_0(y)$ and $H_1(y)$ to the regime above the transition by defining them to be the generating functions for the non-giant out-components in this regime. In that case, $H_0(1)$ is equal to the fraction of all vertices that have a finite out-component. But any vertex A that has only a finite out-component cannot, by definition, belong to the giant *in*-component, i.e., there definitely do not exist an extensive number of vertices that can be reached from A. Thus the size of the giant in-component is simply $S_{\rm in} = 1 - H_0(1)$, which can be calculated as before from Eq. (2.39). Similarly the size of the giant out-component can be calculated from (2.39) with $G_0 \to F_0$ and $G_1 \to F_1$.

To calculate the size of the giant strongly connected component, we observe the following (Dorogovtsev *et al.*, 2001). If at least one of a vertex's outgoing edges leads to anywhere in the giant *in*-component, then one can reach the giant strongly connected component from that vertex. Conversely, if at least one of a vertex's incoming edges leads from anywhere in the giant *out*-component, then the vertex can be reached from the strongly connected component. If and only if both of these conditions are satisfied simultaneously, then the vertex belongs to the giant strongly connected component itself.

Consider then the outgoing edges. The function $H_1(x)$ gives the probability distribution of the sizes of finite out-components reached by following a randomly chosen edge. This implies that $H_1(1)$ is the total probability that an edge leads to a finite out-component (i.e., *not* to the giant in-component) and as before (Eq. (2.39)) $H_1(1)$ is the fixed point of $G_1(x)$, which we denote by v. For a vertex with k outgoing edges, v^k is then the probability that all of them lead to finite components and $1 - v^k$ is the probability that at least one edge leads to the giant in-component. Similarly the probability that at least one incoming edge leads from the giant out-component is $1 - u^j$, where u is the fixed point of $F_1(x)$ and j is the in-degree of the vertex. Thus the probability that a vertex with in- and out-degrees j and k is in the giant strongly connected component is $(1 - u^j)(1 - v^k)$, and the average of this probability over

all vertices, which is also the fractional size of the giant strongly connected component, is

$$\begin{aligned} S_s &= \sum_{jk} p_{jk}(1-u^j)(1-v^k) = \sum_{jk} p_{jk}(1-u^j-v^k+u^j v^k) \\ &= 1 - \mathcal{G}(u,1) - \mathcal{G}(1,v) + \mathcal{G}(u,v), \end{aligned} \tag{2.47}$$

where u and v are solutions of

$$u = F_1(u), \qquad v = G_1(v), \tag{2.48}$$

and we have made use of the definition, Eq. (2.41), of $\mathcal{G}(x,y)$. Noting that $u = v = 1$ below the transition at which the giant in- and out-components appear, and that $\mathcal{G}(1,1) = 1$, we see that the giant strongly connected component also first appears at the transition point given by Eq. (2.46). Thus there are in general two phase transitions in a directed graph: the one at which the giant weakly connected component appears, and the one at which the other three giant components all appear.

Applying the theory of directed random graphs to real directed networks has proved difficult so far, because experimenters rarely measure the joint in- and out-degree distribution p_{jk} that is needed to perform the calculations described above. A few results can be calculated without the joint distribution—see Newman *et al.* (2001) for instance. By and large, however, the theory presented in this section is still awaiting empirical tests.

2.6 Networks with clustering

Far fewer analytical results exist for networks that incorporate clustering than for the non-clustered networks of the previous sections. A first attempt at extending random graph models to incorporate clustering has been made by the present author, who studied the correction to the quantity z_2—the average number of next-nearest neighbours of a vertex—in graphs with a non-zero clustering coefficient C (Newman, 2001d).

Consider a vertex A, with its first and second neighbours in the network arrayed around it in two concentric rings. In a normal random graph, a neighbour of A that has degree m contributes $m-1$ vertices to the ring of second neighbours of A, as discussed in Section 2.2. That is, all of the second neighbours of A are independent; each of them is a new vertex never before seen. This is the reasoning that led to our earlier expression, Eq. (2.6): $z_2 = \langle k^2\rangle - \langle k\rangle$. In a clustered network however, the picture is different. In a clustered network, many of the neighbours of A's neighbour are neighbours of A themselves. This is the meaning of clustering: your friend's friend is also your friend. In fact, by definition, an average fraction C of the $m-1$ neighbours are themselves neighbours of the central vertex A and hence should not be counted as second neighbours. Correspondingly, this reduces our estimate of z_2 by a factor of $1-C$ to give $z_2 = (1-C)(\langle k^2\rangle - \langle k\rangle)$.

But this is not all. There is another effect we need to take into account if we are to estimate z_2 correctly. It is also possible that we are overcounting the second neighbours of A because some of them are neighbours of more than one of the first neighbours. In other words, you may know two people who have another friend in common, whom you personally don't know. Such connections create "squares" in the network, whose density can be quantified by the

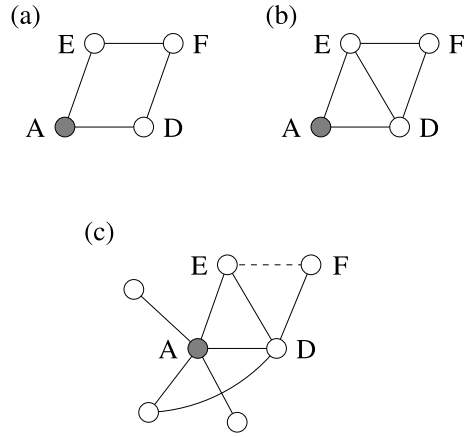


Figure 2.5: (a) An example of a vertex (F) that is two steps away from the center vertex (A, shaded), but is connected to two of A's neighbours (D and E). F should only be counted once as a second neighbour of A, not twice. (b) A similar situation in which D and E are also neighbours of one another. (c) The probability of situation (b) can be calculated by considering this situation. Since D is friends with both E and F, the probability that E and F also know one another (dotted line), thereby completing the quadrilateral in (b), is by definition equal to the clustering coefficient.

so-called **mutuality** M:

$$M = \frac{\text{mean number of vertices two steps away from a vertex}}{\text{mean number of paths of length two to those vertices}}. \tag{2.49}$$

In words, M measures the average number of paths of length two leading to a vertex's second neighbour. As a result of the mutuality effect, our current estimate of z_2 will be too great by a factor of $1/M$, and hence a better estimate is

$$z_2 = M(1-C)(\langle k^2\rangle - \langle k\rangle). \tag{2.50}$$

But now we have a problem. Calculating the mutuality M using Eq. (2.49) requires that we know the mean number of individuals two steps away from the central vertex A. But this mean number is precisely the quantity z_2 that our calculation is supposed to estimate in the first place. There is a partial solution to this problem. Consider the two configurations depicted in Fig. 2.5, parts (a) and (b). In (a) our vertex A has two neighbours D and E, both of whom are connected to F, although F is not itself an neighbour of A. The same is true in (b), but now D and E are friends of one another also. Empirically, it appears that in many networks situation (a) is quite uncommon, while situation (b) is much more common. And we can estimate the frequency of occurrence of (b) from a knowledge of the clustering coefficient.

Consider Fig. 2.5c. The central vertex A shares an edge with D, which shares an edge with F. How many other paths of length two are there from A to F? Well, if A has k_1 neighbours, then by the definition of the clustering coefficient, D will be connected to $C(k_1-1)$ of them on average. The edge between vertices D and E in the figure is an example of one

such. But now D is connected to both E and F, and hence, using the definition of the clustering coefficient again, E and F will themselves be connected (dotted line) with probability equal to the clustering coefficient C. Thus there will on average be $C^2(k_1 - 1)$ other paths of length 2 to F, or $1 + C^2(k_1 - 1)$ paths in total, counting the one that runs through D. This is the average factor by which we will overcount the number of second neighbours of A because of the mutuality effect. As shown by Newman (2001d), the mutuality coefficient is then given by

$$M = \frac{\langle k/[1 + C^2(k-1)]\rangle}{\langle k\rangle}. \tag{2.51}$$

Substituting this into Eq. (2.50) then gives us an estimate of z_2.

In essence what Eq. (2.51) does is estimate the value of M in a network in which triangles of ties are common, but squares that are not composed of adjacent triangles are assumed to occur with frequency no greater than one would expect in a purely random network. It is only an approximate expression, since this assumption will usually not be obeyed perfectly. Nonetheless, it appears to give good results. The author applied Eqs. (2.50) and (2.51) to estimation of z_2 for the two coauthorship networks of Fig. 2.1c, and found that they gave results accurate to within 10% in both cases.

This calculation is certainly only a first step. Ideally we would like to be able to calculate numbers of vertices at any distance from a randomly chosen central vertex in the presence of clustering, and to do it exactly rather than just approximately. If this were possible, then, as in Section 2.2, one could use the ratio of the numbers of vertices at different distances to derive a condition for the position of the phase transition at which a giant component forms on a clustered graph. At present it is not clear if such a calculation is possible.

2.7 Models defined on random graphs

In addition to providing an analytic framework for calculating topological properties of networks, such as typical path lengths or distributions of cluster sizes, random graphs form a useful substrate for studying the behaviour of phenomena that take place on networks. Analytic work in this area is in its infancy; here we describe two examples of recent work on models that use ideas drawn from percolation theory.

2.7.1 Network resilience

As emphasized by Albert and co-workers, the highly skewed degree distributions of Fig. 2.1 have substantial implications for the robustness of networks to the removal of vertices (Albert *et al.*, 2000). Because most of the vertices in a network with such a degree distribution typically have low degree, the random removal of vertices from the network has little effect on the connectivity of the remaining vertices, i.e., on the existence of paths between pairs of vertices, a crucial property of networks such as the Internet, for which functionality relies on connectivity.[8] In particular, removal of vertices with degree zero or one will never have any effect

[8] A few recent papers in the physics literature have used the word "connectivity" to mean the same thing as "degree", i.e., number of edges attaching to a vertex. In this chapter however the word has its standard graph theoretical meaning of existence of connecting paths between pairs of vertices.

on the connectivity of the remaining vertices. (Vertices of degree zero are not connected to anyone else anyway, and vertices of degree one do not lie on any path between another pair of vertices.)

Conversely, however, the specific removal of the vertices in the network with the highest degree frequently has a devastating effect. These vertices lie on many of the paths between pairs of other vertices and their removal can destroy the connectivity of the network in short order. This was first demonstrated numerically by Albert *et al.* (2000) and independently by Broder *et al.* (2000) using data for subsets of the World-Wide Web. More recently however it has been demonstrated analytically also, for random graphs with arbitrary degree distributions, by Callaway *et al.* (2000) and by Cohen *et al.* (2001). Here we follow the derivation of Callaway *et al.*, which closely mirrors some of the earlier mathematical developments of this chapter.

Consider a simple model defined on a network in which each vertex is either "present" or "absent". Absent vertices are vertices that have either been removed, or more realistically are present but non-functional, such as Internet routers that have failed or Web sites whose host computer has gone down. We define a probability b_k of being present which is some arbitrary function of the degree k of a vertex, and then define the generating function

$$F_0(x) = \sum_{k=0}^{\infty} p_k b_k x^k, \tag{2.52}$$

whose coefficients are the probabilities that a vertex has degree k and is present. Note that this generating function is not equal to 1 at $x = 1$; instead it is equal to the fraction of all vertices that are present. By analogy with Eq. (2.14) we also define

$$F_1(x) = \frac{\sum_k k p_k b_k x^{k-1}}{\sum_k k p_k} = \frac{F_0'(x)}{z}. \tag{2.53}$$

Then the distributions of the sizes of connected clusters of present vertices reachable from a randomly chosen vertex or edge are generated respectively by

$$H_0(x) = 1 - F_0(1) + x F_0(H_1(x)), \qquad H_1(x) = 1 - F_1(1) + x F_1(H_1(x)), \tag{2.54}$$

which are logical equivalents of Eqs. (2.25) and (2.26).

Take for instance the case of random failure of vertices. In this case, the probability b_k of a vertex being present is independent of the degree k and just equal to a constant b, which means that

$$H_0(x) = 1 - b + b x G_0(H_1(x)), \qquad H_1(x) = 1 - b + b x G_1(H_1(x)), \tag{2.55}$$

where $G_0(x)$ and $G_1(x)$ are the standard generating functions for vertex degree, Eqs. (2.12) and (2.14). This implies that the mean size of a cluster of connected and present vertices is

$$\langle s \rangle = H_0'(1) = b + b F_0'(1) H_1'(1) = b \left[1 + \frac{b G_0'(1)}{1 - b G_1'(1)} \right], \tag{2.56}$$

and the model has a phase transition at the critical value of b

$$b_c = \frac{1}{G_1'(1)}. \tag{2.57}$$

If a fraction $b < b_c$ of the vertices are present in the network, then there will be no giant component. This is the point at which the network ceases to be functional in terms of connectivity. When there is no giant component, connecting paths exist only within small isolated groups of vertices, but no long-range connectivity exists. For a communication network such as the Internet, this would be fatal. As we would expect from the arguments above however, b_c is usually a very small number for networks with skewed degree distributions. For example, if a network has a pure power-law degree distribution with exponent α, as both the Internet and the World-Wide Web appear to do (see Fig. 2.1a and 2.1b), then

$$b_c = \frac{\zeta(\alpha-1)}{\zeta(\alpha-2)-\zeta(\alpha-1)}, \tag{2.58}$$

where $\zeta(x)$ is the Riemann ζ-function. This expression is formally zero for all $\alpha \leq 3$. Since none of the distributions in Fig. 2.1 have an exponent greater than 3, it follows that, at least to the extent that these graphs can be modelled as random graphs, none of them has a phase transition at all. No matter how many vertices fail in these networks, as long as the failing vertices are selected at random without regard for degree, there will always be a giant component in the network and an extensive fraction of the vertices will be connected to one another. In this sense, networks with power-law distributed degrees are highly robust, as the numerical experiments of Albert *et al.* (2000) and Broder *et al.* (2000) also found.

But now consider the case in which the vertices are removed in decreasing order of their degrees, starting with the highest degree vertex. Mathematically we can represent this by setting

$$b_k = \theta(k_{\max} - k), \tag{2.59}$$

where $\theta(x)$ is the Heaviside step function

$$\theta(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0. \end{cases} \tag{2.60}$$

This is equivalent to setting the upper limit of the sum in Eq. (2.52) to $k_{\max}$.

For this case we need to use the full definition of $H_0(x)$ and $H_1(x)$, Eq. (2.54), which gives the position of the phase transition as the point at which $F_1'(1) = 1$, or

$$\frac{\sum_{k=1}^{\infty} k(k-1)p_k b_k}{\sum_{k=1}^{\infty} k p_k} = 1. \tag{2.61}$$

Taking the example of our power-law degree distribution again, $p_k \propto k^{-\alpha}$, this then implies that the phase transition occurs at a value k_c of $k_{\max}$ satisfying

$$H_{k_c}^{(\alpha-2)} - H_{k_c}^{(\alpha-1)} = \zeta(\alpha-1), \tag{2.62}$$

where $H_n^{(r)}$ is the nth harmonic number of order r:

$$H_n^{(r)} = \sum_{k=1}^{n} \frac{1}{k^r}. \tag{2.63}$$

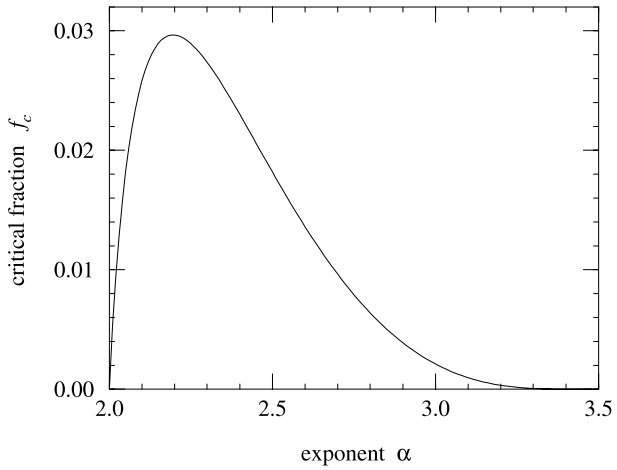


Figure 2.6: The critical fraction, Eq. (2.64), of highest degree vertices that must be removed in order to destroy the giant component in a graph with a power-law degree distribution having exponent α.

This solution is not in a very useful form however. What we really want to know is what fraction f_c of the vertices have been removed when we reach the transition. This fraction is given by

$$f_c = 1 - \frac{H^{(\alpha)}_{k_c}}{\zeta(\alpha)}. \tag{2.64}$$

Although we cannot eliminate k_c from (2.62) and (2.64) to get f_c in closed form, we can solve Eq. (2.62) numerically for k_c and substitute into (2.64). The result is shown as a function of α in Fig. 2.6. As the figure shows, one need only remove a very small fraction of the high-degree vertices to destroy the giant component in a power-law graph, always less than 3%, with the most robust graphs being those around $\alpha = 2.2$, interestingly quite close to the exponent seen in a number of real-world networks (Fig. (2.1)). Below $\alpha = 2$, there is no real solution for f_c: power-law distributions with $\alpha < 2$ have no finite mean anyway and therefore make little sense physically. And $f_c = 0$ for all values $\alpha > 3.4788\ldots$, where the latter figure is the solution of $\zeta(\alpha - 2) = 2\zeta(\alpha - 1)$, because the underlying network itself has no giant component for such values of α (Aiello *et al.*, 2000).

Overall, therefore, our results agree with the findings of the previous numerical studies that graphs with skewed degree distributions, such as power laws, can be highly robust to the random removal of vertices, but extremely fragile to the specific removal of their highest-degree vertices.

2.7.2 Epidemiology

An important application of the theory of networks is in epidemiology, the study of the spread of disease. Diseases are communicated from one host to another by physical contact, and

the pattern of who has contact with whom forms a **contact network** whose structure has implications for the shape of epidemics. In particular, the small-world effect discussed in Section 2.2 means that diseases will spread through a community much faster than one might otherwise imagine.

In the standard mathematical treatments of diseases, researchers use the so-called **fully mixed approximation**, in which it is assumed that every individual has equal chance of contact with every other. This is an unrealistic assumption, but it has proven popular because it allows one to write differential equations for the time evolution of the disease that can be solved or numerically integrated with relative ease. More realistic treatments have also been given in which populations are divided into groups according to age or other characteristics. These models are still fully mixed within each group however. To go beyond these approximations, we need to incorporate a full network structure into the model, and the random graphs of this chapter and the generating function methods we have developed to handle them provide a good basis for doing this.

In this section we show that the most fundamental standard model of disease propagation, the SIR model, and a large set of its generalized forms, can be solved on random graphs by mapping them onto percolation problems. These solutions provide exact criteria for deciding when an epidemic will occur, how many people will be affected, and how the network structure or the transmission properties of the disease could be modified in order to prevent the epidemic.

2.7.3 The SIR model

First formulated (though never published) by Lowell Reed and Wade Hampton Frost in the 1920s, the SIR model (Bailey, 1975; Anderson and May, 1991; Hethcote, 2000) is a model of disease propagation in which members of a population are divided into three classes: susceptible (S), meaning they are free of the disease but can catch it; infective (I), meaning they have the disease and can pass it on to others;[9] and removed (R), meaning they have recovered from the disease or died, and can no longer pass it on. There is a fixed probability per unit time that an infective individual will pass the disease to a susceptible individual with whom they have contact, rendering that individual infective. Individuals who contract the disease remain infective for a certain time period before recovering (or dying) and thereby losing their infectivity.

As first pointed out by Grassberger (1983), the SIR model on a network can be simply mapped to a bond percolation process. Consider an outbreak on a network that starts with a single individual and spreads to encompass some subset of the network. The vertices of the network represent potential hosts and the edges represent pairs of hosts who have contact with one another. If we imagine occupying or colouring in all the edges that result in transmission of the disease during the current outbreak, then the set of vertices representing the hosts infected in this outbreak form a connected percolation cluster of occupied edges. Furthermore, it is easy to convince oneself that each edge is occupied with independent probability. If we denote by τ the time for which an infected host remains infective and by r the probability

[9] In common parlance, the word "infectious" is more often used, but in the epidemiological literature "infective" is the accepted term.

per unit time that that host will infect one of its neighbours in the network, then the total probability of infection is

$$T = 1 - \lim_{\delta t \to 0} (1 - r\,\delta t)^{\tau/\delta t} = 1 - \epsilon^{-r\tau}. \tag{2.65}$$

This quantity we call the **transmissibility**, and it is the probability that any edge on the network is occupied. The size distribution of outbreaks of the disease is then given by the size distribution of percolation clusters on the network when edges are occupied with this probability. When the mean cluster size diverges, we get outbreaks that occupy a finite fraction of the entire network, i.e., epidemics; the percolation threshold corresponds to what an epidemiologist would call the **epidemic threshold** for the disease. Above this threshold, there exists a giant component for the percolation problem, whose size corresponds to the size of the epidemic. Thus, if we can solve bond percolation on our random graphs, we can also solve the SIR model.

In fact, we can also solve a generalized form of the SIR in which both τ and r are allowed to vary across the network. If τ and r instead of being constant are picked at random for each vertex or edge from some distributions $P(\tau)$ and $P(r)$, then the probability of percolation along any edge is simply the average of Eq. (2.65) over these two distributions (Warren *et al.*, 2001; Newman, 2002):

$$T = 1 - \int P(r)P(\tau)\,\epsilon^{-r\tau}\,\delta r\,\delta\tau. \tag{2.66}$$

2.7.4 Solution of the SIR model

The bond percolation problem on a random graph can be solved by techniques very similar to those of Section 2.7.1 (Callaway *et al.*, 2000; Newman, 2002). The equivalent of Eq. (2.55) for bond percolation with bond occupation probability T is

$$H_0(x) = xG_0(H_1(x)), \qquad H_1(x) = 1 - T + TxG_1(H_1(x)), \tag{2.67}$$

which gives an average outbreak size below the epidemic threshold of

$$\langle s\rangle = H_0'(1) = 1 + \frac{TG_0'(1)}{1 - TG_1'(1)}. \tag{2.68}$$

The threshold itself then falls at the point where $TG_1'(1) = 1$, giving a critical transmissibility of

$$T_c = \frac{1}{G_1'(1)} = \frac{\langle k\rangle}{\langle k^2\rangle - \langle k\rangle} = \frac{z_1}{z_2}, \tag{2.69}$$

where we have used Eq. (2.6). The size S of the epidemic above the epidemic transition can be calculated by finding the solution of

$$S = 1 - G_0(v), \qquad v = 1 - T + TG_1(v), \tag{2.70}$$

which will normally have to be solved numerically, since closed form solutions are rare. It is also interesting to ask what the probability is that an outbreak starting with a single carrier will

become an epidemic. This is precisely equal to the probability that the carrier belongs to the giant percolating cluster, which is also just equal to S. The probability that a given infection event (i.e., transmission along a given edge) will give rise to an epidemic is $v \equiv H_1'(1)$.

Newman and co-workers have given a variety of further generalizations of these solutions to networks with structure of various kinds, models in which the probabilities of transmission between pairs of hosts are correlated in various ways, and models incorporating vaccination, either random or targeted, which is represented as a site percolation process (Ancel *et al.*, 2001; Newman, 2002). To give one example, consider the network by which a sexually transmitted disease is communicated, which is also the network of sexual partnerships between individuals. In a recent study of 2810 respondents, Liljeros *et al.* (2001) recorded the numbers of sexual partners of men and women over the course of a year. From their data it appears that the distributions of these numbers follow a power law similar to those of the distributions in Fig. 2.1, with exponents α that fall in the range 3.1 to 3.3. If we assume that the disease of interest is transmitted primarily by contacts between men and women (true only for some diseases), then to a good approximation the network of contacts is bipartite, having two separate sets of vertices representing men and women and edges representing contacts running only between vertices of unlike kinds. We define two pairs of generating functions for males and females:

$$F_0(x) = \sum_j p_j x^j, \qquad F_1(x) = \frac{1}{\mu} \sum_j j p_j x^{j-1}, \tag{2.71}$$

$$G_0(x) = \sum_k q_k x^k, \qquad G_1(x) = \frac{1}{\nu} \sum_k k q_k x^{k-1}, \tag{2.72}$$

where p_j and q_k are the two degree distributions and μ and ν are their means. We can then develop expressions similar to Eqs. (2.68) and (2.69) for an epidemic on this new network. We find, for instance, that the epidemic transition takes place at the point where $T_{mf}T_{fm} = 1/[F_1'(1)G_1'(1)]$ where T_{mf} and T_{fm} are the transmissibilities for male-to-female and female-to-male infection respectively.

One important result that follows immediately is that if the degree distributions are truly power-law in form, then there exists an epidemic transition only for a small range of values of the exponent α of the power law. Let us assume, as appears to be the case (Liljeros *et al.*, 2001), that the exponents are roughly equal for men and women: $\alpha_m = \alpha_f = \alpha$. Then if $\alpha \leq 3$, we find that $T_{mf}T_{fm} = 0$, which is only possible if at least one of the transmissibilities T_{mf} and T_{fm} is zero. As long as both are positive, we will always be in the epidemic regime, and this would clearly be bad news. No amount of precautionary measures to reduce the probability of transmission would ever eradicate the disease. (Similar results have been seen in other types of models also (Pastor-Satorras and Vespignani, 2001; Lloyd and May, 2001).) Conversely, if $\alpha > \alpha_c$, where $\alpha_c = 3.4788\ldots$ is the solution of $\zeta(\alpha - 2) = 2\zeta(\alpha - 1)$, we find that $T_{mf}T_{fm} > 1$, which is not possible. (This latter result arises because networks with $\alpha > \alpha_c$ have no giant component at all, as mentioned in Section 2.7.1 (Aiello *et al.*, 2000).) In this regime then, no epidemic can ever occur, which would be good news. Only in the small intermediate region $3 < \alpha < 3.4788\ldots$ does the model possess an epidemic transition. Interestingly, the real-world network measured by Liljeros *et al.* (2001) appears to fall precisely in this region, with $\alpha \simeq 3.2$. If true, this would be both good and bad news. On the bad side, it means that epidemics can occur. But on the good side, it means that that it is

in theory possible to prevent an epidemic by reducing the probability of transmission, which is precisely what most health education campaigns attempt to do. The predicted critical value of the transmissibility is $\zeta(\alpha-1)/[\zeta(\alpha-2)-\zeta(\alpha-1)]$, which gives $T_c = 0.363\ldots$ for $\alpha = 3.2$. Epidemic behaviour would cease were it possible to arrange that $T_{mf}T_{fm} < T_c^2$.

2.8 Summary

In this chapter we have given an introduction to the use of random graphs as models of real-world networks. We have shown (Section 2.2) how the much studied random graph model of Erdős and Rényi can be generalized to the case of arbitrary degree distributions, allowing us to mimic the highly skewed degree distributions seen in many networks. The resulting models can be solved exactly using generating function methods in the case where there is no clustering (Sections 2.3 and 2.4). If clustering is introduced, then solutions become significantly harder, and only a few approximate analytic results are known (Section 2.6). We have also given solutions for the properties of directed random graphs (Section 2.5), in which each edge has a direction that it points in. Directed graphs are useful as models of the World-Wide Web and food webs, amongst other things. In the last part of this chapter (Section 2.7) we have given two examples of the use of random graphs as a substrate for models of dynamical processes taking place on networks, the first being a model of network robustness under failure of vertices (e.g., failure of routers on the Internet), and the second being a model of the spread of disease across the network of physical contacts between disease hosts. Both of these models can be mapped onto percolation problems of one kind of another, which can then be solved exactly, again using generating function methods.

There are many conceivable extensions of the theory presented in this chapter. In particular, there is room for many more and diverse models of processes taking place on networks. It would also be of great interest if it proved possible to extend the results of Section 2.6 to obtain exact or approximate estimates of the global properties of networks with non-zero clustering.

Acknowledgements

The author thanks Duncan Callaway, Peter Dodds, Michelle Girvan, André Krzywicki, Len Sander, Steve Strogatz and Duncan Watts for useful and entertaining conversations. Thanks are also due to Jerry Davis, Paul Ginsparg, Jerry Grossman, Oleg Khovayko, David Lipman, Heath O'Connell, Grigoriy Starchenko, Geoff West and Janet Wiener for providing data used in some of the examples. The original research described in this chapter was supported in part by the US National Science Foundation under grant number DMS–0109086.

References

Abello, J., Buchsbaum, A. and Westbrook, J., 1998. A functional approach to external graph algorithms. In *Proceedings of the 6th European Symposium on Algorithms*. Springer, Berlin.

Adamic, L. A., 1999. The small world web. In *Lecture Notes in Computer Science*, volume 1696, pp. 443–454. Springer, New York.

Aiello, W., Chung, F. and Lu, L., 2000. A random graph model for massive graphs. In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, pp. 171–180. Association of Computing Machinery, New York.

Albert, R. and Barabási, A.-L., 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97.

Albert, R., Jeong, H. and Barabási, A.-L., 1999. Diameter of the world-wide web. *Nature* **401**, 130–131.

Albert, R., Jeong, H. and Barabási, A.-L., 2000. Attack and error tolerance of complex networks. *Nature* **406**, 378–382.

Amaral, L. A. N., Scala, A., Barthélémy, M. and Stanley, H. E., 2000. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA* **97**, 11149–11152.

Ancel, L. W., Newman, M. E. J., Martin, M. and Schrag, S., 2001. Applying network theory to epidemics: Modelling the spread and control of Mycoplasma pneumoniae. Working paper 01–12–078, Santa Fe Institute.

Anderson, R. M. and May, R. M., 1991. *Infectious Diseases of Humans*. Oxford University Press, Oxford.

Bailey, N. T. J., 1975. *The Mathematical Theory of Infectious Diseases and its Applications*. Hafner Press, New York.

Barabási, A.-L. and Albert, R., 1999. Emergence of scaling in random networks. *Science* **286**, 509–512.

Bender, E. A. and Canfield, E. R., 1978. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory A* **24**, 296–307.

Bollobás, B., 1985. *Random Graphs*. Academic Press, New York.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J., 2000. Graph structure in the web. *Computer Networks* **33**, 309–320.

Callaway, D. S., Newman, M. E. J., Strogatz, S. H. and Watts, D. J., 2000. Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.* **85**, 5468–5471.

i Cancho, R. F. and Solé, R. V., 2001. The small world of human language. *Proc. R. Soc. London B* **268**, 2261–2265.

Cohen, R., Erez, K., ben-Avraham, D. and Havlin, S., 2001. Breakdown of the Internet under intentional attack. *Phys. Rev. Lett.* **86**, 3682–3685.

Dorogovtsev, S. N., Mendes, J. F. F. and Samukhin, A. N., 2001. Giant strongly connected component of directed networks. *Phys. Rev. E* **64**, 025101.

Ebel, H., Mielsch, L.-I. and Bornholdt, S., 2002. Scale-free topology of e-mail networks. Preprint cond-mat/0201476.

Erdős, P. and Rényi, A., 1959. On random graphs. *Publicationes Mathematicae* **6**, 290–297.

Erdős, P. and Rényi, A., 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**, 17–61.

Erdős, P. and Rényi, A., 1961. On the strength of connectedness of a random graph. *Acta Mathematica Scientia Hungary* **12**, 261–267.

Faloutsos, M., Faloutsos, P. and Faloutsos, C., 1999. On power-law relationships of the internet topology. *Computer Communications Review* **29**, 251–262.

Feld, S., 1991. Why your friends have more friends than you do. *Am. J. Sociol.* **96**, 1464–1477.

Fell, D. A. and Wagner, A., 2000. The small world of metabolism. *Nature Biotechnology* **18**, 1121–1122.

Grassberger, P., 1983. On the critical behavior of the general epidemic process and dynamical percolation. *Math. Biosci.* **63**, 157–172.

Hayes, B., 2000. Graph theory in practice: Part I. *American Scientist* **88**(1), 9–13.

Hethcote, H. W., 2000. Mathematics of infectious diseases. *SIAM Review* **42**, 599–653.

Janson, S., Łuczak, T. and Rucinski, A., 1999. *Random Graphs*. John Wiley, New York.

Kleinfeld, J., 2000. Could it be a big world after all? What the Milgram papers in the Yale archives reveal about the original small world study. Working paper, University of Alaska, Fairbanks.

Krzywicki, A., 2001. Defining statistical ensembles of random graphs. Preprint cond-mat/0110574.

Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E. and Åberg, Y., 2001. The web of human sexual contacts. *Nature* **411**, 907–908.

Lloyd, A. L. and May, R. M., 2001. How viruses spread among computers and people. *Science* **292**, 1316–1317.

Milgram, S., 1967. The small world problem. *Psychology Today* **2**, 60–67.

Molloy, M. and Reed, B., 1995. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* **6**, 161–179.

Montoya, J. M. and Solé, R. V., 2002. Small world patterns in food webs. J. Theor. Biol. **214**, 405–412.

Moore, C. and Newman, M. E. J., 2000. Exact solution of site and bond percolation on small-world networks. *Phys. Rev. E* **62**, 7059–7064.

Newman, M. E. J., 2001a. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* **98**, 404–409.

Newman, M. E. J., 2001b. Scientific collaboration networks: I. Network construction and fundamental results. *Phys. Rev. E* **64**, 016131.

Newman, M. E. J., 2001c. Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* **64**, 016132.

Newman, M. E. J., 2001d. Ego-centered networks and the ripple effect. Preprint cond-mat/0111070.

Newman, M. E. J., 2002. Spread of epidemic disease on networks. *Phys. Rev. E* **66**, 016128.

Newman, M. E. J., Strogatz, S. H. and Watts, D. J., 2001. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118.

Pastor-Satorras, R. and Vespignani, A., 2001. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200–3203.

Pastor-Satorras, R., Vázquez, A. and Vespignani, A., 2001. Dynamical and correlation properties of the Internet. *Phys. Rev. Lett.* **87**, 258701.

Pool, I. and Kochen, M., 1978. Contacts and influence. *Social Networks* **1**, 1–48.

Redner, S., 1998. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B* **4**, 131–134.

Stauffer, D. and Aharony, A., 1992. *Introduction to Percolation Theory*. Taylor and Francis, London, 2nd edition.

Strogatz, S. H., 2001. Exploring complex networks. *Nature* **410**, 268–276.

Travers, J. and Milgram, S., 1969. An experimental study of the small world problem. *Sociometry* **32**, 425–443.

Vazquez, A., 2001. Statistics of citation networks. Preprint cond-mat/0105031.

Warren, C. P., Sander, L. M. and Sokolov, I., 2001. Firewalls, disorder, and percolation in epidemics. Preprint cond-mat/0106450.

Watts, D. J., 1999. *Small Worlds*. Princton University Press, Princeton.

Watts, D. J. and Strogatz, S. H., 1998. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442.

3 Emergence of scaling in complex networks

Albert-László Barabási

Abstract

Advances in many fields faced with complex systems, ranging from biology to computer science, are hindered by the limited understanding of the complex webs that characterize the interactions between the system's constituents. These networks assemble and evolve through the addition and removal of nodes and links, dynamical processes that eventually determine their topology. There is increasing evidence that networks appearing in diverse systems share common topological and dynamical features, indicating the existence of robust self-organizing principles and evolutionary laws that govern the interwoven natural and social world around us.

3.1 Introduction

The complexity of many natural and social systems can be attributed to the interwoven web through which the system's constituents interact with each other [1–3]. For example, the cell's ability to sustain its metabolism under often extreme conditions is maintained by a cellular network, whose nodes are substrates and enzymes and links are chemical reactions [4–8]. But equally complex webs describe human societies, whose nodes are individuals and links represent social interactions [9, 10]; the World Wide Web (WWW) [11–14] whose nodes are Web documents connected by URL links; the scientific literature, whose nodes are publications and links are citations [15, 16], or the language, whose nodes are words and links represent various syntaxical or grammatical relationships between them [17–20]. The networks describing these systems constantly evolve by the addition and removal of new nodes and links, such as the accumulation of social links, the appearance of new web pages and links, or the publication of new scientific papers. Due to the diversity and the large number of the nodes and interactions, until recently the topology of these evolving networks was largely unknown and unexplored. Yet, the inability of contemporary science to address the properties of complex networks limited advances in many disciplines, including molecular biology, computer science, ecology and social sciences.

Nonlinear systems have changed our thinking about randomness by demonstrating that relatively simple systems with few degrees of freedom can display unpredictable, chaotic behavior. Recent results on the topology of real networks indicate the emergence of a new paradigm: the apparent randomness of complex systems with many degrees of freedom hides generic mechanisms and order that are crucial to the understanding of the interwoven world surround-

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

ing us. These results have opened new possibilities towards developing a post-reductionist, holistic approach to complex systems, allowing us to follow nature's footsteps and reintegrate the much investigated individual parts into one functional system. By reviewing some advances in the area we hope to convey the potential for understanding complex systems through the evolution and self-assembly of the networks behind them.

3.2 Network models

3.2.1 Random networks

Traditionally the study of complex networks has been the territory of graph theory [21]. While graph theory initially focused on regular graphs, since the 1950's large networks with no apparent design principles were described as random graphs, proposed as the simplest and most straightforward realization of a complex network. Two Hungarian mathematicians, Paul Erdős and Alfréd Rényi, had perhaps the biggest impact on the development of a mathematical theory of random graphs. According to the Erdős-Rényi (ER) model, we start with N nodes and connect every pair of nodes with probability p, creating a graph with approximately $pN(N-1)/2$ edges distributed randomly (Fig. 3.1). This model has guided our thinking about complex networks for decades after its introduction in the late 1950s. But the growing interest in complex systems prompted many scientists to reconsider this modelling paradigm and ask a simple question: are real networks behind such diverse complex systems as the cell or the Internet, fundamentally random? Our intuition offers a clear answer: complex systems must display some organizing principles which should be at some level encoded in their topology as well. But if the topology of these networks indeed deviates from a random graph, we need to develop tools and measures to capture in quantitative terms their underlying organizing principles.

3.2.2 Scale-free networks

Not all nodes in a network have the same number of edges. The spread in the number of edges of the diverse nodes, or a node's degree, is characterized by the degree distribution $P(k)$ which gives the probability that a randomly selected node has exactly k edges. Since in a random graph the edges are placed at random, the majority of nodes have approximately the same degree, close to the average degree $\langle k \rangle$ of the network. Indeed, the degrees in a random graph follow a Poisson distribution with a peak at $\langle k \rangle$.

An unexpected development in our understanding of complex networks was the discovery that for most large networks the degrees do not follow a Poisson distribution. Instead, for a large number of networks, including the World-Wide Web [11], Internet [23], metabolic and protein networks [7, 8], language [17–20] or sexual [22] networks the degree distribution has a power-law tail

$$P(k) \sim k^{-\gamma}. \tag{3.1}$$

Networks with a power-law degree distribution are called scale-free [24]. A non-comprehensive list of scale-free networks reported so far is shown in Table 3.1.

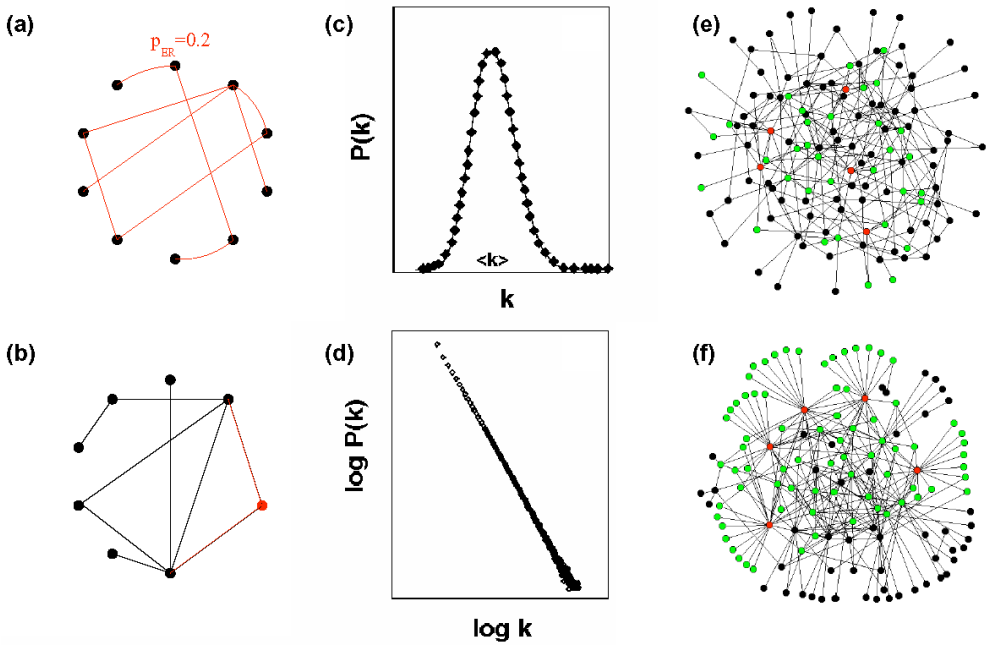


Figure 3.1: (a) The Erdős-Rényi random network model is constructed by laying down N nodes and connecting each pair of nodes with probability p. The figure shows a particular realization of such a network for $N = 10$ and $p = 0.2$. (b) The scale-free model assumes that the network continually grows by the addition of a new nodes. The figure shows the network at time t (black nodes and links) and after the addition of a new node at time $t + 1$ (red). The probability that the new node chooses a node with k links follows (2), favoring highly connected nodes, a phenomenon called preferential attachment. (c) For the random graph generated by the Erdős-Rényi model the degree distribution, $P(k)$, is strongly peaked at $k =< k >$ and decays exponentially for large k. (d) $P(k)$ for a scale-free network does not have a peak, and decays as a power-law, $P(k) \sim k^{-\gamma}$. (e) The random network generated by the Erdős-Rényi model is rather homogeneous, i.e. most nodes have approximately the same number of links. (f) In contrast, a scale-free network is extremely inhomogeneous: while the majority of the nodes have one or two links, a few nodes have a large number of links, guaranteeing that the system is fully connected. To show this, we colored with red the five nodes with the highest number of links, and with green their first neighbors. While in the exponential network only 27% of the nodes are reached by the five most connected nodes, in the scale-free network more than 60% are, demonstrating the key role hubs play in the scale-free network. Note that both networks contain the same number of nodes and links, 130 and 430, respectively. After [47].

There are major topological differences between random and scale-free networks (Fig. 3.1). For the former most nodes have approximately the same number of links, $k \approx \langle k \rangle$, the exponential decay of $P(k)$ guaranteeing the absence of nodes with significantly more links than $\langle k \rangle$. In contrast, the power-law distribution implies that nodes with only a few links are numerous, but a few nodes have a very large number of links. Visually this is similar to the airline routing system: most airports are served only by a few carriers, but there are a few

Table 3.1: The scaling exponents characterizing the degree distribution of several scale-free networks, for which $P(k)$ follows a power-law (3.1). We indicate the size of the network and its average degree $\langle k \rangle$. For directed networks we list separately the indegree (γ_{in}) and outdegree (γ_{out}) exponents, while for the undirected networks, marked with a star, these values are identical. Expanded after Ref. [1].

Network	Size	$\langle k \rangle$	γ_{out}	γ_{in}	Reference
WWW	$325,729$	4.51	2.45	2.1	[11]
WWW	4×10^7	7	2.38	2.1	[55]
WWW	2×10^8	7.5	2.72	2.1	[14]
WWW, site	$260,000$			1.94	[54]
Internet, domain$*$	$3,015$ - $4,389$	3.42 - 3.76	2.1 - 2.2	2.1 - 2.2	[23]
Internet, router$*$	$3,888$	2.57	2.48	2.48	[23]
Internet, router$*$	$150,000$	2.66	2.4	2.4	[56]
Movie actors$*$	$212,250$	28.78	2.3	2.3	[25]
Coauthors, SPIRES$*$	$56,627$	173	1.2	1.2	[57]
Coauthors, neuro.$*$	$209,293$	11.54	2.1	2.1	[53]
Coauthors, math$*$	$70,975$	3.9	2.5	2.5	[53]
Sexual contacts$*$	2810		3.4	3.4	[22]
Metabolic, E. coli	778	7.4	2.2	2.2	[7]
Protein, S. cerev.$*$	1870	2.39	2.4	2.4	[8]
Ythan estuary$*$	134	8.7	1.05	1.05	[58]
Silwood park$*$	154	4.75	1.13	1.13	[58]
Citation	$783,339$	8.57		3	[15]
Phone-call	53×10^6	3.16	2.1	2.1	[59]
Words, conccurence$*$	$460,902$	70.13	2.7	2.7	[20]
Words, synonyms$*$	$22,311$	13.48	2.8	2.8	[19]
Protein, S. Cerev*	$9,85$	1.83	2.5	2.5	[48]
Comic Book Characters	$6,486$	14.9	0.66	3.12	[49]
E-mail	$59,912$	2.88	2.03	1.49	[50]
Protein Domains*	876	9.32	1.6	1.6	[51]
Prot. Dom. (PromDom)*	5995	2.33	2.5	2.5	[52]
Prot. Dom. (Pform)*	2478	1.12	1.7	1.7	[52]
Prot. Dom. (Prosite)*	13.60	0.77	1.7	1.7	[52]

hubs, such as Chicago, New York or Atlanta, from which links emerge to almost all small U.S. airports. Thus, just as the smaller airports, in many real networks the majority of the nodes are "served" only by a few links. These few links are not sufficient to ensure that the network is fully connected, a function guaranteed by a few highly connected hubs, which keep the rest of the nodes together in a single cluster.

3.2.3 Scale-free model

The emergence of the power-law degree distribution has been traced back to two mechanisms that are absent from the classical random network models [24, 25]. First, most networks grow through the addition of new nodes, that link to nodes already present in the system. Indeed, for example, the WWW or the scientific literature, two prototype scale-free networks, continuously expand by the addition of new nodes. Second, most real networks exhibit preferential attachment, i.e. there is a higher probability to link to a node with a large number of connections. Indeed, we link with higher probability to a more connected document on the WWW, or we tend to cite repeatedly much cited papers. These two ingredients, growth and preferential attachment, inspired the introduction of the scale-free model that leads to a network with a power-law degree distribution. The algorithm of the scale-free model is the following [24,25]:

(1) *Growth*: Starting with a small number (m_0) of nodes, at every timestep we add a new node with $m(\leq m_0)$ edges that link the new node to m different nodes already present in the system.

(2) *Preferential attachment*: When choosing the nodes to which the new node connects, we assume that the probability Π that a new node will be connected to node i depends on the degree k_i of node i, such that

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}. \tag{3.2}$$

Numerical simulations indicate that this network evolves into a scale-invariant state with the probability that a node has k edges following a power-law with an exponent $\gamma = 3$ (Fig. 3.2). The scaling exponent is independent of m, the only parameter in the model.

The dynamical properties of the scale-free model can be addressed using various analytic approaches. The continuum theory proposed in [24, 25] focuses on the dynamics of node degrees. Widely used are the master equation approach of Dorogovtsev, Mendes and Samukhin [26] and the rate equation approach introduced by Krapivsky, Redner and Leyvraz [27]. Here we will focus on the continuum theory; for a discussion of the other methods see [1, 2].

Continuum theory: The continuum approach calculates the time dependence of the degree k_i of a given node i. This degree will increase every time a new node enters the system and links to node i, the probability of this process being $\Pi(k_i)$. Assuming that k_i is a continuous real variable, the rate at which k_i changes is proportional to $\Pi(k_i)$. Consequently, k_i satisfies the dynamical equation

$$\frac{\partial k_i}{\partial t} = m\Pi(k_i) = m\frac{k_i}{\sum_{j=1}^{N-1} k_j}. \tag{3.3}$$

The sum in the denominator goes over all nodes in the system except the newly introduced one, thus its value is $\sum_j k_j = 2mt - m$, leading to

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t}. \tag{3.4}$$

The solution of this equation, with the initial condition that every node i at its introduction has

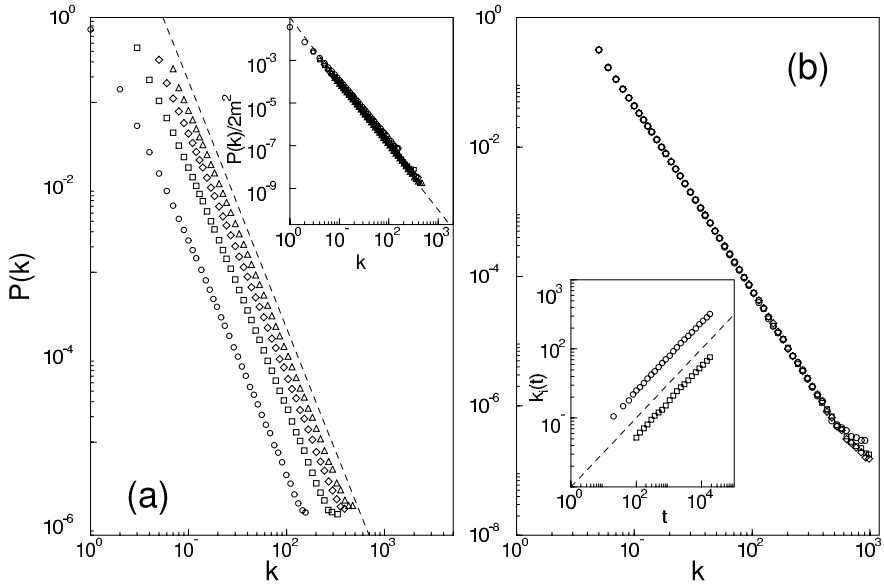


Figure 3.2: (a) Degree distribution of the scale-free model, with $N = m_0 + t = 300,000$ and $m_0 = m = 1$ (circles), $m_0 = m = 3$ (squares), $m_0 = m = 5$ (diamonds) and $m_0 = m = 7$ (triangles). The slope of the dashed line is $\gamma = 2.9$, providing the best fit to the data. The inset shows the rescaled distribution $P(k)/2m^2$ for the same values of m, the slope of the dashed line being $\gamma = 3$. (b) $P(k)$ for $m_0 = m = 5$ and system sizes $N = 100,000$ (circles), $N = 150,000$ (squares) and $N = 200,000$ (diamonds). The inset shows the time-evolution for the degree of two vertices, added to the system at $t_1 = 5$ and $t_2 = 95$. Here $m_0 = m = 5$, and the dashed line has slope 0.5, as predicted by Eq. (3.5). After [25].

$k_i(t_i) = m$, is

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta}, \qquad \text{with} \quad \beta = \frac{1}{2}. \tag{3.5}$$

Equation (3.5) indicates that the degree of all nodes evolves the same way, following a power-law, the only difference being the intercept of the power- law.

Using (3.5), the probability that a node has a degree $k_i(t)$ smaller than k, $P(k_i(t) < k)$, can be written as

$$P(k_i(t) < k) = P(t_i > \frac{m^{1/\beta}t}{k^{1/\beta}}). \tag{3.6}$$

Assuming that we add the nodes at equal time intervals to the network, the t_i values have a constant probability density

$$P(t_i) = \frac{1}{m_0 + t}. \tag{3.7}$$

Substituting this into Eq. (3.6) we obtain that

$$P\left(t_i > \frac{m^{1/\beta}t}{k^{1/\beta}}\right) = 1 - \frac{m^{1/\beta}t}{k^{1/\beta}(t+m_0)}. \tag{3.8}$$

The degree distribution $P(k)$ can be obtained using

$$P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{2m^{1/\beta}t}{m_0 + t}\,\frac{1}{k^{1/\beta+1}}, \tag{3.9}$$

predicting that asymptotically ($t \to \infty$)

$$P(k) \sim 2m^{1/\beta}k^{-\gamma}, \quad \text{with} \quad \gamma = \frac{1}{\beta} + 1 = 3. \tag{3.10}$$

As the power-law observed for real networks describes systems of rather different sizes, it is expected that a correct model should provide a time-independent degree distribution. Indeed, Eq. (3.9) predicts that asymptotically the degree distribution of the scale-free model is independent of time (and, subsequently, independent of the system size $N = m_0 + t$), indicating that despite its continuous growth, the network reaches a stationary scale-free state. Furthermore, Eq. (3.9) also indicates that the coefficient of the power-law distribution is proportional to m^2. All these predictions are confirmed by numerical simulations (see Fig. 3.2).

The scale-free and related models [24–31] view networks as dynamical systems, assuming that they self-assemble and evolve in time through the addition and removal of nodes and links. Such dynamical modeling attempts to capture what nature did when it assembled these networks, expecting that the structural elements and the topology will follow from these. Local decisions about the addition or removal of a link do not aim at global optimization, but try to gain some local advantage to a node such as, for example, enhancing the visibility of a webpage or the content of a scientific paper. The incompleteness of the information available to the local decision maker about the state of the full network, as well as the different interest driving the individual nodes, are the origin of the stochastic component in network evolution.

The scale-free model is the simplest example of an evolving network. In many systems, due to aging and saturation effects that limit the number of links a node can acquire, the preferential attachment function, $\Pi(k_i)$, can be nonlinear, following $\Pi(k_i) = f(k_i)/\sum_j f(k_j)$, where $f(k)$ is an arbitrary function. Nonlinearities in $f(k)$ can result in deviations from the power-law in $P(k)$ [27]. Similarly, the addition and removal of nodes and links can be incorporated by including appropriate terms in $\Pi(k_i)$ [30, 31], changing the exponent γ or the power-law character of $P(k)$. Thus, in contrast with critical phenomena [32], the universal feature of most networks is not reflected by the power-law form of $P(k)$, or the value of the exponent γ. Most complex systems share, however, their dynamical, evolutionary character, captured within the framework provided by evolving networks, indicating that their topology and evolution cannot be divorced from each other.

3.3 Fitness model and Bose-Einstein condensation

In most complex systems nodes vary in their ability to compete for links. For example, some web pages, through a mix of good content and marketing, acquire a large number of links in a

short time, easily leaving behind less popular sites that have been around much longer. A good example is the Google search engine: a relatively latecomer with an excellent product, in less than two years became one of the most connected nodes of the WWW. This competition for links can be incorporated into the scale-free model by adding to each node a fitness, describing its ability to compete for links at the expense of other nodes [33]. For example, a webpage with good and up-to-date content and a friendly interface has a larger fitness than a low quality irregularly updated personal page. Assigning a randomly chosen fitness η_i to each node i modifies the growth rate in the preferential attachment (2) to

$$\Pi(\eta_i, k_i) = \frac{\eta_i k_i}{\sum \eta_j k_j}. \tag{3.11}$$

The competition generated by the different fitnesses leads to multi-scaling: the connectivity of a given node follows $k_i(t) \sim t^{\beta(\eta)}$, where the exponent $\beta(\eta)$ increases with η, allowing fitter nodes with large η to join the network at some later time and overcome the older but less fit nodes.

The competitive fitness models can be mapped exactly into Bose-Einstein condensation, currently one of the most investigated problems in condensed matter physics [34]. Indeed, replacing each node with an energy level of energy (Fig. 3.3)

$$\epsilon_i = \exp(-\beta_B \eta_i), \tag{3.12}$$

where β_B is a dummy variable that plays the role of the Boltzmann constant, and replacing links connected to node i with particles on level ϵ_i, the energy being the quantum mechanical analog of the fitness. As in networks temperature is absent, the behavior of the Bose gas is uniquely determined by the distribution $g(\epsilon)$ from which the random energy levels (fitnesses) are selected. One expects that the functional form of $g(\epsilon)$ is system dependent: the attractiveness of a router for a network engineer comes from a rather different distribution that the fitness of a .com company competing for customers. For a wide class of $g(\epsilon)$ distributions a fits-gets-richer phenomena emerges, in which, while the fittest node acquires more links than its less fit counterparts, there is no clear winner. On the other hand, certain $g(\epsilon)$ distributions can result in Bose-Einstein condensation, which in the network language corresponds to a winner-takes-all phenomenon [34]: the fittest node emerges as a clear winner, developing a condensate by acquiring a finite fraction of the links, independent of the size of the system. While the precise form of the fitness distribution for the WWW or the Internet is not known, it is possible that in the near future $g(\epsilon)$ could be measured, eventually answering the intriguing question: could some complex networks develop a gigantic Bose condensate?

3.4 The Achilles' Heel of complex networks

As the world economy becomes increasingly dependent on the Internet, a much voiced concern arises: Can we maintain its functionality under inevitable failures or frequent hacker attacks? The good news is that so far the Internet has proven rather resilient against failures: While about 0.3% of the routers are down at any moment, we rarely observe major disruptions. A similar question arises in biological systems: despite the frequent errors within the

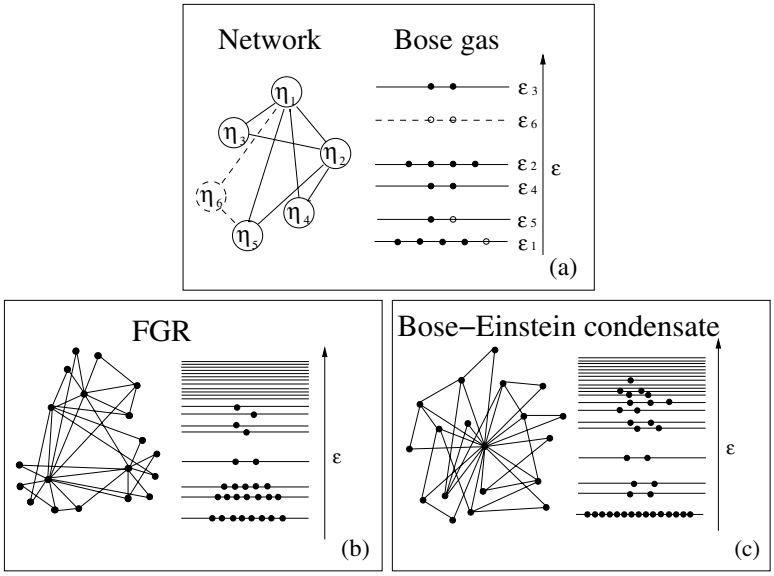


Figure 3.3: Mapping between the network model and a Bose gas. (a) On the left we have a network of five nodes, each characterized by a fitness η_i. Equation (11) assigns an energy ϵ_i to each η_i (right). An edge from node i to node j corresponds to a particle at level ϵ_i and one at ϵ_j. The network evolves by adding a new node (dashed circle, η_6) which connects to $m = 2$ other nodes (dashed lines), chosen following (3.11). In the gas this results in the addition of a new energy level (ϵ_6, dashed) populated by $m = 2$ new particles (open circles), and the deposition of $m = 2$ other particles to energy levels to which the new node is connected to (ϵ_2 and ϵ_5). (b) In the fit-get-rich phase we have a continuous degree distribution, the several high degree nodes linking the low degree nodes together. In the energy diagram this corresponds to a decreasing occupation number with increasing energy. (c) In the Bose-Einstein condensate the fittest node attracts a finite fraction of all edges, corresponding to a highly populated ground level, and sparsely populated higher energies. After [34].

cell, rooted in random mutations or protein misfolding, cells can maintain their basic functions with no or little disruptions. The question is, where does this robustness come from? While there is significant error tolerance built into the protocols that govern packet switching, and into the dynamical loops governing metabolic and genetic networks, lately we are learning that the Internet's scale-free topology also plays a crucial role. In trying to understand the topological component of error tolerance, we get help from percolation, a much studied field of physics [35–37]. Percolation theory tells us that the random removal of nodes from a network will result in an inverse percolation transition: as a critical fraction, f_c, of nodes is removed, the network should fragment into tiny, non-communicating islands of nodes (Figure 4). To our considerable surprise simulations on scale-free networks did not support this prediction: we could remove as many as 80% of the nodes, and the remaining nodes still formed a compact cluster (Figure 5) [38].

The mystery was resolved by Cohen *et. al,* who have shown that as long as the degree exponent γ is smaller than 3 (which is the case for most real networks, including the Internet)

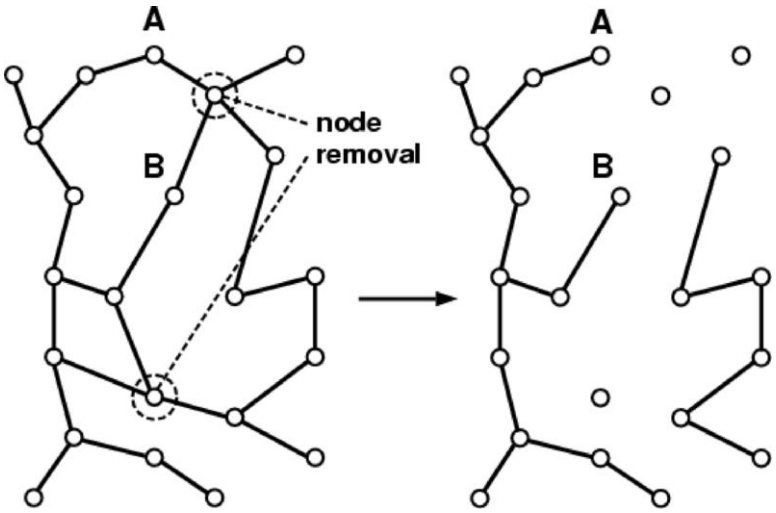


Figure 3.4: Illustration of the effects of node removal on an initially connected network. In the unperturbed state the distance between node A and B is 2, but after two nodes are removed from the system, it increases to 6. In the same time the network breaks into five isolated clusters. After [1].

the critical threshold for fragmentation is $f_c = 1$ [39]. This is a wonderful demonstration that scale-free networks cannot be broken into pieces by the random removal of nodes, a result also supported by the independent calculations of Callaway and collaborators [40]. This extreme robustness to failures is rooted in the inhomogeneous network topology: as there are far more small nodes than hubs, random removal will most likely hit these. But the removal of a small node does not create a significant disruption in the network topology, just like the closure of a small local airport has little impact on international air traffic, explaining the network's robustness against random failures.

The bad news is that the inhomogeneous topology has its drawbacks as well: scale-free networks are rather vulnerable to attacks. Indeed, the absence of a tiny fraction of the most connected nodes will break the network into pieces. Calculations indicate that removing nodes in the order of their degree leads to the rather small critical threshold (f_c) [40, 41]. These findings uncovered the underlying topological vulnerability of scale-free networks: while the Internet is not expected to break under the random failure of its nodes, well informed attackers (crackers) can easily design a scenario to handicap the network.

The error tolerance of complex systems is often attributed to redundancy, guaranteed by multiple paths between most pairs of nodes. Yet, the error tolerance observed in real networks leads to a potential paradigm shift, indicating that redundancy is not sufficient for error tolerance. Indeed, while the number of alternative routes between two nodes in an exponential network is comparable to that in a scale-free network, the former is affected by errors through its increasing diameter, while the latter is not. The robustness of complex networks can be thus attributed to the insignificant role most (non-hub) nodes play in the system, and it is a property that characterizes only scale-free networks, being rooted in the inhomogeneous, hub based topology rather than redundancy.

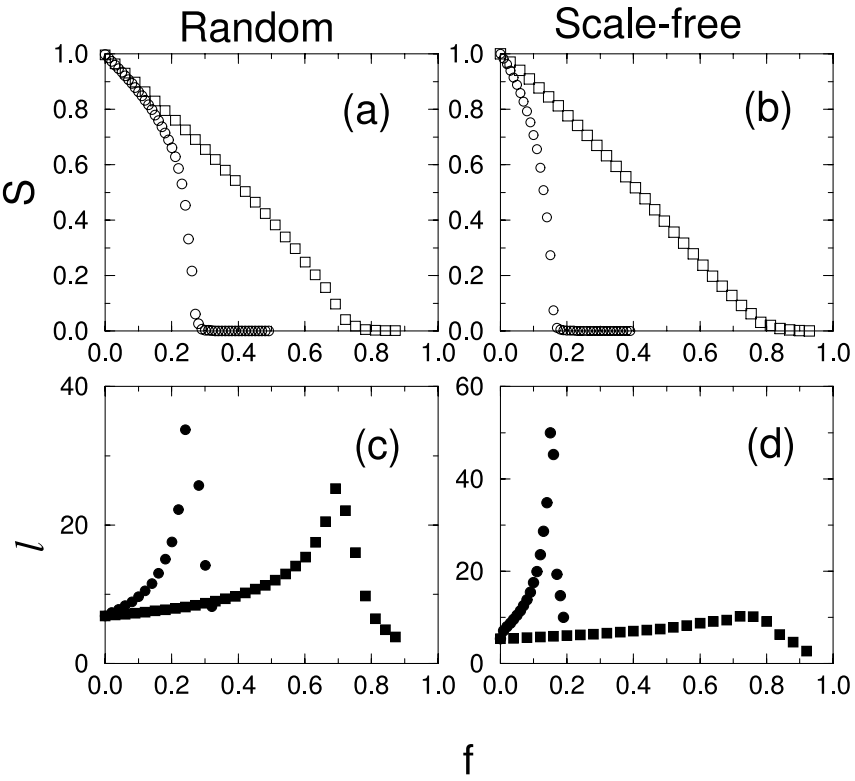


Figure 3.5: The relative size S (a, b) and average path length ℓ (c, d) of the largest cluster in an initially connected network when a fraction f of the nodes are removed. (a, c) Erdős-Rényi random network with $N = 10,000$ and $\langle k \rangle = 4$. (b, d) Scale-free network generated by the BA model with $N = 10,000$ and $\langle k \rangle = 4$. Squares indicate random node removal, while circles correspond to preferential removal of the most connected nodes. After [38]

3.5 A deterministic scale-free model

Stochasticity is a common feature of all network models that generate scale-free topologies. That is, in the scale-free model new nodes use a probabilistic rule to connect to the nodes already present in the system. This randomness, while in line with the major features of networks seen in nature, makes it harder to gain a visual understanding of what makes these networks scale-free, and how do different nodes relate to each other. It would therefore be of major theoretical interest to construct models that lead to scale-free networks in a deterministic fashion. Recently we proposed such a model, demonstrating that we can generate a deterministic scale-free network using a hierarchical construction [42]. Models similar in spirit were proposed and analyzed by several groups [45, 46]. The construction of the model, that follows a hierarchical rule commonly used in deterministic fractals [42, 43], is shown in Figure 3.6. The network is built in an iterative fashion, each iteration repeating and reusing the elements generated in the previous steps as follows:

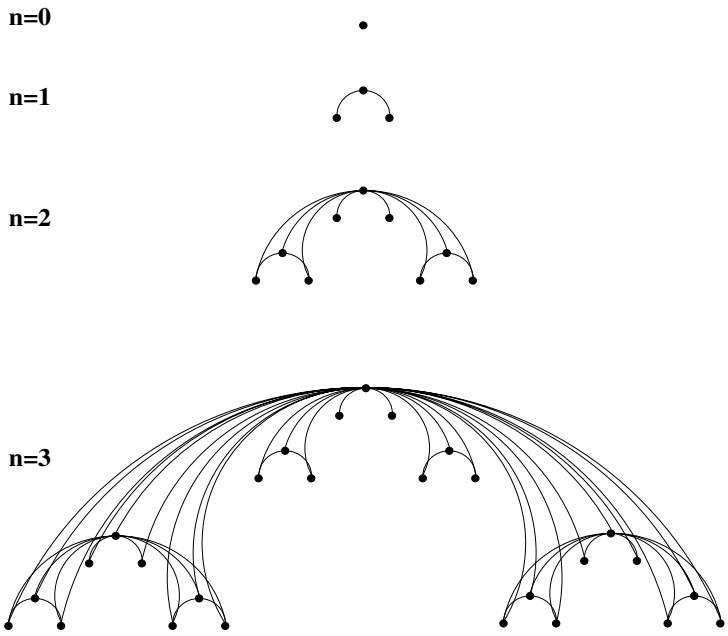


Figure 3.6: Construction of the deterministic scale-free network, showing the first four steps of the iterative process. After [42].

Step 0: We start from a single node, that we designate as the *root* of the graph.

Step 1: We add two more nodes, and connect each of them to the root.

Step 2: We add two units of three nodes, each unit identical to the network created in the previous iteration (step 1), and we connect each of the *bottom* nodes (see Figure 3.6) of these two units to the root. That is, the root will gain four more new links.

Step 3: We add two units of nine nodes each, identical to the units generated in the previous iteration, and connect all eight bottom nodes of the two new units to the root.

These rules can be easily generalized. Indeed, step n would involve the following operation:

Step n: Add two units of 3^{n-1} nodes each, identical to the network created in the previous iteration (step $n-1$), and connect each of the 2^n bottom nodes of these two units to the root of the network.

Thanks to its deterministic and discrete nature, the model described above can be solved exactly. To show its scale-free nature, in the following we concentrate on the degree distribution, $P(k)$.

The tail of the degree distribution is determined by the most connected nodes, or hubs. Clearly the biggest hub is the root, and the next two hubs are the roots of the two units added

to the network in the last step. Therefore, in order to capture the tail of the distribution, it is sufficient to focus on the hubs.

In step i the degree of the most connected hub, the root, is $2^{i+1}-2$. In the next iteration two copies of this hub will appear in the two newly added units. As we iterate further, in the nth step 3^{n-i} copies of this hub will be present in the network. However, the two newly created copies will not increase their degree after further iterations. Therefore, after n iterations there are $(2/3)3^{n-i}$ nodes with degree $2^{i+1}-2$. Since the gap between nodes with consecutive degrees grow with increrasing k, the exponent of the degree distribution can be calculated using the cumulative degree disctribution. The tail of the cumulative degree distribution, determined by the hubs, follows

$$P_{cum}(k) \sim k^{1-\gamma} \sim k^{-\frac{\ln 3}{\ln 2}}.$$

Thus the degree exponent is

$$\gamma = 1 + \frac{\ln 3}{\ln 2}.$$

The origin of this scaling can be understood by inspecting the model's construction. Indeed, at any moment we have a hierarchy of hubs, highly connected nodes which are a common component of scale-free networks. The root is always the largest hub. However, at any step there are two hubs whose connectivity is roughly a half of the root's connectivity, corresponding to the roots of the two units added at step $n-1$. There are six even smaller hubs, with connectivity $2^{n-1}-2$, corresponding to the root of the units added at time $n-2$, and so on. This hierarchy of hubs is responsible for the network's scale-free topology.

3.6 Outlook

The need to understand the evolution and properties of complex networks is not limited to a single discipline. While we briefly touched some important areas, the list of fields that we omitted is probably as prominent as those mentioned: social networks, neural networks (brain), genetic regulatory networks, networks in semantics and linguistics, business networks, river networks, to name only a few [9,44]. Recent advances are the result of highly interdisciplinary studies, that compare networks appearing in often distant fields. Despite the different nature of nodes and links, these networks share a number of topological features, rooted in the inherently similar mechanisms that lead to their assembly, growth and evolution. While it is difficult to identify such universal characteristics from single examples, once they are uncovered, they offer strong support for an emerging theme: networks in nature are far from being random, but they evolve following robust self-organizing principles and evolutionary laws that cross disciplinary boundaries. Progress is possible only if the numerical and analytical work is combined with empirical studies on real networks, potentially opening an unexpectedly revealing window on the structure of complex systems. The results uncovered so far likely represent only the tip of the iceberg, and systematic data driven studies focusing on the topology and evolution of real networks could fundamentally change how we approach the complex world around us.

3.7 Acknowledgments

We thank R. Albert, G. Bianconi, I. Derényi, Z. Dezső, I. Farkas, H. Jeong, Z. Néda, Z.N. Oltvai, E. Ravasz, T. Vicsek and S. H. Yook for comments and suggestions. The National Science Foundation, National Institute of Health, and Department of Energy provided support.

References

[1] R. Albert, A.-L. Barabási, Statistical Mechanics of Complex Networks, Rev. Mod. Phys. 74, 67-97 (2002).

[2] S. N. Dorogotsev, & J. F. F. Mendes, Evolution of Random Networks, Adv. Phys. (2002).

[3] H. G. Schuster, *Complex Adaptive Systems* (Scator Verlag, Saarbrücken, Germany 2002).

[4] L.H. Hartwell, J.J. Hopfield, S. Leibler, A.W. Murray, From molecular to modular cell biology, Nature 402, C47-52 (1999).

[5] R. Overbeek, et al. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction, Nucleic Acids Res. 28, 123-125 (2000).

[6] P.D. Karp, et al, The EcoCyc and MetaCyc databases, Nucleic Acids Res. 28, 56-59 (2000).

[7] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, & A.L. Barabási, The large-scale organization of metabolic networks, Nature 607, 651 (2000).

[8] H. Jeong, S. P. Mason, S. N. Oltvai,& A.-L. Barabási, Nature, London. 411, 41 (2001).

[9] M. Kochen, (ed.), *The Small World* (Ablex, Norwood, NJ,1989).

[10] S. Wasserman, & K. Faust, Social Network Analysis: Methods and Applications, Cambridge University, Cambridge, (1994).

[11] R. Albert, H. Jeong, & A.-L. Barabási, Diameter of the World-Wide Web, Nature 400, 130-131 (1999).

[12] S. Lawrence, and C.L. Giles, Accessibility of information on the web, Nature 400, 107-110 (1999).

[13] J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, & A. Tomkins, The web as a graph: Measurements, models and methods. Proc. of the Int. Conf. on Combinatorics and Computing (1999).

[14] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajalopagan, R. Stata, A. Tomkins, and J. Wiener, Comput. Netw. 33, 309 (2000).

[15] S. Redner, How popular is your paper? An empirical study of the citation distribution, European Phys. J. B4, 131–135 (1998).

[16] D. J. de Solla Price, Networks of Scientific Papers, Science 169, 510-515 (1965).

[17] M. Sigman, G. Cecchi, Global Organization of the Worldnet Lexicon, Proc. Nat. Acad. Sci. 99, 1792 (2002).

[18] S. N. Dorogovtsev, J. F. F. Mendes, Language as an Evolving World Web, Proc. Reg. Soc., London B 258, 2603 (2001).

[19] S. Yook, H. Jeong, & A.-L. Barabási, unpublished (2001).

[20] R., Ferrer i Cancho, and R. V. Solé, The Small World of Human Language, Proc. Roy. Soc., London B 268, 22 61 (2001).

[21] B. Bollobás, *Random Graphs* (Academic, London, 1985).

[22] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, & Y. Aberg, Nature, London 411, 907 (2001).

[23] M. Faloutsos, P. Faloutsos, & C. Faloutsos, On Power-Law Relationships of the Internet Topology. ACM SIGCOMM 99, Comp. Comm. Rev. 29, 251-260 (1999).

[24] A.-L. Barabási, & R. Albert, Emergence of scaling in random networks, Science 286, 509-512 (1999).

[25] A.-L. Barabási, R. Albert, & H. Jeong, Mean-field theory for scale-free random networks, Physica A 272, 173-187 (1999).

[26] S. N. Dorogovtsev, & J. F. F. Mendes, and A. N. Samukhin, Structure of growing networks: Exact solution of the Barabási-Albert, model, Phys. Rev. Lett. 85, 6633 (2000).

[27] P. L. Krapivsky, S. Redner, F. Leyvraz, Phys. Rev. Lett. 86, 5901 (2000).

[28] S.N. Dorogovtsev, & J.F.F. Mendes, Evolution of reference networks with aging, Phys. Rev. E 62, 1862 (2000).

[29] L.A.N. Amaral, A. Scala, M. Barthelemy, & H.E. Stanley, Classes of behavior of small-world networks, Proc. Nat. Acad. Sci. 97, 11169 (2000).

[30] R. Albert, & A.-L. Barabási, Topology of evolving networks: local events and universality, Phys. Rev. Lett. 85, 5234 (2000).

[31] S. N. Dorogovtsev, & J.F.F. Mendes, Scaling Behavior of Developing and Decaying Networks, Europhys. Lett. 52, 33 (2000).

[32] H.E. Stanley, *Introduction to Phase Transitions and Critical Phenomena* (Oxford University Press, New York, 1971).

[33] G. Bianconi, & A.-L., Barabási, Competition and multiscaling in evolving networks, Europhys. Lett. 54, 436 (2001).

[34] G. Bianconi, & A.-L., Barabási, Bose-Einstein coordination in evolving networks, Phy. Rev. Lett. 86, 5632 (2001).

[35] D. Stauffer, & A. Aharony, *Introduction to Percolation Theory* (Taylor & Francis, London 1992).

[36] A. Bunde, & S. Havlin, Eds., *Fractals in Science* (Springer, Berlin, 1994).

[37] A. Bunde, & S. Havlin, Eds., *Fractals and Disordered Systems* (Springer, Berlin, 1996).

[38] R. Albert, H. Jeong, & A.-L., Barabási, Error and attack tolerance of complex networks, Nature, London, 406, 378, 2001; 409, 542 (2000).

[39] R. Cohen, K. Erez, D. ben-Avraham, & S. Havlin, Phys. Rev. Lett. 85, 4626 (2000).

[40] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, & D. J., Watts, Phys. Rev. Lett. 85, 5468 (2000).

[41] R. Cohen, K. Erez, D. ben-Avraham, & S. Havlin, Phys. Rev. Lett. 86, 3682 (2001).

[42] A.-L. Barabási, E. Ravasz, T. Vicsek, Deterministic Scale-Free Networks, Physica A 299, 559 (2001).

[43] B. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1982).

[44] J. R. Banavar, A. Maritan, & A. Rinaldo, Size and form in efficient transportation networks, Nature 399, 130-132 (1999).

[45] S. N. Dorogovtsev, A. V. Goltsev, & J. F. F. Mendes, Pseudofractal Scale-Free Web, *cond-mat* 0112143, (2002).

[46] S. Jang, S. Kim, B. Kahng, A Geometric Fractal Growth Model for Scale-Free Networks, *cond-mat* 0112361, (2002).

[47] A. L. Barabási, The Physics of the Web, Physics World 33-38 (July 2001).

[48] A. Wagner, The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes, Mol. Biol. Evol. 18, 1283 (2001).

[49] R. Alberich, J. Miro-Julia, F. Rosselló, Marvel Universe looks almost like a real social network, cond-mat 0202174 (2002).

[50] H. Ebel, L. I. Mielsch, S. Bornholdt, Scale-Free Topology of E-mail Networks, cond-mat 0201476 (2002), Phys. Rev. E, in press.

[51] J. Park, M. Lappe, A. Teichmann, Mapping Protein Family Interations, J. Mol. Biol. 307, 929-938 (2001).

[52] S. Wuchty, Scale-Free Behavior in Protein Domain Networks, Mol. Biol. Evol. 18, 1699-1702 (2001).

[53] A. L. Barabási, H. Jeong, E. Ravasz, Z. Neda, T. Vicsek, A. Schubert, cond-mat 0104162, Physica A (in press) (2002).

[54] Huberman, B. A.,& Adamic, L. A. Growth dynamics of the World-Wide Web, Nature 401, 131 (1999).

[55] Kumar, R., P. Raghavan, S. Rajalopagan, & A. Tomkins, Proceedings of the 9th ACM Symposium on Principles of Database Systems, p. 1. (1999).

[56] Govindan, R., & H. Tangmunarunkit, Proceedings of IEEE INFOCOM 2000, Tel Aviv, Israel, IEEEE, Piscataway, N.J., Vol. 3, p. 1371 (2000).

[57] M. E. J., Newman, Phys. Rev. E 64, 016131 (2000).

[58] J. M. Montoya and & R. V. Solé, Small World Patterns in Food Webs, J. Theor. Biol. 216, 605 (2002).

[59] W., F. Chung, & L. Lu, Proceedings of the 32nd ACM Symposium on the Theory of Computing, ACM, New York, p. 271 (2000).

4 Structural properties of scale-free networks

Reuven Cohen, Shlomo Havlin, and Daniel ben-Avraham

Abstract

Many networks have been reported recently to follow a scale-free degree distribution in which the fraction of sites having k connections follows a power law: $P(k) = ck^{-\lambda}$. In this chapter we study the structural properties of such networks. We show that the average distance between sites in scale-free networks is much smaller than that in regular random networks, and bears an interesting dependence on the degree exponent λ. We study percolation in scale-free networks and show that in the regime $2 < \lambda < 3$ the networks are resilient to random breakdown and the percolation transition occurs only in the limit of extreme dilution. On the other hand, attack of the most highly connected nodes easily disrupts the nets. We compute the percolation critical exponents and find that percolation in scale-free networks is non-universal, i.e. depends on λ and different from the mean-field behavior in dimensions $d \geq 6$. Finally, we suggest a novel and efficient method for immunization against the spread of diseases in social networks, or the spread of viruses and worms in computer networks.

4.1 Introduction

4.1.1 Random graphs

Graph theory is rooted in the 18th century beginning with the work of Euler. A graph in its mathematical definition is a pair of sets (V, E), where V is a set of vertices (the nodes of the graph), and E is a set of edges, denoting the links between the vertices. In a directed graph, the edges are taken as ordered pairs, *i.e.*, each edge is directed from the first to the second vertex of the pair.

The work on graph theory has mainly dealt with the properties of special graphs. In the 1960s, Paul Erdős and Alfréd Rényi initiated the study of random graphs [1–3]. Random graph theory is, in fact, not the study of graphs (as there is no such thing as a "random graph"), but the study of an ensemble of graphs (or, as mathematicians prefer to call it, a *probability space* of graphs). The ensemble is a class consisting of many different graphs, where each graph has a probability attached to it. A property studied is said to exist with probability P if the total probability of all the graphs in the ensemble of having that property is P. This structure allows the use of probability theory in conjunction with discrete mathematics for the study of graph ensembles.

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

Two well-studied graph ensembles are $G_{N,M}$ — the ensemble of all graphs having N vertices and M edges, and $G_{N,p}$ — consisting of graphs with N vertices, where each possible edge is realized with probability p. These two families, initially studied by Erdős and Rényi themselves, are known to be similar if $M = \binom{N}{2}p$, so long as p is not too close to 0 or 1 [4], and are referred to as ER graphs. Examples of other well-studied ensembles are the family of *regular* graphs, where all nodes have the same number of edges, $P(k) = \delta_{k,k_0}$, and the family of *unlabeled* graphs, where graphs which are isomorphic under permutations of their nodes are considered the same object.

An important attribute of a graph is the average degree, *i.e.*, the average number of edges connected to each node. We shall denote the degree of the ith node by k_i and the average degree by $\langle k \rangle$. N-vertex graphs with $\langle k \rangle = \mathcal{O}(N^0)$ are called sparse graphs. In what follows, we concern ourselves exclusively with sparse graphs.

An interesting characteristic of the ensemble $G_{N,p}$ is that many of its properties have a related threshold function, $p_t(N)$, such that if $p < p_t$ the property exists with probability 0, in the "thermodynamic limit" of $N \to \infty$, and with probability 1 if $p > p_t$. This phenomenon is similar to the physical notion of a phase transition. An example of such a property is the existence of a giant component, *i.e.*, a set of connected nodes, in the sense that a path exists between any two of them, whose size is proportional to N. Erdős and Rényi showed [2] that for ER graphs such a component exists if $\langle k \rangle > 1$. If $\langle k \rangle < 1$ only small components exist, and the size of the largest component is proportional to $\ln N$. Exactly at the threshold, $\langle k \rangle = 1$, a component of size proportional to $N^{2/3}$ emerges. This phenomenon was described by Erdős as the "double jump". Another property is the average path length distance between any two sites, which in almost every graph of the ensemble is of order $\ln N$.

4.1.2 Scale-free networks

The Erdős-Rényi model has been traditionally the dominant subject of study in the field of random graphs. Recently, however, several studies of real world networks have indicated that the ER model fails to reproduce many of their observed properties.

One of the simplest properties of a network that can be measured directly is the degree distribution, or the fraction $P(k)$ of nodes having k connections (degree k). A well-known result for ER networks is that the degree distribution is Poissonian, $P(k) = e^{-z}z^k/k!$, where $z = \langle k \rangle$ is the average degree [4].

Direct measurements of the degree distribution for networks of the Internet [5, 6], WWW (where hypertext links constitute directed edges) [7, 8], e-mail network [9], citations of scientific articles [10], metabolic networks [11, 12], trust network [13], and many more, show that the Poisson law does not apply. Rather, most often these nets exhibit a scale-free degree distribution:

$$P(k) = ck^{-\lambda}, \qquad k = m, ..., K \tag{4.1}$$

where $c \approx (\lambda - 1)m^{\lambda-1}$ is a normalization factor, and m and K are the lower and upper cutoffs for the connectivity of a node, respectively. The divergence of moments higher than $\lceil \lambda - 1 \rceil$ (as $K \to \infty$ when $N \to \infty$) is responsible for many of the special properties attributed to scale-free networks.

All real-life networks are finite (and all their moments are finite). The actual value of the cutoff K plays an important role. It may be approximated by noting that the total probability of nodes with $k > K$ is of order $1/N$ [14, 15]:

$$\int_K^\infty P(k)\, dk \sim 1/N. \tag{4.2}$$

This yields the result

$$K \sim mN^{1/(\lambda-1)}. \tag{4.3}$$

The degree distribution does not characterize the graph or ensemble in full. There are other quantities, such as the degree-degree correlation (between connected sites), the spatial correlations, etc. Several models have been presented for the evolution of scale-free networks, each of which may lead to a different ensemble. The first suggestion was the *preferential attachment* model by Barabási and Albert, which came to be known as the "Barabási-Albert model" [5] . Several variants have been suggested to this model (see, *e.g.*, [16, 17]). In this Chapter we will concentrate on the "Molloy-Reed construction" [18–20], which ignores the evolution and assumes only the degree distribution and no correlations between nodes. Thus, the site reached by following a link is independent of the origin.

Scale-free distributions have been studied in physics, particularly in the context of fractals and of Levyflights@Lévy flights. Fractals are objects which appear similar (at least in some statistical sense) at every lengthscale [21, 22]. Many natural objects, such as mountains, clouds, coastlines and rivers, as well as the cardiovascular and nervous systems are known to be fractals. This is why we find it hard to distinguish between a photograph of a mountain and that of part of the mountain, neither can we ascertain the altitude from which a picture of a coastline had been taken. Diverse phenomena, such as the distribution of earthquakes, biological rhythms and rates of transport of data packets in communication networks, are also known to posses a scale-free distribution. They come in all sizes and rhythms, spanning many orders of magnitude.

Lévy flights were suggested by Paul Lévy [23], who was studying what is now known as Levystable@Lévy stable distributions. The question he asked was, When is the length distribution of a single step in a random walk similar to that of the entire walk? Besides the known result, that of the Gaussian distribution, Lévy found an entire new family — essentially that of scale-free distributions. Stable distributions do not obey the central limit theorem (stating that for large numbers of steps the distribution of the total displacement tends to Gaussian), due to the divergence of the variance of indivual steps. Lévy walks have numerous applications [24, 25]. An interesting observation is that animal foraging patterns which follow stable distributions have been shown to be their most efficient strategy [26, 27]. For recent reviews on complex networks and in particular scale free networks see Refs. [28, 29].

4.2 Small and Ultra-small worlds

Regular lattices are embedded in Euclidean space, of a well-defined dimension, d. This means that $n(r)$, the number of sites within a distance r from an origin, grows as $n(r) \sim r^d$ (for

large r). For fractal objects d in the last relation is non-integer and is replaced by the fractal dimension $d_{\rm f}$. Similarly, the chemical dimension, d_l, is defined by the scaling of the number of sites within l edges or less from a given site (an origin), $n(l) \sim l^{d_l}$. A third dimension, $d_{\rm min}$, relates between the chemical path (the shortest distance along edges) and Euclidean distances, $l \sim r^{d_{\rm min}}$. It satisfies $d_{\rm min} = d_{\rm f}/d_l$ [21, 22, 30].

An example of an object where these concepts fail is the Cayley tree (also known as the Bethe lattice). The Cayley tree is a regular graph, of fixed degree z, and no loops. It has been studied by physicists in many contexts, since its simplicity often allows for exact analyses. An infinite Cayley tree cannot be embedded in a Euclidean space of finite dimensionality. The number of sites at l is $n(l) \sim (z-1)^l$. Since the exponential growth is faster than any power-law, Cayley trees are referred to as infinite-dimensional systems.

In most random network models the structure is locally tree-like (since most loops occur only for $n(l) \sim N$), and, since the number of sites grows as $n(l) \sim \langle k-1 \rangle^l$, they are also infinite-dimensional. As a consequence, the diameter of such graphs (*i.e.*, the minimal path between the most distant nodes) scales like $D \sim \ln N$ [4]. This small diameter is to be contrasted with that of finite-dimensional lattices, where $D \sim N^{1/d_l}$.

Recently, a model has been suggested by Watts and Strogatz [31, 32] which retains the local high clustering of lattices while reducing the diameter to $D \sim \ln N$. This, so called, small world network is achieved by replacing a fraction ϕ of the links in a regular lattice with random links, to random distant neighbors. (In other variants of the small world model the "long range" links are simply added on, without prior removal of lattice links.) A study of scale-free networks embedded in Euclidean space (at the obvious price of a cutoff in k) which exhibit finite dimensions can be found in [33].

4.2.1 Diameter of scale-free networks

We now aim to show that scale-free networks with degree exponent $2 < \lambda < 3$ possess a diameter $D \sim \ln \ln N$, smaller even than that of ER and small world networks. If the network is fragmented, we will only be interested in the diameter of the largest cluster (assuming there is one). Our analysis of the diameter of the Molloy-Reed scale-free networks is based on [34].

We adopt a different definition of diameter: the *average* distance between any two sites on the graph. We find it easier still to focus on the radius of a graph, $L \equiv \langle l \rangle$: the average distance of all sites from the site of highest degree in the network (if there is more than one, we pick one arbitrarily). The diameter of the graph, D, is restricted to:

$$L \leq D \leq 2L, \tag{4.4}$$

and thus essentially scales like L.

4.2.2 Minimal graphs and lower bound

We begin by showing that the radius of any scale-free graph with $\lambda > 2$ has a rigorous lower bound that scales as $\ln \ln N$. It is easy to convince oneself that the smallest diameter of a graph, of a given degree distribution, is achieved by the following construction: Start with the highest degree site, then connect to each successive layer the extant sites of highest degree, until the layer is full. By construction, loops will occur only in the last layer.

Let the number of links outgoing from the lth shell (layer) be χ_l. Let K_l denote the highest degree of a site not yet reached by the lth layer. Then, for the graph of minimal diameter described above,

$$\chi_l = N \int_{K_{l+1}}^{K_l} P(k)dk \approx m^{\lambda-1} N K_{l+1}^{1-\lambda}. \tag{4.5}$$

The number of links outgoing from layer $l+1$ equals the total number of links in all the sites between K_l and K_{l+1} minus one link for each site — the one used to connect to the previous layer:

$$\chi_{l+1} = N \int_{K_{l+1}}^{K_l} (k-1)P(k)dk \approx \frac{\lambda-1}{\lambda-2} m^{\lambda-1} N K_{l+1}^{2-\lambda}. \tag{4.6}$$

Solving these recursion relations, with the initial conditions $K_0 = N^{1/(\lambda-1)}$ and $\chi_0 = K_0$, leads to:

$$\chi_l = a^{(\lambda-1)(1-u^l)} N^{1-u^{l+1}}, \tag{4.7}$$

where $a = (\lambda-1)/(\lambda-2)m$, $u = (\lambda-2)/(\lambda-1)$, and

$$K_l = m(\chi_l/N)^{\frac{1}{1-\lambda}} \ . \tag{4.8}$$

To bound the radius L of the graph, we will assume that the low degree sites are connected randomly to the giant cluster. We pick a site of degree $1 \ll k^* \ll (\ln\ln N)^{1/(\lambda-1)}$. Using Eq. (4.8) we can show that if $l_1 \approx \ln\ln N/\ln(\lambda-2)$ then $K_{l_1} < k^*$, so, with probability 1 all sites of degree $k \geq k^*$ lie within l_1 layers from the site we picked. On the other hand, if we start uncovering the graph from any site — provided it belongs to the giant component — then within a distance l_2 from this site there are at least l_2 bonds. The probability that none of those bonds leads to our site (of degree k^*) is $(1-P(k^*)k^*/\langle k\rangle)^{l_2}$. That is, if $l_2 k^* P(k^*)/\langle k\rangle \gg 1$, at least one bond will lead to our site. Thus, taking $k^{*\lambda-1} \ll l_2 \ll \ln\ln N$, we will definitely reach a site of at least degree k^* in the l_2th layer from almost any site. Since $l = l_1 + l_2$, all sites are at a distance of order $\ln\ln N$ from the highest degree site, and $L \sim \ln\ln N$ is a rigorous lower bound for the diameter of scale-free networks with $\lambda > 2$.

4.2.3 The general case of random scale-free networks

We now argue that the scaling of $D \sim \ln\ln N$ is actually realized in the general case of *random* scale-free graphs with $2 < \lambda < 3$. One can view the process of uncovering the network as actually building it, by following the links one at a time. For simplicity, let us start with the site of highest degree, $K \sim N^{1/(\lambda-1)}$ (guaranteed to belong to the giant component). Next, we expose the layers one at a time. We view the graph as built from one large developing cluster, and sites which have not yet been reached (they might or might not belong to the giant component), see Fig. 4.1. A similar consideration has been used by Molloy and Reed [19].

After l layers are explored the distribution of the yet unreached sites changes (since most high-degree sites are exposed first) to $P'(k) \approx P(k)exp(-k/K_l)$ [19].

Let us now consider layer $l+1$. A threshold function emerges: the new distribution of unvisited sites behaves like a step function — almost $P(k)$ for $k < K_{l+1}$, and 0 for $k > K_{l+1}$. The reason for this is as follows. A site with degree k has a probability of $p = k/(N\langle k\rangle)$ to be reached by following a link[1]. If there are χ_l outgoing links then for $p\chi_l > 1$ we can assume that, in the limit $N \to \infty$, the site will be reached in the next level with probability 1. Therefore, all unvisited sites with degree $k > N/\chi_l$ will be surely reached in the next chemical layer. On the other hand, almost all the unvisited sites with degree $k < N/\chi_l$ will remain unvisited in the next layer and their distribution will remain virtually unchanged. From these considerations, the highest degree of the unexplored sites in layer $l+1$ is determined by:

$$K_{l+1} \approx N/\chi_l. \tag{4.9}$$

In layer $l+1$ all sites with degree $k > N/\chi_l$ will be exposed. Since the probability of reaching a site via a link is proportional to $kP(k)$, the average degree of sites reached by following a link is $\kappa \equiv \langle k^2\rangle/\langle k\rangle$ [14]. κ for layer l can be computed from the general formula [14], valid for scale-free distributions,

$$\kappa = \left(\frac{\lambda-2}{\lambda-3}\right)\left(\frac{K^{3-\lambda}-m^{3-\lambda}}{K^{2-\lambda}-m^{2-\lambda}}\right), \tag{4.10}$$

but with the layer cutoff K of (4.9). That is, $\kappa_l \sim K_{l+1}^{3-\lambda}$.

Using the above consideration, the number of outgoing links from layer $l+1$ can be computed. Consider the total degree of all sites reached in the $l+1$ level. This includes all sites with degree k, $K_{l+1} < k < K_l$, as well as other sites with average degree proportional to $\kappa - 1$ (the -1 is due to one link going into the shell). Thus, the value of κ (the avergae number of links for a site reached via a link) is calculated using the cutoff K_{l+1}. (Loops within a layer, and multiple links connecting a site in layer $l+1$, can be neglected as long as the number of sites in the layer is less than order N, $N \to \infty$.) The two contributions can be written as the sum of two terms:

$$\chi_{l+1} \approx N\int_{K_{l+1}}^{K_l}(k-1)P(k)dk + \chi_l\left[\kappa(K_{l+1})-1\right]. \tag{4.11}$$

Noting that $P(k) \propto k^{-\lambda}$ and that $\kappa \propto K^{3-\lambda}$ [14], it follows that $\chi_{l+1} \propto NK_{l+1}^{2-\lambda}$ (note that both terms in Eq. (4.11) scale similarly). This results in a second recurrence equation:

$$\chi_{l+1} = ANK_l^{2-\lambda}, \tag{4.12}$$

where $A = \frac{\lambda-1}{\lambda-2}m^{\lambda-1} + \frac{\lambda-2}{3-\lambda}m^{\lambda-2}$.

Solving the equations (4.9) and (4.12) yields

$$\chi_l \sim A^{\frac{(\lambda-2)^l-1}{\lambda-3}}N^{1-\frac{(\lambda-2)^{l+1}}{\lambda-1}}, \tag{4.13}$$

where χ_l is the number of outgoing links from the lth layer. Eq. (4.9) then leads to:

$$K_l \sim A^{\frac{(\lambda-2)^{l-1}-1}{3-\lambda}}N^{\frac{(\lambda-2)^l}{\lambda-1}}. \tag{4.14}$$

[1] We assume that $\langle k\rangle$ for the unvisited sites is fixed, since it is dominated by the low-degree nodes, whose distribution is unchanged.

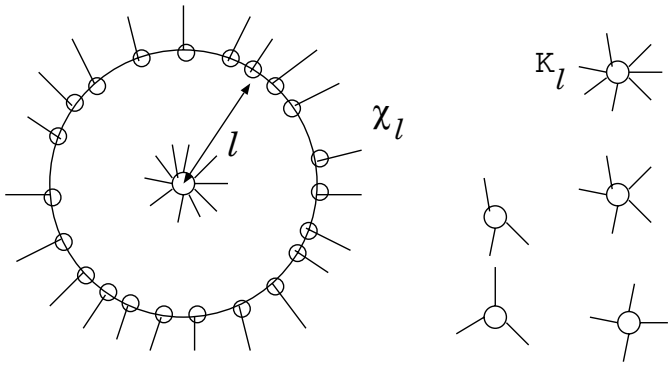


Figure 4.1: Illustration of the exposure process. The large circle denotes the exposed fraction of the giant component, while the small circles denote individual sites. The sites on the right have not been reached yet. After [34]

Using the same considerations that follow Eq. (4.8), one can deduce that also here

$$D \sim \ln \ln N. \tag{4.15}$$

Our result that $D \sim \ln \ln N$ is consistent with the observations that the distance in the Internet network is extremely small, and that the distance in metabolic scale-free networks is almost independent of N [11].

For $\lambda > 3$ and $N \gg 1$, κ is independent of N, and since the second term of Eq. (4.11) is dominant, Eq. (4.11) reduces to $\chi_{l+1} = (\kappa - 1)\chi_l$, where κ is a constant depending only on λ. This leads to the known result $\chi_l \approx C(N, \lambda)(\kappa - 1)^l$ and the radius of the network [35] is

$$L \sim \ln N. \tag{4.16}$$

For $\lambda = 3$, Eq. (4.11) reduces to $\chi_{l+1} = \chi_l \ln \chi_l$. Taking the logarithm of this equation one obtains $\ln \chi_{l+1} - \ln \chi_l = \ln \ln \chi_l$. Defining $g(l) = \ln \chi_l$ one obtains a difference equation which may be approximated (in the continuum limit) by $g' = \ln g$. Substituting $u = \ln g$, the equation reduces to

$$L = \int_{\ln \ln \sqrt{N}}^{\ln \ln N} e^{u - \ln u} du. \tag{4.17}$$

The lower bound is obtained from the highest degree site for $\lambda = 3$, having degree $K = m\sqrt{N}$. Thus, $\chi_0 = m\sqrt{N}$. The upper bound results from that l for which $\chi_l \sim N$ (and lower order corrections). The integral in Eq. (4.17) can be approximated by the steepest descent method, leading to

$$L \sim \ln N / (\ln \ln N), \tag{4.18}$$

(assuming $\ln \ln N \gg 1$).

The result of Eq. (4.18) has been obtained rigorously for the maximum distance in the Barabasi-Albert (BA) model [5], where $\lambda = 3$ (for $m \geq 2$) [36]. Although the result in [36] applies to the largest distance between two sites, their derivation leaves no doubt that the average distance would behave similarly. For $m = 1$, the graphs in the Barabasi-Albert model turn into trees, and the behavior of $D \sim \ln N$ is obtained [36, 37]. It should be noted that for $m = 1$ the giant component in the random model contains only a fraction of the sites (while for $m \geq 2$ it contains all sites — at least to leading order). This might explain why exact trees and BA trees are different from Molloy-Reed random graphs.

4.3 Percolation

Since the 1940s percolation has been the subject of intense studies among physicists and mathematicians. In site percolation, usually defined on lattices, the sites (nodes) are present (or *occupied*, or *wet*) with probability q, or equivalently, removed (blocked) with probability $p = 1 - q$. The (infinite) network undergoes a sharp phase transition at a critical threshold q_c, from a connected, or percolating phase, where a spanning cluster runs across the entire size of the system, for $q > q_c$, to a fragmented phase, where only finite clusters exist, for $q < q_c$. An introduction to the general subject of percolation can be found in [22, 30, 38][2]

The percolation transition is continuous (second order), and near the transition point many properties behave as power laws. For example, the probability for a site to be in the spanning cluster (for $q > q_c$) grows as $P_\infty \sim (q - q_c)^\beta$, and the number of clusters of size s, at $q = q_c$, is $n_s \sim s^{-\tau}$. The critical exponents β, τ, and their likes, are universal, depending only upon the dimensionality d of the lattice. For $d \geq d_c = 6$ the lattice dimension no longer plays a significant role and the critical exponents assume their mean-field values (*e.g.*, $\beta = 1$, $\tau = 5/2$). The mean-field case can be conviniently obtained from the exact solution of the percolation problem on Cayley trees (whose effective dimension is infinite). Percolation in ER graphs also follows mean-field behavior. The percolation threshold is $q_c = 1 - p_c = 1/(z-1)$ for Cayley trees, and $q_c = 1/\langle k \rangle$ for ER graphs.

The problem of percolation on scale-free networks has important practical applications. Below we explore consequences to the resilience of the Internet in the face of random breakdown of servers as well as under intentional attack, and to immunization strategies against the spread of contagious epidemics in population and computer networks.

4.3.1 Random breakdown

For a graph having degree distribution $P(k)$ to have a spanning cluster, a site j which is reached by following a link (from site i on) the giant cluster must have at least one other link, on average, to allow the cluster to exist. For this to happen the average degree of site j must be at least 2 (one incoming and one outgoing link), given that site i is connected to j:

$$\langle k_i | i \leftrightarrow j \rangle = \sum_{k_i} k_i P(k_i | i \leftrightarrow j) = 2. \tag{4.19}$$

[2] We have exchanged here the traditional roles of p and q in percolation theory, $p \leftrightarrow q$, to conform with what seems to be the norm in papers on scale-free graphs.

Using Bayes' rule we get

$$P(k_i|i \leftrightarrow j) = P(k_i, i \leftrightarrow j)/P(i \leftrightarrow j) = P(i \leftrightarrow j|k_i)P(k_i)/P(i \leftrightarrow j), \tag{4.20}$$

where $P(k_i, i \leftrightarrow j)$ is the *joint* probability that node i has degree k_i and that it is connected to node j. For randomly connected networks (neglecting loops) $P(i \leftrightarrow j) = \langle k \rangle/(N-1)$ and $P(i \leftrightarrow j|k_i) = k_i/(N-1)$, where N is the total number of nodes in the network. Using the above criterion Eq. (4.19) reduces to [14, 18]:

$$\kappa \equiv \frac{\langle k^2 \rangle}{\langle k \rangle} = 2, \tag{4.21}$$

at the critical point. A spanning cluster exists for graphs with $\kappa > 2$, while graphs with $\kappa < 2$ contain only small clusters whose size s is negligible compared to the entire network, $\lim_{N\to\infty} s/N = 0$. The criterion (4.21) and its range of validity was derived rigorously by Molloy and Reed [18], using somewhat different arguments.

Neglecting the loops is justified below the transition, since the probability for a bond to form a loop in an s-node cluster is proportional to $(s/N)^2$ (*i.e.*, proportional to the probability of choosing two sites in that cluster). An estimate of the fraction of loops P_{loop} in the network yields

$$P_{loop} \propto \sum_i \frac{s_i^2}{N^2} < \sum_i \frac{s_i S}{N^2} = \frac{S}{N}, \tag{4.22}$$

where the sum is over all clusters in the system (s_i is the size of the ith cluster), and S is the size of the biggest cluster. Since $S \sim \ln N$ below the transition, P_{loop} is negligible when $N \to \infty$.

4.3.2 Percolation critical threshold

The above reasoning can be applied to the problem of percolation in a generalized random network [14]. If we randomly remove a fraction p of the sites (along with the emanating links), the degree distribution of the remaining sites will change. For instance, sites with initial degree k_0 will have, after the random removal of nodes, a different number of connections k, depending on the number of removed neighbors. The initial degree distribution, $P_0(k_0)$, becomes, following dilution

$$P(k) = \sum_{k_0=k}^{\infty} P_0(k_0) \binom{k_0}{k} (1-p)^k p^{k_0-k}. \tag{4.23}$$

The first two moments of the new $P(k)$ are

$$\langle k \rangle = \sum_{k=0}^{\infty} P(k)k = (1-p)\langle k_0 \rangle, \tag{4.24}$$

and

$$\langle k^2 \rangle = \sum_{k=0}^{\infty} P(k)k^2 = (1-p)^2\langle k_0^2 \rangle + p(1-p)\langle k_0 \rangle, \tag{4.25}$$

where $\langle k_0 \rangle$ and $\langle k_0^2 \rangle$ are computed with respect to the original distribution $P_0(k_0)$. On substituting the new moments in Eq. (4.21) we obtain the criterion for criticality, following dilution:

$$\kappa \equiv \frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{(1-p)^2 \langle k_0^2 \rangle + p(1-p)\langle k_0 \rangle}{(1-p)\langle k_0 \rangle} = 2. \tag{4.26}$$

This can be rearranged, to yield the critical threshold for percolation [14]:

$$1 - p_{\rm c} = \frac{1}{\kappa_0 - 1}, \tag{4.27}$$

where $\kappa_0 \equiv \langle k_0^2 \rangle / \langle k_0 \rangle$ is calculated using the original distribution, before the random removal of sites.

Eqs. (4.21) and (4.27) are applicable to random graphs of arbitrary degree distribution. For example, using (4.27) for Cayley trees yields the well-known threshold [22, 38] $q_{\rm c} = 1 - p_{\rm c} = 1/(z-1)$. Another example is ER graphs. Their edges are distributed randomly and the resulting degree distribution is Poissonian [4]. Applying the criterion from Eq. (4.21) to the Poisson distribution of ER graphs yields

$$\kappa \equiv \frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{\langle k \rangle^2 + \langle k \rangle}{\langle k \rangle} = 2, \tag{4.28}$$

which reduces to the known result [4] $\langle k \rangle = 1$.

Evidently, the key parameter governing the threshold, according to (4.27), is the ratio of second- to first-moment, κ_0. This may be estimated by approximating (4.1) to a continuous distribution (the approximation becomes exact for $1 \ll m \ll K$, and it preserves the essential features of the transition even for small m):

$$\kappa_0 = \left(\frac{2-\lambda}{3-\lambda} \right) \frac{K^{3-\lambda} - m^{3-\lambda}}{K^{2-\lambda} - m^{2-\lambda}}. \tag{4.29}$$

In the limit of $K \gg m$, we have

$$\kappa_0 \to \left| \frac{2-\lambda}{3-\lambda} \right| \times \begin{cases} m, & \lambda > 3; \\ m^{\lambda-2} K^{3-\lambda}, & 2 < \lambda < 3; \\ K, & 1 < \lambda < 2. \end{cases} \tag{4.30}$$

We see that for $\lambda > 3$ the ratio κ_0 is finite and there is a percolation transition at $1 - p_{\rm c} \approx \left(\frac{\lambda-2}{\lambda-3} m - 1 \right)^{-1}$: for $p > p_{\rm c}$ the spanning cluster is fragmented and the network is destroyed. However, for $\lambda < 3$ the ratio κ_0 diverges with K and so $p_{\rm c} \to 1$ when $K \to \infty$ (or $N \to \infty$). The percolation transition does not take place: a spanning cluster exists for arbitrarily large fractions of dilution, $p < 1$. In *finite* systems a transition is always observed, though for $\lambda < 3$ the transition threshold is exceedingly high. For the case of the Internet ($\lambda \approx 5/2$), we have $\kappa_0 \approx K^{1/2} \approx N^{1/3}$. Considering the enormous size of the Internet, $N > 10^6$, one needs to destroy over 99% of the nodes before the spanning cluster collapses. For $\lambda > 4$ calculation of κ shows that it is lower than 2 even before the breakdown occurs. For $\lambda > 4$ and $m = 1$ the network consists of only finite clusters and no spanning cluster to begin with (this is reminiscent of the result for $\lambda > 3.478...$ found in [20], where the different threshold stems from rigorous consideration of the *discrete* nature of (4.1)). Note that if $m \geq 2$, a spanning cluster exists for all values of λ.

4.3.3 Generating functions

A general method for studying the size of the infinite cluster and the residual network for a graph with an arbitrary degree distribution was first found by Molloy and Reed [19]. They consider the infinite cluster as it is being exposed, layer by layer, and develop differential equations relating the number of unexposed links and unvisited sites in subsequent shells (see Section 4.2).

An alternative, and very powerful approach based on generating functions was advanced by Newman, Watts and Strogatz [35]. Their method is beautifully reviewed in this book, in the Chapter by Newman. Here, we follow closely in their footsteps. Our ultimate goal is to compute the size of the giant component as well as the critical exponents associated with the percolation transition in scale-free networs, resulting from random dilution (Section **4.3.5**).

In [35, 39] a generating function is constructed for the degree distribution:

$$G_0(x) = \sum_{k=0}^{\infty} P(k) x^k. \tag{4.31}$$

The probability of reaching a site with degree k by following a specific link is $kP(k)/\langle k\rangle$ [14, 18, 35], and its corresponding generating function is

$$G_1(x) = \frac{\sum kP(k)x^{k-1}}{\sum kP(k)} = \frac{d}{dx} G_0(x)/\langle k\rangle \ . \tag{4.32}$$

Let $H_1(x)$ be the generating function for the probability of reaching a branch of a given size by following a link. When a fraction $p = 1 - q$ of the sites are diluted, H_1 satisfies the self-consistent equation:

$$H_1(x) = 1 - q + qxG_1(H_1(x)) \ . \tag{4.33}$$

Given that $G_0(x)$ is the generating function for the degree of a site, the generating function for the probability of a site to belong to an n-site cluster is:

$$H_0(x) = 1 - q + qxG_0(H_1(x)) \ . \tag{4.34}$$

Below the transition, all clusters are finite and $H_0(1) = 1$. However, above the transition $H_0(1)$ is no longer normalized, since it excludes the probability of the incipient infinite cluster, P_∞. In other words, $P_\infty = 1 - H_0(1)$. It follows that

$$P_\infty(q) = q\big(1 - \sum_{k=0}^{\infty} \tilde{P}(k) u^k\big) \ , \tag{4.35}$$

where $u \equiv H_1(1)$ is the smallest positive root of

$$\langle k\rangle u = 1 - q + \frac{q}{\langle k\rangle} \sum_{k=0}^{\infty} kP(k)u^{k-1} \ . \tag{4.36}$$

This equation can be solved numerically and the solution can be substituted into Eq. (4.35) to compute the size of the infinite cluster in a random graph of arbitrary degree distribution. (The analysis neglects the presence of loops, which is well justified near criticality.)

4.3.4 Intentional attack

Another model of interest, suggested in [40], is that of intentional attack on the most highly connected nodes of the network. In this model an attacker (*e.g.*, computer hackers trying to cause damage to the network, or doctors trying to disrupt a contagious epidemic) succeeds in knocking off a fraction p of the most highly connected sites in the network. As might be expected, such a strategy is far more effective than *random* dilution. We shall see, in fact, that a *small* threshold p suffices to disrupt the net (for all λ).

Next we consider analytically the consequence of such an attack, or sabotage, on scale-free networks [41]. A different approach was given independently in [39]: (a) the cutoff degree K reduces to some new value $\tilde{K} < K$, and (b) the degree distribution of the remaining sites is no longer scale-free, but is changed due to the removal of many of their links. Recall that the upper cutoff K, before the attack, may be estimated from

$$\sum_{k=K}^{\infty} P(k) = \frac{1}{N} \, . \tag{4.37}$$

Similarly, the new cutoff $\tilde{K}$, after the attack, follows from

$$\sum_{k=\tilde{K}}^{K} P(k) = \sum_{k=\tilde{K}}^{\infty} P(k) - \frac{1}{N} = p \, . \tag{4.38}$$

If the size of the system is large, $N \gg 1/p$, the original cutoff K may be safely ignored. We can then obtain $\tilde{K}$ approximately by replacing the sum with an integral:

$$\tilde{K} = mp^{1/(1-\lambda)} \, . \tag{4.39}$$

We estimate the impact of the attack on the distribution of the remaining sites as follows. The removal of a fraction p of the sites with the highest degree results in a random removal of links from the remaining sites — links that had connected the removed sites with the remaining sites. The probability $\tilde{p}$ for a link to lead to a deleted site equals the ratio of the number of links belonging to deleted sites to the total number of links:

$$\tilde{p} = \sum_{k=\tilde{K}}^{K} \frac{kP(k)}{\langle k_0 \rangle}, \tag{4.40}$$

where $\langle k_0 \rangle$ is the initial average degree. With the usual continuous approximation, and neglecting K, this yields

$$\tilde{p} = \left(\frac{\tilde{K}}{m} \right)^{2-\lambda} = p^{(2-\lambda)/(1-\lambda)} \, , \tag{4.41}$$

for $\lambda > 2$. For $\lambda = 2$, $\tilde{p} \to 1$, since just a few nodes of very high degree control the entire connectedness of the system. Indeed, consider a finite system of N sites and $\lambda = 2$. The upper cutoff $K \approx N$ must then be taken into account, and approximating Eq. (4.40) by an

integral yields $\tilde{p} = \ln(Np/m)$. That is, for $\lambda = 2$, a very small value of p is needed to destroy an arbitrarily large fraction of the links as $N \to \infty$.

With the above results we can compute the effect of intentional attack, using the theory previously developed for random removal of sites [14]. Essentially, the network after attack is equivalent to a scale-free network with cutoff $\tilde{K}$, that has undergone random removal of a fraction $\tilde{p}$ of its sites. Since the latter effect influences the probability distribution as described in Eq. (4.23), the result in Eqs. (4.27) and (4.29) can be used, but with $\tilde{p} = (\tilde{K}/m)^{2-\lambda}$ and $\tilde{K}$ replacing p_c and K, respectively. This yields the equation:

$$(\tilde{K}/m)^{2-\lambda} - 2 = \frac{2-\lambda}{3-\lambda} m[(\tilde{K}/m)^{3-\lambda} - 1] \,, \tag{4.42}$$

which can be solved numerically to obtain $\tilde{K}(m, \lambda)$, and then $p_c(m, \lambda)$ can be retrieved from Eq. (4.39). In Fig. 4.4 we plot p_c (there denoted f_c) — the critical fraction of sites needed to be removed in the sabotage strategy to disrupt the network — computed in this fashion, and compared to results from numerical simulations. A phase transition exists (at a finite p_c) for all $\lambda > 2$. The decline in p_c for large λ is explained from the fact that as λ increases the spanning cluster becomes smaller in size, even before attack. (Furthermore, for $m < 2$ the original network is disconnected for λ large enough.) The decline in p_c as $\lambda \to 2$ results from the critically high degree of just a few sites: their removal disrupts the whole network. This was first argued in [40]. We note that for infinite systems $p_c \to 0$ as $\lambda \to 2$. The critical fraction p_c is rather sensitive to the lower degree cutoff m. For larger m the networks are more robust, though they still undergo a transition at a finite p_c. (Fig. 4.4 illustrates the case of $m = 1$.)

4.3.5 Critical exponents

The generating functions method has the advantage of turning a combinatorial problem into an algebraic one, concerning power series. The algebraic problem is often simpler.

Here we use generating functions for obtaining the percolation critical exponents [42]. These can be extracted from the resulting power series by means of appropriate Abelian and Tauberian methods [43, 44].

First, we compute the order parameter critical exponent β. Near criticality the probability of belonging to the spanning cluster behaves as $P_\infty \sim (q - q_c)^\beta$. For infinite-dimensional systems (such as a Cayley tree) it is known that $\beta = 1$ [22, 30, 38]. This regular mean-field result is not always valid, however, for scale-free networks. Eq. (4.35) has no special behavior at $q = q_c$; the singular behavior comes from u. At criticality, $P_\infty = 0$ and Eq. (4.35) imply that $u = 1$. We therefore examine Eq. (4.36) for $u = 1 - \epsilon$ and $q = q_c + \delta$:

$$1 - \epsilon = 1 - q_c - \delta + \frac{(q_c + \delta)}{\langle k \rangle} \sum_{k=0}^{\infty} kP(k)(1-\epsilon)^{k-1}. \tag{4.43}$$

The sum in (4.43) has the asymptotic form

$$\begin{aligned} \sum_{k=0}^{\infty} kP(k)u^{k-1} \;\sim\; & \langle k \rangle - \langle k(k-1) \rangle \epsilon \\ & + \tfrac{1}{2}\langle k(k-1)(k-2) \rangle \epsilon^2 + \cdots + c\Gamma(2-\lambda)\epsilon^{\lambda-2} \,, \end{aligned} \tag{4.44}$$

where the highest-order analytic term is $\mathcal{O}(\epsilon^n)$, $n = \lfloor \lambda - 2 \rfloor$. Using this in Eq. (4.43), with $q_c = 1/(\kappa - 1) = \langle k \rangle / \langle k(k-1) \rangle$, we get

$$\frac{\langle k(k-1) \rangle^2}{\langle k \rangle} \delta = \tfrac{1}{2} \langle k(k-1)(k-2) \rangle \epsilon + \cdots + c\Gamma(2-\lambda)\epsilon^{\lambda-3} \ . \tag{4.45}$$

The divergence of δ as $\lambda < 3$ confirms the vanishing threshold for the phase transition in that regime. Thus, limiting ourselves to $\lambda > 3$, and keeping only the dominant term as $\epsilon \to 0$, Eq. (4.45) implies

$$\epsilon \sim \begin{cases} \left(\frac{\langle k(k-1) \rangle^2}{c \langle k \rangle \Gamma(2-\lambda)} \right)^{\frac{1}{\lambda-3}} \delta^{\frac{1}{\lambda-3}} & 3 < \lambda < 4, \\ \frac{2\langle k(k-1) \rangle^2}{\langle k \rangle \langle k(k-1)(k-2) \rangle} \delta & \lambda > 4. \end{cases} \tag{4.46}$$

Returning to P_∞, Eq. (4.35), we see that the singular contribution in ϵ is dominant only for the irrelevant range of $\lambda < 2$. For $\lambda > 3$, we find $P_\infty \sim q_c \langle k \rangle \epsilon \sim (q - q_c)^\beta$.

We see that the order parameter exponent β attains its usual mean-field value only for $\lambda > 4$. Moreover, for $\lambda < 4$ the percolation transition is higher than 2nd-order: for $3 + \frac{1}{n-1} < \lambda < 3 + \frac{1}{n-2}$ the transition is of the nth-order.

For networks with $\lambda < 3$ the transition still exists, though at a vanishing threshold, $q_c = 0$. The sum in Eq. (4.43) becomes:

$$\sum_{k=0}^{\infty} k P(k) u^{k-1} \sim \langle k \rangle + c\Gamma(2-\lambda)\epsilon^{\lambda-2} \ . \tag{4.47}$$

Using this in conjunction with Eq. (4.36), and remembering that here $q_c = 0$ and therefore $q = \delta$, leads to

$$\epsilon = \left(\frac{-c\Gamma(2-\lambda)}{\langle k \rangle} \right)^{\frac{1}{3-\lambda}} \delta^{\frac{1}{3-\lambda}} \ . \tag{4.48}$$

The results can be summarized by

$$\beta = \begin{cases} \frac{1}{3-\lambda} & 2 < \lambda < 3, \\ \frac{1}{\lambda-3} & 3 < \lambda < 4, \\ 1 & \lambda > 4. \end{cases} \tag{4.49}$$

In other words, the transition in $2 < \lambda < 3$ is a mirror image of the transition in $3 < \lambda < 4$. An important difference is that $q_c = 0$ is not λ-dependent in $2 < \lambda < 3$, and the amplitude of P_∞ diverges as $\lambda \to 2$ (but remains finite as $\lambda \to 4$). Some of the results for β have been reported before in [41], and also found independently in a different but related model of virus spreading [45–47]. The existence of an infinite-order phase transition at $\lambda = 3$ for growing networks of the Albert-Barabási model has been reported in [48, 49]. These examples suggest that the critical exponents are not model-dependent but depend only on λ.

In [35] it was shown that for a random graph of arbitrary degree distribution the finite clusters (of size s) follow the usual scaling form:

$$n_s \sim s^{-\tau} e^{-s/s^*} \ . \tag{4.50}$$

At criticality $s^* \sim |q - q_c|^{-\sigma}$ diverges and the tail of the distribution behaves as a power law. We now derive the exponent τ. The probability that a site belongs to an s-cluster is $p_s = sn_s \sim s^{1-\tau}$, and is generated by H_0:

$$H_0(x) = \sum p_s x^s \ . \tag{4.51}$$

The singular behavior of $H_0(x)$ stems from $H_1(x)$, as can be seen from Eq. (4.34). $H_1(x)$ itself can be expanded from Eq. (4.33), by using the asymptotic form (4.44) of G_1. We let $x = 1-\epsilon$, as before, but work at the critical point, $q = q_c$. With the notation $\phi(\epsilon) = 1 - H_1(1-\epsilon)$, we finally get (note that at criticality $H_1(1) = 1$):

$$-\phi = -q_c + (1-\epsilon)q_c[1 - \frac{\phi}{q_c} + \frac{\langle k(k-1)(k-2)\rangle}{2\langle k\rangle}\phi^2 + \cdots + c\frac{\Gamma(2-\lambda)}{\langle k\rangle}\phi^{\lambda-2}] \ . \tag{4.52}$$

>From this relation we extract the singular behavior of H_0: $\phi \sim \epsilon^y$. Then, using Tauberian theorems [43] it follows that $p_s \sim s^{-1-y}$, hence $\tau = 2 + y$.

For $\lambda > 4$ the term proportional to $\phi^{\lambda-2}$ in (4.52) may be neglected. The linear term $\epsilon\phi$ may be neglected as well, due to the factor ϵ. This leads to $\phi \sim \epsilon^{1/2}$ and to the usual mean-field result

$$\tau = \frac{5}{2} \ , \qquad \lambda > 4 \ . \tag{4.53}$$

For $\lambda < 4$, the terms proportional to $\epsilon\phi$, ϕ^2 may be neglected, leading to $\phi \sim \epsilon^{1/(\lambda-2)}$ and

$$\tau = 2 + \frac{1}{\lambda-2} = \frac{2\lambda-3}{\lambda-2} \ , \qquad 2 < \lambda < 4 \ . \tag{4.54}$$

Note that for $2 < \lambda < 3$ the percolation threshold is strictly $q_c = 0$. In that case we work at $q = \delta$ small but fixed, taking the limit $\delta \to 0$ at the very end. For the case $2 < \lambda < 3$, τ in Eq. (4.54) represents the singularity of the distribution of branch sizes. For the distribution of cluster sizes in this range one has to consider the singularity of x in Eq. (4.34), leading to $\tau = 3$.

For growing networks of the Albert-Barabási model with $\lambda = 3$, it has been shown that $sn_s \propto (s \ln s)^{-2}$ [49]. This is consistent with $\tau = 3$ plus a logarithmic correction. Related results for scale-free trees have been presented in [50].

At the transition point the size of the largest cluster, S, can be obtained from the finite cluster distribution by taking the integral over the tail of the distribution to equal $1/N$. This results in

$$S \propto N^{\tau-1} = N^{(\lambda-2)/(\lambda-1)}. \tag{4.55}$$

For $\lambda = 4$ this reduces to the known $N^{2/3}$ dependence [4]. For $\lambda \to 3$, $S \propto N^{1/2}$. It is not yet clear whether the results have a meaningful interpretation for $\lambda < 3$.

The critical exponent σ, for the cutoff cluster size, may be also derived directly. Finite-size scaling arguments predict [38] that

$$q_c(\infty) - q_c(N) \sim N^{-\frac{1}{d\nu}} = N^{-\frac{\sigma}{\tau-1}} \ , \tag{4.56}$$

where N is the number of sites in the network, ν is the correlation length critical exponent: $\xi \sim (q - q_c)^{-\nu}$, and d is the dimensionality of the embedding space. Using a continuous approximation of the distribution (4.1) one obtains [14]

$$\kappa \approx \left(\frac{2-\lambda}{3-\lambda}\right)\frac{K^{3-\lambda} - m^{3-\lambda}}{K^{2-\lambda} - m^{2-\lambda}} , \tag{4.57}$$

where, as usual $K \sim N^{1/(\lambda-1)}$. For $3 < \lambda < 4$, this and Eq. (4.27) yield

$$q_c(\infty) - q_c(N) \sim \Delta\kappa \sim K^{3-\lambda} \sim N^{\frac{3-\lambda}{\lambda-1}} , \tag{4.58}$$

which in conjunction with Eq. (4.56) leads to

$$\sigma = \frac{\lambda-3}{\lambda-2} , \qquad 3 < \lambda < 4 . \tag{4.59}$$

For $\lambda > 4$ we recover the regular mean-field result $\sigma = 1/2$. Note that Eqs. (4.56), (4.49), (4.54) are consistent with the known scaling relation: $\sigma\beta = \tau - 2$ [22,30,38]. For $2 < \lambda < 3$, $q_c(\infty) = 0$ and $q_c(N) \sim K^{\lambda-3} \sim N^{(\lambda-3)/(\lambda-1)}$. Therefore

$$\sigma = \frac{3-\lambda}{\lambda-2} , \qquad 2 < \lambda < 3 , \tag{4.60}$$

again consistent with the scaling relation $\sigma\beta = \tau - 2$ (cf Eq. (4.49)).

4.3.6 Fractal dimension

It is well known that on a random network in the well connected regime, the average distance between sites is of order $\log_k N$ [4, 51]. This has also been shown to hold for general networks [35] and may be even lower for scale-free networks [34]. However, the diluted case is essentially the same as infinite-dimensional percolation. In this case, there is no notion of geometrical distance (since the graph is not embedded in an Euclidean space), but only of chemical distance (the smallest number of edges connecting any two nodes). It is known from infinite-dimensional percolation theory that the fractal dimension at criticality is $d_f = 2$ [22]. Therefore the average (chemical) distance $\langle l \rangle$ between pairs of sites on the spanning cluster at criticality behaves as

$$\langle l \rangle \sim \sqrt{M} , \tag{4.61}$$

where M is the number of sites in the spanning cluster. This is analogous to percolation in finite dimensions, where in lengthscales smaller than the correlation length the cluster is a fractal with dimension d_f and above the correlation length the cluster is homogeneous and has the dimension of the embedding space. In our infinite-dimensional case, the crossover between these two behaviors occurs around the correlation length $\xi \approx |p_c - p|^{-1}$.

For $3 < \lambda < 4$ the situation is somewhat different. Below the transition all clusters are finite and almost all finite clusters are trees. The correlation length can be defined using the formula [22]:

$$\xi_l^2 = \frac{\sum l^2 g(l)}{\sum g(l)} . \tag{4.62}$$

The number of sites in the l shell can be seen to be approximately $\langle k\rangle(\kappa-1)^{l-1}$ [35]. Since $\kappa-1=(\kappa_0-1)q$ and $q_c=1/(\kappa_0-1)$ we get $g(l)=c(1-\delta)^l$, where $\delta=q-q_c$. This leads to $\xi_l\sim(q-q_c)^{-1}$, *i.e.*, $\nu_l=1$. Above the threshold, the finite clusters can be seen as a random graph with the residual degree distribution of sites not included in the infinite cluster [19]. That is, the degree distribution for sites in the finite clusters is

$$P_r(k)=P(k)u^k, \tag{4.63}$$

where u is the solution of Eq. (4.36). Using this distribution, we can define κ_r for the finite clusters. This adds a term proportional to $\epsilon^{\lambda-3}$ to the expansion of ξ_l. But, since $\delta\propto\epsilon^{\lambda-3}$ (4.46), this leads again to $\nu_l=1$.

Using ν, the dimension of the network at criticality can be found. The chemical dimension $d_l=1/\sigma\nu_l$, therefore

$$d_l=\frac{\lambda-2}{\lambda-3}. \tag{4.64}$$

Since every path, when embedded in a space above the critical dimension, can be seen as a random walk, it follows that $\nu=\nu_l/2$ [22]. Therefore, the fractal dimension is,

$$d_f=\frac{1}{\nu\sigma}=2\frac{\lambda-2}{\lambda-3}. \tag{4.65}$$

The dimension of the embedding space is,

$$d_c=\frac{1}{\nu\sigma(\tau-1)}=2\frac{\lambda-1}{\lambda-3}. \tag{4.66}$$

d_l, d_f, and d_c of the embedding space reduce to the known values of 2, 4, and 6, respectively, when $\lambda=4$. Some of the above results have been obtained also by Burda *et al.*, [50].

4.4 Percolation in directed networks

Many of the large complex networks of interest, such as the Internet, WWW, electric power grid, cellular, and social networks are directed [28, 29, 52]. For example, in social and economical networks if node A gains information or acquires physical goods from node B, it does not necessarily mean that node B gets similar input from node A. Likewise, most metabolic reactions are one-directional, thus changes in the concentration of molecule A affect the concentration of its product B, but the reverse is not always true. Despite the directedness of many real networks, the modeling literature, with few notable exceptions [35,53], has focused mainly on undirected networks.

Important aspects of directed networks are captured by their degree distribution, $P(j,k)$, or the probability that an arbitrary node has j incoming and k outgoing edges. Many naturally occurring directed networks, such as the WWW, metabolic networks, citation networks, etc., exhibit a power-law, or *scale-free* degree distribution for the incoming or outgoing links:

$$P_{in(out)}(l)=cl^{-\lambda_{in(out)}},\quad m\le l\le K\ , \tag{4.67}$$

similar to that of Eq. (4.1) [7, 8].

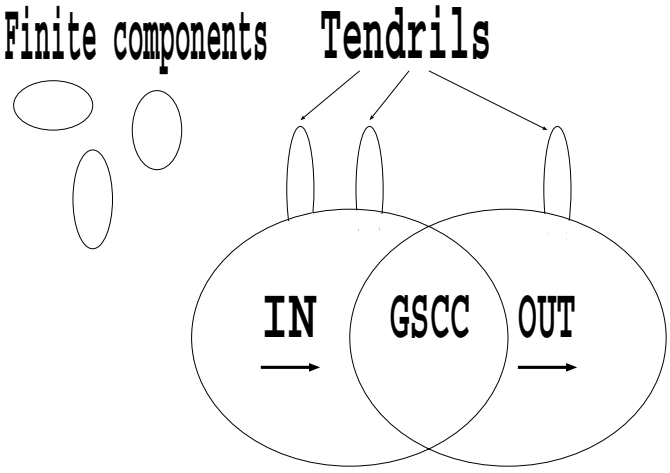


Figure 4.2: Structure of a general directed graph.

The structure of a directed graph has been characterized in [35, 53], and in the context of the WWW in [8, 54]. In general, a directed graph consists of a giant weakly connected component (GWCC) and several finite components. In the GWCC every site is reachable from every other, provided that the links are treated as bi-directional. The GWCC is further divided into a giant strongly connected component (GSCC), consisting of all sites reachable from each other following directed links. All the sites reachable from the GSCC are referred to as the giant OUT component, and the sites from which the GSCC is reachable are referred to as the giant IN component. The GSCC is the intersection of the IN and OUT components. Sites in the GWCC, but not in the IN and OUT components, constitute the "tendrils" (see Fig. 4.2.).

Here we repeat the analysis of Section 4.3 for the case of directed networks. We limit ourselves mostly to final results and conclusions. A detailed derivation can be found in [55].

4.4.1 Threshold

The condition for the existence of a giant component in a directed random network of arbitrary degree distribution can be deduced in a manner similar to [14]. If a site is reached following a link pointing to it, then it must have at least one outgoing link, on average, in order to be part of a giant component. This condition can be written as

$$\langle k_j | i \to j \rangle = \sum_{k_i,k_j} k_j P(k_i, k_j | i \to j) = 1. \tag{4.68}$$

Reasoning as in Section 4.3.2, the above criterion reduces to [35, 53]

$$\langle jk \rangle \geq \langle k \rangle. \tag{4.69}$$

Suppose a fraction $p = 1 - q$ of the nodes is removed from the network, then the original degree distribution, $P(j,k)$, becomes

$$P'(j,k) = \sum_{j_0,k_0}^{\infty} P(j_0,k_0) \binom{j_0}{j} (1-p)^j p^{j_0-j} \binom{k_0}{k} (1-p)^k p^{k_0-k} \ . \tag{4.70}$$

In view of this new distribution, Eq. (4.69) yields the percolation threshold

$$q_c = 1 - p_c = \frac{\langle k \rangle}{\langle jk \rangle} \ , \tag{4.71}$$

where averages are computed with respect to the original distribution before dilution, $P(j,k)$. Eq. (4.71) indicates that in directed scale-free networks if $\langle jk \rangle$ diverges then $q_c \to 0$ and the network is resilient to random breakdown of nodes and bonds.

The term $\langle jk \rangle$ may be dramatically influenced by the appearance of correlations between the *in*- and *out*-degrees of the nodes. This effect has been discussed in [53]. Our own studies [55] of uncorrelated and correlated distributions reveal that In the former case the threshold is simply $q_c = 1/\langle k \rangle$, while in the latter case $\langle jk \rangle$ diverges whenever

$$(\lambda_{out} - 2)(\lambda_{in} - 2) \leq 1 \ , \tag{4.72}$$

causing the percolation threshold to vanish. The various regimes resulting from this observation are summarized in Fig. 4.3.

Percolation of the GWCC can be seen to be similar to percolation in the non-directed graph created from the directed graph by ignoring the directionality of the links. The threshold is obtained from the criterion (cf Eq. (4.27))

$$q_{\rm c} = \frac{\langle k \rangle}{\langle k(k-1) \rangle}. \tag{4.73}$$

Here the connectivity distribution is the convolution of the *in* and *out* distributions,

$$P'(k) = \sum_{l=0}^{k} P(l, k-l). \tag{4.74}$$

Whether the distribution is correlated or not, $P'(k)$ is always dominated by the slower decay-exponent, therefore percolation of the GWCC is the same as in non-directed scale-free networks, with $\lambda_{eff} = min(\lambda_{in}, \lambda_{out})$. Note that the percolation threshold of the GWCC may differ from that of the GSCC and the IN and OUT components [53].

4.4.2 Critical exponents

Percolation of the GSCC and IN and OUT components may be analyzed with the formalism of generating functions [35,39,44,53] (see, also, the Chapter by Newman in this book). We have computed the critical exponents β and τ, following the approach of Section 4.3.5. The results are the same as for the non-directed case, Eqs. (4.49) and (4.54), but where λ is replaced by an

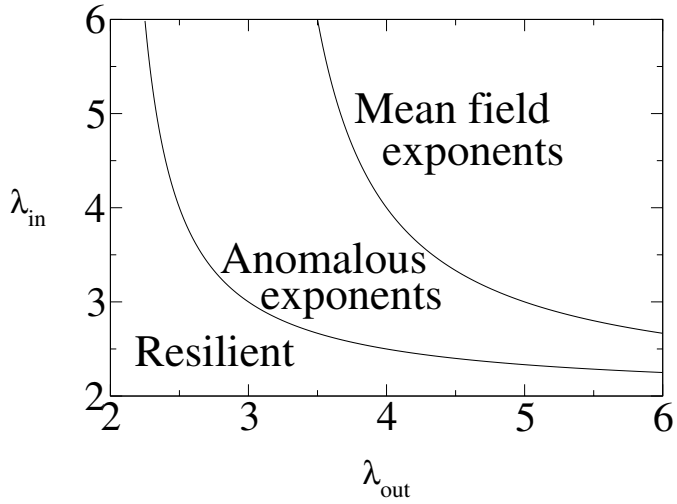


Figure 4.3: Phase diagram of the different regimes for the IN component of scale-free correlated directed networks. The boundary between Resilient and Anomalous exponents is derived from Eq. (4.72) while that between Anomalous exponents and Mean field exponents is given in table 4.1 for $\lambda^\star = 4$. For the diagram of the OUT component λ_{in} and λ_{out} change roles

	uncorrelated	*correlated*
GWCC	$min(\lambda_{out}, \lambda_{in}) + 1$	$min(\lambda_{out}, \lambda_{in})$
IN	$\lambda_{out} + 1$	$\lambda_{out} + \frac{\lambda_{in} - \lambda_{out}}{\lambda_{in} - 1}$
OUT	$\lambda_{in} + 1$	$\lambda_{in} + \frac{\lambda_{out} - \lambda_{in}}{\lambda_{out} - 1}$
GSCC	$min(\lambda_{out}, \lambda_{in}) + 1$	$min(\lambda^*_{out}, \lambda^*_{in})$

Table 4.1: Values of $\lambda^\star$ for the different network components for both correlated and uncorrelated cases.

effective λ^* whose value differs for uncorrelated and correlated distributions. The value of λ^* for the various components in both the uncorrelated and correlated scenario are summarized in table 4.1. Our findings [55] indicate that even the tiniest amount of correlation results in behavior typical to the correlated case. We may therefore conclude that in practical situations only the correlated case counts, for it is expected, to some extent, in most naturally occurring directed networks.

4.5 Efficient immunization strategies

It is well established that random immunization fails to prevent epidemics of diseases that spread upon contact between infected individuals. On the other hand, targeted immunization requires global knowledge of the topology of the social network in question, rendering it

impractical. We propose an effective strategy, based on the immunization of a small fraction of *random acquaintances* of randomly selected individuals, that prevents epidemics without requiring global knowledge of the network [56].

Social networks are known to possess a broad distribution of the number of links (contacts) k, emanating from a node (an individual) [10, 57–59]. [3] Studies of percolation on broad distribution networks show that a large fraction f_c of the nodes need to be removed (immunized) before the integrity of the network is compromised. This is particularly true for scale-free networks with $2 < \lambda < 3$ — the case of most known networks [6, 9, 13] — where the percolation threshold $f_c \to 1$, and the network remains connected (contagious) even after immunization of most of its nodes [14, 39, 45, 60–63]. In other words, with a random immunization strategy most of the population needs to be immunized before an epidemic is arrested (see Fig. 4.4).

When the most highly connected nodes are targeted first, removal of just a small fraction of the nodes results in the network's disintegration [39, 41]. This has led to the suggestion of targeted immunization of the HUBs (the most highly connected nodes in the network) [64,65]. The main shortcoming of this approach is that it requires a complete, or at least fairly good knowledge of the connectivity of each node in the network. Such global information often proves hard to gather, and may not even be well-defined (as in social networks, where the number of social relations depends on subjective judging). Here we propose an effective immunization strategy that works at low immunization rates f, and obviates the need for global information.

4.5.1 Acquaintance immunization

In our approach, we choose a random fraction p of the population (of size N) and ask each individual to point at an acquaintance with whom they are in contact. The acquaintances, rather than the individuals themselves, are the ones immunized. The fraction f_c needed to be immunized in order to stop the epidemic can be computed analytically.

In each immunization event the probability that a node with k contacts is selected is $kP(k)/(N\langle k\rangle)$. Let $n_l(k)$ be the number of individuals in chemical shell l who are susceptible (not immunized). In the next chemical shell, $l+1$, each of those sites connects to $k-1$ neighbors (excluding the one connecting to shell $l-1$). To find out $n_{l+1}(k')$, we multiply the number of links going out of the lth layer by the probability of reaching a site of connectivity k' by following a link from a *susceptible* site, $p(k'|k \wedge s_k)$, and the probability that this site is also susceptible, $p(s_{k'}|k' \wedge k \wedge s_k)$. This gives

$$n_{l+1}(k') = \sum_k n_l(k)(k-1)p(k'|k \wedge s_k)p(s_{k'}|k' \wedge k \wedge s_k) \; . \tag{4.75}$$

From Bayes' rule,

$$p(k'|k \wedge s_k) = \frac{p(s_k|k \wedge k')p(k'|k)}{p(s_k|k)} \; . \tag{4.76}$$

$p(k'|k) = k'P(k')/\langle k\rangle$ is independent of k. $p(s_k|k \wedge k') = e^{-p/k'} \times \langle e^{-p/k}\rangle^{k-1}$, where the average is taken with respect to $p(k)$ as defined before. $p(s_k|k) = \langle \exp(-p/k)\rangle^k$, since no

[3] Often this is the scale-free distribution $P(k) = ck^{-\lambda}$. Our results apply, however, to broad distributions in general.

knowledge exists on its neighbors. Using all these relations one obtains:

$$p(k'|k \wedge s_k) = \frac{p(k')e^{-p/k'}}{\langle e^{-p/k} \rangle} \ . \tag{4.77}$$

The above results, along with (4.75) yield

$$n_{l+1}(k') = v_p^{k'-2} p(k') e^{-p/k'} \sum_k n_l(k)(k-1)e^{-p/k} \ , \tag{4.78}$$

where $v_p = \langle \exp(-p/k) \rangle$. This leads to the stable distribution of connectivity in a chemical layer: $n_l(k) = a v_p^{k-2} p(k) e^{-p/k}$, for some a. Putting this back into (4.78) results in:

$$n_{l+1}(k') = n_l(k') \sum_k p(k)(k-1) v_p^{k-2} e^{-2p/k} \ . \tag{4.79}$$

Therefore, if the sum in (4.79) is larger than 1 the population is above the percolation threshold and the epidemics would propagate, while it would be arrested if the sum is smaller than 1. Thus,

$$\sum_k p(k)(k-1) v_{p_c}^{k-2} e^{-2p_c/k} = 1 \ , \tag{4.80}$$

is the condition for criticality. The desired immunization fraction then follows:

$$f_c = \sum_k P(k) v_{p_c}^k \ . \tag{4.81}$$

A related immunization strategy calls for the immunization of acquaintances referred to by at least n individuals. (Above, we specialized to $n = 1$.) The threshold is lower the larger n is, and may justify, under certain circumstances, this somewhat more involved protocol.

In Fig. 4.4, we show the immunization threshold f_c needed to stop an epidemic in networks with $2 < \lambda < 3.5$ (this covers all known cases). Plotted are curves for the (inefficient) random strategy, and the strategy advanced here, for the cases of $n = 1$ and 2. Note the dramatic decrease of f_c with the suggested strategy. Improvements can be achieved for any broad distribution.

Various immunization strategies have been proposed earlier, mainly for the case of an already spread disease and are based on tracing the chain of infection towards the super-spreaders of the disease [66]. Our approach can be used even before the epidemic starts spreading, and therefore does not require any knowledge of the chain of infection.

4.6 Summary and outlook

The main goal of this chapter has been to study the effect of the special nature of scale-free distribution on the properties of random network models. Some general methods have been presented for the study of generalized random networks. Those include methods for the study of the layer structure of the graph, the percolation threshold and the critical exponents.

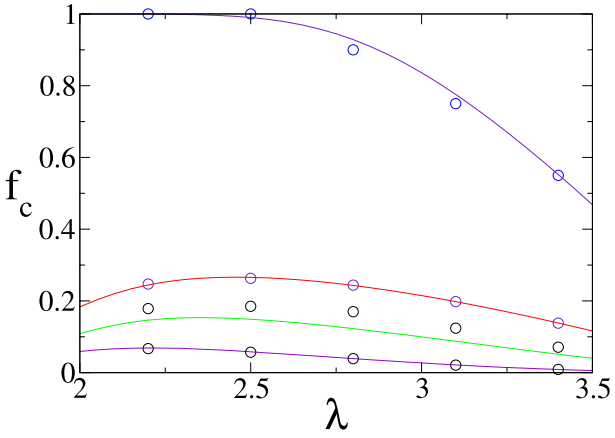


Figure 4.4: Critical probability, f_c, as a function of λ, for the random immunization (top), acquaintance immunization (middle), double acquaintance immunization (lower middle) and attack (bottom) strategies. Curves represent analytical results (approximate for double acquaintace), while data points represent simulation data, for a population $N = 10^6$.

The special properties of scale-free networks, in conjunction with the general method presented for the study of scale-free and other networks, might prove useful for applications such as the design of more robust networks [40], the improvement of routing [67] and search algorithms [68], and the predicting and arresting of computer and human viruses [45, 65].

Acknowledgments

We would like to thank Keren Erez, Nehemia Schwartz, Alejandro Rozenfeld and Albert-Lazslo Barabási for their collaboration, help and insights on many of the topics covered in this chapter. DbA thanks the support of the National Science Foundation (USA).

References

[1] P. Erdős and A. Rényi, On random graphs. *Publicationes Mathematicae* **6**, 290–297 (1959).

[2] P. Erdős and A. and Rényi, On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**, 17–61 (1960).

[3] P. Erdős, and A. Rényi, On the strength of connectedness of a random graph. *Acta Mathematica Scientia Hungary* **12**, 261–267 (1961).

[4] B. Bollobás, *Random Graphs*. Academic Press, New York (1985).

[5] A.-L. Barabási and R. Albert, Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

[6] M. Faloutsos, P. Faloutsos and C. Faloutsos, On power-law relationships of the internet topology. *Computer Communications Review* **29**, 251–262 (1999).

[7] A.-L. Barabási, R. Albert and H. Jeong, Scale-free characteristics of random networks: the topology of the World-Wide Web *Physica A*, **281**, 69–77 (2000).

[8] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins J. Wiener, Graph structure in the web. *Computer Networks* **33**, 309–320 (2000).

[9] H. Ebel, L.-I. Mielsch and S. Bornholdt, Scale-free topology of e-mail networks. *Preprint* cond-mat/0201476 (2002), Phys. Rev. E, in press.

[10] S. Redner, How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B* **4**, 131–134 (1998).

[11] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A.-L. Barabási, The large-scale organization of metabolic networks *Nature*, **407**, 651, (2000).

[12] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).

[13] X. Guardiola, R. Guimera, A. Arenas, A. Diaz-Guilera, D. Streib and L. A. N. Amaral, Macro- and micro-structure of trust networks. *Preprint* cond-mat/0206240 (2002).

[14] R. Cohen, K. Erez, D. ben-Avraham and S. Havlin, Resilience of the Internet to Random Breakdown. *Phys. Rev. Lett.* **85**, 4626–4628 (2000).

[15] S. N. Dorogovtsev and J. F. F. Mendes, Natural Scale of Scale-free Networks. *Phys. Rev. E* **63**, 62101 (2001).

[16] G. Bianconi and A.-L. Barabási, Bose-Einstein condensation in complex networks. *Phys. Rev. Lett.* **86**, 5632–5635 (2001).

[17] P. L. Krapivsky, S. Redner and F. Leyvraz, Connectivity of Growing Random Networks. *Phys. Rev. Lett.* **85**, 4629–4632 (2000).

[18] M. Molloy and B. Reed, A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* **6**, 161–179 (1995).

[19] M. Molloy and B. Reed, The size of the giant component of a random graph with a given degree sequence. *Combin. Probab. Comput.* **7**, 295–305 (1998).

[20] W. Aiello, F. Chung and L. Lu, A random graph model for massive graphs. In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, pp. 171–180. Association of Computing Machinery, New York (2000).

[21] A. Bunde and S. Havlin (eds.), *Fractals in Science*. Springer, New York (1994).

[22] A. Bunde and S. Havlin (eds.), *Fractals and Disordered System*. Springer, New York (1996).

[23] P. Lévy, *Calcul des Probabilities*. Gauthier Villars, Paris (1925).

[24] M. F. Shlesinger and J. Klafter, Accelerated Diffusion in Josephson Junctions and Related Chaotic Systems. *Phys. Rev. Lett.* **54**, 2551 (1985).

[25] J. Klafter, M. F. Shlesinger and G. Zumofen, Beyond Brownian motion. *Phys. Today* **49**, 33 (1996).

[26] G. M. Viswanathan, S. V. Buldyrev, S. Havlin, M. G. E. da Luz, E. P. Raposo and H. E. Stanley, Optimizing the success of random searches. *Nature* **401**, 911 (1999).

[27] J. M. Kleinberg, Navigation in a Small World. *Nature* **406**, 845 (2000).

[28] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).

[29] S. N. Dorogovtsev and J. F. F. Mendes, Evolution of networks. *Adv. in Phys.* **51**, 1079–1187 (2002).

[30] D. ben-Avraham and S. Havlin, *Diffusion and Reactions in Fractals and Disordered Systems*. Cambridge University Press, Cambridge (2000).

[31] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).

[32] D. J. Watts, *Small Worlds*. Princton University Press, Princeton (1999).

[33] A. F. Rozenfeld, R. Cohen, D. ben-Avraham and S. Havlin, Scale-free Networks on Lattices. *Preprint* cond-mat/0205613 (2002).

[34] R. Cohen and S. Havlin, Ultra Small World in Scale-Free Networks. *Preprint* cond-mat/0205476 (2002).

[35] M. E. J. Newman, S. H. Strogatz and D. J. Watts, Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118, (2001).

[36] B. Bollobás and O. Riordan, The diameter of a scale-free random graph process *Preprint* (2001).

[37] G. Szabó, M. Alava and J. Kertész, Shortest paths and load scaling in scale-free trees *Preprint* cond-mat/0203278 (2002).

[38] D. Stauffer and A. Aharony, *Introduction to Percolation Theory*. Taylor and Francis, London, 2nd edition (1992).

[39] D. S. Callaway, M. E. J. Newman, S. H. Strogatz and D. J. Watts, Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.* **85**, 5468–5471 (2000).

[40] R. Albert, H. Jeong and A.-L. Barabási, Attack and error tolerance of complex networks. *Nature* **406**, 378–382 (2000).

[41] R. Cohen, K. Erez, D. ben-Avraham and S. Havlin, Breakdown of the Internet under Intentional Attack. *Phys. Rev. Lett.* **86**, 3682–3685 (2001).

[42] R. Cohen, D. ben-Avraham and S. Havlin, Percolation Critical Exponents in Scale-Free Networks. Cond-mat/0202259 (2002); *Phys. Rev. E* (in press, 2002).

[43] G. H. Weiss, *Aspects and Applications of the Random Walk*. North-Holland, Amsterdam, (1994).

[44] H. S. Wilf, *Generatingfunctionology* 2nd ed. Academic Press, London, (1994).

[45] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200–3203 (2001).

[46] R. Pastor-Satorras and A. Vespignani. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E* **63**, 066117 (2001).

[47] Y. Moreno, R. Pastor-Satorras and A. Vespignani, Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. B* **26**, 521–529 (2002).

[48] D. S. Callaway, J. E. Hopcroft, J. M. Kleinberg, M. E. J. Newman and S. H. Strogatz, Are randomly grown graphs really random? *Phys. Rev. E*, **64**, 041902 (2001).

[49] S. N. Dorogovtsev and J. F. F. Mendes, Anomalous percolating properties of growing networks *Phys. Rev. E* **64**, 066110 (2001).

[50] Z. Burda, J. D. Correia and A. Krzywicki, Statistical ensemble of scale-free random graphs. *Phys. Rev. E* **64**, 046118 (2001).

[51] F. Chung, and L. Y. Lu, The diameter of sparse random graphs. *Adv. Appl. Math.*, **26**, 257, (2001).

[52] S. H. Strogatz, Exploring complex networks. *Nature* **410**, 268–276 (2001) .

[53] S. N. Dorogovtsev, J. F. F. Mendes and A. N. Samukhin, Giant strongly connected component of directed networks. *Phys. Rev. E* **64**, 025101 (2001).

[54] R. Albert, H. Jeong and A.-L. Barabási, Diameter of the world-wide web. *Nature* **401**, 130–131 (1999).

[55] N. Schwartz, R. Cohen, D. ben-Avraham, A.-L. Barabaśi and S. Havlin, Percolation in Directed Scale-Free Networks. *Phys. Rev. E* **66**, in press (2002).

[56] R. Cohen, D. ben-Avraham and S. Havlin, Efficient immunization of populations and computers. *Preprint* cond-mat/0207387 (2002).

[57] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley and Y. Åberg, The web of human sexual contacts. *Nature* **411**, 907–908 (2001).

[58] M. E. J. Newman, The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* **98**, 404–409 (2001).

[59] R. V. Sole and J. M. Montoya, Complexity and fragility in ecological networks *Proc. Roy. Soc. Lond. B Bio.* **268**, 2039–2045, (2001).

[60] R. M. Anderson, and R. M. May, *Infectious Diseases of Humans*. Oxford University Press, Oxford (1991).

[61] A. L. Lloyd and R. M. May, How viruses spread among computers and people. *Science* **292**, 1316–1317 (2001).

[62] C. P. Warren, L. M. Sander, I. Sokolov, C. Simon and J. Koopman, Percolation on disordered networks as a model for epidemics. *Math. Biosci.,* in press (2002).

[63] M. E. J. Newman, Exact solutions of epidemic models on networks. Preprint cond-mat/0201433 (2002).

[64] Z. Dezsö and A.-L. Barabási, Halting viruses in scale-free networks, *Preprint* cond-mat/0107420 (2001).

[65] R. Pastor-Satorras and A. Vespignani, Immunization of complex networks. *Phys. Rev. E* **65**, 036104 (2001).

[66] W. H. Wethcote and J. A. Yorke, Gonorrhea transmission dynamics and control. *Lecture notes in Biomathematics* **56**, (Springer-Verlag, 1984).

[67] K.-I. Goh, B. Kahng, D. Kim, Universal Behavior of Load Distribution in Scale-free Networks. *Phys. Rev. Lett.* **87**, 278701 (2001).

[68] L. A. Adamic, R. M. Lukose, A. R. Puniyani, B. A. Huberman, Search in Power-Law Networks. *Phys. Rev. E* **64**, 046135 (2001).

5 Epidemics and immunization in scale-free networks

Romualdo Pastor-Satorras and Alessandro Vespignani

Abstract

In this chapter we want to provide a review of the main results obtained in the modeling of epidemic spreading in scale-free networks. In particular, we want to show the different epidemiological framework originated by the lack of any epidemic threshold and how this feature is rooted in the extreme heterogeneity of the scale-free networks' connectivity pattern.

5.1 Introduction

Epidemic models are heavily affected by the connectivity patterns characterizing the population in which the infective agent spreads [1–3]. Many models map this pattern in terms of networks, in which nodes represent individuals and links represent the possible contacts along which the epidemic diffuses. In this perspective scale-free (SF) networks [4–6] represents a very interesting cases since they exhibit a power-law connectivity distribution

$$P(k) \sim k^{-\gamma} \tag{5.1}$$

for the probability $P(k)$ that a node of the network has k connections to other nodes. For connectivity exponents in the range $2 < \gamma \leq 3$ this fact implies that each node (element of the population) has a statistically significant probability of having a very large number of connections compared to the average connectivity $\langle k \rangle$ of the network. In mathematical terms, the implicit divergence of $\langle k^2 \rangle$ is signalling the extreme heterogeneity of the connectivity pattern, and it easy to foresee that this property is going to change drastically the behavior of epidemic outbreaks in SF networks. The interest in the study of epidemic models in SF networks is enhanced by the evidence that both the Internet [7–11] and the maps of human sexual contacts [12] are characterized by scale-free connectivity properties. The Internet and the web of sexual contacts are, in fact, the natural environment in which cyber viruses and sexually transmitted diseases (STD), respectively, live and proliferate. The study and characterization of epidemics in scale-free networks is therefore of potential importance for those seeking to control and arrest human and electronic plagues. These considerations motivate the analysis of the effect of complex network topologies in standard epidemic models leading to several interesting and novel results [13–15].

Perhaps, the most surprising result, first originated by the inspection of the susceptible-infected-susceptible (SIS) model, is that the spread of infections is tremendously strengthened on SF networks [15, 16]. Opposite to standard models, epidemic processes in these networks

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

do not possess any epidemic threshold below which the infection cannot produce a major epidemic outbreak or an endemic state. In principle, SF networks are prone to the persistence of diseases whatever infective rate they may have. The same peculiar absence of an epidemic threshold has been readily confirmed in other epidemic models such as the susceptible-infected-removed (SIR) model [17–19] and appears as a general feature of epidemic spreading in SF networks. This feature reverberates also in the choice of immunization strategies [20,21] and changes radically the standard epidemiological framework usually adopted in the description and characterization of disease propagation.

In this chapter we want to provide a review of the main results obtained in the modeling of epidemic spreading in SF networks. In particular, we want to show the different epidemiological framework originated by the lack of any epidemic threshold and how this feature is rooted in the extreme heterogeneity of the SF networks' connectivity pattern. As a real world example of epidemic spreading occurring on SF connectivity patterns, we shall consider the diffusion of computer viruses. Computer virus spreading, in fact, can be characterized by simple population models that do not consider properties such as gender, sex, or age, that must be included in the modeling of STD and other kinds of epidemics. On the other hand, computer viruses proliferate in the Internet, that is a capital example of SF network, and it is natural to include this topology in their modeling. Finally, many real data are available in computers epidemiology and we can use them to show experimentally the failure of the standard epidemic framework and support the new picture arising for SF networks.

We also present how the scale-free nature of the network calls for different immunization strategies in order to eradicate infections. Opposite to standard models, it is found that SF networks do not acquire global immunity from major epidemic outbreaks even in the presence of unrealistically high densities of randomly immunized individuals. Successful immunization strategies, therefore, can be developed only by taking advantage of the inhomogeneous connectivity properties of the scale-free connectivity patterns. Finally we consider the effect of the network finite size, referring to real systems which are actually made up by a finite number of individuals. The presented results provide a general view of the novel features of epidemic models in SF networks that, besides the application to computer viruses, prompt to the relevant implications of these studies in human and animal epidemiology.

5.2 Computers and epidemiology

In a classic paper [22], it is described the Domain-Name-Server (DNS) cache corruption spreading as a *natural computer virus* proliferating on the Internet. Computers on the Internet rely upon DNS servers to translate Internet protocol addresses into computer names and vice-versa. On their turn, DNS servers communicate with their DNS peers to share and update these informations. The updating is periodic in time and in the meanwhile, translation tables are "cached" and eventually transmitted to the other DNS peers. If any portion of this cache is corrupted, the DNS server will provide incorrect addresses not only to requesting computers but to DNS peers as well, propagating the error. At the same time, any DNS server can get "cured" by an updating with an error-free DNS peer. The same kind of processes can occur with routing tables exchanged by routers. This propagation of errors occurring on routers and servers that are physically linked is a typical example of epidemic process, in which the corruption (virus) is transmitted from infected to healthy individuals.

From a more familiar point of view, however, computer viruses are usually referred to as little programs that can reproduce themselves by infecting other programs [23, 24]. The basic mechanism of infection is as follows: When the virus is active inside the computer, it is able to copy itself, by different ways, into the code of other, clean, programs. When the newly infected program is run into another computer, the code of the virus is executed first, becoming active and being able to infect other programs. Apart from reproducing themselves, computer viruses perform threatening tasks that range from flashing innocuous messages on the screen to seriously corrupt data stored in the computer. These deleterious effects render most computer viruses as dangerous as their biological homonyms, and explain the interest, both commercial and scientific, arisen around their study.

Computer viruses have evolved in time (driven of course by their programmers' skills), adopting different strategies that take advantage of the different weak points of computers and software. Computer viruses can be classified into three main classes, or *strains* [24]. The first strain includes *file viruses* that infect application programs. A second and more harming family contains the *boot-sector viruses* that infect the boot sector of floppy disks and hard drives, a portion of the disk containing a small program in charge of loading the operating system of the computer. A third and nowadays prevailing strain is formed by the *macro viruses*. These viruses are independent of the platform's hardware and infect data files, such as documents produced with spreadsheets or word processors. They are coded using the *macro* instructions that are appended in the document, instructions used to perform a set of automatic actions, such as formatting the documents or typing long sequences of characters. In addition, with the ever more efficient deployment of antivirus software, more harmful viruses combining together the properties of the main strains have been developed.

Noticeably, however, the nowadays dominant and most aggressive type of cyber organisms is represented by the *worms* family. Worms are actually viruses infecting the computer with mechanisms similar to usual viruses and making a particularly effective use of the e-mail for infecting new computers. In fact, by using the instructions of some commercial mail software applications, worms are capable of sending themselves to all the e-addresses found in the address-book of the person receiving the infected mail. This possibility renders worms the most effective viruses, especially in terms of the velocity at which they can propagate starting from a single infection.

The spreading of computer viruses has been studied for long years, in close analogy with the models developed for the study of the transmission of biological diseases (for a review see Refs. [25, 26]). In this biological framework, the key point is the description of the epidemic process in terms of *individuals* and their *interactions*. In this simplified formalism, individuals can only exist in a discrete set of states, such as susceptible (or healthy), infected (and ready to spread the disease), immune, dead (or removed), etc. On the other hand, the interactions among individuals are schematized in the structure of the contacts along which the epidemics can propagate. Within this formalism, the system can be described as a *network* or graph [27], in which the nodes represent the individuals and the links are the connections along which the epidemics propagates.

Standard epidemiological models usually consider *homogeneous* networks, which are those that have a connectivity distribution peaked at an average connectivity $\langle k \rangle$, and decaying exponentially fast for $k \ll \langle k \rangle$ and $k \gg \langle k \rangle$. A typical example of deterministic homogeneous network is the standard hypercubic lattice, while among the random homogeneous network we

can count the Erdős-Rényi model [28] and the Watts-Strogatz model [29]. On the other hand, as we shall see in the following, computer viruses and worms spread in environments characterized by scale-free connectivities. This will lead to the failure of the standard epidemic picture and will naturally introduce the scale-free connectivity as an essential ingredient for the understanding of computer viruses.

5.3 Epidemic spreading in homogeneous networks

The simplest epidemiological model one can consider is the susceptible-infected-susceptible (SIS) model [2, 3]. In the SIS model, individuals can only exist in two discrete states, namely, susceptible and infected. These states completely neglect the details of the infection mechanism within each individual. The disease transmission is also described in an effective way. At each time step, each susceptible node is infected with probability ν if it is connected to one or more infected nodes. At the same time, infected nodes are cured and become again susceptible with probability δ, defining an effective *spreading rate*

$$\lambda = \frac{\nu}{\delta}. \tag{5.2}$$

Without lack of generality, we set can set $\delta = 1$, since it only affects at the definition of the time scale of the disease propagation. Individuals thus run stochastically through the cycle

$$\text{susceptible} \rightarrow \text{infected} \rightarrow \text{susceptible},$$

and hence the name of the model. The SIS model does not take into account the possibility of individuals removal due to death or acquired immunization which would lead to the so-called susceptible-infected-removed (SIR) model [2,3]. It is mainly used as a paradigmatic model for the study of infectious disease leading to an endemic state with a stationary and constant value for the density of infected individuals, i.e. the degree to which the infection is widespread in the population. The SIS has been adopted in the modeling of computer viruses and worms since, also in the presence of antiviruses, computer immunization statistically depends upon the user concerns in not skipping the antivirus control when opening e-mail attachments or new files.

The analytical study of the SIS model can be undertaken in terms of a dynamical mean-field (MF) theory. For homogeneous networks, in which the connectivity fluctuations are very small, we can approach the MF theory by means of a reaction equation for the total prevalence $\rho(t)$, defined as the density of infected nodes present at time t. That is, we can consider all the nodes as equivalent, irrespective of their corresponding connectivity. The reaction equation for $\rho(t)$ can be written as

$$\partial_t \rho(t) = -\rho(t) + \lambda \langle k \rangle \rho(t) [1 - \rho(t)]. \tag{5.3}$$

The MF character of this equation stems from the fact that we have neglected the density correlations among the different nodes. In Eq. (5.3) we have also ignored all higher order corrections in $\rho(t)$, since we are interested in the onset of the infection close to the point $\rho(t) \ll 1$. The first term on the right-hand-side in Eq. (5.3) considers infected nodes becoming healthy with unit rate. The second term represents the average density of newly infected nodes

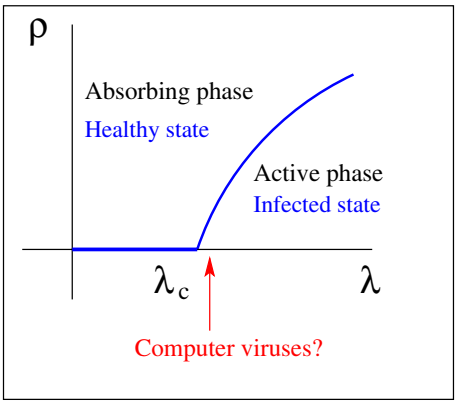


Figure 5.1: Schematic phase diagram for the SIS model in homogeneous networks. The epidemic threshold λ_c separates an active or infected phase, with finite prevalence, from an absorbing or healthy phase, with null prevalence. The very small prevalence and long lifetimes observed in computer virus data are only compatible with a value of λ infinitesimally close to the epidemic threshold.

generated by each active node. This is proportional to the infection spreading rate, λ, to the number of links emanating from each node, and to the probability that a given link points to a healthy node, $[1-\rho(t)]$. In the homogeneous networks we are considering here, connectivity has only very small fluctuations ($\langle k^2 \rangle \sim \langle k \rangle$) and as a first approximation we have considered that each node has the same number of links, $k \simeq \langle k \rangle$. This is equivalent to an homogeneity assumption for the system's connectivity. In writing this last term of the equation we are also assuming the *homogeneous mixing hypothesis* [3], which asserts that the force of the infection (the per capita rate of acquisition of the disease for the susceptible individuals) is proportional to the density of infected individuals $\rho(t)$. The homogeneous mixing hypothesis is indeed equivalent to a mean-field treatment of the model, in which one assumes that the rate of contacts between infectious and susceptibles is constant, and independent of any possible source of heterogeneity present in the system. Another implicit assumption of this model is that the time scale of the disease is much smaller than the lifespan of individuals; therefore we do not include in the equations terms accounting for the birth or natural death of individuals.

After imposing the stationarity condition $\partial_t \rho(t) = 0$, we obtain the equation, valid for the behavior of the system at large times,

$$\rho\left[-1+\lambda\langle k\rangle(1-\rho)\right]=0 \tag{5.4}$$

for the steady state density ρ of infected nodes. This equation defines an epidemic threshold $\lambda_c = \langle k \rangle^{-1}$, and yields:

$$\rho = 0 \qquad \text{if } \lambda < \lambda_c, \tag{5.5}$$

$$\rho = (\lambda - \lambda_c)/\lambda \qquad \text{if } \lambda \geq \lambda_c. \tag{5.6}$$

The most significant prediction of this model is the presence of a nonzero *epidemic threshold* λ_c [2, 30]. If the value of λ is above the threshold, $\lambda \geq \lambda_c$, the infection spreads and be-

comes persistent. Below the threshold, $\lambda < \lambda_c$, the infection dies exponentially fast. From the point of view of nonequilibrium phase transitions, the SIS model exhibits an *absorbing-state phase transition* [30] at the threshold λ_c, separating an active or infected phase, with finite prevalence, from an absorbing or healthy phase, with null prevalence. A qualitative picture of the phase diagram of this transition is depicted in Figure 5.1. It is easy to recognize that the SIS model is a generalization of the contact process model, widely studied in this context as the paradigmatic example of an absorbing-state phase transitions to a unique absorbing state [30].

To summarize, the main prediction of the SIS model in homogeneous networks is the presence of a *positive* epidemic threshold, proportional to the inverse of the average number of neighbors of every node, $\langle k \rangle$, below which the epidemics always dies, and endemic states are impossible.

5.4 Real data analysis

The statistical properties of computer virus data have been analyzed by several authors, in close analogy with the classical epidemiology of biological diseases [1–3]. Within this framework, studies have focused specially in the measurement of the virus *prevalence*, defined as the average fraction of computers infected with respect to the total number of computers present. From these studies [24, 26, 31] two main conclusions have been drawn. First, viruses which are able to survive in order to produce a significant outbreak usually reach an *endemic* or metastable steady state, with a stationary prevalence. The second empirical observation is that these endemic viruses do attain in general a very small average prevalence, that can be of the order of one out of 1000 computers or less.

More recently, other studies [15] have focused in the dynamics of the spreading process, measuring the *surviving probability* of homogeneous groups of viruses, classified according to their infection mechanism (strains). In these studies one considers the total number of viruses of a given strain that are born and die within a given observation window. The surviving probability $P_s(t)$ of the strain is defined as the fraction of viruses still alive at time t after their birth. Figure 5.2 reproduces the results reported in Ref. [15], obtained from prevalence data from the *Virus Bulletin*[1] in the period February 1996 to March 2000, covering a time interval of 50 months.

Figure 5.2 shows that the surviving probability suffers a sharp drop in the first two months of a virus' life. On the other hand, Figure 5.2 also shows for larger times a clean exponential tail,

$$P_s(t) \sim \exp(-t/\tau), \tag{5.7}$$

where τ represents the characteristic life-time of the virus strain. The numerical fit of the data [15] yields $\tau \simeq 14$ months for boot and macro viruses and $\tau \simeq 6 - 9$ months for file viruses.

When comparing the theoretical picture delivered by the SIS model on homogeneous networks with the behavior observed in real computer viruses, one is faced with an unexpected

[1] Virus prevalence data publicly available at the web site `http://www.virusbtn.com`.

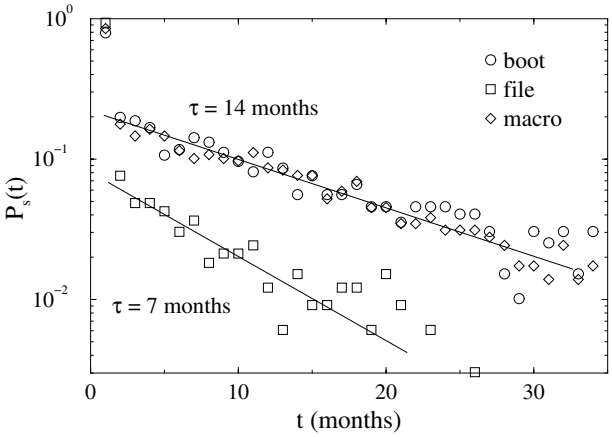


Figure 5.2: Surviving probability for the three main strands of computer viruses. After a sharp initial drop, it is clear the presence of an exponential decay, with an associated characteristic time τ that depends on the given strand.

and paradoxical conclusion. First of all, the extremely low prevalence shown by endemic viruses is only compatible with the phase diagram sketched in Figure 5.1 in the very *unlikely* chance that all surviving viruses are constructed such that their respective spreading rate λ is tuned infinitesimally close to λ_c, above the epidemic threshold. On the other hand, the characteristic life times observed in the analysis of the surviving probability of the different virus strains are impressively large if compared with the interval in which anti-virus software is available on the market (usually within days or weeks after the first incident report) and corresponds to the occurrence of metastable endemic states. Such a long lifetime on the scale of the typical spread/recovery rates would suggest an effective spreading rate larger than the epidemic threshold, which is in contradiction with the always low prevalence levels of computer viruses but in the case of an unrealistic tuning of all viruses to the system epidemic threshold. In summary, the comparison with the known experimental data points out that the view obtained so far with the modeling of computer viruses is very instructive, but fails to represent, even at a qualitative level, the nature of the real phenomenon. The explanation of this discrepancy has been claimed to be one of the most important open problems in computer virus epidemiology [32].

The key point to elucidating the riddle posed by computer viruses resides in the capacity of many of them to propagate via data exchange with communication protocols (FTP, e-mails, etc.) [24]. Viruses will spread preferentially to computers which are highly connected to the outer world and are thus proportionally exchanging more data and information. It is thus rather intuitive to consider the scale-free Internet connectivity as the effective one on which the spreading occurs. For instance, this is the case of *natural computer viruses* which spread on the topology identified by routers and servers [7–9]. Apart from the Internet, scale-free properties emerge also in the world-wide-web [4,33] and in social networks [34]. The fact that all virus strains show the same statistical features indicates that very likely all of them spread

on scale-free connectivity patterns. Further and strong support for this conclusion comes from the recent study of a social network of e-mail exchange within a community of users [35], which was proven to have a scale-free connectivity, with an exponent close to 2. This finding has an immediate repercussion on the modeling of worms, whose spreading environment is in fact given by this kind of network.

The conclusion from the above arguments is that computer viruses spread in a scale-free network, in which, even though the average connectivity is well defined, the connectivity fluctuations are unbounded; i.e. there is always a finite probability that a node has a number of neighbors much larger than the average value. These fluctuations in the connectivity are the key difference with respect to the epidemic models discussed in homogeneous graphs, and they must be included in a correct characterization of the system.

5.5 Epidemic spreading in scale-free networks

In order to fully take into account connectivity fluctuations in a analytical description of the SIS model, we have to relax the homogeneity assumption used for homogeneous networks, and work instead with the relative density $\rho_k(t)$ of infected nodes with given connectivity k; i.e. the probability that a node with k links is infected. The dynamical mean-field equations can thus be written as [15, 16]

$$\frac{d\rho_k(t)}{dt} = -\rho_k(t) + \lambda k \left[1 - \rho_k(t)\right] \Theta[\{\rho_k(t)\}], \tag{5.8}$$

where also in this case we have considered a unitary recovery rate and neglected higher order terms ($\rho_k(t) \ll 1$). The creation term considers the probability that a node with k links is healthy $[1 - \rho_k(t)]$ and gets the infection via a connected node. The probability of this last event is proportional to the infection rate λ, the real number of connections k, and the probability $\Theta[\{\rho_k(t)\}]$ that any given link points to an infected node. We make the assumption that Θ is a function of the partial densities of infected nodes $\{\rho_k(t)\}$. In the steady (endemic) state, the ρ_k are functions of λ. Thus, the probability Θ becomes also an implicit function of the spreading rate, and by imposing the stationarity condition $\partial_t \rho_k(t) = 0$, we obtain

$$\rho_k = \frac{k\lambda\Theta(\lambda)}{1 + k\lambda\Theta(\lambda)}. \tag{5.9}$$

This set of equations show that the higher the node connectivity, the higher the probability to be in an infected state. This inhomogeneity must be taken into account in the computation of $\Theta(\lambda)$. The exact calculation of Θ for a general network is a very difficult task. However, we can exactly compute its value for the case of a *random* SF network, in which there are no correlations among the connectivities of the different nodes [15, 16]. Indeed, the probability that a link points to a node with s connections is equal to $sP(s)/\langle k\rangle$, which yields an average probability of a link pointing to an infected node

$$\Theta(\lambda) = \frac{1}{\langle k\rangle} \sum_k kP(k)\rho_k. \tag{5.10}$$

Since ρ_k is on its turn a function of $\Theta(\lambda)$, we obtain a self-consistency equation that allows to find $\Theta(\lambda)$ and an explicit form for Eq. (5.9). Finally, we can evaluate the order parameter (persistence) ρ using the relation

$$\rho = \sum_k P(k)\rho_k. \tag{5.11}$$

The self-consistent Eqs. (5.9) and (5.10) can be approximately solved, in the limit of small Θ, for any scale-free connectivity distribution [16]. However, we can very easily calculate the epidemic threshold by just noticing that λ_c is the value of λ above which it is possible to obtain a nonzero solution for Θ. In fact, from Eqs. (5.9) and (5.10), we obtain the self-consistent relation

$$\Theta = \frac{1}{\langle k\rangle}\sum_k kP(k)\frac{\lambda k\Theta}{1+\lambda k\Theta}, \tag{5.12}$$

where Θ is now a function of λ alone [15, 16]. The solution $\Theta = 0$ is always satisfying the consistency equation. A non-zero stationary prevalence ($\rho_k \neq 0$) is obtained when the right-hand-side and the left-hand-side of Eq. (5.12), expressed as function of Θ, cross in the interval $0 < \Theta \leq 1$, allowing a nontrivial solution. It is easy to realize that this corresponds to the inequality

$$\frac{d}{d\Theta}\left(\frac{1}{\langle k\rangle}\sum_k kP(k)\frac{\lambda k\Theta}{1+\lambda k\Theta}\right)\bigg|_{\Theta=0} \geq 1 \tag{5.13}$$

being satisfied. The value of λ yielding the equality in Eq. (5.13) defines the critical epidemic threshold λ_c, that is given by

$$\frac{\sum_k kP(k)\lambda_c k}{\langle k\rangle} = \frac{\langle k^2\rangle}{\langle k\rangle}\lambda_c = 1 \quad \Rightarrow \quad \lambda_c = \frac{\langle k\rangle}{\langle k^2\rangle}. \tag{5.14}$$

This results implies that in SF networks with connectivity exponent $2 < \gamma \leq 3$, for which $\langle k^2\rangle \to \infty$ in the limit of a network of infinite size, we have $\lambda_c = 0$.

5.5.1 Analytic solution for the Barabási-Albert network

In order to discuss in detail a specific example, it is simpler to consider a toy model of SF network, which is easy to generate for simulation purposes and shows the correct connectivity properties. The paradigmatic example of SF network is the Barabási and Albert (BA) model [4, 5, 36]. The construction of the BA graph starts from a small number m_0 of disconnected nodes; every time step a new vertex is added, with m links that are connected to an old node i with probability

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}, \tag{5.15}$$

where k_i is the connectivity of the i-th node. This algorithm implements the so-called "rich-get-richer" paradigm [4], that implies that highly connected nodes have always larger chances

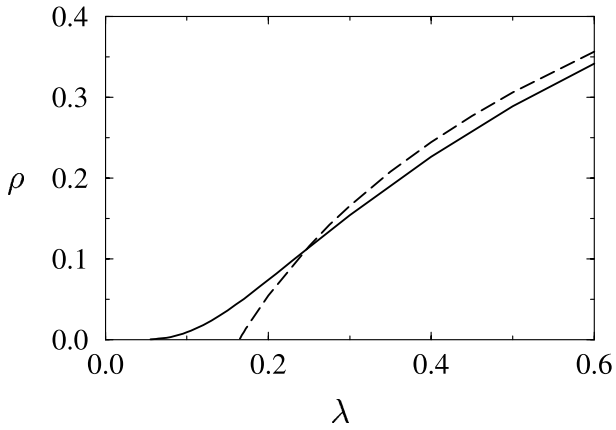


Figure 5.3: Total prevalence ρ for the SIS model in a BA network (full line) as a function of the spreading rate λ, compared with the theoretical prediction for a homogeneous network (dashed line).

to become even more connected. The networks generated this way have a connectivity distribution $P(k) \sim k^{-3}$.

In the explicit calculations for the BA model, we use a continuous k approximation that allows the practical substitution of series with integrals [4]. The full connectivity distribution is thus given by $P(k) = 2m^2k^{-3}$. By noticing that the average connectivity is $\langle k \rangle = \int_m^\infty kP(k)dk = 2m$, Eq. (5.10) gives

$$\Theta(\lambda) = m\lambda\Theta(\lambda)\int_m^\infty \frac{1}{k}\frac{dk}{1+k\lambda\Theta(\lambda)} = m\lambda\Theta(\lambda)\log\left(1+\frac{1}{m\lambda\Theta(\lambda)}\right), \tag{5.16}$$

which yields the solution

$$\Theta(\lambda) = \frac{e^{-1/m\lambda}}{\lambda m}(1-e^{-1/m\lambda})^{-1}. \tag{5.17}$$

In order to find the behavior of the density of infected nodes we have to solve Eq. (5.11), that reads as

$$\rho = 2m^2\lambda\Theta(\lambda)\int_m^\infty \frac{1}{k^2}\frac{dk}{1+k\lambda\Theta(\lambda)} = 2m^2\lambda\Theta(\lambda)\left[\frac{1}{m}+\lambda\Theta(\lambda)\log\left(1+\frac{1}{m\lambda\Theta(\lambda)}\right)\right]. \tag{5.18}$$

By substituting the obtained expression for $\Theta(\lambda)$ we find at lowest order in λ

$$\rho \sim 2e^{-1/m\lambda} \tag{5.19}$$

This result shows the absence of any epidemic threshold or critical point in the model; i.e., $\lambda_c = 0$, in agreement with the result from Eq. (5.14) for a scale-free network with $\langle k^2 \rangle = \infty$.

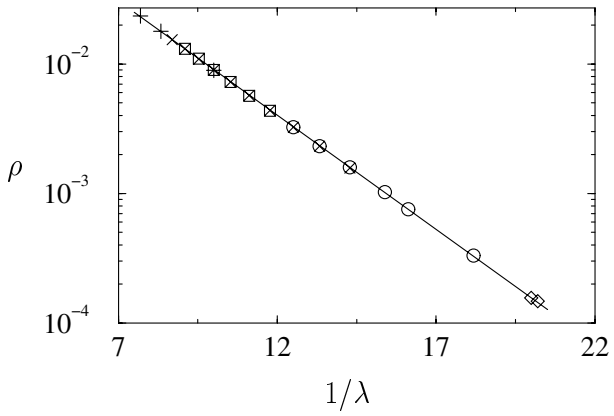


Figure 5.4: Persistence ρ as a function of $1/\lambda$ for BA networks of different sizes: $N = 10^5$ (+), $N = 5 \times 10^5$ (□), $N = 10^6$ (×), $N = 5 \times 10^6$ (○), and $N = 8.5 \times 10^6$ (◇). The linear behavior on the semi-logarithmic scale proves the stretched exponential behavior predicted for the persistence. The full line is a fit to the form $\rho \sim \exp(-C/\lambda)$.

Numerical simulations of the SIS model performed on a BA network confirm the analytical picture extracted from the mean-field analysis. Figure 5.3 shows the total prevalence ρ in the steady state as a function of the spreading rate λ [16]. As we can observe, it approaches zero in a continuous and smooth way, compatible with the presence of a vanishing epidemic threshold (see for comparison the behavior expected for a homogeneous network, also drawn in Figure 5.3). On the other hand, Figure 5.4 represents ρ in a semilogarithmic plot as a function of $1/\lambda$, which shows that $\rho \sim \exp(-C/\lambda)$, where C is a constant independent of the size N of the network.

The spreading dynamical properties of the model can also be studied by means of numerical simulations [16]. For example, the surviving probability $P_s(t)$ for a fixed value of λ and different network sizes N is represented in Figure 5.5. In this case, we recover an exponential behavior in time, that has its origin in the finite size of the network. In fact, for any finite system, the epidemic will eventually die out because there is a finite probability that all individuals cure the infection at the same time. This probability is decreasing with the system size and the lifetime is infinite only in the thermodynamic limit $N \to \infty$. However, the lifetime becomes virtually infinite (the metastable state has a lifetime too long for our observation window) for large enough sizes that depend upon the spreading rate λ. This is a well-known feature of the survival probability in finite size absorbing-state systems poised above the critical point [30]. In our case, this picture is confirmed by numerical simulations that show that the average lifetime of the survival probability is increasing with the network size for all the values of λ.

The outcome of the analysis presented in this section is that the SIS model in a BA scale-free network, with connectivity distribution $P(k) \sim k^{-\gamma}$ and connectivity exponent $\gamma = 3$, yields the absence of any epidemic threshold or critical point, $\lambda_c = 0$. It is worth remarking that the present framework can be generalized to networks with $2 < \gamma \leq 3$, recovering

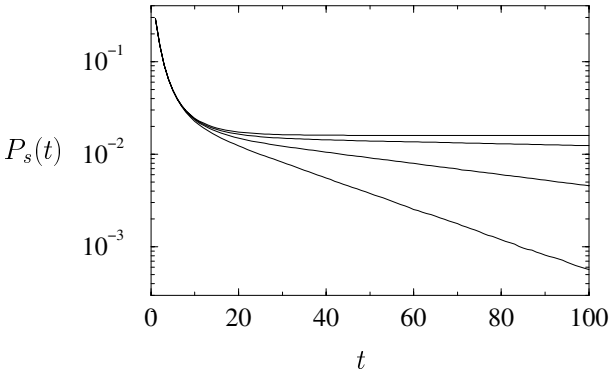


Figure 5.5: Surviving probability $P_s(t)$ as a function of time in supercritical spreading experiments in the BA network. Spreading rate $\lambda = 0.065$. Network sizes ranging from $N = 6.25 \times 10^3$ to $N = 5 \times 10^5$ (bottom to top).

qualitatively the same results [16]. Only for $\gamma > 4$, epidemics on SF networks have the same properties as on homogeneous networks. The emerging picture for epidemic spreading in scale-free networks emphasizes the role of topology in epidemic modeling. In particular, the absence of epidemic threshold and the associated critical behavior in a wide range of scale-free networks provide an unexpected result that radically changes many standard conclusions on epidemic spreading. This indicates that infections can proliferate on these scale-free networks whatever spreading rates they may have. These very bad news are, however, balanced by the exponentially small prevalence for a wide range of spreading rates ($\lambda \ll 1$). This picture fits perfectly with the observations from real data, and solve the long-standing mystery of the generalized low prevalence of computer viruses without assuming any global tuning of the spreading rates. In addition, the model explains successfully the exponential time decay of the virus surviving probability, with an average lifetime of viral strains that appears to be related to an effective spreading rate and the network size.

5.5.2 Finite size scale-free networks

Real systems are actually made up by a finite number of individuals which is far from the thermodynamic limit. This finite population introduces a maximum connectivity k_c, depending on N, which has the effect of restoring a bound in the connectivity fluctuations, inducing in this way an effective nonzero threshold. More generally, we can consider bounded scale-free networks in which the connectivity distribution has the form $P(k) \sim k^{-\gamma} f(k/k_c)$, where the function $f(x)$ decreases very rapidly for $x > 1$ [37,38]. The cut-off k_c can be due to the finite size of the network or to the presence of constraints limiting the addition of new links in an otherwise infinite networks. In both cases, $\langle k^2 \rangle$ assumes a finite value in bounded SF networks, defining from Eq. (5.14) an effective nonzero threshold due to finite size effects as usually encountered in nonequilibrium phase transitions [30]. This epidemic threshold, however, is not an *intrinsic* quantity as in homogeneous systems and it vanishes for a increasing network size or connectivity cut-off. Explicit calculations can be performed for the SIS model [39] in

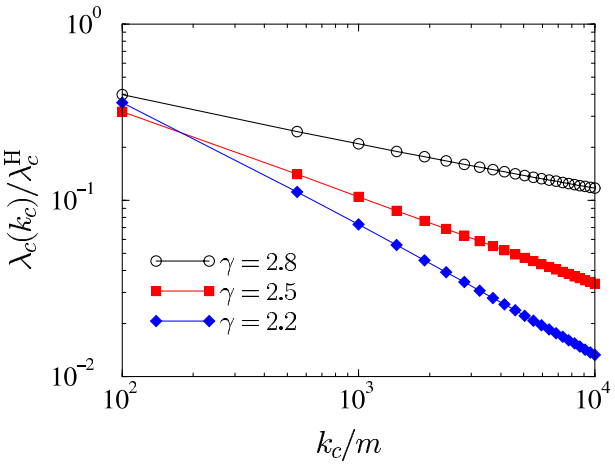


Figure 5.6: Ratio between the effective epidemic threshold in bounded SF networks with exponential cut-off k_c and the intrinsic epidemic threshold of homogeneous networks with the same average connectivity, for different values of γ.

SF networks with exponentially bounded connectivity, $P(k) \sim k^{-\gamma} \exp(-k/k_c)$, obtaining that the effective nonzero epidemic threshold $\lambda_c(k_c)$ induced by the cut-off k_c behaves as

$$\lambda_c(k_c) \simeq (k_c/m)^{\gamma-3}, \tag{5.20}$$

where m is the smallest connectivity in the graph. The limit $\gamma \to 3$, on the other hand, corresponds to a logarithmic divergence, yielding at leading order $\lambda_c(k_c) \simeq (m \ln(k_c/m))^{-1}$. In all cases we have that the epidemic threshold vanishes when increasing the characteristic cut-off. It is thus interesting to compare the intrinsic epidemic threshold obtained in homogeneous networks with negligible fluctuations and the nonzero effective threshold of bounded SF networks. The intrinsic epidemic threshold of homogeneous networks with constant node connectivity $\langle k \rangle$ is given by $\lambda_c^{\rm H} = \langle k \rangle^{-1}$ [2]. In Fig. 5.6 we report the ratio obtained by using the full expression for $\lambda_c(k_c)$. It is striking to observe that, even with relatively small cut-offs ($k_c \sim 10^2 - 10^3$), for $\gamma \approx 2.5$ the effective epidemic threshold of finite size SF networks is smaller by a factor close to $1/10$ than the intrinsic threshold obtained on homogeneous networks. This implies that the SF networks weakness to epidemic agents is also present in finite-size and connectivity-bounded networks. Using the homogeneity assumption in the case of SF networks will lead to a serious over-estimate of the epidemic threshold even for relatively small networks.

5.6 Immunization of scale-free networks

As we have seen in Section 5.5, epidemic processes in SF networks do not possess, in the limit of an infinitely large network, an epidemic threshold below which diseases cannot set into an endemic state. SF networks are in this sense very prone to the spreading and persistence of

infections, whatever virulence (parametrized by the spreading rate λ) the infective agent might possess. In view of this weakness, it becomes a major task to find optimal immunization strategies oriented to minimize the risk of epidemic outbreaks in SF networks.

5.6.1 Uniform immunization

The simplest immunization procedure one can consider consists in the random introduction of immune individuals in the population [3], in order to get a uniform immunization density. In this case, for a fixed spreading rate λ, the relevant control parameter in the density of immune nodes present in the network, the immunity g. At the mean-field level, the presence of a uniform immunity will have the effect of reducing the spreading rate λ by a factor $1-g$; i.e. the probability of finding and infecting a susceptible and nonimmune node will be $\lambda(1-g)$. For homogeneous networks we can easily see that, for a constant λ, the stationary prevalence is given in this case by

$$\rho_g = 0 \qquad \text{if } g > g_c \,, \tag{5.21}$$

$$\rho_g = (g_c - g)/(1-g) \qquad \text{if } g \leq g_c \,, \tag{5.22}$$

where g_c is the critical immunization value above which the density of infected individuals in the stationary state is null and depends on λ as

$$g_c = 1 - \frac{\lambda_c}{\lambda}. \tag{5.23}$$

Thus, for a uniform immunization level larger than g_c, the network is completely protected and no large epidemic outbreaks are possible. On the contrary, uniform immunization strategies on SF networks are totally ineffective. The presence of uniform immunization is able to locally depress the infection's prevalence for any value of λ, but it does so too slowly, and it is impossible to find any critical fraction of immunized individuals that ensures the infection eradication. After a moment's reflection, one can convince oneself of the reason of this failure: With the uniform immunization strategy we are giving the same weight to very connected nodes (with the largest infection potential) and to nodes with a very small connectivity (which are relatively safe). Due to the large fluctuations in the connectivity, heavily connected nodes, which are statistically very significant, can overcome the effect of the immunization and maintain the endemic state. On the other hand, the absence of an epidemic threshold ($\lambda_c = 0$) in the thermodynamic limit implies that whatever rescaling $\lambda \to \lambda(1-g)$ of the spreading rate does not eradicate the infection except the case $g = 1$. Indeed, by inserting Eq. (5.14) into Eq. (5.23) we have that the immunization threshold is given by

$$1 - g_c = \frac{1}{\lambda} \frac{\langle k \rangle}{\langle k^2 \rangle}. \tag{5.24}$$

In SF networks with $\langle k^2 \rangle \to \infty$ only a complete immunization of the network (i.e. $g_c = 1$) ensures an infection-free stationary state. The fact that uniform immunization strategies are less effective has been noted in several cases of spatial heterogeneity [3]. In SF networks we face a limiting case due to the extremely high (virtually infinite) heterogeneity in the connectivity properties. Specifically, it follows from Eq. (5.19) that the SIS model on the BA network

shows for $g \simeq 1$ and any λ the prevalence

$$\rho_g \simeq 2\exp[-1/m\lambda(1-g)]. \tag{5.25}$$

In other words, the infection always reaches an endemic state if the network size is large enough (see Fig. 5.7(a)). This fact points out the absence of an immunization threshold; SF networks are weak in face of infections, also after massive uniform vaccination campaigns.

5.6.2 Targeted immunization

We have seen in Section 5.6.1 that the very peculiar nature of SF networks hinders the efficiency of naive uniform immunization strategies. However, we can take advantage of the heterogeneity of SF networks, by devising an immunization strategy that takes into account the inherent hierarchy in the network's nodes. In fact, it has been shown that SF networks posses a noticeable resilience to *random* connection failures [40–42], which implies that the network can resist a high level of damage (disconnected links), without loosing its global connectivity properties; i.e. the possibility to find a connected path between almost any two nodes in the system. At the same time, SF networks are strongly affected by *selective* damage; if a few of the most connected nodes are removed, the network suffers a dramatic reduction of its ability to carry information [40–42]. Applying this argument to the case of epidemic spreading, we can devise a *targeted* immunization scheme in which we progressively make immune the most highly connected nodes, i.e., the ones more likely to spread the disease. While this strategy is the simplest solution to the optimal immunization problem in heterogeneous populations [3], its efficiency is comparable to the uniform strategies in homogeneous networks with finite connectivity variance. In SF networks, on the contrary, it produces an arresting increase of the network tolerance to infections at the price of a tiny fraction of immune individuals.

We can make an approximate calculation of the immunization threshold in the case of a random SF network [20]. Let us consider the situation in which a fraction g of the individuals with the highest connectivity have been successfully immunized. This corresponds, in the limit of a large network, to the introduction an upper cut-off k_t—which is obviously an implicit function of the immunization g—, such that all nodes with connectivity $k > k_t$ are immune. The introduction of immune nodes implies at the same time the elimination of all the links emanating from them, which translates, in a mean-field approximation, into a probability $p(g)$ of deleting any link in the network. This elimination of links yields a new connectivity distribution, for which all moments can be computed. Recalling Eq. (5.14), we can then compute the critical fraction g_c of immune individuals needed to eradicate the infection. An explicit calculation for the BA network [20] yields the approximate solution for the immunization threshold in the case of targeted immunization as

$$g_c \simeq \exp(-2/m\lambda). \tag{5.26}$$

This clearly indicates that the targeted immunization program is extremely convenient in SF networks where the critical immunization is exponentially small in a wide range of spreading rates λ.

In order to assess the efficiency of the targeted immunization scheme we show in Fig. 5.7 the results from numerical simulations of the SIS model on BA networks, together with the

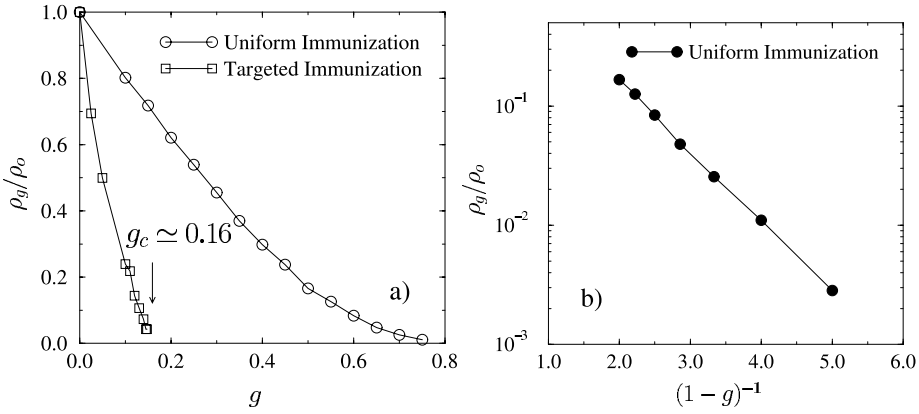


Figure 5.7: a) Reduced prevalence ρ_g/ρ_0 from computer simulations of the SIS model in the BA network with uniform and targeted immunization, at a fixed spreading rate $\lambda = 0.25$. A linear extrapolation from the largest values of g yields an estimate of the threshold $g_c \simeq 0.16$ in BA networks with targeted immunization. b) Check of the predicted functional dependence $\rho_g \sim \exp(-1/m\lambda(1-g))$ for the SIS model in the BA network with uniform immunization.

results from simulations with uniform immunization [20]. In particular the plot shows the reduced prevalence ρ_g/ρ_0, where ρ_0 is the prevalence in the nonimmunized network, as a function of the fraction of immunized nodes g, at a fixed spreading rate $\lambda = 0.25$. Fig. 5.7(a) indicates that, for uniform immunization, the prevalence decays slowly when increasing g, and will be effectively null only for $g = 1$, as predicted by Eq. (5.25). In fact, the plot in Fig. 5.7(b) recovers the theoretical predicted behavior Eq. (5.25). On the other hand, for the targeted immunization, the prevalence shows a very sharp drop and exhibits the onset of an immunization threshold above which the system in infection-free. A linear regression from the largest values of g yields an approximate estimation $g_c \simeq 0.16$, that definitely proves that SF networks are very sensitive to the targeted immunization of a very small fraction of the most connected nodes.

This result can be readily extended to SF networks with arbitrary γ values, and it is also possible to devise alternative immunization schemes which take advantage of the SF connectivity patterns in order to achieve a high level of tolerance to infections [20]. Other strategies have been put forward in Ref. [21] by proposing to cure with proportionally higher rates the most connected nodes. Also in this case it is possible to reintroduce a threshold in the network with this hub-biased policy for the administered cures.

The present results indicate that the SF networks' susceptibility to epidemic spreading is reflected also in an intrinsic difficulty in protecting them with local—uniform—immunization. On a global level, uniform immunization policies are not satisfactory and only targeted immunization strategies successfully lower the vulnerability of SF networks. This evidence radically changes the usual perspective of the regular epidemiological framework. Spreading of infectious or polluting agents on SF networks, such as food or social webs, might be contrasted only by a careful choice of the immunization procedure. In particular, these procedures

should rely on the identification of the most connected individuals. The protection of just a tiny fraction of these individuals raises dramatically the tolerance to infections of the whole population. The computer virus case is once again providing support to this picture. Despite deployment of antivirus software is timely and capillary, viruses' lifetimes are extremely long; in other words, very high levels of immunization are not able to eradicate the epidemic. In the standard epidemic framework this would be possible only in the case of very high spreading rate for the virus that is in contradiction with the always small prevalence of epidemic outbreaks. These empirical findings are, however, in good agreement with the picture obtained for the immunization of SF networks. In fact, the antivirus deployment is not eradicating the epidemics on the global scale since it is alike to a random immunization process where file scanning and antivirus updating are statistically left to the good will of users and system managers. Needless to say, from the point of view of the single user, antiviruses are extremely important, being the only way to ensure local protection for the computer.

5.7 Conclusions

The topology of the network has a great influence in the overall behavior of epidemic spreading. The connectivity fluctuations of the network play a major role by strongly enhancing the infection's incidence. This issue assumes a particular relevance in the case of SF networks that exhibit connectivity fluctuations diverging with the increasing size N of the web. Here we have reviewed the new epidemiological framework obtained in population networks characterized by a scale-free connectivity pattern. SF networks are very weak in face of infections, presenting an effective epidemic threshold that is vanishing in the limit $N \to \infty$. In an infinite population this corresponds to the absence of any epidemic threshold below which major epidemic outbreaks are impossible. SF networks' susceptibility to epidemic spreading is reflected also in an intrinsic difficulty in protecting them with local—uniform—immunization policies. Only targeted immunization procedures achieve the desired lowering of epidemic outbreaks and prevalence.

The present picture qualitatively fits the observations from real data of computer virus spreading, and could solve the long standing problem of the generalized low prevalence and long lifetime of computer viruses without assuming any global tuning of the spreading rates. Moreover, recent findings on the web of human sexual contacts [12] prompt that the presented results could have potentially interesting implications also in the case of human sexual disease control.

In order to illustrate the new features of epidemic spreading in SF networks, we used the SIS model. It is important to stress, however, that the analysis on SF networks of different models, such as the SIR model, confirm the presented epidemiological picture [17–19]. Yet, many other ingredients concerning the infection mechanisms should be considered in a more realistic representation of real epidemics [2,3]. In addition, simple rules defining the temporal patterns of the networks, such as the frequency of forming new connections, the actual time that a connection exists, or different types of connections, should be included in the modeling. These dynamical features are highly valuable experimental inputs which are necessary ingredients in the use of complex networks theory in epidemic modeling.

Acknowledgements

This work has been partially supported by the European Network Contract No. ERBFMRXCT980183 and by the European Commission - Fet Open project COSIN IST-2001-33555. R.P.-S. acknowledges financial support from the Ministerio de Ciencia y Tecnología (Spain).

References

[1] N. T. J. Bailey, *The mathematical theory of infectious diseases*, (Griffin, London, 1975). 2nd edition.

[2] O. Diekmann and J.A.P Heesterbeek, *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, (John Wiley & Sons, New York, 2000).

[3] R. M. Anderson and R. M. May, *Infectious diseases in humans*, (Oxford University Press, Oxford, 1992).

[4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science* **286**, 509–511 (1999).

[5] S. N. Dorogovtsev, J.F.F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett.* **85**, 4633–4636 (2000).

[6] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationship of the Internet topology. *Comput. Commun. Rev.* **29**, 251–263 (1999).

[8] G. Caldarelli, R. Marchetti, and L. Pietronero. The fractal properties of Internet. *Europhys. Lett.* **52**, 386 (2000).

[9] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the Internet. *Phys. Rev. Lett.* **87**, 258701 (2001).

[10] S.-H. Yook, H. Jeong, and A.-L. Barabási. Modeling the Internet's large-scale topology, (2001). e-print cond-mat/0107417.

[11] K.-I. Goh, B. Kahng, and D. Kim. Fluctuation-driven dynamics of the internet topology. *Phys. Rev. Lett* **88**, 108701 (2002).

[12] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg. The web of human sexual contacts. *Nature* **411**, 907–908 (2001).

[13] G. Abramson and M. Kuperman. Small world effect in an epidemiological model. *Phys. Rev. Lett.* **86**, 2909–2912 (2001).

[14] C. Moore and M. E. J. Newman. Epidemics and percolation in small-world networks. *Phys. Rev. E* **61**, 5678 (2000).

[15] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200–3203 (2001).

[16] R. Pastor-Satorras and A. Vespignani. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E* **63**, 066117 (2001).

[17] R. M. May and A. L. Lloyd. Infection dynamics on scale-free networks. *Phys. Rev. E* **64**, 066112 (2001).

[18] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. B* **26**, 521–529 (2002).

[19] M. E. J. Newman. Exact solutions of epidemic models on networks, (2001). e-print cond-mat/0201433.

[20] R. Pastor-Satorras and A. Vespignani. Immunization of complex networks. *Phys. Rev. E* **65**, 036104 (2001).

[21] Z. Dezsö and A.-L. Barabási. Halting viruses in scale-free networks, (2001). e-print cond-mat/0107420.

[22] S. M. Bellovin. Packets found on an Internet. *Comput. Commun. Rev.* **23**, 26–31 (1993).

[23] C. D. Harley, R. Slade, D. Harley, E. H. Spafford, and U E. Gattiker, *Viruses Revealed*, (McGraw-Hill, New York, 2001).

[24] J. O. Kephart, G. B. Sorkin, D. M. Chess, and S. R. White. Fighting computer viruses. *Scientific American* **277**(5), 56–61 November (1997).

[25] J. O. Kephart and S. R. White. Directed-graph epidemiological models of computer viruses. In *Proceedings of the 1991 IEEE computer society symposium on research in security and privacy (SSP '91)*, 343–361 (IEEE, Washington - Brussels - Tokyo, 1991).

[26] J. O. Kephart, S. R. White, and D. M. Chess. Computers and epidemiology. *IEEE Spectrum* **30**, 20–26 (1993).

[27] G. Chartrand and L. Lesniak, *Graphs & digraphs*, (Wadsworth & Brooks/Cole, Menlo Park, 1986).

[28] P. Erdös and P. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–60 (1960).

[29] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).

[30] J. Marro and R. Dickman, *Nonequilibrium phase transitions in lattice models*, (Cambridge University Press, Cambridge, 1999).

[31] J. O. Kephart and S. R. White. Measuring and modeling computer virus prevalence. In *Proceedings of the 1993 IEEE computer society symposium on security and privacy (SSP '93)*, 2–15 (IEEE, Washington - Brussels - Tokyo, 1993).

[32] S. R. White. Open problems in computer virus research. Virus Bulletin Conference, Munich, October (1998). Available on-line at http://www.av.ibm.com/ScientificPapers/White/Problems/Problems.html.

[33] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world-wide web. *Nature* **401**, 130–131 (1999).

[34] S. Wasserman and K. Faust, *Social network analysis*, (Cambridge University Press, Cambridge, 1994).

[35] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks, (2002). e-print cond-mat/0201476, Phys. Rev. E, in press.

[36] A.-L. Barabási, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A* **272**, 173–187 (1999).

[37] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA* **97**, 11149–11152 (2000).

[38] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics* **51**, 1079–1187 (2002).

[39] R. Pastor-Satorras and A. Vespignani. Epidemic dynamics in finite size scale-free networks. *Phys. Rev. E* **65**, 035108 (2002).

[40] R. A. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).

[41] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Network robustness and fragility: percolation on random graphs. *Phys. Rev. Lett.* **85**, 5468–5471 (2000).

[42] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Breakdown of the Internet under intentional attack. *Phys. Rev. Lett.* **86**, 3682–3685 (2001).

6 Cells and genes as networks in nematode development and evolution

Ralf J. Sommer

6.1 Introduction

During the last decades, the concept of networks became very useful in many areas of modern biology. Biological systems are organized in a hierarchical fashion, generating networks of interacting modules at many levels of organization. Not only ecosystems, populations of species and individual organisms, but also organs, tissues and cells within one organism, as well as genetic networks and signaling cascades, build networks of interacting units within a hierarchical system [26]. At each level, multiple units (modules) exist that interact with one another to form complex networks. Therefore, the concept of networks can be applied to many, if not to all biological phenomena. In the following article, I would like to describe a case study in evolutionary developmental biology that indicates how the concept of networks can be used in developmental and evolutionary biology.

Developmental biology is amongst the most active research fields in modern biology, in particular the model systems Drosophila melanogaster, Caenorhabditis elegans, Xenopus laevis, zebrafish and mouse. The general question in developmental biology is to understand the mechanisms and principles governing the ontogeny of individual organisms. What are the processes by which an unicellular egg gives rise to the adult? How do cells interact during development? Which genes control the proliferation, determination and differentiation of cells? Although developmental biology has been a descriptive discipline for most of the last century, the use of genetic and molecular techniques during the last 20 to 30 years, resulted in the detailed mechanistic understanding of selected developmental processes in a handful of model systems [42].

Evolutionary biology has a somewhat overlapping, but nonetheless different intention. Evolutionary biology tries to understand the patterns and processes that gave rise to the origin and diversification of life. How many species are there? What are their phylogenetic relationships? How does speciation occur? What are the mechanisms of evolution? And what type of alterations in the developmental program causes new evolutionary phenotypes? One common theme in evolutionary and developmental biology is the diversification of structures - during ontogeny of an individual (developmental biology) and during phylogeny (evolutionary biology). Since both disciplines use a holistic approach, the concept of networks is of elementary importance in both of them.

There has been a long search for connections between developmental processes and evolutionary patterns. For example, observations from developmental biology were with the first

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

ones to support Darwins theory of common descent. However, only with the major breakthroughs in molecular developmental genetics, a new interdisciplinary research field, "evolutionary developmental biology", has been initiated that now tries to reveal the relationship between both fields [13, 24]. It is now possible to analyze how developmental systems evolve and how developmental modules are modified over evolutionary timescales, thereby providing the unique opportunity to combine experimental approaches to developmental biology with the comparative research tradition of evolutionary biology. In the following, I will introduce one such case study in evolutionary developmental biology, indicating how the network concept can be used in developmental and evolutionary biology and also, how developmental networks change during evolution.

6.2 Nematode developmental biology: studying processes at a cellular level

The free-living soil nematode Caenorhabditis elegans is an important model system in developmental biology (Fig. 1A). Free-living nematodes like C. elegans have "invariant" cell lineages and the adult organisms consists only of around 1000 somatic cells [37]. The term "Invariability" means that the cell division pattern during development is highly reproducible and nearly identical between all individuals of a species. Given the invariability and the simplicity of free-living nematodes, all cells of the worms body are known, i.e. the self-fertilizing hermaphrodites of C. elegans consists of 959 somatic cells, the males have 1031 cells [25,43]. The existence of invariable cell lineages in nematodes allows developmental processes to be studied at the level of individual cells and the interactions among cells to be analyzed. When complemented with genetic and molecular analyses, an integrated understanding of developmental processes at the cellular, genetic and molecular level can be achieved. Thereby, networks of interacting modules in adjacent hierarchical levels of organization are generated.

6.3 Nematode Vulva formation as a case study

A developmental process that has been studied in great detail in C. elegans is the formation of the vulva, the egg-laying structure and copulatory organ of the hermaphrodite. The vulva is generated by three precursor cells, which divide within five hours to give rise to the 22 cells eventually forming the complete organ. The ease by which this cellular system can be analyzed, attracted researchers already in the seventies, when the postembryonic lineage was originally described by Sulston and Horvitz [37]. Over the years, experimental studies on C. elegans vulva development revealing the cell-cell interactions during cell fate specification were complemented by genetic and molecular approaches and finally gave insight into the mechanisms of this pattern formation process. The vulva is a derivative of the ventral epidermis, a part of the worm that consists of 12 precursor cells, called P1.p to P12.p. These cells form a linear array between the pharynx and the rectum and differentiate in a position-specific manner (Fig. 2A, B) [37].

P(1,2,9-11).p remain epidermal and fuse with the hypodermal syncytium, whereas P(3-8).p in the central body region remain unfused and form the so-called vulva "equivalence

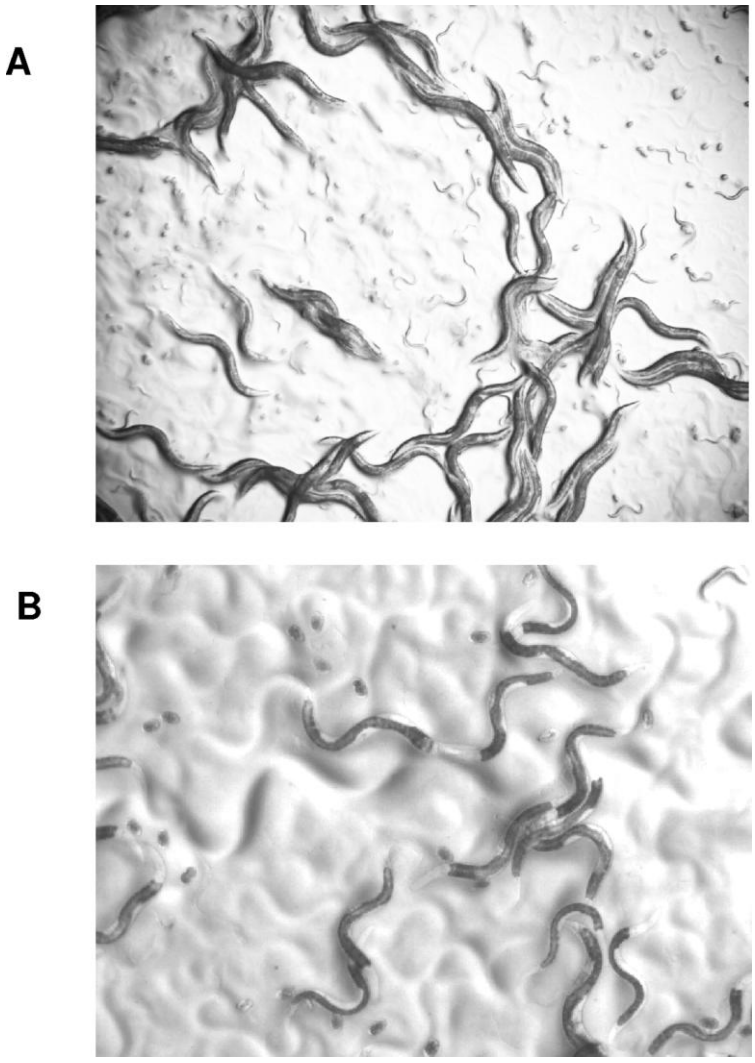


Figure 6.1: Laboratory cultures of nematodes using E. coli as food source. Comparison of the morphology of Caenorhabditis elegans (A) and Pristionchus pacificus (B) Adult worms are appr. 1 mm in size.

group" (Fig. 2B). In principle, all six cells of the vulva equivalence group can participate in vulva formation. Therefore, these cells have been designated as "vulval precursor cells" (VPC). In a wild-type animal however, only the three central cells, P(5-7).p divide to form vulval tissue, whereas P(3,4,8).p remain epidermal. The latter cell fate has been designated as 3° fate [38, 39]. P(5,7).p generate 7 progeny, which form the outer part of the vulva, respectively (Fig. 2C). P6.p generates 8 progeny forming the central part of the vulva which connects

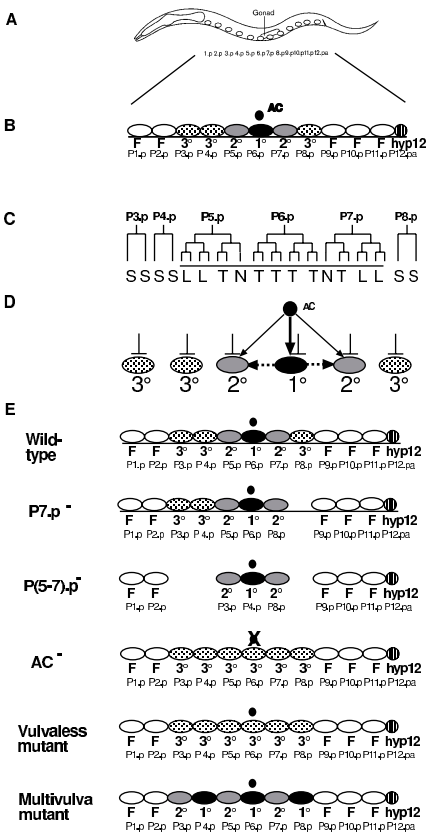


Figure 6.2: Schematic summary of vulva formation in C. elegans. (A) During the L1 stage, the 12 ventral epidermal cells, P(1-12).p, are equally distributed between pharynx and rectum. (B) P(1,2,9-11).p fuse with the hypodermal syncytium hyp7 (F, white ovals). P(3-8).p form the vulva equivalence group and adopt one of three alternative cell fates. P6.p has a 1° fate (black oval), and P(5,7).p have a 2° fate (grey ovals). P(3,4,8).p have a 3° fate and remain epidermal (dotted ovals). The anchor cell (AC, black circle) provides an inductive signal for vulva formation. (C) Cell lineage pattern of the vulval precursor cells. P(3,4,8).p divide once and then fuse with hyp7 (S). P(5,7).p generate seven progeny each. The first two cell divisions occur along the anteroposterior axis, the third division can be longitudinal (L), transversal (T) or can be absent (N). P6.p generates eight progeny. (D) Schematic summary of signaling interactions during vulva formation in C. elegans. An inductive EGF-like signal originates from the AC (black arrows). P6.p signals its neighbors to adopt a 2° fate via 'lateral signaling' (dotted arrows). Negative signaling (bars) prevents inappropriate vulva differentiation. See text for details. (E) Summary of cell ablation experiments. After ablation of P7.p, P8.p adopts a 2° fate and forms part of the vulva. After ablation of P(5-7).p, P(3,4,8).p can form a functional vulva. After ablation of the AC, all precursor cells adopt a 3° fate. Vulvaless mutants have a similar phenotype as AC-ablated animals. Multivulva mutants show a phenotype opposite to vulvaless mutants, namely the ectopic proliferation of P(3,4,8).p.

the uterus to the outside environment (Fig. 2C). The fate of P(5,7).p has been designated as the 2° fate, the one of P6.p as the 1° fate [38, 39] (Fig. 2B). These hierarchical cell fate designations resulted from combinatorial cell ablation experiments, which indicated that cells with a lower fate (i.e. 3° cells) can replace cells with a higher fate (2° or 1° cells) (Fig. 2E). Another important result of the classical cell ablation experiments was that vulva formation is induced by a signal from the gonadal anchor cell (AC) (Fig. 2D, E) [21]. The AC is morphologically distinct from the surrounding cells of the somatic gonad and eventually makes contact with the progeny of P6.p. When the AC was ablated, P(5-7).p had an epidermal fate, like P(3,4,8).p and no vulva was formed (Fig. 2E). To understand the cell replacement capabilities among the six vulval precursor cells (VPCs) and the inductive interaction with the AC at the molecular level, intensive genetic studies were carried out and many vulva-defective mutants were isolated [11]. The general idea behind carrying out genetic screens in order to understand the molecular nature of a developmental process is that the proteins involved in pattern formation are encoded by genes, which can be individually altered by introducing single mutations in the genome. When such gene "knock-out" approaches were carried out, many mutants were identified that change the development of the vulva in a similar way as the physical ablation experiments discussed above. Thus, the molecular mechanisms of vulva formation or any other developmental process can be understood by studying genes, that when mutated, result in a specific (vulva) phenotype related to the structure of interest. Most vulva mutants fall into one of two phenotypic classes. In "vulvaless" mutants, no vulva is formed and in many such mutants all VPCs have a 3° cell fate, resembling AC-ablated animals (Fig. 2E). In contrast, "multivulva" mutants show ectopic vulva differentiation by P(3,4,8).p (Fig. 2E). Molecular studies revealed that these mutants constitute at least four different signaling systems that have to interact with one another during vulva formation. The inductive signal for vulva development is an epidermal growth factor-like molecule, encoded by the gene lin-3 and is specifically expressed in the AC [15]. This signal is transmitted by an EGFR-RAS/MAPK pathway within the VPCs (Fig. 2D) (for review see [40]). This inductive signal acts in a redundant fashion with a lateral signal, involving the Notch-like molecule lin-12 of C. elegans (for review see [14]). Both pathways act together to induce the proper 2°-1°-2° pattern (Fig. 2D). Negative signaling prevents inappropriate and precocious vulva differentiation and consists itself of two redundant signaling systems [12]. Recent work indicated that lin-35 and lin-53 encode an Rb-like molecule and its binding protein RbAp48 [22] and that the NURD complex is also involved in negative signaling (Fig. 2D) [2, 28]. Also, canonical Wnt signaling was shown to play a role in the regulation of the transcription factor lin-39, an important regulator of vulval cell fate specification [5]. In summary, intensive genetic and molecular studies over the last 20 years have provided a detailed mechanistic understanding of vulva formation in C. elegans. Networks of interacting units can be defined at the cellular and genetic level. Communication between cells is required during vulva induction, cell fate determination and during the final stages of differentiation and morphogenesis to form a perfect vulva. Genes are involved in all of these processes and are organized in at least four characterized distinct pathways. These detailed studies in C. elegans also provide a platform for the evolutionary analysis of this process.

6.4 Nematode collections

Nematodes are one of the most prolific animal phyla with an estimated ten to hundred million species in all types of ecosystems [1]. Besides animal and plant parasitic species, free-living species are found in aquatic and terrestrial systems. Free-living nematodes can be isolated from soil samples around the world and many strains can be easiliy cultured in the laboratory on lawns of E. coli as a food source [32]. In recent years, species collections of free-living nematodes have been established in several laboratories. Comparative studies on vulva development were initiated by cell lineage and cell ablation studies in different species of the Rhabditidae, the family to which C. elegans belongs to [31, 33]. By now, these studies have been extended to the family Diplogastridae [29, 34] and members of four different families of the Cephalobina [8–10]. In the following, the major cellular changes during nematode vulva evolution are reviewed.

6.5 Cellular networks: how cells change their function

6.5.1 Evolution of vulva position

All nematodes analyzed so far, have 12 ventral epidermal cells that can be homologized based on their position along the anteroposterior body axis. One important difference between nematodes is the position of the vulva within the animal. Whereas most species form their vulvae in the central body region, others form the vulva in a species-specific position in the posterior body region. In some species, like in members of the genus Teratorhabditis, the vulva forms at 95% body length, just anterior to the rectum [31, 32]. Cell lineage analysis in members of the genera Mesorhabditis, Teratorhabditis, Cruznema [31] and Brevibucca [10] revealed that in all analyzed species with a posterior vulva, a similar mechanism has been used: the centrally born cells migrate towards the posterior, stop in a species-specific position and undergo vulva differentiation (Fig. 3A, B). In all cases, except for Cruznema, vulva formation occurs in a gonad-independent way; i.e. a proper vulva is formed even after the ablation of the somatic gonad, including the AC (Fig. 3C) [10, 31]. Thus, posterior vulva formation involves two important deviations from C. elegans: first, in all species the central cells constitute the vulva equivalence group and the VPCs migrate towards posterior prior to differentiation. Second, a shift in vulva position is associated with a different mode of vulval patterning, which no longer relies on gonadal signaling. Projecting these vulval character states on a phylogenetic tree of nematodes indicates that posterior vulva formation represents a derived character that evolved several times independently [30].

6.5.2 Evolution of vulval cell fate specification

Besides vulva position, cell fate specification of the VPCs changed enormously during nematode evolution. One species, in which the ventral epidermal cells differ from C. elegans is P. pacificus (Fig. 1B). Many of the characters seen in P. pacificus are unique to this species or its close relatives. Generally, P. pacificus has a four day life-cycle and is a hermaphroditic species, like C. elegans [35]. Four major differences were observed by comparing vulva formation between P. pacificus and C. elegans. First, seven of the 12 ventral epidermal cells die

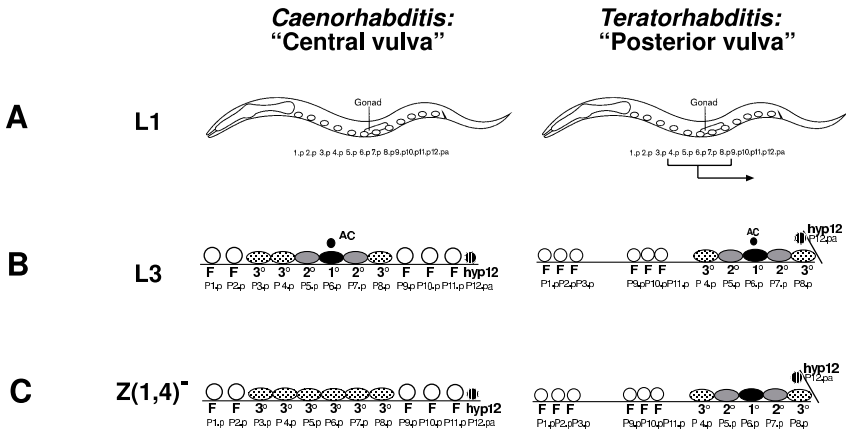


Figure 6.3: Evolutionary variation in vulva position. C. elegans, like most other free-living nematodes forms the vulva in the central body region. Other species, like Teratorhabditis palmarum form their vulvae in the posterior body region. (A) The 12 Pn.p cells are equally distributed in the ventral region during the first larval stage (L1). (B) During the second larval stage, P(4-8).p migrate all the way to the posterior in T. palmarum and stop in the region of the rectum. P(5-7).p form vulval tissue with a 2°-1°-2° pattern. (C) After ablation of the precursor cells of the somatic gonad, Z(1,4), no vulva is formed in C. elegans, whereas a normal vulva forms in T. palmarum indicating that vulva formation does not rely on gonadal signaling.

of apoptosis, shortly after these cells are born during late embryogenesis (Fig. 4A) [34]. The vulva itself is built from P(5-7).p with a cell lineage pattern very similar to the one in C. elegans. Second, vulva induction is a continuous process, involving several cells of the somatic gonad rather than a single AC, as in C. elegans [27] (Fig. 4E). Cell ablation experiments of the somatic gonad, at different stages during postembryonic development indicated that somatic cells start to provide the inductive signal long before the AC is born. However, to form a proper vulva with a 2°-1°-2° pattern, the AC is still strictly required. Ablating only the AC, P(5-7).p all have a 2° fate [27]. Third, among the four surviving cells in the central body region, P8.p, the only cell not participating in wild-type vulva formation, represents a new cell type with characteristics unknown in VPCs from other analyzed nematode species [17]. P8.p provides a lateral inhibitory signal that influences the cell fate decision of P(5,7).p but not P6.p upon gonadal induction (Fig. 4B-D). If P(6-8).p are ablated, P5.p has the highest potential cell fate, the 1° fate (Fig. 4C). However, if only P(6,7).p are ablated, P5.p will in the presence of an epidermal P8.p adopt a 2° fate (Fig. 4B). These results indicate that P8.p, influences the cell fate decision that P5.p takes in response to gonadal induction over a distance. No such interaction has been observed in C. elegans or any other nematode studied before. This new type of lateral inhibition also requires the mesoblast M, the precursor of all postembryonically derived mesodermal tissue (Fig. 4D). In addition, P8.p also provides a negative signaling, the molecular nature of which remains unknown (Fig. 4E). Given all of these differences, Pristionchus pacificus has been selected as a satellite species, to investigate vulva formation using genetic and molecular tools [7, 30, 35].

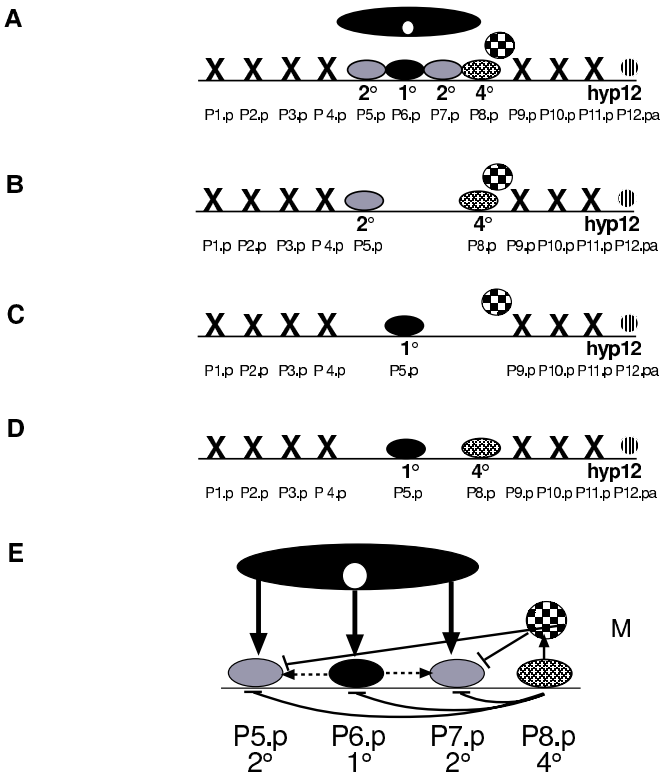


Figure 6.4: Schematic summary of vulva formation in P. pacificus. (A) Cell fate specification of the 12 ventral epidermal cells. P(1-4,9-11).p die of programmed cell death during late embryogenesis. P(5-7).p form the vulva with a 2°-1°-2° pattern. P8.p (shaded oval) has a special fate designated as 4°. (B) After ablation of P(6,7).p, P5.p has a 2° fate in the presence of P8.p. (C) After ablation of P(6-8).p, P5.p has a 1° fate, indicating that the presence of P8.p can influence the cell fate decision of P5.p, a phenomenon that has been designated as “lateral inhibition". (D) The ablation experiments described above were in the presence of the mesoblast M (chequed circle). If P(6,7).p and M are ablated together, P8.p is unable to inhibit P5.p from adopting the 1° indicating that P8.p and M interact during lateral inhibition. (E) Model for cell-cell interactions during vulva development in P. pacificus. P8.p provides a lateral inhibition to P(5,7).p, mediated by the mesoblast M (chequed circle). Lateral inhibition influences the 1° vs. 2° cell fate decision of P(5,7).p. P8.p also provides a negative signal (black bars), which influences the vulva vs. non-vulval cell fate decision. For clarity, negative signaling is shown here as an interaction between P8.p and P(5-7).p. It is possible that indirect interactions involving other cells could exist. Inductive signaling from the somatic gonad is a continuous process (black arrows). Lateral signaling occurs between P6.p and P8.p (not indicated) and perhaps also between P6.p and P(5,7).p (dotted arrows). The gonad is shown in black with the AC as a white circle.

Taken together, the cellular analysis of vulva formation in representative nematodes of several different families indicates that the vulva in itself is a homologous organ, because it is formed by homologous precursor cells in all analyzed species. Nonetheless, the cells forming the vulva change their function and behavior during evolution. Vulva formation in the posterior rather than the central body region represents an example of how modifications at one level of organization, for instance the introduction of cell migration, can result in evolutionary novelty. Also, the cell-cell interactions during vulva formation change greatly during nematode evolution, even if the final outcome is conserved. It is the same players (modules) that form the vulva, but it is the manner in which they play and interact that changes during evolution. Thus, cellular networks are both evolutionarily stable and changeable networks. They are stable networks, because the group of cells forming vulval tissue is basically conserved between nematode species. They are changeable networks, because their behavior and interactions change during evolution.

6.6 Genetic networks: how genes change their function

Given all of the differences during vulval cell fate specification between P. pacificus and C. elegans, large scale mutagenesis screens have been carried out in order to identify important genes for P. pacificus vulva formation [7]. Many vulva-defective mutants have been isolated and the molecular characterization of several of them provides an insight into the molecular mechanisms underlying vulva formation in P. pacificus. Comparing the function of homologous genes between P. pacificus and C. elegans allows the investigation of the evolvability of developmental networks in vulva formation to be taken one step further, to the level of gene function.

6.6.1 Evolution of lin-39 function

The first gene to be studied in detail was lin-39, a homeotic gene that specifies the vulva equivalence group in C. elegans, early in development. lin-39 shows highest sequence similarity to the Drosophila gene Deformed. Cel-lin-39 mutants have a generation-vulvaless phenotype, because the vulva equivalence group is not formed and P(3-8).p fuse with the surrounding hypodermis, like their anterior and posterior counterparts (Fig. 5A, C) [4, 41]. Thus, LIN-39 provides positional information that makes the central body region different from the anterior and posterior body region. In lin-39 mutant animals, this positional information is no longer provided so that cells do not become different from one another. Using a temperature-sensitive allele of Cel-lin-39, Maloof and Kenyon [23] could show that Cel-lin-39 has a second important function during vulva induction (Fig. 5E). LIN-39 acts as a transcription factor downstream of the EGF/RAS/MAPK signaling pathway, providing specificity to vulval signaling [23]. If Cel-lin-39 is not provided during vulva induction, P(5-7).p have a 3° fate. Thus, Cel-lin-39 has two distinct functions during C. elegans vulva formation.

In P. pacificus, several generation-vulvaless mutants have been isolated, in which P(5-8).p die of programmed cell death, like their anterior and posterior counterparts. Molecular analysis indicated that mutations in Ppa-lin-39 cause this phenotype (Fig. 5B, D) [6]. Thus, Cel-lin-39 and Ppa-lin-39 have similar functions early in development in setting up the vulva

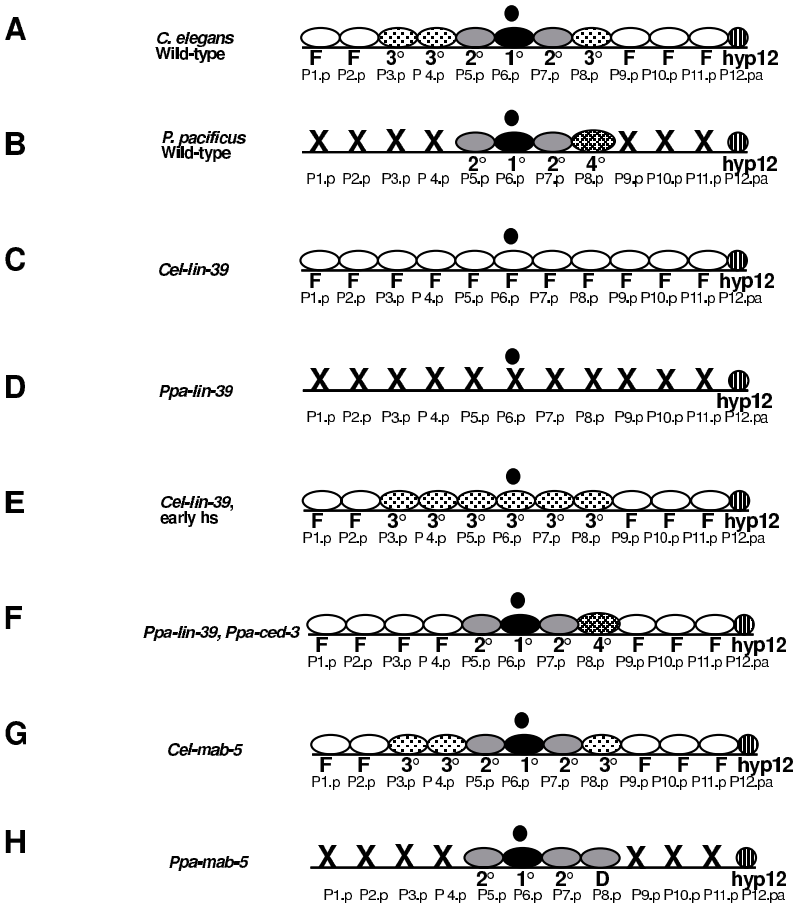


Figure 6.5: Schematic summary of the function of the homeotic genes lin-39 and mab-5 in C. elegans and P. pacificus. (A, B) C. elegans and P. pacificus wild-type vulva development as described in Figs. 2 and 4, respectively. (C) In Cel-lin-39 mutant animals, the central body region shows a homeotic transformation and P(3-8).p fuse with the surrounding hypodermis like their anterior and posterior lineage homologs. (D) Ppa-lin-39 mutant animals also show a homeotic transformation and P(5-8).p die of programmed cell death. (E) If the first function of Cel-lin-39 is rescued by providing lin-39 under the control of a heat-shock promoter, P(3-8).p have a 3° fate because Cel-lin-39 is required during vulva induction. (F) The first function of Ppa-lin-39 can be overcome by generating a Ppa-lin-39 Ppa-ced-3 double mutant. Ppa-CED-3 is a general regulator of programmed cell death and mutations in Ppa-ced-3 result in animals unable to undergo apoptosis. Such double mutants form a normal vulva indicating that in contrast to Cel-lin-39, Ppa-lin-39 is not required during vulva induction. (G) Cel-mab-5 mutant animals show no vulval patterning defects. (H) Ppa-mab-5 mutant animals show a homeotic transformation and P8.p forms an ectopic vulva-like structure. X, programmed cell death; F, cell fusion; D, ectopic vulva differentiation.

equivalence group. Cel-lin-39 prevents the fusion of P(3-8).p, whereas Ppa-lin-39 prevents the programmed cell death of P(5-8).p.

Is Ppa-lin-39 also required during vulva induction, as is Cel-lin-39? To study a potential second role of Ppa-lin-39 during vulva formation, Ppa-lin-39, Ppa-ced-3 double mutants have been generated [36]. Ppa-ced-3 encodes a Cysteine-protease that is one of the key regulators of programmed cell death in C. elegans and P. pacificus. In Ppa-ced-3 mutants, programmed cell death is not executed, resulting in animals with 12 Pn.p cells in the ventral epidermis. In Ppa-lin-39, Ppa-ced-3 double mutants, the early function of Ppa-lin-39 - the inhibition of programmed cell death of P(5-8).p - is bypassed because cell death cannot be executed at all [36]. Surprisingly, such double mutants form a normal vulva by P(5-7).p, indicating that Ppa-lin-39 has no second role during vulva induction, such as Cel-lin-39(Fig. 5F). Thus, the comparison of Ppa-lin-39 and Cel-lin-39 during vulva formation shows both conservation and change of gene function: the first function is conserved between both nematodes, whereas the second function is only present in C. elegans.

6.6.2 Evolution of mab-5 function

Besides Ppa-lin-39, Ppa-mab-5, a second homeotic gene has been shown to play an important role during cell fate specification of P(5-8).p. mab-5 is the Antennapedia-homolog of nematodes and has been originally described based on its abnormalities in C. elegans males [20]. Ppa-mab-5 was isolated based on the ectopic differentiation of P8.p in the hermaphrodite [16]. In most Ppa-mab-5 mutant animals, P(5-7).p form a normal vulva, but P8.p also differentiates and forms a small pseudovulva (Fig. 5H). The differentiation of P8.p is not dependent on gonadal induction but rather depends on a signal from the lineage of the mesoblast M. In Ppa-mab-5 mutants, the M lineage undergoes overproliferation, resulting in an egg-laying defective phenotype and an inappropriate signaling to P8.p. If the M cell is ablated in Ppa-mab-5 mutant animals, the ectopic differentiation of P8.p can be significantly reduced indicating a new type of cell-cell interaction [17,19]. Although Cel-MAB-5 plays a role in the specification of P(7,8).p, Cel-mab-5 mutants have no obvious vulva phenotype and no ectopic proliferation can be seen, as in Ppa-mab-5 (Fig. 5G) [3]. Thus in contrast to lin-39, mab-5 plays a more important role during vulva formation in P. pacificus. Besides lin-39 and mab-5, other important genes like the even-skipped homolog vab-7 have also been compared between C. elegans and P. pacificus. Again, when compared in detail, similarities and differences in gene function were observed [18].

Taken together, the genetic and molecular analysis of vulva formation between C. elegans and P. pacificus indicates that genetic networks - like celleluar networks - are stable developmental units. However, while some functions of genes are conserved, other functions clearly differ between the two species. It is this interplay between conservation and change that ensures the continuity of a homologous developmental system on the one hand and the evolvability of the structure on the other hand.

6.7 Conclusion

In free-living nematodes, developmental processes can be studied at the cellular, genetic and molecular level. Besides C. elegans, the diplogasterid species P. pacificus has been shown to be amenable to genetic analysis. Comparative studies on vulva development indicate that cellular and genetic networks are stable developmental entities. Nonetheless, individual modules in these networks (cells or genes) can change their function during evolution generating novelty, for example in the position of vulva formation. Thus, both cells and genes are evolutionarily stable and changeable developmental modules. It is the integrative comparative approach at the cellular, the genetic, and the mechanistic level, which makes the analysis of nematode vulva development a fruitful case study, indicating how developmental networks work and evolve.

Acknowledgements

I thank J. Srinivasan for critically reading the manuscript and members of the lab and my nematode colleagues for discussion.

References

[1] Blaxter, M. L., De Ley, P., Garey, J. R. et al. 1998. A molecular evolutionary framework for the phylum Nematoda. Nature 392: 71-75.

[2] Ceol, C. J. and Horvitz, H.. R.. 2001. dpl-1 DP and efl-1 E2FF act with lin-35 Rb to antagonize Ras signaling in C. elegans vulval development. Mol. Cell 7: 461-473.

[3] Clandinin, T. R., Katz, W. S. and Sternberg, P. W. 1997. Caenorhabditis elegans HOM-C genes regulate the response of vulval precursor cells to inductive signal. Dev. Biol. 182: 150-161.

[4] Clark, S. G., Chisholm, A. D. and Horvitz, H. R. 1993. Control of cell fates in the central body region of C. elegans by the homeobox gene lin-39. Cell 74: 43-55.

[5] Eisenmann, D. M., Maloof, J. N., Simske, J. S., Kenyon, C. and Kim, S. K. 1998. The β-catenin homolog BAR-1 and LET-60 Ras coordinately regulate the Hox gene lin-39 during Caenorhabditis elegans vulva development. Development 125: 3667-3680.

[6] Eizinger, A. and Sommer, R. J. 1997. The homeotic gene lin-39 and the evolution and nematode epidermal cell fates. Science 278: 452-455.

[7] Eizinger, A., Jungblut B. and Sommer, R. J. 1999. Evolutionary change in the functional specificity of genes. Trends Genet. 15: 191-196.

[8] Félix, M.-A. and Sternberg, P. W. 1997. Two nested gonadal inductions of the vulva in nematodes. Development 124: 253-259.

[9] Félix, M.-A. and Sternberg, P. W. 1998. A gonad-derived survival signal for vulval precursor cells in two nematode species. Current Biology 8: 287-290.

[10] Félix, M. A., De Ley, P., Sommer, R. J., Frisse, L., Nadler, S. A., Thomas, K., Vanfleteren, J. and Sternberg, P. W. 2000. Evolution of vulva development in the Cephalobina (Nematoda). Dev. Biol. 221: 68-86.

[11] Ferguson, E. L., Sternberg, P. W. and Horvitz, H. R. 1987. A genetic pathway for the specification of the vulval cell lineages of Caenorhabditis elegans. Nature 326: 2559-267.

[12] Ferguson, E. L. and Horvitz, H. R. 1989. The multivulva phenotype of certain C. elegans mutants results from defects in two functionally redundant pathways. Genetics 123: 109-121.

[13] Gerhard, J. and Kirschner, M. 1997. Cells, embryos and evolution. Oxford: Blackwell Science.

[14] Greenwald, I. 1997. Development of the vulva. In: C. elegans II (eds.: Riddle, D. L., Blumenthal, T., Meyer, B. J. & Priess, J. R.) pp. 519-542. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

[15] Hill, R. J. and Sternberg, P. W. 1992. The gene lin-3 encodes an inductive signal for vulval development in C. elegans. Nature 358: 470-476.

[16] Jungblut, B. and Sommer, R. J. 1998. The Pristionchus pacificus mab-5 gene is involved in the regulation of ventral epidermal cell fates. Current Biology 8: 775-778.

[17] Jungblut, B. and Sommer, R. J. 2000. Novel cell-cell interactions during vulva development in Pristionchus pacificus. Development, 127: 3295-3303.

[18] Jungblut, B. and Sommer, R. J. 2001. The nematode even-skipped homolog vab-7 regulates gonad and vulva position in Pristionchus pacificus. Development , 128: 253-261.

[19] Jungblut, B., Pires-daSilva, A. and Sommer, R. J. 2001. Formation of the egg-laying system in Pristionchus pacificus requires complex interactions between gonadal, mesodermal and epidermal tissues and does not rely on single cell inductions. Development, 128: 3395-3404.

[20] Kenyon, C. 1986. A gene involved in the development of the posterior body region of C. elegans. Cell 46: 477-487.

[21] Kimble, J. 1981. Lineage alterations after ablation of cells of the somatic gonad of Caenorhabditis elegans. Dev. Biol. 87: 286-300.

[22] Lu, X. and Horvitz, H. R. 1998. lin-35 and lin-53, two genes that antagonize a C. elegans Ras pathway, encode proteins similar to Rb and its binding protein RbAp48. Cell 95: 981-991.

[23] Maloof, J. N. and Kenyon, C. 1998. The HOX gene lin-39 is required during C. elegans vulval induction to select the outcome of Ras signaling. Development 125: 181-190.

[24] Raff, R. A. 1996. The shape of life. Chicago: The University of Chicago Press.

[25] Riddle, D. L., Blumenthal, T., Meyer, B. J. & Priess, J. R. (eds.) 1997. C. elegans II . Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

[26] Riedl, R. 1975. Die Ordnung des Lebendigen. Hamburg: Paul Parey Verlag.

[27] Sigrist, C. B. and Sommer, R. J. 1999. Vulva formation in Pristionchus pacificus relies on continuous gonadal induction. Dev. Genes Evol. 209: 451-459.

[28] Solari, F. and Ahringer, J. 2000. NURD-complex genes antagonize Ras-induced vulval development in Caenorhabditis elegans. Curr. Biol. 10: 223-226.

[29] Sommer, R. J. 1997. Evolutionary change of developmental mechanisms in the absence of cell lineage alterations during vulva formation in the Diplogastridae. Development 124: 243-251.

[30] Sommer, R. J. 2000. Evolution in worms. Curr. Opin. Gen. & Dev. 10: 443-448.

[31] Sommer, R. J. and Sternberg, P. W. 1994. Changes of induction and competence during the evolution of vulva development in nematodes. Science, 265: 114-118.

[32] Sommer, R. J., Carta, L. K. and Sternberg, P. W. 1994. The evolution of cell lineage in nematodes. Development 1994 Supplement: 85-95.

[33] Sommer, R. J. and Sternberg, P. W. 1995. Evolution of cell lineage and pattern formation in the vulval equivalence group of rhabditid nematodes. Dev. Biol. 167: 61-74.

[34] Sommer, R. J. and Sternberg, P. W. 1996. Apoptosis and change of competence limit the size of the vulva equivalence group in Pristionchus pacificus: a genetic analysis. Current Biology 6: 52-59.

[35] Sommer, R. J., Carta, L. K., Kim, S. Y. and Sternberg, P. W. 1996. Morphological, genetic and molecular description of Pristionchus pacificus sp. n. (Nematoda: Neodiplogastridae). Fund. Appl. Nemat. 19: 511-521.

[36] Sommer, R. J., Eizinger, A., Lee, K. Z., Jungblut, B., Bubeck, A. and Schlak, I. 1998. The Pristionchus Hox gene Ppa-lin-39 inhibits programmed cell death to specifiy the vulva equivalence group and is not required during vulval induction. Development 125: 3865-3873.

[37] Sulston, J. E. and Horvitz, H. R. 1977. Postembryonic cell lineages of the nematode Caenorhabditis elegans. Dev. Biol. 56: 110-156.

[38] Sulston, J. E. and White, J. G. 1980. Regulation and cell autonomy during postembryonic development in Caenorhabditis elegans. Dev. Biol. 78: 577-597

[39] Sternberg, P. W. and Horvitz, H. R. 1986. Pattern formation during vulval development in C. elegans. Cell 44: 761-772.

[40] Sternberg, P. W. and Han, M. 1998. Genetics of RAS signaling in C. elegans. Trend Genet. 14: 466-472.

[41] Wang, B. B., Müller-Immerglück, M. M., Austin, J., Robinson, N. T., Chisholm, A. and Kenyon, C. 1993. A homeotic gene cluster patterns the anteroposterior body axis of C. elegans. Cell 74: 29-42.

[42] Wolpert, L. 1998. Principles of Development. Oxford: Oxford Univ. Press.

[43] Wood, W. 1988. The nematode C. elegans. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

7 Complex networks in genomics and proteomics

Ricard V. Solé and Romualdo Pastor-Satorras

7.1 Introduction

Complex multicellular organisms contain large genomes in which each structural gene is associated with at least one regulatory element and each regulatory element integrates the activity of at least two other genes. The nature of such regulation started to be understood from the analysis of small prokaryotic regulation subsystems and the current picture indicates that the webs that shape cellular behavior are very complex. Actually, integration of extracellular signals often involves the crosstalk between signal cascades that has been suggested to share some common traits with neural networks [1]. In a related context, detailed analyses of subsets of interacting genes reveal that cell biology is highly modular [2]. Here "modules" are made up of many species of interacting molecules and the functional relevance of these subnets is highlighted by the observation that they are conserved through evolution.

In many cases, proteins composed by multiple subunits behave as switch-like elements that can flip, for example, from an active to an inactive state and back. The switching behavior of these complexes, together with the underlying information processing that takes place at the network level, allows for a computational description of intracellular signaling. In this context, one might consider some key features of standard computational systems that should apply here. One particularly important aspect is the resilience of the signaling network under different sources of perturbation. The analysis of mutational robustness in different organisms revealed an extraordinary level of homeostasis: in many cases the total suppression of a given gene in a given organism leads to a small phenotypic effect or even to no effect at all [3, 4].

Following the analogy with engineered systems, the immediate explanation for such robustness would come from the presence of a high degree of redundancy. Under mutation, additional copies of a given gene might compensate the failure of the other copy. However, the analysis of redundancy in genome data indicates that redundant genes are rapidly lost and that redundancy is not the leading mechanism responsible for mutational robustness [4].

The origins of robustness against mutations is particularly well highlighted by the analysis of genome-wide scale data of the budding yeast *Saccharomyces cerevisiae* [4]. The main conclusion of this study is that the major cause of robustness comes from the interactions among unrelated genes. This mechanism would be illustrated by the following example: given a metabolic network, completely unrelated enzymes can catalyse different reactions but contribute to a pathway whose goal is to sustain an optimal flux of metabolites. Under these conditions, mutations in genes encoding those enzymes will have little or mild effects. Additionally, it is interesting to see that many examples of experimental biotechnology manipulations involving the tinkering of one or two genes fail to reach the expected goals: very often, counterintuitive outcomes are obtained.

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

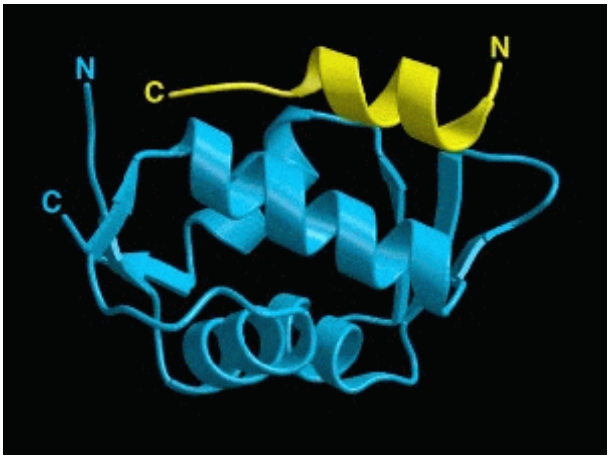


Figure 7.1: The domain of molecular interaction among p53 and MDM2 is shown in this 3D reconstruction [5]. MDM2 (here in cyan) binds a specific domain of p53 in a region (here shown in yellow) important for the interaction of p53 with components of the transcription machinery.

On the other hand, mutations involving some key genes can have very important consequences. This is the case, in particular, of the p53 tumor suppressor gene, Figure 7.1, which is known to play a critical role in genome stability and integrates many different signals related to cell-cycle or apoptosis (cell death) [6]. This and other tumor-suppressor genes prevent cell proliferation (thus keeping cell numbers under control) but can also promote apoptosis. The example of p53 is particularly important because it is mutated or there is a functional defect in the p53 pathway in approximately half of human cancers. The p53 network (partially shown in Figure 7.2) is quite well-known in mammals and involves genes that control, for example, apoptosis, the development of blood vessels, or cell differentiation. The core of this net is defined by the feedback loop existing between p53 and its negative regulator, the MDM2 oncoprotein. In invertebrates (such as *Drosophila*) homologues of p53 are known to be active throughout early development [7].

The fact that many mutations have little or no effect seems to be consistent with the presence of genes that either cannot propagate their failure or whose function can be replaced by other parts of the net. The presence of some genes that integrate multiple signals and can trigger widespread changes under their failure shows that the underlying network includes some highly-connected hubs. It seems to be a compromise between integration and homeostasis that should be observable when looking at the map of interactions within the cellular net.

Although a complete description of cellular networks would require the explicit consideration of dynamics, topological approaches—in which only the static architecture of the net is considered—are often successful in providing insight into biological complexity. This is the case, for example, of some models of ecological networks: in spite that populations fluctuate in time and changes in biomass or productivity take place at different scales, some of the fundamental regularities exhibited by food webs can be fairly well explained by means of static approaches [9]. Besides, the comparison of a wide range of complex networks (both natural

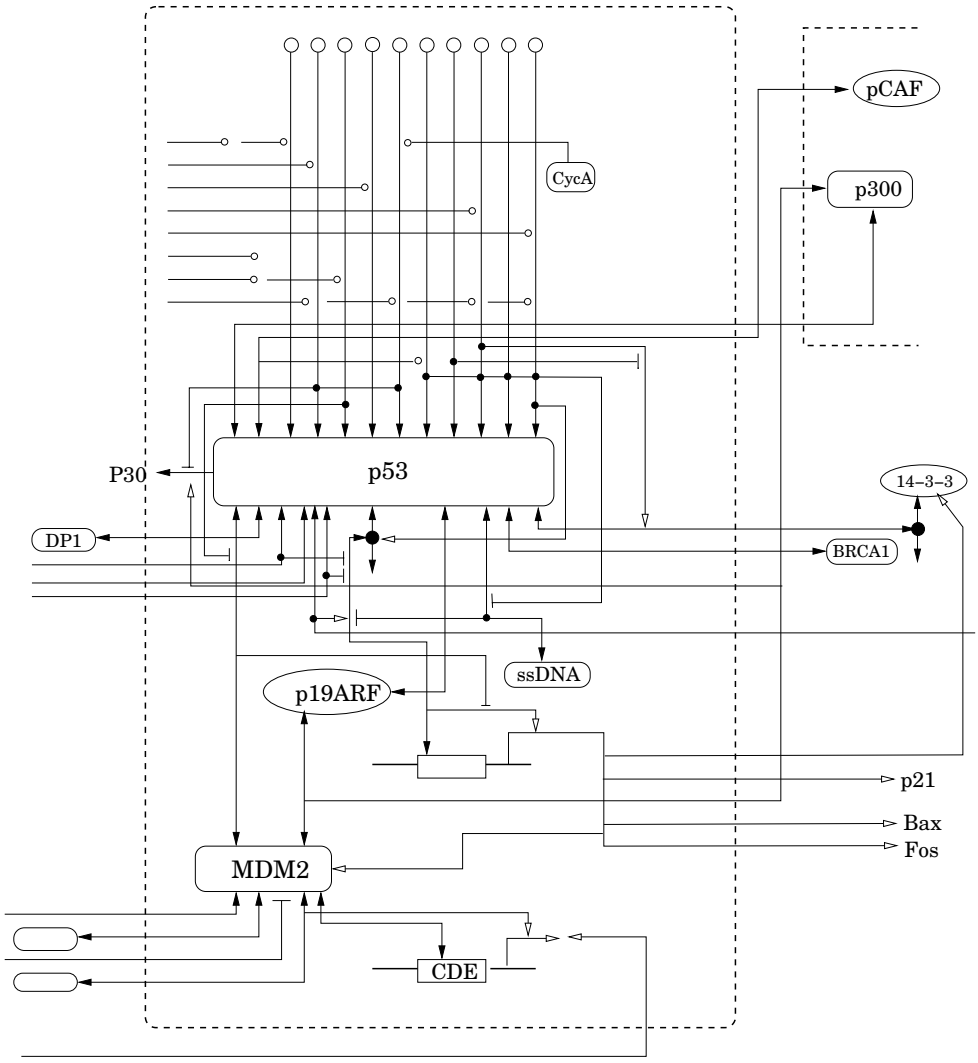


Figure 7.2: Schematic architecture of the p53 network. The p53 node integrates information from very different parts of the system. Only part of the cell circuitry is shown here. For a detailed presentation, see Ref. [8].

and artificial) reveals that strong regularities are shared by them, in spite that their underlying components, the nature of their interactions, and their time scales are very different. In this chapter the topological patterns displayed by these networks will be explored. As will be shown, the compromise between stability and integration can be made explicit by looking at the large-scale organization of cellular networks.

7.2 Cellular networks

The molecular basis of genetic control in cells, particularly in eukaryotic cells (i. e. cells with nucleus) is one of the most basic active areas of molecular cell biology. Of particular interest is the understanding of the regulation mechanisms involved in the development of multicellular organisms. In most well-known case studies, such as in the fruit fly *Drosophila melanogaster*, it has been shown that regulation among the genes that control early development (such as fushi tarazu, Figure 7.3) takes place at the transcription level [10]. The web of interactions can be very complex, and an example of a sub-web of the genetic net associated to *Drosophila* early development is shown in Figure 7.3(b). Mutations in genes associated to early stages of development have typically a strong effect and sometimes, as it occurs with the so-called homeotic genes [11], they result in important morphological changes.

Models of gene regulation have a long history in theoretical biology [12, 13]. The discovery of the mechanisms of transcription regulation in the Lac operon of *E. coli* was followed by the formulation of some simple mathematical models [14]. Inspired in early models of neural networks, a standard formulation of gene regulation can be introduced by means of a dynamical system:

$$\frac{dg_i(t)}{dt} = \Phi_\mu^i[\mathbf{g}] - \gamma_i g_i, \qquad i = 1, ..., n, \tag{7.1}$$

where a set of n different genes is defined. Here $\mathbf{g} = (g_1, ..., g_n)$ gives the activity state of each gene. Degradation is introduced by the last term $\gamma_i g_i$. The function $\Phi_\mu^i[\mathbf{g}]$ introduce the nature and extent of the interactions among components. An example of such type of model is:

$$\frac{dg_i(t)}{dt} = \Phi_\mu^i \left(\sum_{j=1}^{n} W_{ij} g_j(t) - \theta_i \right) - \gamma_i g_i(t), \tag{7.2}$$

where $\Phi_\mu^i(x)$ is a sigmoidal function of the local field $h_i = \sum_j W_{ij} g_j$, θ_i is a threshold, and the weights W_{ij} give the sign and strength of the gene-gene interactions. Usually the set $\mathbf{W} = \{W_{ij}\}$ is generated from a given distribution $\rho(W)$ that is assumed to be symmetric and with zero mean. This type of net can display a huge variety of dynamical patterns, including oscillations and chaos [15]. But the really interesting behavior (see below) comes from the statistical properties derived from the presence of phase transitions [16] when the connectivity is tuned.

Why to consider this type of mathematical approximations? Some attempts of building large-scale models of cellular nets based on near-realistic descriptions have failed to reproduce the whole spectrum of dynamical patterns displayed by even simple controlled systems. On the other hand, some key questions can find powerful answers in the generic properties exhibited by simple representations of real nets [16]. As an example, a striking feature of multicellular diversity is the surprisingly small repertoire of cell types, given the potentially astronomic diversity of cell states that would be obtained from the combinatorics of gene states [17]. Assuming that the number of genes in a multicellular organism is $N \approx 10^4$, 2^N different possible states are available. Yet, if cell types are considered as indicators of gene expression states, only $200 - 300$ states are actually realized.

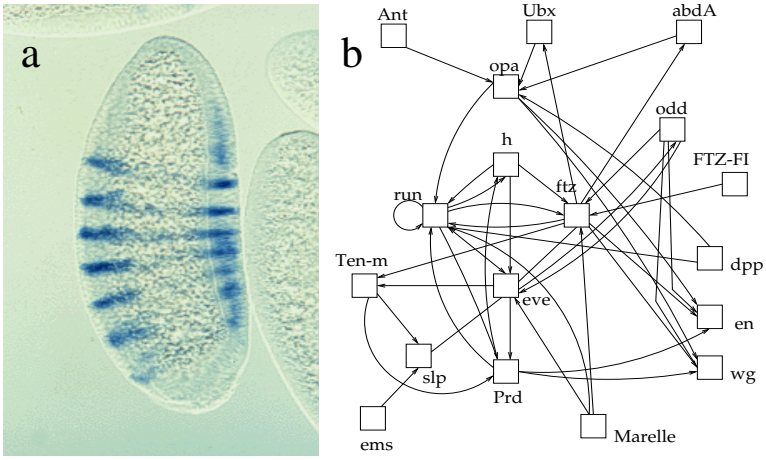


Figure 7.3: (a) Spatial pattern of activity of a given gene involved in *Drosophila* development (the so-called fushi tarazu gene (FTZ); see also Figure 7.2). The darker areas correspond to higher levels of activity of FTZ, indicating what cells are expressing it. Cell-to-cell interactions generate this set of stripes with a characteristic length. In (b) an example of a real gene network is shown. It includes some part (i.e. a directed subgraph) of the genetic net involved in the development of *Drosophila* fly. The names of the genes involved are indicated, such as FTZ=fushi tarazu. Only the connections are shown, not their sign.

In this section we will summarize some key features of this type of dynamical systems by considering the richness of their attractors when low-dimensional nets are used. Afterwards, the general scenario involving a large number of genes (i.e. large networks) will be considered.

7.2.1 Two-gene networks

The minimal number of genes needed in order to obtain a rich spectrum of behavioral patterns is given by two elements in interaction, although single-gene models with the appropriate nonlinearities can also display complex dynamic behavior [18]. Two-gene models allow to understand particularly important problems, such as the dynamics of virus-cell interactions in bacteria [19]. An example is the following two-gene system with no self-interaction, described by the equations:

$$\frac{dg_1}{dt} = \delta \frac{W_{21} g_2}{1 + W_{21} g_2} - g_1 \tag{7.3}$$

$$\frac{dg_2}{dt} = \delta \frac{W_{12} g_1}{1 + W_{12} g_1} - g_2. \tag{7.4}$$

The fixed points are easily found; together with the trivial fixed point, $P_0^* = (0, 0)$ we get a second nontrivial point $P_1^* = (g_1^*, g_2^*)$ given by:

$$g_1^* = \frac{\alpha}{W_{12} + \Omega} \quad g_2^* = \frac{\alpha}{W_{21} + \Omega}, \tag{7.5}$$

whose stability can be easily determined. Here $\Omega = \delta W_{21} W_{12}$ and $\alpha = \delta^2 W_{21} W_{12} - 1$. The eigenvalues associated to the Jacobi matrix for this system for $P_0^* = (0, 0)$ are

$$\lambda_{\pm} = -1 \pm \delta \sqrt{W_{12} W_{21}}, \tag{7.6}$$

and thus this point will be stable if $\delta \sqrt{W_{12} W_{21}} < 1$. There is an exchange of stability and P_1^* becomes stable when the previous condition does not hold (i. e. a transcritical bifurcation takes place) [22].

When self-interactions are also considered (i. e. $W_{ii} \neq 0$) several attractors can be present as a consequence of the competition between positive feedbacks and mutual inhibition. One particular case is given by networks such that the matrix of connections $\mathbf{W}$ is symmetric, of the form:

$$\mathbf{W} = \begin{pmatrix} \gamma & \beta \\ \beta & \gamma \end{pmatrix}, \tag{7.7}$$

with $\beta \in \mathbb{R}$ and $\gamma > 0$. In other words, when there is self-activation by both genes plus cross interactions which can be positive or negative. The later is a very common situation in real morphogenetic processes and is strongly related with the process of competition between species in ecosystems.

The stability analysis of this general problem can be performed by using the general Jacobi matrix:

$$\mathbf{L} = \begin{pmatrix} \alpha\delta/\epsilon_{12}^2 - 1 & \alpha\beta/\epsilon_{12}^2 \\ \alpha\beta/\epsilon_{21}^2 & \alpha\delta/\epsilon_{21}^2 - 1 \end{pmatrix} \tag{7.8}$$

where $\epsilon_{ij} \equiv 1 + \alpha g_i^* + \beta g_j^*$. For $\beta > 0$, the mutual reinforcement between both genes leads to the same state (indicated as *homogeneous* in Figure 7.4). Here $g_1^* = g_2^* = [\delta(\alpha + \beta) - 1]/(\alpha + \beta)$ and it is stable (this point disappears at $\beta = (1 - \delta\gamma)/\delta$). For $\beta < -1$, the self-interaction is unable to sustain gene activity and it decays to zero. Finally, an interesting domain is observed for $(1 - \delta\gamma)/\delta > \beta > 0$, where three attractors are present (the previous one, where both coexist, and two exclusion points). In Figure 7.4(b) we show an example of the flow field for the 3-attractor domain. We can see that there are three basins of attraction associated to each possible final state (fixed point).

These results, in particular the presence of multiple attractors for some parameter ranges, are specially important within the context of development [20,21]. In many cases the behavior of cells that become differentiated is very similar to that of a switch. By depending on initial conditions or external perturbations, which might emerge from some other genes in the networks, the system can reach one or another basin of attraction and thus a different final state. More importantly, it has been shown that some well-defined, small sets of interacting genes (so-called modules), are responsible for specific spatial patterns emerging in morphogenetic processes [20, 21]. As a consequence, not only single genes, but modules, can be the target of selection.

7.2.2 Random networks

Beyond the specific wiring diagrams that can be considered in small-sized genetic nets, the study of large-N nets has been dominated by randomly-wired systems [16]. Here genes are

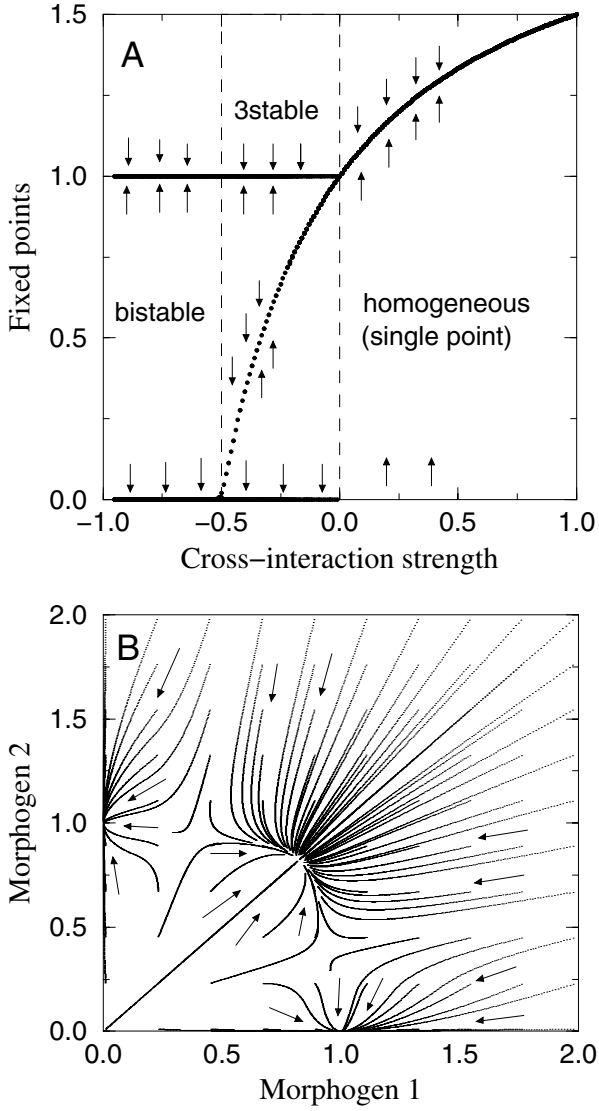


Figure 7.4: Multistability in gene network models: (a) bifurcation diagram for the two-gene network model with a symmetric matrix. Here $\delta = 2$ and $\gamma = 1$. Three basic domains are involved (see text); (b) flow diagram of the model for $\beta = -0.15$, in the three-attractor domain, indicated as 3stable in (a).

connected at random, with an average number of z connections per gene. An extensive literature on random Boolean networks has shown that a number of generic features are characteristic of these nets as a consequence of the presence of phase transition phenomena in random graphs [23].

In order to illustrate this idea, let us consider a graph $\Omega_{n,p}$ that consists of n nodes joined by links with some probability p. Specifically, each possible link between two given nodes occurs with a probability p. The average number of links (also called the average *degree*) of a given node will be $z = np$, and it can be easily shown that the probability $P(k)$ that a node has a degree k (it is connected to k other nodes) follows a Poisson distribution,

$$P(k) = e^{-z}\frac{z^k}{k!}. \tag{7.9}$$

This so-called Erdős-Rényi (ER) random graph [24] will be fairly well characterized by an average degree

$$\langle k \rangle = \sum_k kP(k) = z, \tag{7.10}$$

where $P(k)$ shows a peak. The distribution $P(k)$ is in this sense a single-scaled distribution [25] and an example is shown in Figure 7.5(a).

The ER model displays a phase transition at a given critical average degree $z_c = 1$ [23,26]. At this critical point, a *giant component* forms: for $z > z_c$ a large fraction of the nodes are connected in a single cluster, whereas for $z < z_c$ the system is fragmented into small subwebs. This type of random model has been used in different contexts, including ecological, genetic, metabolic, and neural networks [26]. The importance of this phase transition is obvious in terms of the collective properties that arise at the critical point: communication among the whole system becomes possible, and thus information can flow from the units to the whole system and back. Besides, the transition occurs suddenly and implies an innovation. No less important, it takes place at a low cost in terms of the number of required links ($\sim N$).

The ER model can be extended to directed graphs and has been analyzed by Kauffman within the context of genetic regulatory networks [16]. In the language presented in section 7.2, this will correspond to a network in which genes are randomly connected, and regulated by an average of z other genes. This means that the $N \times N$ matrix $\mathbf{W} = \{W_{ij}\}$ will have zN^2 nonzero elements, distributed at random. The probability that a gene is regulated by exactly k other genes will be then given by the distribution (7.9). Beyond the specific time-dependent features associated to the particular model chosen, one important characteristic of these systems is the presence of the percolation threshold: once a critical average connectivity $z_c = 1$ (the ratio of directed links to genes) is reached, the system becomes suddenly connected. Below the critical threshold the system is essentially disconnected and thus changes in a given gene cannot propagate to the rest of the system. The presence of the percolation threshold allows the system to exhibit a complex dynamical behavior, including deterministic chaos, Figure 7.5(b).

One consequence of these models (but strongly tied to the topological properties of sparse random graphs) is that a high diversity of attractors compatible with a high degree of homeostasis seems to naturally emerge close to the percolation threshold. However, early evidence indicated that the degree distributions that characterize *real* genetic nets are far from Poissonian. Actually, as we will see in section 7.4, the topology of real networks strongly departs from the Erdős-Rényi scenario.

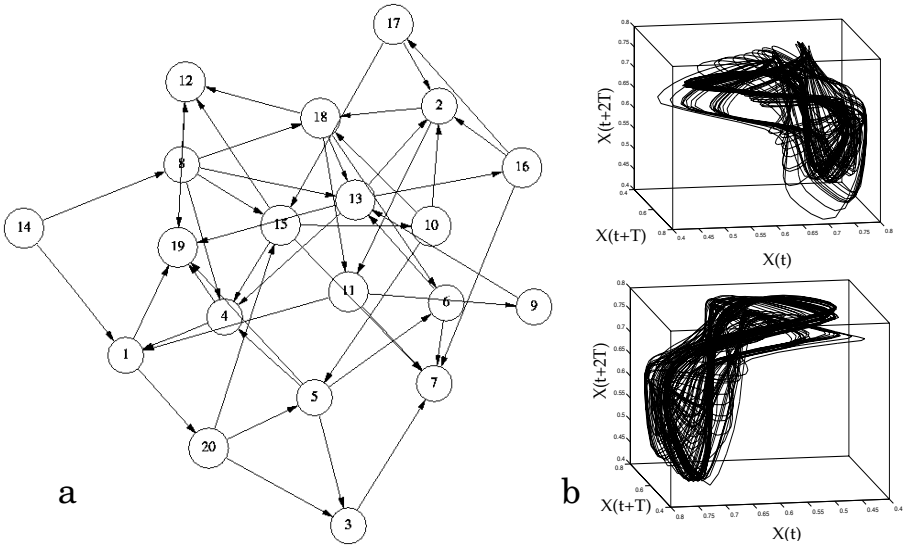


Figure 7.5: (a) An example of a directed random network with Poissonian structure. Here each node is a gene in a model gene network and arrows indicate the regulatory connections. This type of graph is characterized by an average degree z; together with the appropriate nonlinear coupling among genes, it can generate different types of dynamical patterns, including deterministic chaos. An example of the strange attractors obtained from these nets is shown in (b) in two different views.

7.3 Three interconnected levels of cellular nets

Gene regulation takes place at different levels and involves the participation of proteins. The whole cellular network includes three levels of integration:

- The genome, and the regulation pathways defined by interactions among genes;
- The proteome, defined by the set of proteins and their interactions; and
- The metabolic network, also under the control of proteins that operate as enzymes.

Unlike the relatively unchanging genome, the dynamic proteome changes through time in response to intra- and extracellular environmental signals. The proteome is particularly important. Proteins unify genome structure on the one hand and functional biology on the other: they are both the products of genes and regulate reactions or pathways.

Complicating the study of gene function is the fact that multiple proteins can arise from a single gene. In eukaryotic organisms, genes appear fragmented into pieces (exons) separated by non-coding domains (introns). After transcription, the resulting messenger RNA (mRNA) is generated by the excision and elimination of introns followed by the joining of exons. This process is called *splicing*. Once the mRNA is formed, it will be translated into a protein by the translation machinery.

A very important feature is that splicing can occur in different ways so that different sets of exons are joined together. In this way, different mRNA's (and thus different proteins) are

produced. The combinatorial potential of this so-called *alternative splicing* is obvious. In some cases, thousands of different proteins are potentially available for a given gene.

Alternative splicing expands genome complexity in an extraordinary fashion. In this context, although the genomes of complex organisms might not strongly differ in terms of their number of genes, the underlying proteome complexity can be very different. As will be discussed in the next sections, the actual structure of protein networks is shown to be strongly heterogeneous and shares several previously unsuspected traits with many different systems.

7.4 Small world graphs and scale-free nets

The analysis of the topological structure of protein interaction maps (in the budding yeast *Saccharomyces cerevisiae* and other simple organisms) revealed a surprising result: the protein-protein interaction net shares some universal features with the topological organization of other complex nets, both natural and artificial, ranging form technological networks [25, 27–29], neural networks [30], metabolic pathways [31–33], and food webs [34, 35] to the human language graph [36]. These studies actually offer the first global view of the proteome map and show that it strongly departs from the simple Erdős-Rényi scenario.

The first feature characteristic of the proteome map is that the probability $P(k)$ that a given protein interacts with other k proteins has a *scale-free* (SF) nature, i.e. it follows a power law, $P(k) \sim k^{-\gamma}$, with a sharp exponential cut-off for large k. Thus most proteins have a small number of links with other proteins and a few of them are highly connected (*hubs*). Those last ones are likely to be very important to cell function [6, 37, 38].

The second feature is the presence of the so-called *small world* (SW) property [30, 39, 40]. Small world graphs have a number of surprising features that make them specially relevant to understand how interactions among individuals, metabolites, or species lead to the robustness and homeostasis observed in nature. The SW pattern can be detected from the analysis of two basic statistical properties of the network[1]: (a) the *clustering coefficient* C and (b) the *average path length* $\bar{\ell}$.

The proteome graph (see Figure 7.6) is defined by a pair $\Omega_p = (W_p, E_p)$, where $W_p = \{p_i\}, (i = 1, ..., N)$ is the set of N proteins (nodes) and $E_p = \{\{p_i, p_j\}\}$ is the set of edges/connections between proteins. The *adjacency matrix* ξ_{ij} indicates that an interaction exists between proteins $p_i, p_j \in W_p$ ($\xi_{ij} = 1$) or that the interaction is absent ($\xi_{ij} = 0$). Two connected proteins are thus called *adjacent* and the *degree* k_i of a given protein is the number of edges that connect it with other proteins. Let us consider the adjacency matrix and indicate by $\Gamma_i = \{p_j \,|\, \xi_{ij} = 1\}$ the set of nearest neighbors of a protein $p_i \in W_p$. The clustering coefficient for this protein is defined as the ratio between the actual number of connections between the proteins $p_j \in \Gamma_i$, and the total possible number of connections, $k_i(k_i - 1)/2$ [30] (see Figure. 7.6). Denoting

$$\mathcal{L}_i = \sum_{j=1}^{N} \xi_{ij} \left[\sum_{k \in \Gamma_i} \xi_{jk} \right], \tag{7.11}$$

[1] Since the proteome map is a disconnected network, these quantities are actually defined on the *giant component*, defined as the largest cluster of connected nodes in the network [23].

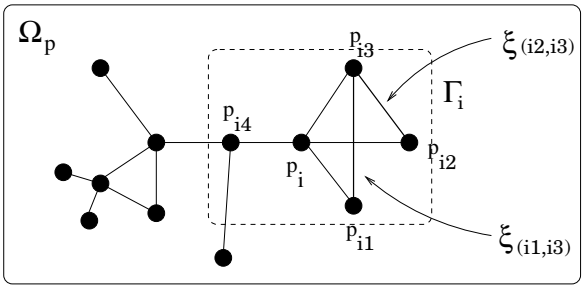


Figure 7.6: Measuring the clustering from a proteome graph Ω_p. Here each node (black circles) is a protein and physical interactions are indicated by means of edges connecting nodes.

we define the clustering coefficient of the i-th protein as

$$C(i) = \frac{2\mathcal{L}_i}{k_i(k_i-1)}, \tag{7.12}$$

where k_i is the degree of the i-th protein. The clustering coefficient is defined as the average of $C(i)$ over all the proteins,

$$C = \frac{1}{N}\sum_{i=1}^{N} C(i). \tag{7.13}$$

The average path length $\bar{\ell}$ is defined as follows: Given two proteins $p_i, p_j \in W_p$, let ℓ_{ij} be the length of the shortest path connecting these two proteins, following the links present in the network. The average path length $\bar{\ell}$ is defined as:

$$\bar{\ell} = \frac{2}{N(N-1)}\sum_{i<j}^{N} \ell_{ij}. \tag{7.14}$$

For the ER graph, we have a clustering coefficient inversely proportional to the network size, $C_{\rm ER} \approx z/N$; this is a very small quantity, that tends to zero for large networks. The average path length, on the other hand, is proportional to the logarithm of the network size $\bar{\ell}_{ER} \approx \log(N)/\log(z)$. At the other extreme, regular lattices with only nearest-neighbor connections among units exhibit a long average path length. Graphs with SW structure are characterized by a high clustering, $C \gg C_{\rm ER}$, while possessing an average path comparable to an ER graph with the same average connectivity and number of nodes, $\bar{\ell} \approx \bar{\ell}_{\rm ER}$.

The experimental observations on the proteome map can be summarized as follows:

1. The proteome map is a sparse graph, with a small average number of links per protein. In Ref. [41] an average connectivity $z \sim 1.9 - 2.3$ was reported for the proteome map of *S. cerevisiae*. This observation is also consistent with the study of the global organization of the *E. coli* gene network from available information on transcriptional regulation [42].

2. It exhibits a SW pattern, different from the properties displayed by purely random (ER) graphs. In particular, Ref. [41] reported the values $C = 2.2 \times 10^{-2}$ and $\bar{\ell} = 7.14$, to

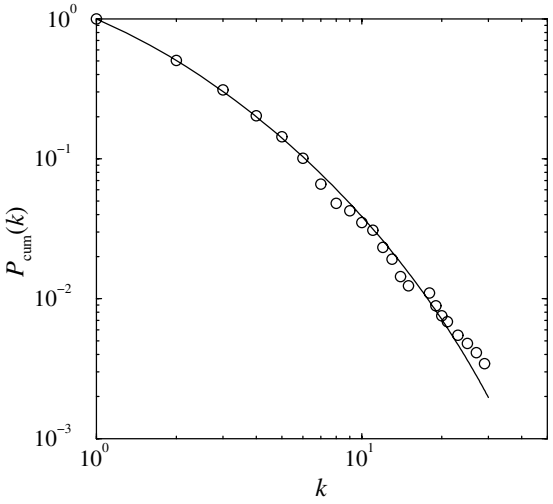


Figure 7.7: (b) Cumulated degree distribution for the yeast proteome map from Ref. [37]. The degree distribution has been fitted to the scaling behavior $P(k) \approx (k_0 + k)^{-\gamma} e^{-k/k_c}$, with an exponent $\gamma \simeq 2.6$ and a sharp cut-off $k_c \simeq 15$.

be compared with the values corresponding to an ER network with comparable size and average connectivity, $C_{\rm ER} = 1 \times 10^{-3}$ and $\bar{\ell}_{ER} = 8.0$.

3. The degree distribution of links follows a power-law with a well-defined cut-off. To be more precise, Jeong *et al.* [37] reported a functional form for the degree distribution of *S. cerevisiae*

$$P(k) \simeq (k_0 + k)^{-\gamma} e^{-k/k_c}. \tag{7.15}$$

Parameters reported in Ref. [37] are $k_0 \simeq 1$, $\gamma \approx 2.4$ and a cut-off $k_c \approx 20$. In Figure 7.7 we check this functional dependence on the cumulated degree distribution of the protein map[2] used in Ref. [37]. A fit to the form (7.15) yields the values $k_0 \simeq 1.1$, $k_c \simeq 15$, and $\gamma = 2.6 \pm 0.2$, compatible with the results found in [37, 41]. This particular form of the degree distribution could have adaptive significance as a source of robustness against mutations.

The highly heterogeneous character of these maps has important consequences within the context of molecular cell biology [6, 32]. It indicates that the evolution of proteome/genome complexity has been driven towards a well-defined topological pattern that provides the substrate for an extraordinary homeostatic stability against random mutational events.

[2] Data available at the web site `http://www.nd.edu/~networks/database/index.html`.

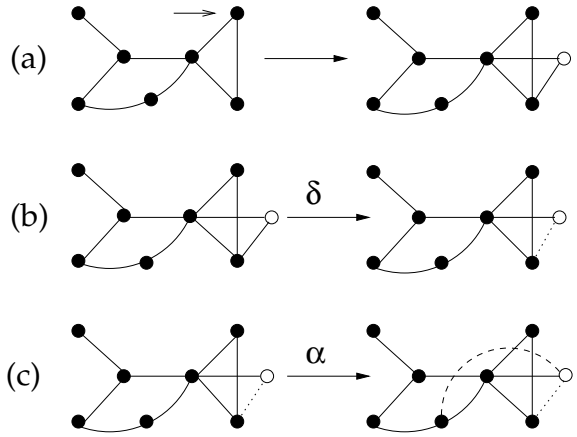


Figure 7.8: Growing network by duplication of nodes. First (a) duplication occurs after randomly selecting a node (arrow). The links from the newly created node (white) now can experience deletion (b) and new links can be created (c); these events occur with probabilities δ and α, respectively.

7.5 Scale-free proteomes: gene duplication models

Several models have been proposed in order to explain the regularities displayed by the proteome map [44–46]. These models of proteome evolution are based on a gene duplication plus rewiring process that includes the basic ingredients of proteome growth and intends to reproduce the previous set of observations. The first component of the models allows the system to grow by means of the copy process of previous units (together with their wiring). The second introduces novelty by means of changes in the wiring pattern, usually constrained to the newly created genes. This constraint is required if we assume that conservation of gene (protein) interactions is due to functional restrictions and that further changes in the regulation map are limited. Such constraint would be strongly relaxed when involving a newly created (and redundant) unit. The models proposed so far are intended to capture the *topological* properties of the proteome map. No explicit functionality is included in the description of the proteins and this is certainly a drawback. But by ignoring the specific features of the protein-protein interactions and the underlying regulation dynamics, one can explore the question of how much the network topology is due to the duplication and diversification processes.

In this chapter we will focus in particular in the model described in Refs. [45, 46]. This model considers single-gene duplications, which occur in most cases due to unequal crossover [47], plus re-wiring. Multiple duplications should be considered in future extensions of these models: molecular evidence shows that even whole-genome duplications have actually occurred in *S. cerevisiae* [48] (see also Ref. [49]). Re-wiring has also been used in dynamical models of the evolution of robustness in complex organisms [50].

The proteome graph at any given step t (i.e. after t duplications) will be indicated as $\Omega_p(t)$. The rules of the model, summarized in Figure 7.8, are implemented as follows. Each time step: (a) one node in the graph is randomly chosen and duplicated; (b) the links emerging from the

new generated node are removed with probability δ; (c) finally, new links (not previously present) can be created between the new node and all the rest of the nodes with probability α. Step (a) implements gene duplication, in which both the original and the replicated proteins retain the same structural properties and, consequently, the same set of interactions. The rewiring steps (b) and (c) implement the possible mutations of the replicated gene, which translate into the deletion and addition of interactions, with different probabilities.

7.5.1 Mean-field rate equation for the average connectivity

Since the model just presented has two free parameters, namely the deletion probability δ and the addition probability α, one preliminary task is to constrain their possible values by using the available empirical data. One average property that can be determined is the evolution of the average number of interactions per protein/gene through time, which can be compared with the evidence from real proteomes [37,41], as well as recent analysis of large-scale perturbation experiments [51].

Let us indicate by z_N and L_N the average connectivity of the system and its number links, respectively, when it is composed by N proteins. These magnitudes satisfy the relation $L_N = z_N N/2$. It is easy to check (see also Ref. [44]) that, at a mean-field level, that number of links L_N fulfill the following rate equation

$$L_{N+1} = L_N + z_N + \alpha(N - z_N) - \delta z_N, \tag{7.16}$$

where the last two terms correspond to the addition of links to a fraction α to the $N - z_N$ units not connected to the duplicated node, plus the deletion of any of the new z_N links, with probability δ. Using the continuous approximation

$$\frac{dz_N}{dN} \simeq z_{N+1} - z_N, \tag{7.17}$$

Eq. (7.16) can be written

$$\frac{dz_N}{dN} = \frac{1}{N}\left[z_N + 2\alpha(N - z_N) - 2\delta z_N\right], \tag{7.18}$$

whose solution is

$$z_N = \frac{\alpha}{\alpha + \delta} N + \left(z_1 - \frac{\alpha}{\alpha + \delta}\right) N^{\Gamma}, \tag{7.19}$$

where $\Gamma = 1 - 2(\alpha + \delta)$ and z_1 is the initial connectivity at $N = 1$. For any constant value of α and δ this model leads to an increasing connectivity through time. In order to have a finite z in the limit of large N, one possible solution is to impose an addition rate α that is a function of the size of the network, with the form

$$\alpha(N) = \frac{\beta}{N}, \tag{7.20}$$

where β is a constant. That is, the rate of addition of new links (the establishment of new viable interactions between proteins) is inversely proportional to the network size, and thus

much smaller than the deletion rate δ, in agreement with the rates observed in [41]. In this case, for large N, the differential rate equation (7.18) equation takes the form

$$\frac{dz_N}{dN} = \frac{1}{N}(1-2\delta)z_N + \frac{2\beta}{N}. \tag{7.21}$$

The solution of this equation is

$$z_N = \frac{2\beta}{2\delta - 1} + \left(z_1 - \frac{2\beta}{2\delta - 1}\right) N^{1-2\delta}. \tag{7.22}$$

For $\delta > 1/2$ a finite connectivity is reached in the limit of a large network,

$$z \equiv \lim_{N\to\infty} z_N = \frac{2\beta}{2\delta - 1}. \tag{7.23}$$

In order to reduce the number of independent parameters of the model, Ref. [45] used the available experimental data to estimate the average degree z and the ratio of addition and deletion rates in the yeast proteome, α/δ [41] to find a relation between β and δ, which, together with Eq. (7.23), yields a numerical estimate of β and δ. Since it is clear that this estimate is strongly dependent on the assumed value α/δ, Ref. [46] followed a more pragmatical approach, considering a δ-dependent model and fixing the actual value of δ by comparing numerical simulations with experimental data.

7.5.2 Rate equation for the node distribution n_k

The rate equation approach to evolving networks [52] can be fruitfully applied to the proteome model under consideration [46]. This approach focuses on the time evolution of the number $n_k(t)$ of nodes in the network with exactly k links at time t. Defining our network by means of the set of numbers $n_k(t)$, we have that the total number of nodes N is given by

$$N = \sum_k n_k, \tag{7.24}$$

while the total number of links is given by

$$L = \frac{1}{2}\sum_k k n_k. \tag{7.25}$$

Time is divided into periods. In each period, $t \to t+1$, one node is duplicated at random, so that $N \to N+1$. If, after each duplication, there is a probability δ to delete each link from the just-duplicated node, the probability of increasing the number of nodes at degree k, by direct duplication without link deletion, is given by

$$\Pr_{\text{self,dup}}\left[n_k \to n_k + 1\right] = \frac{n_k}{N}(1 - k\delta). \tag{7.26}$$

On the other hand, a node of degree k can be created from the duplication of a node of degree $k+1$ in which a link is deleted, contributing with a probability

$$\Pr_{\text{above,dup}}\left[n_k \to n_k + 1\right] = \frac{n_{k+1}}{N}(k+1)\delta. \tag{7.27}$$

The probability of degree change, from duplication of a node connected to a degree-k node, is given by:

$$\Pr_{\text{other,dup}}\left[(n_{k-1}, n_k) \to (n_{k-1} - 1, n_k + 1)\right] = \frac{n_{k-1}}{N}(k-1)(1-\delta). \tag{7.28}$$

Finally, in the same period, we proceed to add $N - k_d$ links with probability $\alpha = \beta/N$, where k_d is the connectivity of the just duplicated node. In the limit $N \gg k_d$, we can simply consider the addition of $N\alpha = \beta$ new links to the graph. When this last step is performed with the *correlated* prescription given for the model (i.e. adding links from the duplicated node to the rest of the nodes in the graph), it leads to a nonlocal rate equation for the functions n_k [46]. For the sake of simplicity, we will consider now the simpler case of a *uncorrelated* addition of links (new links created between any two nodes in the graph). However, it can be proved that both prescriptions lead qualitatively to similar results [46].

The case of uncorrelated addition of links can be represented as the distribution of $2\alpha N$ new link ends among the N nodes in the network. This event contributes with a probability

$$\Pr_{\text{add}}\left[(n_k, n_{k+1}) \to (n_k - 1, n_{k+1} + 1)\right] = \frac{n_k}{N} 2\alpha N = \frac{n_k}{N} 2\beta, \tag{7.29}$$

The probabilities (7.26), (7.27), (7.28), and (7.29) define the rate equation for the connectivity distribution

$$\begin{aligned}\frac{dn_k(t)}{dt} &= \frac{n_k}{N} + \frac{\delta}{N}\left[(k+1)n_{k+1} - kn_k\right] + \frac{1-\delta}{N}\left[(k-1)n_{k-1} - kn_k\right] \\ &+ \frac{2\beta}{N}\left[n_{k-1} - n_k\right].\end{aligned} \tag{7.30}$$

Since each time step a new node is added, Eq. (7.30) satisfies the condition

$$\frac{dN}{dt} = \sum_k \frac{dn_k(t)}{dt} = 1, \tag{7.31}$$

that yields the expected result $N(t) = N_0 + t$, where N_0 is the initial number of nodes in the network. In order to solve Eq. (7.30), we impose the homogeneous condition on the population number

$$n_k(t) = N(t)p_k \simeq tp_k, \tag{7.32}$$

where p_k is the probability of finding a node of connectivity k, which we assume to be independent of time. With this approximation, the rate equation reads

$$(k+1)\delta p_{k+1} - (k + 2\beta)p_k + [(k-1)(1-\delta) + 2\beta]p_{k-1} = 0. \tag{7.33}$$

Eq. (7.33) can be solved using the generating functional method [53]. Let us define the the generating functional

$$\phi(x) = \sum_k x^k p_k. \tag{7.34}$$

Introducing this definition into Eq. (7.33), we obtain an equation for $\phi(x)$, whose solution is

$$\phi(x) = \left(\frac{\delta - x(1-\delta)}{2\delta - 1} \right)^{-2\beta/(1-\delta)}. \tag{7.35}$$

Knowing $\phi(x)$ we can compute immediately the average connectivity

$$z = \sum_k k p_k \equiv x \frac{d\phi(x)}{dx}\bigg|_{x=1} = \frac{2\beta}{2\delta - 1}, \tag{7.36}$$

in agreement with the mean-field prediction of Eq. (7.23). On the other hand, performing a Taylor expansion of $\phi(x)$ around $x = 0$ we can obtain p_k as

$$p_k = \frac{1}{k!} \frac{d\phi(x)}{dx}\bigg|_{x=0}. \tag{7.37}$$

Applying this formula to the function (7.35), and using Stirling's approximation for large k, we can obtain the asymptotic behavior of p_k, given by:

$$p_k \sim (k_0 + k)^{-\gamma} e^{-k/k_c}, \tag{7.38}$$

with

$$\gamma = -k_0 = 1 - \frac{2\beta}{1-\delta}, \qquad k_c = \frac{1}{\ln\left(\frac{\delta}{\delta - 1}\right)}. \tag{7.39}$$

As we can observe from the previous result, we recover the same functional form experimentally observed in [37]. However, it is important to notice that for all the parameter range in which the exponential cut-off k_c is well-defined, we obtain a value of the degree exponent, as given by Eq. (7.39), that is $\gamma \leq 1$. The same result holds when considering the rate equation for the correlated model, in which the link addition is fully correlated with the new duplicated node [46]. This result is unsatisfactory, because it does not correspond with the results from numerical simulations of the model [46]. This discrepancy is explained by the fact that the $N \to \infty$ solution presented has only meaning for $\delta > 1/2$ (see Eq. (7.36)). Yet the master equation was defined on the basis of an independent-event approximation that only makes sense for $\delta \ll 1$. The master equation itself should become valid for $\delta \to 0$, but then the convergence results assumed at $N \to \infty$ seem questionable, as indicated by the fact that we get an analytic, but negative, z.

There is, however, something qualitative still to be learned from these equations, in the neighborhood of $\delta \sim 1/2$, small β. This is a neighborhood where the convergence results at large N still give sensible answers, even if they are not quantitatively correct due to marginal approximations in the underlying master equation. Yet at the same time, since this is the smallest value of δ where we can get answers, it is the one where the master equation we have constructed is likely to be the best approximation to the much more complicated true equation (one with frequent coupled events).

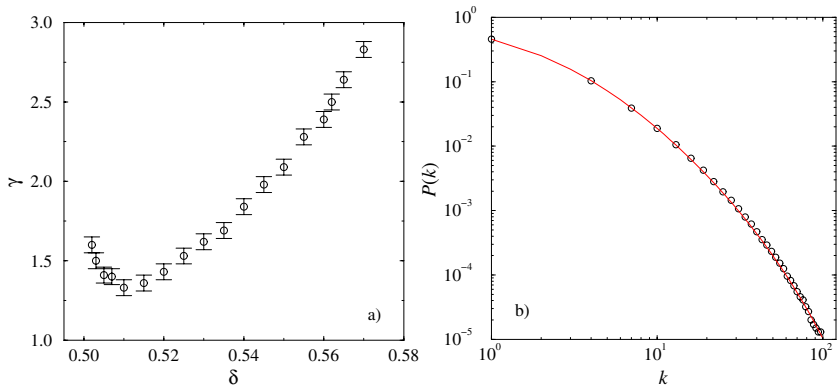


Figure 7.9: a) Degree exponent γ as a function of the deletion rate δ from computer simulations of the proteome model with average connectivity $z = 2.5$. Network size $N = 2 \times 10^3$. The degree distributions is averaged over 1000 different network realizations. b) Degree distribution for the same model, $\delta = 0.562$, averaged over 10000 different network realizations. The distribution can be fitted to the form $P(k) \approx (k_0 + k)^{-\gamma} e^{-k/k_c}$, with an exponent $\gamma = 2.5 \pm 0.1$ and a sharp cut-off $k_c \simeq 37$.

7.5.3 Numerical simulations

The proteome model defined in section 7.5 depends effectively on two independent parameters: the average connectivity of the network z and the deletion rate of newly created links δ, being the addition rate β computed from Eq. (7.23). The average connectivity can be estimated from the experimental results from real proteome maps. The data analyzed in Ref. [37] yields a value $z \simeq 2.40$. As a safe estimate, one can impose the value $z = 2.5$ [46], and consider values $\delta > 1/2$, in accordance with Eq. (7.23). In spite of the drawbacks of the analytical study of the model, section 7.5.2, one should expect the model to yield, for each value of δ, the functional form Eq. (7.38) for the degree distribution, with a degree exponent γ which is a function of δ (for a fixed average connectivity $z = 2.5$). From numerical simulations of the model one can then compute the function $\gamma(\delta)$ and select the value of δ that yields a degree exponent in agreement with the experimental observations. Fig 7.9(a) shows values of γ estimated from the functional form (7.38) for the degree distribution obtained from computer simulations of model, averaging over 1000 network of size $N = 2 \times 10^3$ nodes, of the same order of those found in the maps analyzed in Ref. [37]. Apart form a concave region for δ very close to $1/2$, γ is an increasing function of δ. The value of δ yielding the degree exponent closest to the experimentally observed one is then

$$\delta = 0.562. \tag{7.40}$$

In Figure 7.10(a) we show the topology of the giant component of a typical realization of the network model of size $N = 2 \times 10^3$. This Figure clearly resembles the giant component of a real yeast networks, as we can see comparing with Figure 7.10(b)[3]; we can appreciate the

[3] Figure kindly provided by W. Basalaj (see `http://www.cl.cam.uk/~wb204/GD99/#Mewes`).

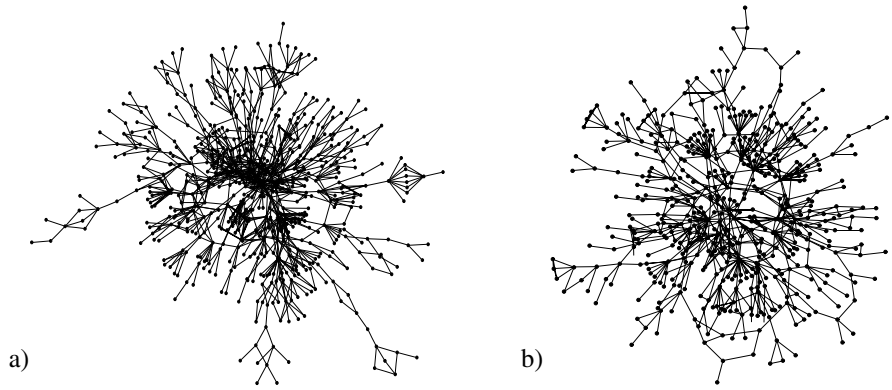


Figure 7.10: a) Topology of the giant component of the map obtained with the proteome model with parameters $\langle k \rangle = 2.5$ and $\delta = 0.565$. Network size $N = 2 \times 10^3$. b) Topology of a real yeast proteome map obtained from the MIPS database [43].

presence of a few highly connected hubs plus many nodes with a relatively small number of connections, in close resemblance of the real proteome map. On the other hand, Figure 7.9(b) shows the connectivity $P(k)$ obtained for networks of size $N = 2 \times 10^3$, averaged of 10000 realizations, for $\delta = 0.562$. In this Figure we observe that the resulting connectivity distribution can be fitted to a power-law with an exponential cut-off, of the form given by Eq. (7.38), with parameters $\gamma = 2.5 \pm 0.1$ and $k_c \simeq 37$, in fair agreement with the measurements reported by [41] and [37].

Finally, Table 7.1 reports the values of z, γ, C, and $\bar{\ell}$ obtained for the proteome model, compared with the values reported for the yeast *S. cerevisiae* by Ref. [41], those calculated for the map used in Ref. [37], and those corresponding to an ER random graph with size and average connectivity comparable with both the model and the real data. All the magnitudes displayed by the model compare quite well with the values measured for the yeast, and represent a further confirmation of the SW conjecture for the protein networks advanced by [41].

Table 7.1: Comparison between the observed regularities in the yeast proteome reported by Wagner [41], those calculated from the proteome map used by Jeong *et al.* [37], the model predictions with $N = 2000$, $\delta = 0.562$ and $z = 2.50$, and a ER network with the same size and connectivity as the model.

	Wagner's data	Jeong's data	Proteome model	ER model
z	1.83	2.40	2.4 ± 0.6	2.50 ± 0.05
γ	2.5	2.4	2.5 ± 0.1	—
C	2.2×10^{-2}	7.1×10^{-2}	1.0×10^{-2}	1×10^{-3}
$\bar{\ell}$	7.14	6.81	5.5 ± 0.7	8.0 ± 0.2

7.6 Discussion

Simple models of complex biological interactions have been used through the last decades as powerful metaphors of natural complexity. Networks pervade biology and there is little doubt that the untangling biological complexity demands a considerable degree of simplification. This view works well when generic mechanisms are at work: percolation close to criticiality in random graphs would be a perfect example in this context. Since information transfer (network communication) is a key property to all biosystems, reaching a threshold in connectivity allows information to propagate in a very effective way under a low wiring cost.

Similar principles might be operating in technology graphs [28, 54] and the striking similarities between man-made networks (such as electronic circuits or software graphs) and natural webs suggests that an organizing principle involving optimal communication might be at work in both types of systems. This seems a reasonable possibility, since the cost of wiring is an important constraint in both cases. For technology graphs, however, random failure typically leads to collapse and thus there is no robustness associated to the scale-free architecture. Biological systems might have taken advantage of the SF patterns that arise from optimization of path length under low cost [55] and make use of the source of robustness *for free* that might be generated.

As it occurs with many other aspects of biological complexity, historic constraints play an important role in shaping network topology. Not surprisingly, some of the oldest actors in the metabolic scene seem to be highly connected, thus suggesting a leading role of preferential attachment [26] at least at early stages of the evolution of metabolism. But the proteome map is a very large web incorporating a large amount of plasticity that might have been tuned through evolution in order to reach optimaly wired pathways. Future research will provide a new perspective on how biological nets get organized through evolution and what are the contributions of emergence, selection, and tinkering to network biocomplexity.

Acknowledgements

We thank the members of the Complex Systems Lab for useful discussions. This work has been partially supported by the European Network Contract No. ERBFMRXCT980183, the European Commission - Fet Open project COSIN IST-2001-33555, a grant PB97-0693 and by the Santa Fe Institute (R. V. S.). R.P.-S. acknowledges financial support from the Ministerio de Ciencia y Tecnología (Spain).

References

[1] D. Bray. Protein molecules as computational elements in living cells. *Nature* **376**, 307–312 (1995).

[2] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).

[3] P. Ross-Macdonald, P. S. R. Coelho, T. Roemer, S. Agarwal, A. Kumar, R. Jansen, K. H. Cheung, A. Sheehan, D. Symoniatis, L. Umansky, M. Heldtman, F. K. Nelson,

H. Iwasaki, K. Hager, M. Gerstein, P. Miller, G. S. Roeder, and M. Snyder. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413–418 (1999).

[4] A. Wagner. Robustness against mutations in genetic networks of yeast. *Nature Genet.* **24**, 355–361 (2000).

[5] P. H. Kussie, S. Gorina, V. Marechal, B. Elenbaas, J. Moreau, A. J. Levine and N. P. Pavletich. Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. *Science* **274**, 948–953 (1996) .

[6] B. Vogelstein, D. Lane, and A. J. Levine. Surfing the p53 network. *Nature* **408**, 307–310 (2000).

[7] S. Jin. Identification and characterization of a p53 homologue in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **97**, 7301–7306 (2000).

[8] K. W. Kohn, Molecular interaction map of the mammalian cell cylce control and DNA repair systems. *Mol. Biol. Cell* **10**, 2703-2734 (1999).

[9] R. J. Williams ans N. D. Martinez. Simple rules yield complex food webs. *Nature* **404**, 180–183 (2000).

[10] H. Lodish, A. Berk, S. L. Zipursky, and P. Matsudaira, *Molecular Cell Biology*, (W. H. Freeman, New York, 2000). 4th edition.

[11] W. J. Gehring, *Master Control Genes in Development and Evolution*, (Yale University Press, New Haven, 1998).

[12] J, Hasty, D. McMillen, F. Isaacs and J. J. Collins, Computational studies of gene regulatory networks: in numero molecular biology. *Nature Reviews Genet.* **2**, 268–279 (2001).

[13] P. Smolen, D. A. Baxter and J. H. Byrne, Mathematical modeling of gene networks. *Neuron* **26**, 567–580 (2000).

[14] B. Goodwin, *Temporal organization in cells*, (Academic Press, New York, 1963).

[15] J. E. Lewis and L. Glass, Steady states, limit cycles and chaos in models of complex biological networks. *Int. J. Bif. Chaos* **1**, 477–483 (1991).

[16] S. A. Kauffman, *Origins of Order*, (Oxford University Press, New York, 1993).

[17] S. B. Carroll. Chance and necessity: the evolution of morphological complexity and diversity. *Nature* **409**, 1102–1109 (2000).

[18] M. Laurent and N. Kellershohn, Multistability: a major means of differentiation and evolution in biological systems. *Trends Biochem. Sci.* **24**, 418–422 (1999).

[19] M. Ptashne, *A Genetic Switch* (Blackwell, Cambridge, 1992).

[20] I. Salazar, J. Garcia-Fernandez and R. V. Solé, Gene networks capable of pattern formation: from induction to reaction-diffusion. *J. Theor. Biol.* **205**, 587–603 (2000).

[21] R. V. Solé, I. Salazar and J. Garcia-Fernandez, Common Pattern Formation, Modularity and Phase Transitions in a Gene Network Model of Morphogenesis. *Physica A* **305**, 640-647 (2002).

[22] J. M. T. Thompson and H. B. Stewart, *Nonlinear dynamics and chaos*, (John Wiley & Sons, New York, 1986).

[23] B. Bollobás, *Random Graphs*, (Academic Press, London, 1985).

[24] P. Erdös and P. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–60 (1960).

[25] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA* **97**, 11149–11152 (2000).

[26] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).

[27] R. A. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).

[28] R. Ferrer i Cancho, C. Janssen, and R. V. Solé. The topology of technology graphs: small world pattern in electronic circuits. *Phys. Rev. E* **63**, 32767 (2001).

[29] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the internet. *Phys. Rev. Lett.* **87**, 258701 (2001).

[30] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).

[31] D. Fell and A. Wagner. The small world of metabolism. *Nature Biotech.* **18**, 1121 (2000).

[32] H. Jeong, B. Tombor, R. Albert, Z. N.Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2001).

[33] J. Podani, Z. Oltvai, H. Jeong, B. Tombor, A.-L. Barabási, and E. Szathmáry. Comparable system-level organization of Archaea and Eukaryotes. *Nature Genetics* **29**, 54–56 (2001).

[34] J. M. Montoya and R. V. Solé. Small world patters in fodd webs. *J. Theor. Biol.* **214**, 405–412 (2002).

[35] R. J. Williams, N. D. Martinez, E. L. Berlow, J. A. Dunne, and A.-L. Barabási. Two degrees of separation in complex food webs, (2001). Santa Fe working paper 01-07-036.

[36] R. Ferrer i Cancho, C. Janssen, and R. V. Solé. The small world of human language. *Procs. Roy. Soc. London B* **268**, 2261–2266 (2001).

[37] H. Jeong, S. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature* **411**, 41 (2001).

[38] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science* **296**, 910-913 (2002).

[39] D. J. Watts, *Small Worlds*, (Princeton University Press, Princeton, 1999).

[40] M. E. J. Newman. Models of the Small World. *J. Stat. Phys.* **101**, 819–841 (2000).

[41] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* **18**, 1283–1292 (2001).

[42] D. Thieffry, A. M. Huerta, E. Pérez-Rueda, and J. Collado-Vives. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *escherichia coli*. *BioEssays* **20**, 433–440 (1998).

[43] H. W. Mewes, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, and D. Frishman. Mips: a database for genomes and protein sequences. *Nucleic Acids Res.*, **27**,44–48 (1999).

[44] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani. Modelling of protein interaction networks, (2001). cond-mat/0108043.

[45] R. V. Solé, R. Pastor-Satorras, E. Smith, and T. Kepler. A model of large-scale proteome evolution. *Adv. Complex. Syst.* **5**, 43–54 (2002).

[46] R. Pastor-Satorras, E. Smith, and R. V. Solé. Evolving protein interaction networks through gene duplication, (2002). Santa Fe working paper 02-02-008.

[47] S. Ohno, *Evolution by gene duplication*, (Springer, Berlin, 1970).

[48] K. H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).

[49] A. Wagner. Evolution of gene networks by gene duplications: A mathematical model and its implications on genome organization. *Proc. Natl. Acad. Sci. USA* **91**, 4387–4391 (1994).

[50] S. Bornholdt and K. Sneppen. Robustness as an evolutionary principle. *Proc. Roy. Soc. Lond. B* **267**, 2281–2286 (2000).

[51] A. Wagner. Estimating coarse gene network structure from large-scale gene perturbation data, (2001). Santa Fe working paper 01-09-051.

[52] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Phys. Rev. Lett.* **85**, 4629 (2000).

[53] C. W. Gardiner, *Handbook of stochastic methods*, (Springer, Berlin, 1985). 2nd edition.

[54] S. Valverde, R. Ferrer Cancho and R. V. Solé. Scale-free networks from optimal design. Santa Fe working paper 02-04-019.

[55] R. Ferrer Cancho and R. V. Solé. Optmization in Complex Networks. Santa Fe working paper 01-11-068.

8 Correlation profiles and motifs in complex networks

Sergei Maslov, Kim Sneppen, and Uri Alon

8.1 Introduction

Networks have recently emerged as a unifying theme in complex systems research [1]. It is in fact no coincidence that networks and complexity are so heavily intertwined. Any future definition of a complex system should reflect the fact that such systems consist of many mutually interacting components. These components are far from being identical as say electrons in systems studied by condensed matter physics. In a truly complex system each of them has a unique identity allowing one to separate it from the others. The very first question one may ask about such a system is which other components a given component interacts with? This information systemwide can be visualized as a graph, whose nodes correspond to individual components of the complex system in question and edges to their mutual interactions. Such a network can be thought of as a backbone of the complex system. Of course, system's dynamics depends not only on the topology of an underlying network but also on the exact form of interaction of components with each other, which can be very different in various complex systems. However, the underlying network may contain clues about the basic design principles and/or evolutionary history of the complex system in question. The goal of this article is to provide readers with a set of useful tools that would help to decide which features of a complex network are there by pure chance alone, and which of them were possibly designed or evolved to their present state.

Living organisms provide us with a paradigm for a complex system. Therefore, it should not be surprising that in biology networks appear on many different levels. All biochemical processes taking place in a single cell constitute its metabolic network, where nodes are individual metabolites, and edges are metabolic reactions converting them to each other. Virtually every one of these reactions is catalyzed by an enzyme and the specificity of its catalytic function is ensured by the key and lock principle of its physical interaction with the substrate. Often the functional enzyme is formed by several mutually interacting proteins. Thus the structure of the metabolic network is shaped by the network of physical interactions of cell's proteins with their substrates and each other. Another way in which the network of physical interactions contributes to the complex dynamics of a living cell is through regulation of activity of individual proteins e.g. by phosphorylation or allosteric regulation. This constitutes a major mechanism for propagation of various biochemical signals in the cell. Hence a more complete version of the physical interaction network in addition to substrates and proteins should also include all of their functionally modified forms. The production and degradation of each of the proteins in the physical interaction network in turn is controlled by the regulatory network of the cell. In Fig. 8.1 we show a part of such network in the bacterium

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

Escherichia coli (*E. coli*) corresponding to positive or negative transcriptional regulation of its proteins by transcription factors. More generally regulatory network in the cell in addition to transcriptional regulation includes translational regulation, RNA editing, specific targeting of individual proteins for degradation, etc. On yet higher level individual cells of a multicellular organism exchange signals with each other. This gives rise to several new networks such as e.g. nervous, hormonal, and immune systems of an animal. The inter-cellular signaling network stages the development of a multicellular organism of a given species from the fertilized egg. Finally, on even larger scale interactions between individual species form the food web of an ecosystem.

By no means complex networks are unique to living organisms: in fact they lie at the foundation of an increasing number of artificial systems. The most prominent example of this is the Internet and the World Wide Web (WWW) being the "hardware" and the "software" of the network of communications between computers. While the Internet is formed by "physical" connections between constituent computers, or on a more coarse-grained scale, between so-called Autonomous Systems (AS), which are large domains of computers managed by the same organization such as e.g a university, or a business enterprise. The World Wide Web, on the other hand, is a much larger network whose nodes are individual webpages, and directed edges are links between them.

Networks are also ubiquitous in systems studied by social sciences. To name just a few, scientists are connected by a network of collaborations defined as co-authorship of scientific articles, while articles themselves are linked through a directed network of citations. Examples of networks in economics include that of customers and their choices of products, or economies of individual countries connected by the volume of direct foreign investments. This last example illustrates one important notion about complex networks: in some cases it appears that every component of a complex system is connected to every other component, which makes the concept of a network useless. Indeed, in this case the network is just a fully connected graph, which contains no information about the complexity of the underlying system. However, a meaningful network can be constructed in this case if one chooses to include only interactions stronger than a certain threshold, which has to be selected to maximize the information content of the resulting graph. For interacting economies that corresponds to including edges only for the strongest coupled pairs of countries, while for physical interaction networks in biological systems – for pairs of molecules with the binding constant above a certain threshold.

The above mentioned complex networks in biological, technological, and social systems for the most part lack the top-down design. Instead they grow and evolve as a result of the bottom-up stochastic dynamics of their individual nodes. It makes an Erdős-Rényi (ER) random network [3] the first null model to which topological properties of these networks can be compared. An interesting unifying feature of many complex networks that clearly distinguishes them from ER random networks is an extremely broad distribution of connectivities (defined as the number of immediate neighbors) of individual nodes [4]. While the majority of nodes in such a network are each connected to just a handful of neighbors, there exist some nodes, which will be referred to as "hubs", that have a disproportionately large number of interaction partners. The connectivity of the highest connected hub is usually several orders of magnitude larger than the average connectivity of the network. This property stands in sharp contrast with ER networks, in which connectivities of individual nodes are Poisson-

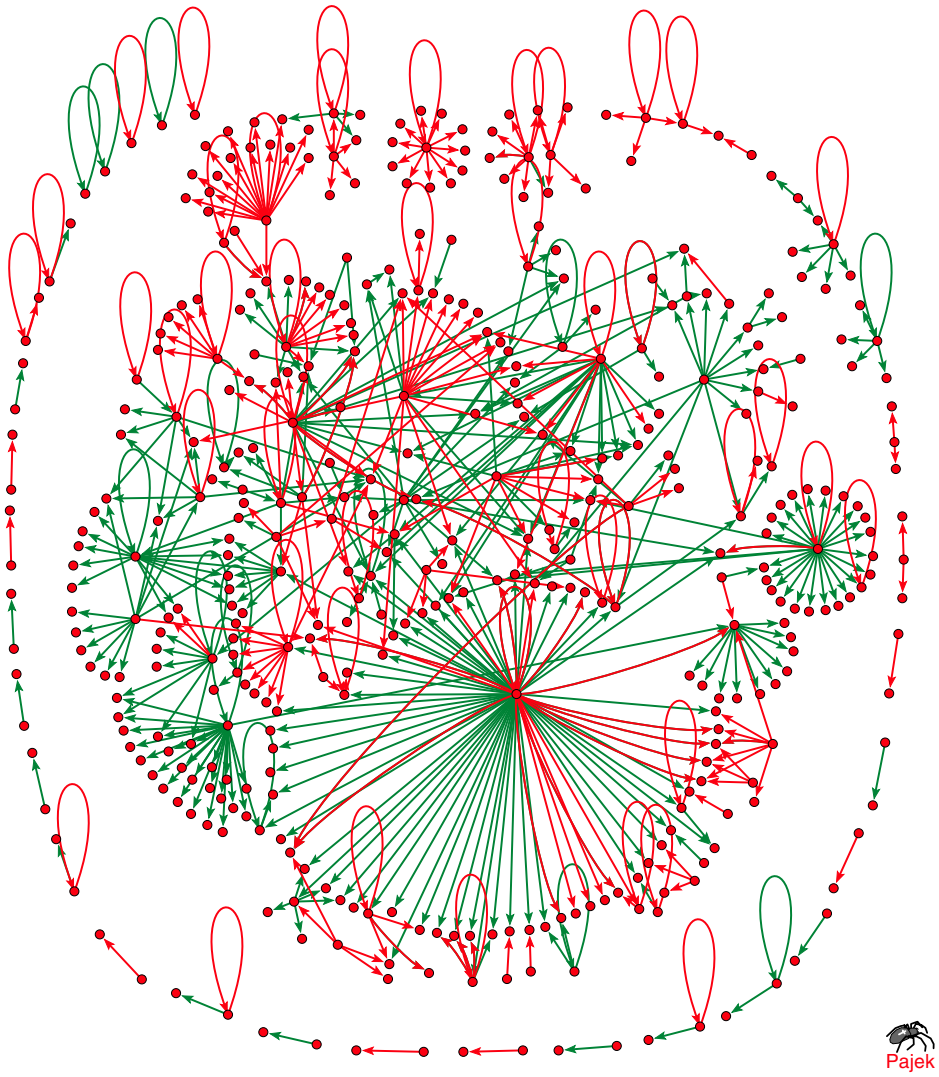


Figure 8.1: The transcription regulatory network of *E. coli*. Nodes in this network represent operons (groups of genes transcribed onto a single mRNA) and arrows (edges) – direct transcriptional regulation of a protein encoded in the downstream operon by a regulatory protein encoded in the upstream operon. Red and green arrows refer to respectively negative and positive regulations in a living *E. coli* cell. In the present text we discuss how one can extract characteristic topological properties of networks such as the one shown here. The network was displayed using the Pajek software [2].

distributed and thus the number of nodes with a connectivity significantly above average is negligibly small. Often the connectivity distribution in complex networks can be approximated by a scale-free power law form [4]. Prominent examples of this are the Internet [5] and the WWW [6], where in the last case the power law extends for up to four orders of magnitude. Among biological networks histograms of node connectivities in metabolic [7] and protein interaction [8] networks can be reasonably approximated by scale-free distributions extending for about two orders of magnitude.

The set of connectivities of individual nodes is an example of a low-level topological property of a network. While it answers the question about how many neighbors a given node has, it gives no information about the identity of those neighbors. It is clear that most of non-trivial properties of networks are defined at a higher level in the exact pattern of connections of nodes to each other. However, such multi-node connectivity patterns are rather difficult to quantify and measure. By just looking at many complex networks one gets the impression that their components are linked to each other in a completely haphazard way. One may wonder which connectivity patterns are indeed random, while which arose due to evolution or fundamental design principles and limitations? Such non-random features can be used to identify a given complex network and better understand the underlying complex system.

In this work we describe a universal recipe for how such information can be extracted. To this end we first construct a proper randomized version (null model) of a given network. As was pointed out by Newman and collaborators [9], broad distributions of connectivities observed in most real complex networks indicate that the connectivity is an important individual characteristic of their nodes and as such it should be preserved in any meaningful random counterpart. In addition to connectivities of its nodes one may choose to preserve some other low-level topological properties of the network. Any higher level topological property, such as e.g. the total number of edges connecting pairs of nodes with given connectivities, the number of loops of a certain type, the number and sizes of components, the diameter of the network, can then be measured in the real complex network and separately in an ensemble of its randomized versions. Dealing with an ensemble allows one to put error bars on any quantity measured in the randomized network. One then concentrates only on those topological properties of the complex network that significantly deviate from the null model, and, therefore, are likely to reflect its design principles and/or evolutionary history.

The idea of comparing topological properties of a network to its randomized counterpart is not new. For example, in [10] the level of clustering in some real world networks was compared to its value in ER random networks with the same number of edges and nodes. In the field of sociology there exists a rich history of testing hypotheses about social networks by comparison to randomized null-model networks [11]. A more recent twist on this idea was put in Ref. [9]. Authors of this work derived a number of useful analytical results for random networks with an *arbitrary distribution of connectivities* and compared certain topological properties of a number of real-life complex networks to these analytical expressions. In Ref. [12] it was demonstrated that when constructing a random network with a broad distribution of connectivities it is important to take into account the constraint of having no multiple edges between the same pair of nodes. This constraint applicable to most real-life complex networks modifies topological properties of their random counterparts, especially around their highly connected (hub) nodes. One may also select to conserve some other low-level topological properties in addition to connectivities of individual nodes [12]. In the absence of

analytical results in this case one has to resort to numerical simulation of such randomized networks Basic algorithms generating an ensemble of such random networks were applied to studies of complex networks in [12–14]. Earlier on these algorithms were actively studied in mathematical literature [15]. In these works a number of important results concerning their ergodicity were rigorously proven.

The plan of this review is as follows: In the next section we introduce the local rewiring algorithm for generation of an ensemble of randomized networks [12, 13, 15] and compare it with global rewiring algorithms studied in [9, 16]. We also propose several modifications of this algorithm, which in addition to node connectivities conserve some other low-level topological properties of the complex network in question [12]. In the section 3 we use these random ensembles to measure correlation profiles of several complex networks, namely those of physical interactions and transcriptional regulation between proteins in yeast *Saccharomyces cerevisiae* [13], and that of the Internet defined on the level of Autonomous Systems (AS) [12]. In the section 4 the comparison to a randomized network reveals the set of ubiquitous network motifs in the genetic regulatory network of *Escherichia coli* bacterium [14]. The potential meaning of these empirically detected elements of design is discussed in the last section. The set of MATLAB numerical algorithms used to generate some of the results described in this work can be found at [17].

8.2 Randomization algorithm: Constructing the proper null model

One may generate a random version of a given network using various algorithms. They differ from each other by which low-level topological features of the original network are preserved in its randomized counterpart. Below are the first three representatives of such randomization algorithms listed in the order of increasing number of constraints:

1. Randomly rewire all edges in the network. This algorithm only conserves the *average* connectivity of all nodes in the network.

2. Randomly rewire edges in the network while preserving the number of edges emanating from each individual node (node's connectivity). This algorithm conserves all "single-node" topological properties of a network, while completely randomizes multi-node connection patterns. In a directed network one may rewire edges in such a way that both the number of outgoing and incoming edges are separately conserved for each node.

3. When nodes in a network are divided into several mutually exclusive subgroups one may rewire its edges in such a way that the total number of neighbors from each of these subgroups is separately conserved for every node. This algorithm may prove useful if some subgroups are known to preferentially connect to some other subgroups and one wants to preserve this preferential linking in a randomized network.

The first rewiring scheme from the list above irrespective of the original form of this distribution generates an Erdős-Rényi (ER) random network characterized by a narrow Poisson distribution $p(k) = \langle k \rangle^k \exp(-\langle k \rangle)/k!$ of node connectivities k. As both percolation properties and the abundance of most topological patterns in a network are very sensitive to the

exact form of the distribution of connectivities [9, 18], they would be dramatically modified as a result of the randomization algorithm 1. A much more informative comparison is to a randomized network generated by algorithms 2-3, where connectivities of individual nodes (and hence their distribution) are strictly conserved.

The rewiring algorithm giving rise to such random network was proposed in [13, 15]. In its most general formulation (for a directed network whose nodes are divided into several subgroups) it consists of multiple repetition of the following switch move (rewiring step) (see illustration in Fig. 8.2):

Randomly select a pair of directed edges A→B and C→D where A and C belong to the same subgroup (marked blue in Fig. 8.2), and B and D belong to the same subgroup (marked red in Fig. 8.2). The two edges are then rewired in such a way that A becomes connected to D, while C to B, provided that none of these edges already exist in the network, in which case the rewiring step is aborted and a new pair of edges is selected.

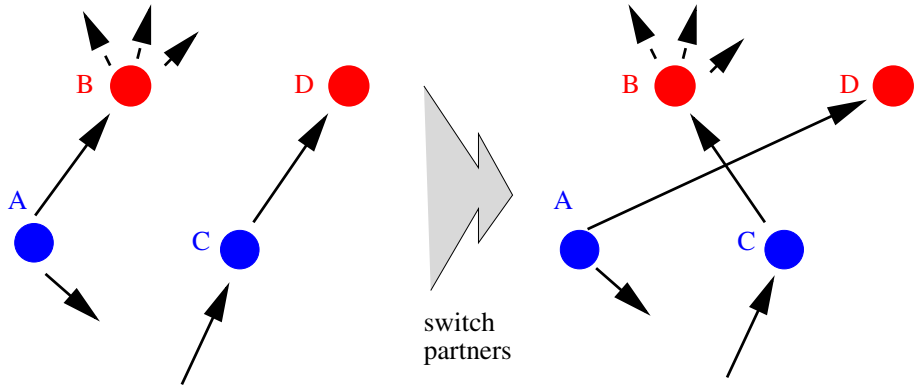


Figure 8.2: One step of the random local rewiring algorithm. A pair of directed edges A→B and C→D such that nodes A and C belong to the same subgroup (marked black), and B and D both belong to another subgroup (marked grey) are selected. The two edges are then rewired in such a way that A becomes connected to D, while C to B, provided that none of these edges already exist in the network, in which case the rewiring step is aborted and a new pair of edges is selected. An independent random network is obtained when the network is rewired by the above local switch a large number of times, say several times in excess of the total number of edges in the system. The above rewiring algorithm conserves both the in- and out- connectivity of each individual node as well as the exact distribution of its interaction partners among subgroups of nodes.

The last restriction prevents the appearance of multiple edges connecting the same pair of nodes. A repeated application of the above rewiring step leads to a randomized version of the original network. The set of MATLAB programs generating such a randomized version of any complex network can be downloaded from [17].

A number of nice analytical results for random networks with an arbitrary (in general non-Poisson) probability distribution of connectivities were recently reported in [9, 18]. Also, in Refs. [9, 16] a "stub reconnecting" numerical algorithm allowing one to construct such networks was proposed. The basic idea of this algorithm was to first generate the set of connectivities k_i for every node in the system, and create k_i "edge stubs" sticking out of every node, which are not yet connected to other nodes. A random network is then generated by randomly picking two such edge stubs and joining them together to form an edge between the two nodes they emanate from. As this process goes on the number of disconnected stubs diminishes until finally all stubs are used up.

The stub reconnecting algorithm explicitly allows for multiple edges to form between the same pair of nodes. On the other hand, the construction principles of complex networks discussed in this paper prohibit the appearance of such multiple edges, and hence they should not be allowed in their random counterparts as well. If one explicitly forbids the formation of multiple edges during the stub reconnecting algorithm [9, 16], for sufficiently broad distribution of node connectivities the algorithm would normally get stuck in a "frozen" configuration in which all nodes with remaining unconnected stubs are already connected to each other. The probability to reach such a frozen configuration increases with both the size and the fraction of highly connected nodes in the network. In this case it becomes computationally impossible to avoid double edges by aborting and restarting the algorithm every time a frozen configuration is reached. We would also like to point out that the set of analytical results obtained in [9, 18] apply to an ensemble of random networks generated by the stub reconnecting algorithm in which multiple edges are allowed. Therefore, they have to be modified for an ensemble of random networks in which such edges are forbidden.

The above mentioned limitations of the stub reconnecting algorithm forced us to use the *local* rewiring algorithm [12, 13, 15] described above. Instead of completely deconstructing a given complex network and then creating the corresponding random network *de novo* this algorithm modifies it through multiple simple local rearrangements and hence avoids frozen configurations. Later on in this section we will show how the basic rules of our local rewiring algorithm can be recursively modified to analyze patterns present at higher levels of network architecture, while maintaining other established low-level topological properties of the complex network in addition to connectivities of its nodes.

The difference in frequencies of appearance of any particular topological pattern j in a given complex network and its properly randomized counterpart can be quantified by the following set of *correlation profiles*. In the first profile one computes the ratio

$$R(j) \; = \; \frac{N(j)}{\overline{N_r(j)}} \tag{8.1}$$

where $N(j)$ is the number of times the pattern j is seen in the real network, and $\overline{N_r(j)}$ is the average number of occurrences of the pattern in an ensemble of random networks generated by the appropriate null model. Patterns selected by design or evolution of the complex network in question would manifest themselves by $R(j) > 1$, while suppressed patterns correspond to $R(j) < 1$. While $R(j)$ determines the magnitude of the suppression/enhancement it tells

nothing about the statistical significance of the effect. This latter quantity is given by the Z-score of the deviation defined as

$$Z(K_0, K_1) \;=\; \frac{N(j) \;-\; \overline{N_r(j)}}{\sigma_r(j)} \tag{8.2}$$

where $\sigma_r(j)$ is the standard deviation of $N_r(j)$ measured in a sufficiently large ensemble of randomized networks.

Alternatively the statistical significance of the difference between real and randomized networks can be quantified in terms of its P-value. The P-value is defined as the probability that the number of patterns $N_r(j)$ in a randomized network is larger or equal (or smaller or equal in case when $N(j) < \overline{N_r(j)}$) than $N(j)$. For patterns that are highly statistically significant it is often impossible to directly evaluate the P-value in a reasonable number of realizations of random networks. In this case one reports an upper bound on such a P-value given by the inverse size of the ensemble studied numerically. If one can verify that N_r is a Gaussian-distributed random variable the Z-score can be easily converted to the P-value.

Examples of patterns discussed in this work include:

- Correlations between connectivities of neighboring nodes quantified as $N(K_0, K_1)$ – the number of edges connecting nodes of connectivity K_0 to those of connectivity K_1 [12, 13];
- The number of small network motifs [14] such as e.g. the triangular loop (Fig. 8.3A), or the double triangle (Fig. 8.3B).

In case of more complex topological patterns like the double triangle in Fig. 8.3B the calculation of R and Z becomes somewhat more involved. Indeed since this pattern contains two simple triangles and a square among its sub-patterns, its statistical over- or under- representation in the real network may be caused simply by over- or under- representation of these more elementary sub-patterns. One strategy is to deal with this problem recursively and analytically [14]. In this case one has to start computing Z and R from the simplest "irreducible" patterns and gradually work it up toward more complicated composite patterns, each time renormalizing out trivial "reducible" correlations. Another strategy numerically determines the statistical significance of a given high-level topological motif [12]. To this end one first generates an ensemble of random networks that have both the same set of connectivities and the same number of low-level motifs (such as e.g. triangle loops) as the original network. This can be done in several different ways:

1. One may consider only those local rewiring steps that strictly preserve both node connectivities and the number of sub-motifs. One example of such specific move is shown in Fig. 8.4, which conserves both node connectivities and the total number of simple triangular loops.
2. One may employ the switch move in Fig. 8.2, without constraint, but then limit the ensemble generated by the simple rewiring algorithm to include only networks which by accident have the observed number of sub-motifs. In most cases this algorithm is

prohibitively numerically expensive, in particular when the sub-motifs in the real network are hugely overrepresented on underrepresented relative to a typical random network.

3. A way to remedy the numerical inefficiency of the previous algorithm is to use biased sampling, which favors the correct number of sub-motifs. This can be done by supplementing the simple switch move with a Metropolis acceptance/rejection criterion based on an artificial energy function H_j that favors the same number of sub-motifs j that was observed in the real network [12]:

$$H_j = \frac{(N_r(j) - N(j))^2}{N(j)} \quad . \tag{8.3}$$

That is to say for a given temperature T, the algorithm accepts any local rewiring step that lowers the energy H_j or leaves it unchanged, while those steps that lead to a ΔH increase in H_j are accepted with a probability $\exp(-\Delta H/T)$. The temperature should be selected low enough to favor the desired number of low-level sub-motifs j yet high enough to ensure an ergodic sampling of the phase space. Obviously the $T = 0$ version of this algorithm is equivalent to the algorithm #1.

The Metropolis algorithm #3 can be easily extended to take care of several independent sub-motifs by using the composite energy function $H = \sum_j H_j$. Such sampling of networks has the advantages of being simple to implement and generalizes easily to several sub-patterns. When counting the statistics one can limit the ensemble to include only the random networks that have exactly the same number of sub-motifs as the original complex network.

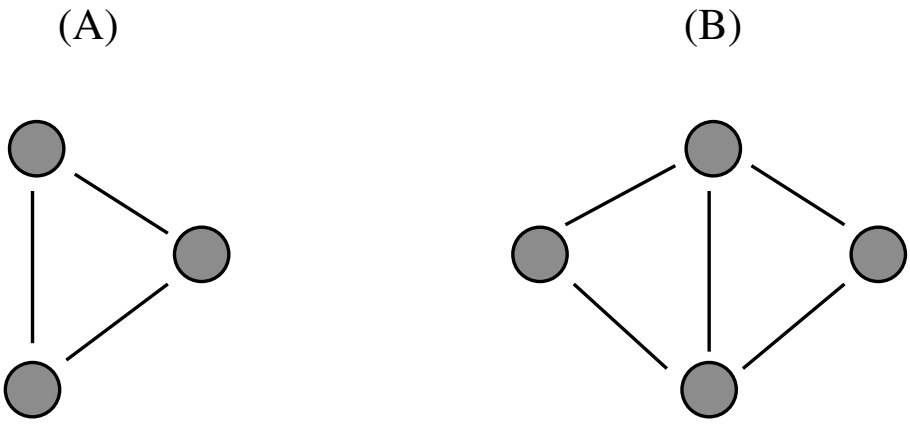


Figure 8.3: Example of motifs: (A) a simple triangle whose abundance quantifies the level of clustering in the network. (B) a somewhat more complex pattern. It contains two triangles and a square among its sub-patterns.

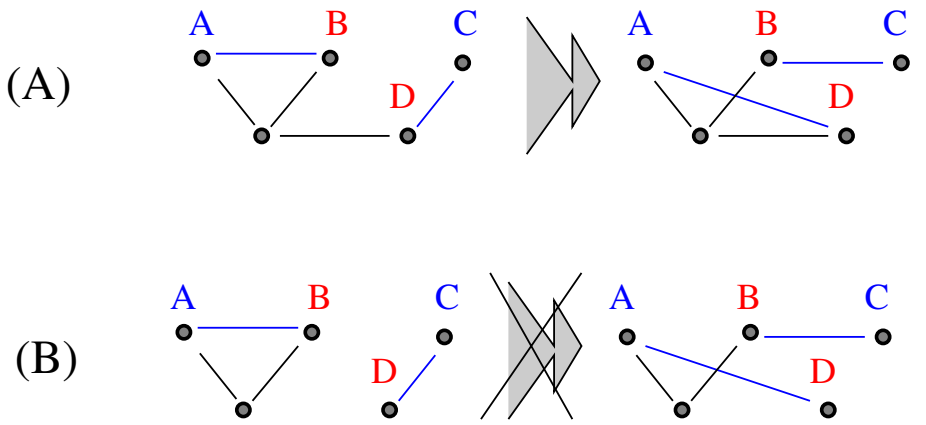


Figure 8.4: The local rewiring step shown in the panel A is allowed since it preserves the total number of triangles in the network, while that in the panel B is forbidden since it decreases the number of triangles by one. In the Metropolis algorithm both moves would be allowed albeit with different probabilities: move in panel B would be accepted with a lower probability due to the increase in the energy function (8.3).

8.3 Correlation profiles: Yeast molecular networks and the Internet

Methods described in the previous section allow us to define and measure the *correlation profile* of a complex network. The correlation profile quantifies correlations between connectivities of neighboring nodes in the network. We have applied these numerical tools to two levels of molecular networks operating in yeast *Saccharomyces cerevisiae*, which at present is perhaps the best characterized biological model organism:

1. The *protein interaction network* used in this work consists of 4475 physical interactions between 3279 yeast proteins as measured in the most comprehensive high-throughput yeast two-hybrid screen [19]. To answer the question if proteins A and B interact with each other the two-hybrid experimental technique uses a pair of artificially prepared hybrid proteins A^b and B^p, which are referred to as the bait and the prey hybrid correspondingly. In order to better visualize the protein interaction network in Fig. 8.5 we plotted a small part of it using the software package Pajek developed by Vladimir Batagelj and Andrej Mrvar [2]. The subset used in this figure consists of all proteins known to be localized in the yeast nucleus [20] and to interact with at least one other nuclear protein in the full set of Ref. [19].

2. The most general definition of the *regulatory network* operating in a living cell includes all cases when production or degradation of one of its proteins is *directly* controlled by another. Edges of this network correspond to transcription and translational regulation, RNA editing, specific targeting of individual proteins for degradation, etc. The YPD

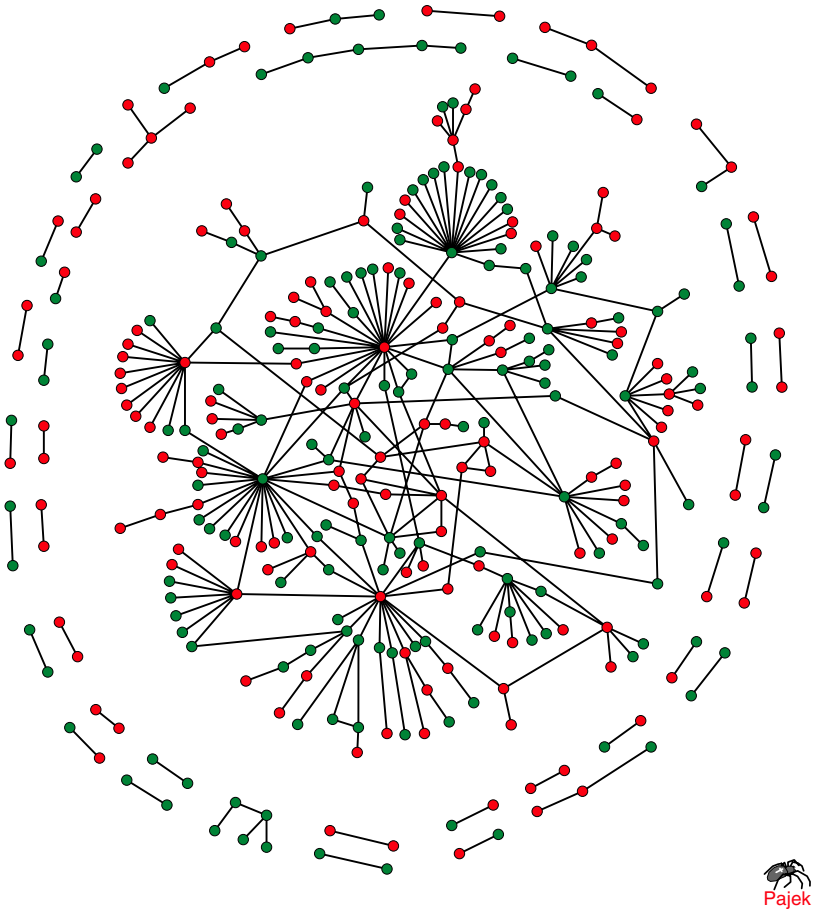


Figure 8.5: Network of physical interactions between nuclear proteins in yeast. Here we show the subset of the yeast protein interaction network reported in the full set of Ref. [19]. The subset consists of 318 interactions among 329 proteins, which are known to be localized in the yeast nucleus [20], and to interact with at least one other nuclear protein [19]. Note that most neighbors of highly connected nodes have rather low connectivity. This feature will be later quantified in the correlation profile of this network (Figs 8.7, 8.9). Nodes are color coded according to how essential they are for the survival of yeast cells under laboratory conditions [20]. Green nodes correspond to viable and red ones to non-viable null-mutants lacking the corresponding protein.

database [20] contains 1750 such regulations among 848 yeast proteins. To narrow down the range of possible regulatory mechanisms and make the network more homogeneous we have constructed correlation profiles of the *transcription regulatory network*, which is the subset of the general regulatory network formed by all positive and negative direct transcription regulations. This network shown in Fig. 8.6 consists of 1289 (1047 positive

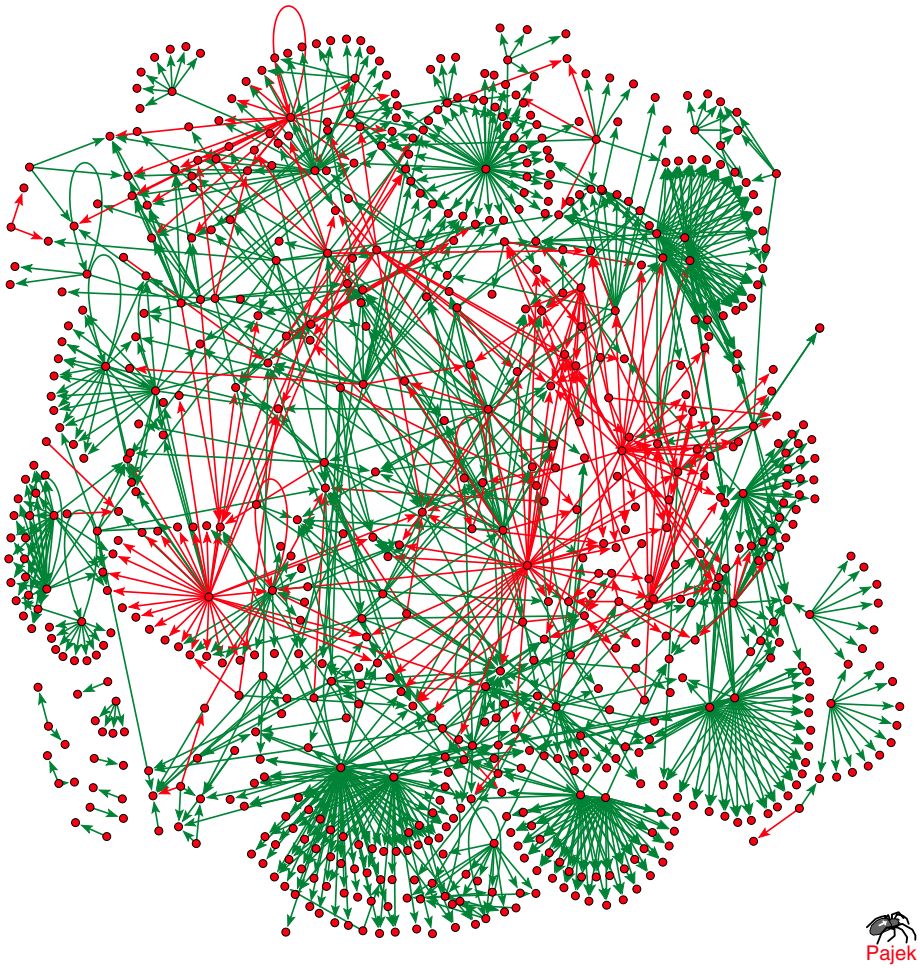


Figure 8.6: Transcription regulatory network in yeast. Apart from an overall apparent lack of modularity, one notices several striking features related to hub proteins that each regulate many other proteins: 1) they tend to avoid to regulate each other, 2) each hubs is either a predominantly positive regulator or a predominantly negative regulator, and 3) it is much more frequent for a protein to regulate many other proteins, than to be regulated by many. It is the first of these features, the separation of hubs from each other, that is quantified with the correlation profile of this network (Figs 8.8, 8.10.)

and 242 negative) regulations by 125 transcription factors [20] within the set of 682 proteins.

While the regulatory network is naturally directed, the network of physical interactions among proteins in principle lacks directionality. However, for poorly understood reasons all high-throughput two-hybrid experimental data [19, 22] defining pairs of physically interacting

proteins have a significant asymmetry between baits and preys, with bait hybrids being more likely to be highly connected than their prey counterparts. This can be seen e.g. in the fact that the average connectivity of baits with at least one interaction partner is close to 3, whereas the same quantity measured for preys is only 1.8. Since each reported interaction involves exactly one bait and one prey protein, this asymmetry needs to be taken into account when selecting a proper "null" model for the interaction network. For this purpose in our randomization procedure we would treat the two-hybrid data as a directed network with an arrow on each edge pointing away from the bait hybrid towards the prey hybrid.

Randomized versions of these two molecular networks were constructed by randomly rewiring their directed edges, while preventing "unphysical" multiple connections between a given pair of nodes as described in the previous section. By construction this algorithm separately conserves the in- and out-connectivities of each node. Therefore, in a randomized version of the regulatory network each protein has the same numbers of regulators and regulated proteins as in the original network. Similarly, in a random counterpart of the interaction network numbers of interaction partners of the bait-hybrid and the prey-hybrid of every protein are individually conserved. The set of MATLAB programs for both the randomization and the correlation profile detection and visualization in any complex network are available at [17].

The topological property of the network giving rise to its correlation profile is the number edges $N(K_0, K_1)$ connecting pairs of nodes with connectivities K_0 and K_1. To find out if in a given complex network connectivities of interacting nodes are correlated, $N(K_0, K_1)$ should be compared to its value $N_r(K_0, K_1) \pm \Delta N_r(K_0, K_1)$ in a randomized network, generated by the edge rewiring algorithm. When normalized by the total number of edges E, $N(K_0, K_1)$ defines the joint probability distribution $P(K_0, K_1) = N(K_0, K_1)/E$ of connectivities of interacting nodes. Any correlations would manifest themselves as systematic deviations of the ratio

$$R(K_0, K_1) = P(K_0, K_1)/P_r(K_0, K_1) \tag{8.4}$$

from 1. Statistical significance of such deviations is quantified by their Z-score

$$Z(K_0, K_1) = (P(K_0, K_1) - P_r(K_0, K_1))/\sigma_r(K_0, K_1) \ , \tag{8.5}$$

where $\sigma_r(K_0, K_1) = \Delta N_r(K_0, K_1)/N$ is the standard deviation of $P_r(K_0, K_1)$ in an ensemble of randomized network.

Figs. 8.7 and 8.8 show the ratio $R(K_0, K_1)$ as measured in yeast interaction and transcription regulatory networks, respectively. In the interaction network K_0 and K_1 stand for the total number of neighbors of two interacting proteins, while in the regulatory network K_0 is the out-connectivity of the regulatory protein and K_1 – the in-connectivity of its regulated partner. Thus by the very construction $P(K_0, K_1)$ is symmetric for the physical interaction network but not for the regulatory network. Fig. 8.9 and Fig. 8.10 plot the statistical significance $Z(K_0, K_1)$ of deviations from 1 visible in Fig. 8.7 and Fig 8.8 correspondingly. To arrive to these Z-scores 100 randomized networks were sampled with connectivities logarithmically binned in two bins per decade. The combination of R- and Z-profiles reveals the regions on the $K_0 - K_1$ plane, where connections between proteins in the real network are significantly enhanced or suppressed, compared to the null model. In particular, the blue/green

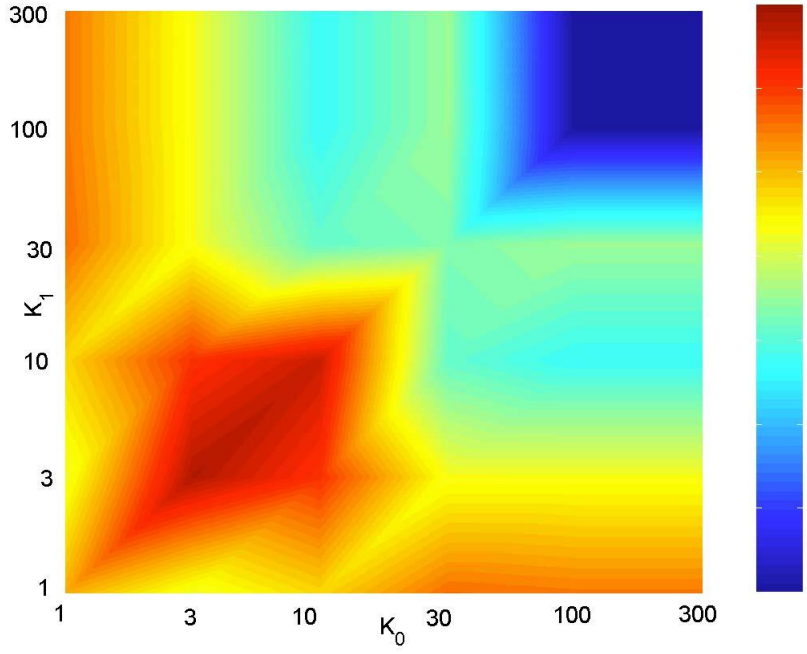


Figure 8.7: Correlation profile of the protein interaction network in yeast. The ratio $R(K_0, K_1) = P(K_0, K_1)/P_r(K_0, K_1)$, where $P(K_0, K_1)$ is the probability that a pair of proteins with total numbers of interaction partners given by K_0, K_1 correspondingly, directly interact with each other in the full set of Ref. [19], while $P_r(K_0, K_1)$ is the same probability in a randomized version of the same network, generated by the random rewiring algorithm described in the text. Note the logarithmic scale of both axes.

region in the upper right corner of Figs. 8.7-8.10 reflects the reduced likelihood that two hubs are directly linked to each other, while red regions in the upper left and the lower right corners of these figures reflect the tendency of hubs to associate with nodes of low connectivity. One should also note a prominent feature on the diagonal of the Fig. 8.7 and 8.9 corresponding to an enhanced affinity of proteins with between 4 and 9 physical interaction partners towards each other. This feature can be tentatively attributed to members of multi-protein complexes interacting with other proteins from the same complex. The above range of connectivities thus correspond to a typical number of neighbors of a protein in a multi-protein complex. When we studied pairs of interacting proteins in this range of connectivities we found 39 of such pairs to belong to the same complex in the recent high-throughput study of yeast protein complexes [21]. This is about 4 times more than one would expect to find by pure chance alone.

When analyzing molecular networks one should consider possible sources of errors in the underlying data. Two-hybrid experiments give rise to false positives of two kinds. In one case the interaction between proteins is real but it never happens in the course of the normal life

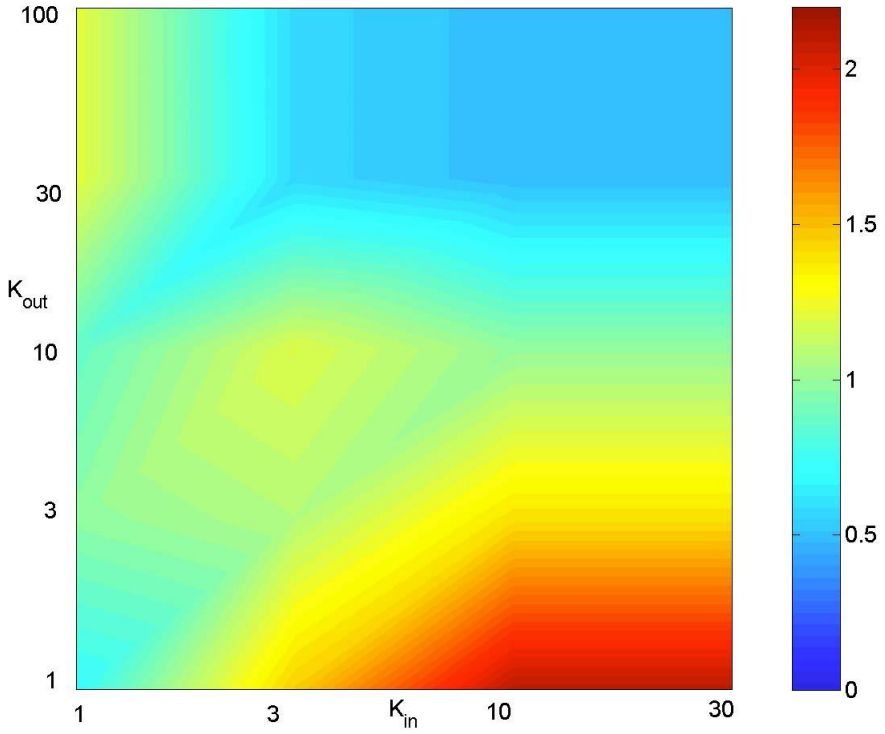


Figure 8.8: Correlation profile of the transcription regulatory network in yeast. The ratio $R(K_{out}, K_{in}) = P(K_{out}, K_{in})/P_r(K_{out}, K_{in})$, where $P(K_{out}, K_{in})$ is the probability that a protein node with the out-connectivity K_{out} transcriptionally regulates the protein node with the in-connectivity K_{in} in the network from the YPD database [20], while $P_r(K_{out}, K_{in})$ is the same probability in a randomized version of the same network, generated by the random rewiring algorithm described in the text. Note the logarithmic scale of both axes.

cycle of the cell due to spatial or temporal separation of participating proteins. In another case an indirect physical interaction is mediated by one or more unknown proteins localized in the yeast nucleus. Reversely, in a high throughput two-hybrid screens one should expect a sizable number of false negatives. Primarily a binding may not be observed if the conformation of the bait or prey heterodimer blocks relevant interaction sites or if the corresponding heterodimer altogether fails to fold properly. In addition to this 391 proteins out of the potential 5671 baits in [19] were not tested as possible bait hybrids because they were found to activate transcription of the reporter gene in the absence of any prey proteins.

Fortunately, the qualitative features of the correlation profile are very robust with respect to an unbiased set of false positives and false negatives. Indeed, as previously undetected edges are added to the network (or falsely detected edges are removed from it) the average connec-

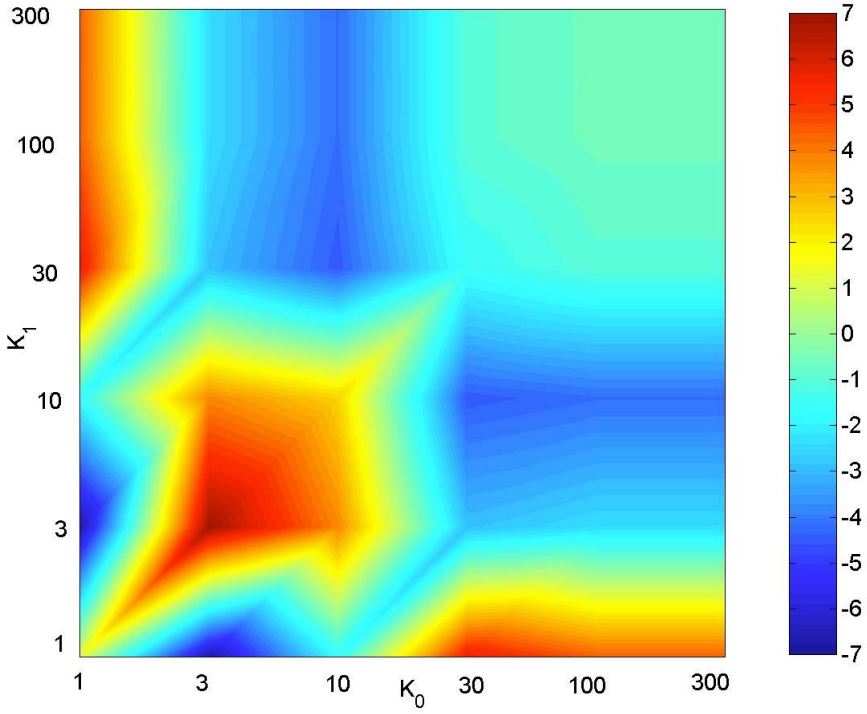


Figure 8.9: Statistical significance of correlations present in the protein interaction network in yeast. The Z-score of correlations $Z(K_0, K_1) = (P(K_0, K_1) - P_r(K_0, K_1))/\sigma_r(K_0, K_1)$, where $P(K_0, K_1)$ is the probability that a pair of proteins with total numbers of interaction partners given by K_0, K_1 correspondingly, directly interact with each other in the full set of Ref. [19], while $P_r(K_0, K_1)$ is the same probability in a randomized version of the same network, generated by the random rewiring algorithm described in the text, and $\sigma_r(K_0, K_1)$ is the standard deviation of $P_r(K_0, K_1)$ measured in 1000 realizations of a randomized network. Note the logarithmic scale of both axes.

tivity of its nodes changes. As a result correlation features visible in its correlation profiles may shift their positions and intensity, but are likely to preserve their qualitative characteristics up to a very high level of false positives or false negatives.

The data for the protein interaction network used in this work come from a high-throughput experiment performed in one lab using a unique experimental technique [19]. This fact makes it a perfect candidate for correlation profiling. Indeed, since almost all pairs of yeast proteins were tested as potential interacting partners, the statistical information contained in the resulting network contains no anthropomorphic bias. On the other hand, when the information about edges in a network is obtained from a database, combining results of many experimental groups using various techniques, one should worry about a hidden anthropomorphic factor: some proteins just constitute more attractive subjects of research and are, therefore, relatively

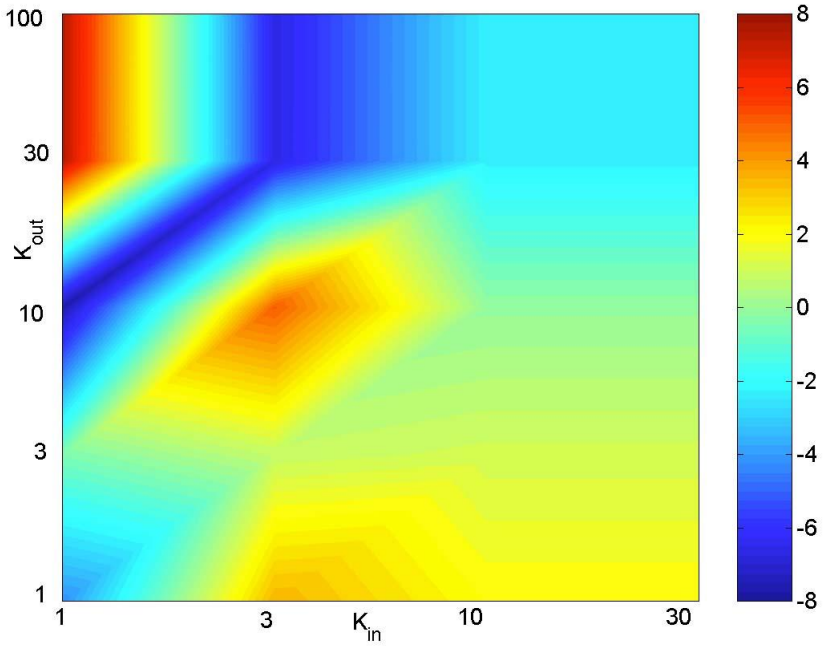


Figure 8.10: Statistical significance of correlations present in the transcription regulatory network in yeast. The ratio $Z(K_{out}, K_{in}) = (P(K_{out}, K_{in}) - P_r(K_{out}, K_{in}))/\sigma_r(K_{out}, K_{in}))$, where $P(K_{out}, K_{in})$ is the probability that a protein node with the out-connectivity K_{out} transcriptionally regulates the protein node with the in-connectivity K_{in} in the network from the YPD database [20], while $P_r(K_{out}, K_{in})$ is the same probability in a randomized version of the same network, generated by the random rewiring algorithm described in the text, and $\sigma_r(K_{out}, K_{in})$ is the standard deviation of $P_r(K_{out}, K_{in})$ measured in 1000 realizations of a randomized network. Note the logarithmic scale of both axes.

better studied than the others. The level of clustering in networks based on the database data may be overestimated due to several reasons: 1) With the exception of systemwide experiments such as high-throughput two-hybrid screens in yeast [19,22], experimentalists are more likely to check for interactions between pairs of proteins within the same functional group. 2) A complete analysis of all possible pairwise interactions within a small group of proteins would influence the level of clustering in the network. In this case this group would manifest itself by a relatively dense pattern of interactions with other members of the same group compared to interactions outside of the group.

A good example to illustrate the danger of indiscriminately using the database data is given by the network of physical interactions among yeast protein listed in [23]. This database contains contributions from several high throughput two-hybrid experiments [19, 22] as well as protein interactions determined by other methods such as co-immunoprecipitation technique,

and mass spectroscopy of protein complexes. Statistical properties of this network were analyzed in a recent preprint [24], where it was shown that it has a remarkably high clustering coefficient. The clustering coefficient of a network is given by the number of triangles normalized by the total number of places in the network, where a triangle can be formed. When we constructed a correlation profile of this network we found that it contains a pronounced red linear region all along the diagonal of the K_0-K_1 plane. Such a region corresponds to an increased affinity of proteins of a given connectivity to others with approximately the same value of connectivity. However, a closer analysis has revealed that this feature is an artifact of the way that interactions among complex-forming proteins were reported in this particular database. Apparently, an interaction was reported between *any two proteins* which were found to belong to the same multi-protein complex. Hence, all members of a given multi-protein complex consisting of N_c proteins have connectivity close to N_c. Needless to say this artificial feature would lead to a gross over counting of the clustering coefficient reported in Refs. [24]. In reality protein in a large multi-protein complex directly interact with no more than a few members of the same complex, which gives rise to a red spot on the diagonal of $R(K_0, K_1)$ for intermediate values of K_0 and K_1 (see Fig. 8.7).

Correlation profiles similar to those presented in Figs. 8.7-8.10 can be constructed for any network. In what follows we measure them in the Internet connectivity network on the level of so-called Autonomous Systems (AS). An Autonomous System is a group of workstations, servers, and routers belonging to one organization such as e.g. a university, a company, or an Internet Service Provider. Connections between such Autonomous Systems are achieved by the virtue of the Border Gateway Protocol (BGP), which establishes which other AS a given AS directly communicates with and what kind of routing information is exchanged in the course of these communications. Daily data about connections between individual Autonomous Systems are available at the website of the National Laboratory for Applied Network Research (NLANR) [25]. In our analysis we used the information about the network of Autonomous Systems collected on January 2, 2000. This data set consists of 12572 symmetric connections between 6474 AS. It is a scale-free network in which the power law connectivity distribution $p(k) \sim k^{-\gamma}$ with $\gamma = 2.2 \pm 0.1$ spans over 3 orders of magnitude in k [5].

An ensemble of 1000 randomized networks with the same connectivities of individual nodes was generated by the random rewiring algorithm described in the previous section. The corresponding R- and Z- correlation profiles are shown in Figs 8.11-8.12. From these figures one infers that the Internet is characterized by the following set of correlations:

1. Strong suppression of edges between nodes of low connectivity $3 \geq K_0, K_1 \geq 1$.

2. Suppression of edges between nodes that both are of intermediate connectivity $100 > K_0, K_1 \geq 10$,

3. Strong enhancement of the number of edges connecting nodes of low connectivity $3 \geq K_0 \geq 1$ to those of intermediate connectivity $100 > K_1 \geq 10$.

On the other hand any pair among 5 hub nodes with $K_0, K_1 > 300$ was found to be connected by an edge, both in the real network, and in a typical random sample. Hence $R(K_0, K_1)$ is close to 1 in the upper right corner of Fig. 8.11. The strong suppression of connections between pairs of nodes of low connectivity can in part be attributed to the constraint that all

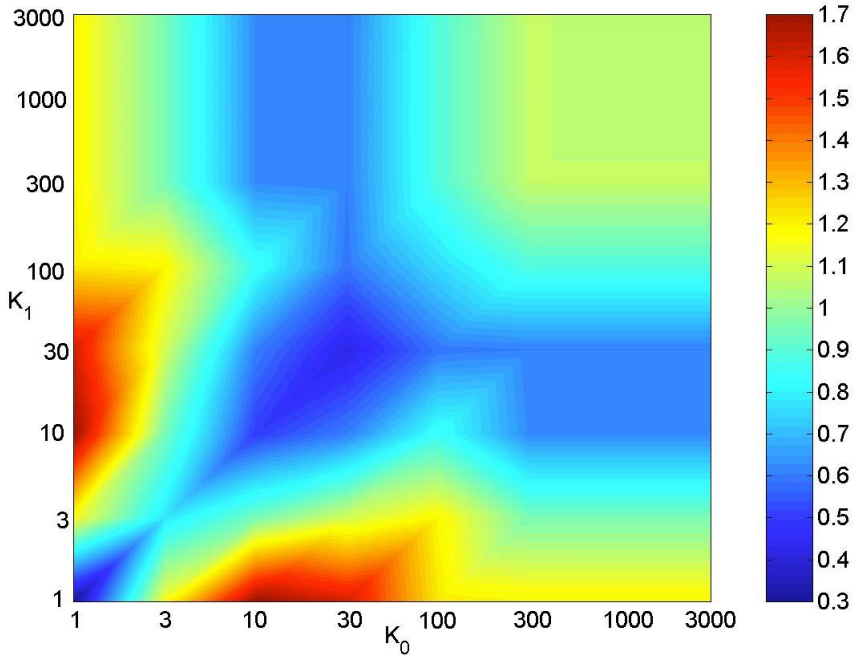


Figure 8.11: Correlation profile of the Internet. The ratio $R(K_0, K_1) = P(K_0, K_1)/P_r(K_0, K_1)$, where $P(K_0, K_1)$ is the probability that a pair of AS with connectivities K_0 and K_1 to be nearest neighbors of each other in the Internet, while $P_r(K_0, K_1)$ is the same probability in a randomized version of the same network, generated by the random rewiring algorithm described in the text. Note the logarithmic scale of both axes.

nodes on the Internet have to be connected to each other by at least one path. We have explicitly checked that there are indeed no isolated clusters in our data for the Internet. However, when we used an ensemble of random networks in which the formation of isolated clusters was prevented at every rewiring step, we found little change in the observed correlation profile.

The pattern of correlations observed in the Internet is consistent with a picture of multilevel hierarchy among its nodes. Indeed, based on their connectivity Autonomous Systems can be loosely separated into several hierarchical levels [26], which starts from "user level" AS of very low connectivity connected to regional, national, and international Internet Services Providers with increasing ranges of connectivity. From the correlation profile described above one infers that user level AS tend to connect to intermediate level AS (regional ISP) and that the really large ISP are all linked to each other by peer-to-peer connections.

The Internet is a convenient example to illustrate the difference between randomized version of the network generated by local randomization algorithms [12,13,15] in which multiple edges are forbidden, and the truly uncorrelated randomized version of the network generated by the stub reconnection algorithm [9, 16], which inevitably has multiple edges. In complex

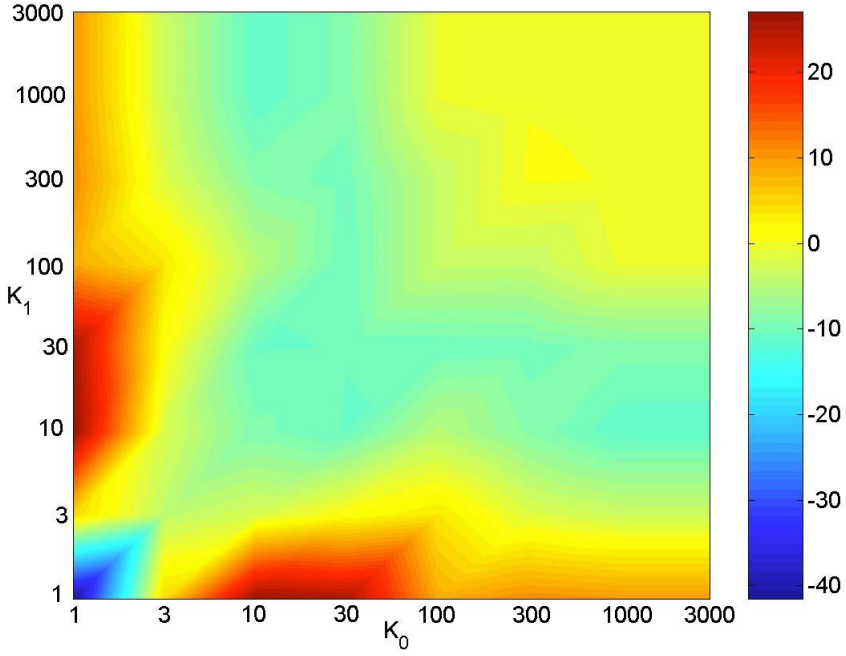


Figure 8.12: Statistical significance of correlations present in the Internet. The Z-score of correlations $Z(K_0, K_1) = (P(K_0, K_1) - P_r(K_0, K_1))/\sigma_r(K_0, K_1)$, where $P(K_0, K_1)$ is the probability that a pair of AS with connectivities K_0 and K_1 are nearest neighbors of each other in the Internet, while $P_r(K_0, K_1)$ is the same probability in a randomized version of the same network, generated by the random rewiring algorithm described in the text, and $\sigma_r(K_0, K_1)$ is the standard deviation of $P_r(K_0, K_1)$ measured in 1000 realizations of a randomized network. Note the logarithmic scale of both axes.

networks characterized by a broad connectivity distribution the stub reconnection algorithm usually results in multiple edges between hub nodes. In a network with E edges in total, the probability that the stub reconnecting algorithm would create a multiple edge between a pair of nodes with connectivities K_0 and K_1 becomes substantial if $K_0 K_1/(2E) > 1$. Since in scale-free networks characterized by a power law distribution of node connectivities $p(k) \sim k^{-\gamma}$ the connectivity of a few highest connected nodes in the system scales as $N^{1/(\gamma-1)}$, the expected number of edges between a pair of such hub nodes scales as $N^{2/(\gamma-1)}/E \sim N^{2/(\gamma-1)-1}$ becomes significant for the large number of nodes N provided that $\gamma < 3$. For example, in a randomized version of the Internet generated by this algorithm the expected number of edges connecting the two highest connected hubs of respectively $K_0 = 1458$ and $K_1 = 750$ is a swooping $K_0 K_1/(2E) = 1458 \cdot 750/(2 \cdot 12572) = 43.5$! This means that in a randomized version of the Internet with no multiple connections between nodes the connectivity between these hub nodes would be suppressed by a factor of 43 relative to a random network allowing for multiple edges. Thus the ban on multiple connections between a given pair of nodes gives

rise to an effective "repulsion" between hubs in such a randomized network. To quantify the level of this repulsion we measured the average connectivity $\langle K_1 \rangle_{K_0}$ of neighbors of sites with connectivity K_0 as a function of K_0 in the real Internet network (squares in Fig. 8.13) as well as in an ensemble of random networks with no multiple connections between nodes generated by the local rewiring algorithm (circles in Fig. 8.13.) From this figure it is clear that most of the $\langle K_1 \rangle_{K_0} \propto K_0^{-0.5}$ dependence reported in Ref. [27] is reproduced in our random ensemble and hence can be attributed to the above mentioned effective repulsion between hubs due to the constraint of having no more than one edge directly connecting them to each other. It is worthwhile to note that in the random network generated by the stub reconnection algorithm $\langle K_1 \rangle_{K_0} = \langle {K_1}.^2 \rangle / \langle K_1 \rangle \simeq 165$ would be independent of K_0.

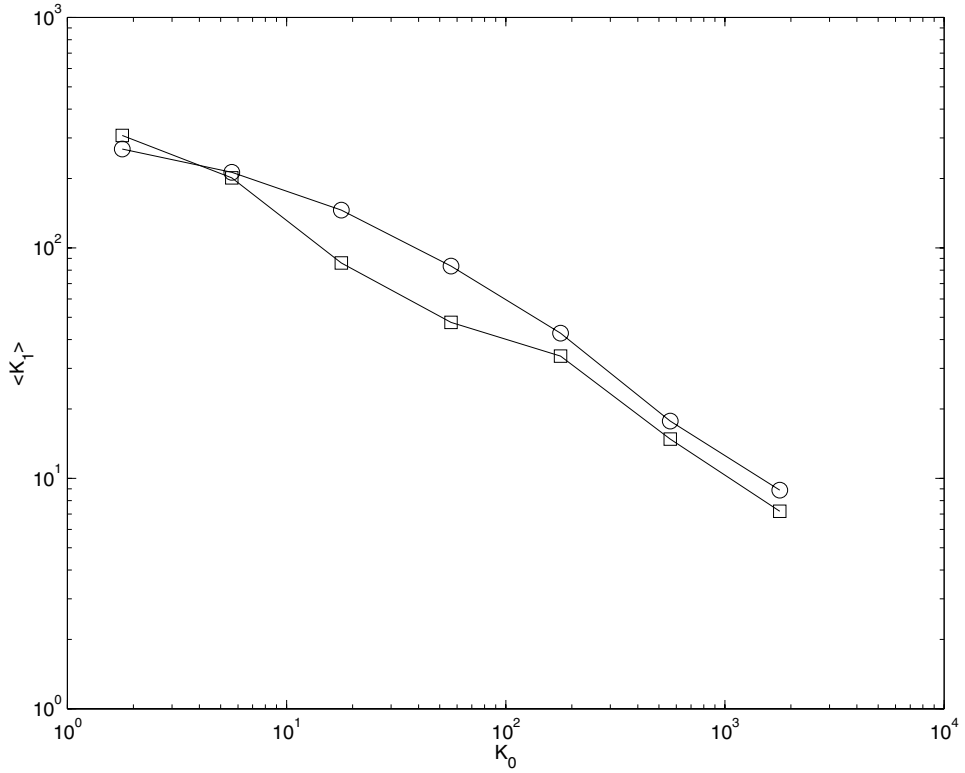


Figure 8.13: The average connectivity of a neighbor $\langle K_1 \rangle$ vs the connectivity of a node K_0 in the Internet (squares) and its randomized version with no multiple edges (circles). The error bars in multiple realizations of the randomized network are smaller than the size of the symbol.

Let $p(K)$ be the probability distribution of connectivities in the complex network and let us consider its random counterpart generated by stub reconnection algorithm [9,16]. Since in this algorithm each of the two nodes is independently selected to form a connection through one of its edge stubs, the probability to pick a node with connectivity K is given by $Kp(K)/\langle K \rangle$,

and the conditional probability distribution $P_r^{stub}(K_1|K_0)$ is independent of K_0 and equal to

$$P_r^{stub}(K_1|K_0) = K_1 p(K_1)/\langle K\rangle \ . \tag{8.6}$$

On the other hand, in an ensemble of random scale-free networks with no multiple edges the conditional probability distribution $P(K_1|K_0)$ crosses over between $K_1/p(K_1)$ functional form for $K_1 \ll K_1^* = 2E/K_0$ to $p(K_1)$ for $K_1 \gg K_1^*$. We have confirmed numerically that $P(K_1|K_0)$ in our randomized ensemble has a very similar shape to that observed in the real Internet [28] thus once more verifying that most of correlation effects visible in the internet network can be attributed to the effective repulsion between hubs due to the constraint of no multiple connections. The remaining correlations are quantified in the correlation profile in Figs. (8.11,8.12).

8.4 Network motifs: Transcriptional regulation in E. coli

A fundamental question in understanding a complex network is whether it can be decomposed into building blocks. Ideally, such building blocks would be well separated from each other, and thus the entire network dynamics could be approximated by a combination of dynamics of these basic elements.

A natural place to examine this question further is the best characterized biological regulation network, that of transcription regulation in the bacterium *E. coli* [29, 30]. In this network, the nodes are operons (groups of genes transcribed form a single mRNA). Some of the operons encode for regulatory proteins, which regulate the transcription rates of certain other operons. Thus, each edge is directed from an operon encoding a regulatory protein to an operon regulated by that protein.

The transcription network of *E. coli* has a broad scale-free-like distribution of outgoing edges, and a compact distribution of incoming edges. The resulting directed graph shown in Fig. 8.1 appears quite complex, as seen by plotting it with a standard graph display algorithm [2] (the dataset is available at www.weizmann.ac.il/mcb/UriAlon). We note that transcription graphs have an additional color for each edge: each regulatory protein can be a positive or a negative regulator, termed activator or repressor respectively (in rare cases a dual regulation is found). To detect recurring patterns in this graph that are likely to have a functional role, a new approach based on **network motifs** was recently presented [14]. The approach is simple to define: one enumerates the appearance of all types of subgraphs in the graph. The number of appearances of each subgraph is then compared to an ensemble of randomized networks, generated as discussed in section 2, such that each of the randomized networks preserves the incoming and outgoing edge degrees for each node. Network motifs are subgraphs that satisfy a statistical-significance criterion: the probability that they appear in a randomized graph more often than in the real graph is smaller than a threshold P-value (eg $P < 0.01$). The profile of the network given by the set of such statistically significant network motifs nicely compliments the correlation profile described in the previous section.

Here we review the representation of 3 particular kinds of patterns:

a. Type I patterns: Patterns with no free structural parameters. This involves full enumeration of small subgraphs: all 13 types of 3-node connected, directed subgraphs, and 199 types of 4-node subgraphs were enumerated in Ref [14]. The complexity rises exponentially for

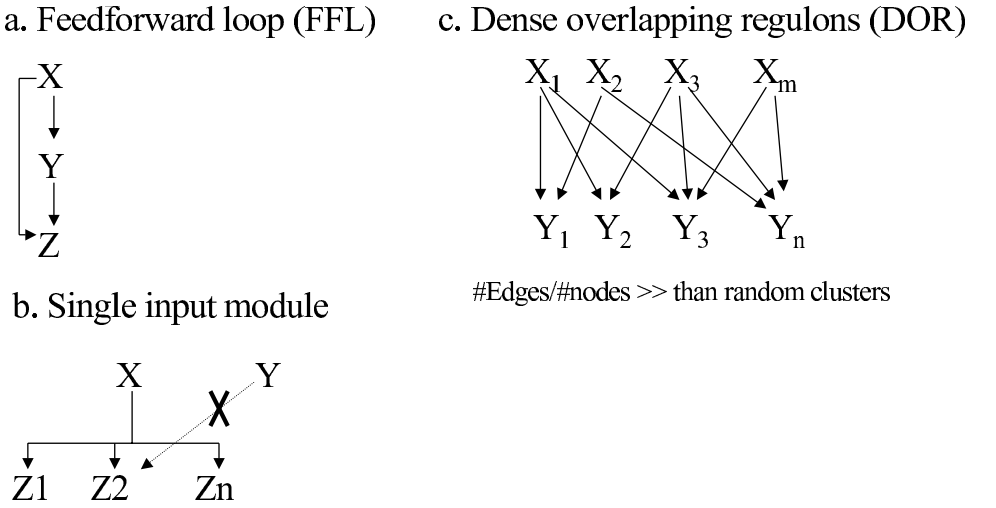


Figure 8.14: The three network motifs in the *E. coli* transcription network. In this figure, the edge colors (regulation signs) are not shown. **a)** Feed-Forward Loop (FFL). **b)** Single Input Module (SIM) where X regulates n output nodes. The output nodes have no other incoming edges. All regulations are of the same sign. **c)** Dense Overlapping Regulon (DOR). A set of input nodes X1... Xm regulates a set of output nodes Z1... Zn, with a resulting cluster of edges that is much denser than those found in randomized networks. In [14], a clustering algorithm that clusters nodes according to the number of overlapping inputs was presented, and used to compare the DOR structures to the randomized ensemble.

larger subgraphs, and efficient Monte-Carlo sampling methods still need to be devised for completely enumerating all types of n-node directed subgraphs. The algorithms for type-I-pattern detection can either take into account or ignore the edge colors. As shown below, the dynamical behavior of each subgraph may critically depend on the edge coloring.

b. Type II patterns: Patterns with a free structural parameter. Visual examination of the transcription network showed a recurring pattern: a set of nodes with only one incoming edge, all from the same master node. In Ref. [14] such patterns were termed Single Input Modules (SIM) since no node other than the master node regulates any of the nodes in this pattern. This pattern has a free structural parameter, the number of nodes in the SIM. An algorithm that searches for this pattern (and in general for identical rows in the connectivity matrix of the graph) was employed for comparison with randomized graphs. Taking edge color into account, the definition of SIMs is further restricted to the case where all edges have the same color (that is all negative or all positive regulation).

c. Type III patterns: Classes of patterns defined by an extensive characteristic such as edge density. In the present case, a clustering approach was defined for detecting regions in the graph that are more dense then in randomized graphs. This defined Dense Overlapping Regulons (DOR). A regulon is a biological term for a set of operons all regulated by the same regulatory protein, not necessarily exclusively [31]. Naturally, other pattern classes can be defined in this way.

First let us summarize the overall findings of this detailed analysis of network motifs: In regards to type I patterns it was found that out of 13 possible 3-node subgraphs only one is statistically significant. This motif was termed the Feed-Forward Loop (FFL) (Fig. 8.14), and the real network had 40 of these as compared to the 7 ± 5 found in an ensemble of randomized networks. Of the 4-node subgraphs, only one was significant, a pattern with 4 edges representing overlapping regulation $X \rightarrow W, Z$ and $Y \rightarrow W, Z$. This 4-node motif hints that dense overlapping regions are extant in the network. This was indeed found to be the case, but the dense region broke down into 6 weakly overlapping dense clusters, the type III patterns termed DORs (Fig. 8.14c). Finally, among the the type II patterns, large SIMs were found to be highly significant (Fig. 8.14b).

To further validate the motifs it was tested [14] whether they are sensitive to data errors/incomplete data. As in the case of correlation profiles discussed in the previous section, it was found that the statistical significance of the network motif is highly robust with respect to data errors. For example, when over 30% of the connections are removed at random, or added at random, all 3 motifs remain significant and no new motifs appear in the network. This robustness may fail, of course, if the dataset contains systematic errors with a bias for certain kinds of patterns. For example, well-known regulatory proteins are investigated by many labs, which may result in a tendency for an increased number of known connections for these nodes (an effect similar to searching for the coin under the streetlamp). This may exaggerate the number of SIMs detected. The existence of SIMs was also hinted at by the correlation profile of yeast regulatory network discussed in the previous section. Indeed, the abundance of SIMs must be at least partially responsible for the observed preference of highly connected proteins (hubs) to connect to neighbors with lower than average connectivity in this network. The fact that the preference of proteins with low and high connectivity to connect to each other was also observed in both the protein interaction and regulatory networks in yeast indicate that, perhaps, SIMs are a significant feature in all types of bio-molecular networks.

Over 80% of the nodes in the *E. coli* transcription network belong to one of the three motifs defined above, FFL, SIM or DOR. The remaining nodes usually belong to tiny disjoint components of 1-3 nodes. Thus the decomposition of the network into recurrent motifs allow us an alternative way to present the data, in terms of its structure, and of the relative position of the various motifs (Fig. 8.15). A small set of nodes with an outgoing edge degree much larger than average (global regulators) complicates the graph image. An important step in visualizing the network is to allow the nodes with high output degrees to appear multiple times in the image, acting as inputs to the various DOR structures in which they participate. This preserves all of the information but removes many complicating edges. It is seen that the transcription network of *E. coli* is mostly a two-layer feed-forward network. The FFLs and SIMs are often at the outputs of the DORs. These two motifs are therefore integrated into the DOR structures.

Dynamical behavior of network motifs: Two of the motifs, FFL and SIM, have been shown to carry out distinct information processing functions, using numerical simulations [14]. The SIM motif allows the operons to be turned on in a particular temporal order and turned off in the reverse order, akin to the "First-In Last-Out" (FILO) pipeline. The temporal order is encoded in the relative **strengths** of the edges to each node (mechanistically, in the strength and position of regulatory protein and RNA polymerase binding sites in the regulatory DNA region that precedes each operon). To understand the dynamics of the FFL motif, the

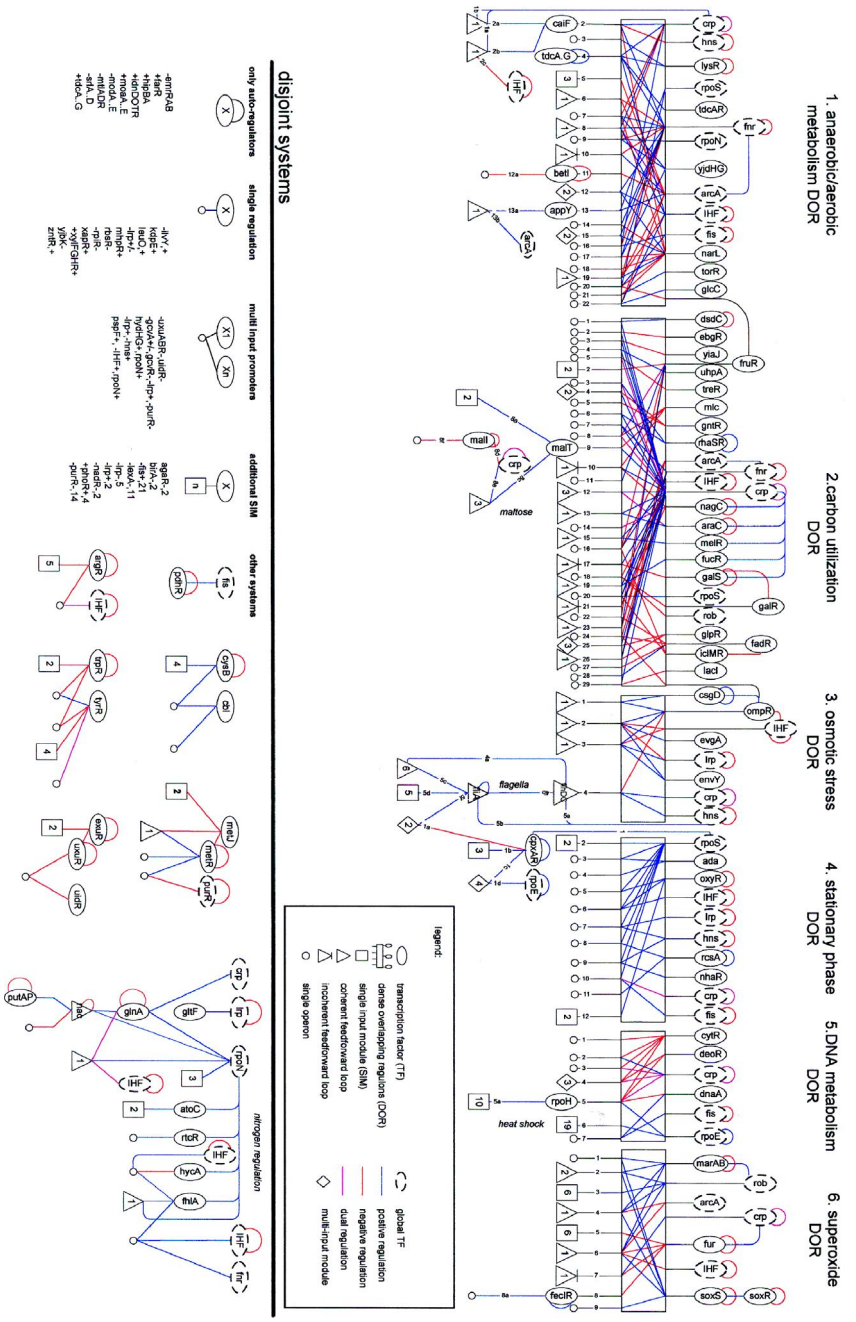


Figure 8.15: The complete known *E. coli* transcription network displayed using network motifs. This version has several corrected edges and several new nodes as compared to the images in [14]. The complete dataset is available at www.weizmann.ac.il/mcb/UriAlon

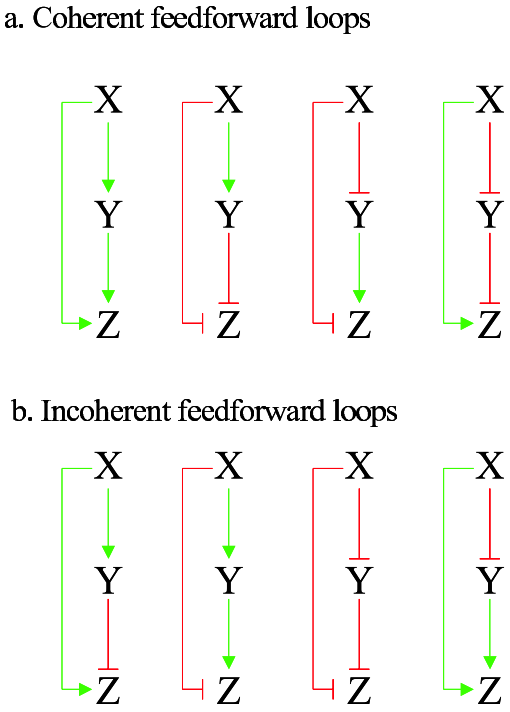


Figure 8.16: Coherent (a) and Incoherent (b) feed-forward loops. Arrows represent positive regulation, and $\dashv$ symbols represent negative regulation (repression).

edge coloring becomes important. When taking into account the two edge colors, there are 8 possible colorings of the FFLs three edges. However, it is sufficient to consider two types of FFLs. The FFL is composed of a direct path from node X to node Z, and an indirect path through node Y. A coherent FFL has the same sign on the direct path as the net sign of the indirect path (Fig.8.16a), while an incoherent FFL has opposing signs in the direct and indirect paths (Fig 8.16b). The coherent FFL is the dominant form in *E. coli* ($P < 0.001$), while the incoherent FFL is only marginally significant ($P \sim 0.03$). The two types of FFL motifs can show very different dynamical behavior. For simplicity we consider here the case where the two inputs act as an AND gate to control the output Z, a typical case in transcription systems. The coherent feed-forward loop acts to reject rapid input pulses of X that go from OFF to ON, responding only to persistent inputs. However, there is a strong response even to short reverse pulses from ON to OFF. Thus the coherent FFL can act as a **sign-sensitive filter**. The condition for this is that the level of Y in the OFF state of X is below the activation threshold for Z. The typical width of pulses first passed by this filter is given by the time it takes Y to cross Z's activation threshold.

The incoherent FFL can act as a sign-sensitive differentiator (pulse generator). In a step where X goes from OFF to ON, Z is rapidly activated. Then, Y levels build up to cross the repression threshold for Z. Thus after a delay, Z becomes inactivated, given that the negative

effect of Y is strong enough. On the other hand, no response is seen in a step from ON to OFF, where Z remains suppressed.

These two functions, sign-sensitive filtering and temporal transcription programs, may be basic tasks in the information processing performed by *E. coli*. Indeed, it has recently been found experimentally that temporal programs exist in systems such as flagella bio-synthesis [32] and SOS DNA repair (Ronen, Rosenberg, Shraiman, Alon, submitted 2002), where the temporal order within groups of operons controlled by the same regulator corresponded to the functional order of these genes. The SIM mechanism is likely to be at play.

These findings can point the way to experiments designed to understand the functions of each motif. Once these functions are understood, one may check whether the dynamics of the entire network can be well approximated as a combination of the dynamics of its separate motifs.

8.5 Discussion: What it may all mean?

The large scale organization of molecular networks deduced from correlation profiles of protein interaction and transcription regulatory networks in yeast, and the set of statistically significant network motifs in the regulatory network of *E. coli* is consistent with compartmentalization and modularity characteristic of many cellular processes [33]. Indeed, the suppression of connections between highly connected proteins and the abundance of DOR network motifs both suggest the picture of semi-independent modules centered around or regulated by individual hubs. On the other hand, the very fact that these molecular networks do not separate into many isolated components but are dominated by one "giant component" suggests that this tendency towards modularity is not taken to its logical end. It can in fact be described as "soft modularity", in which interactions between individual modules are suppressed but not completely eliminated. Thus on sufficiently large scale molecular networks exhibit system properties making their behavior different from that of a set of mutually independent modules. Two recent observations independently hint at global interrelations in the overall connectivity pattern of molecular networks:

1. Elena and Lenski [34] studied the cooperativity of regulation in *E. coli* by comparing changes of the cell cycle length in single-gene null mutants with those in double null mutants. They concluded that about 30% of gene pairs exhibited more than additive effects on cell cycle length, and thus at least 30% of protein pairs are functionally interconnected. Such level of cooperativity would be impossible in a regulatory network consisting of a large number of independent modules.

2. C.K. Stover et al. [35] found that the number of transcription factors (N_{tr}) in procaryotic organisms grows as a *square* of the number of genes N: $N_{tr} \propto N^2$. Hence, each additional gene (or gene module/regulon) appears to be regulated with respect to all genes that are already present. This indicates an overall regulation pattern that on sufficiently large scale is neither modular, nor hierarchic.

On the other hand, in this work we demonstrated that already on the level of the correlation profile (the two point correlation function) these networks exhibit a certain degree of modularity. A further implication of those modular features manifested by the deficit of connections

between highly connected proteins (Figs. 8.7, 8.8) is in the suppression of propagation of deleterious perturbations over the network. It is reasonable to assume that certain perturbations such as e.g. a significant change in the concentration of a given protein (including it vanishing altogether in a null-mutant cell) with a ceratin probability can affect its first, second, and sometimes even more distant neighbors in the corresponding network. While the number of immediate neighbors of a node is by definition equal to its own connectivity K_0, the average number of its second neighbors is bound from above by $K_0\langle(K_1-1)\rangle_{K_0}$ and thus depends on the correlation profile of the network. In addition it is sensitive to higher order correlation patterns of the network. For example, in the presence of a significant level of clustering the number of second neighbors can fall well below the above mentioned upper bound. Since highly connected nodes serve as powerful amplifiers for the propagation of deleterious perturbations it is especially important to suppress this propagation beyond their immediate neighbors. It was argued that scale-free networks in general are very vulnerable to attacks aimed at hubs [36, 37]. The deficit of edges directly connecting hubs to each other reduces the branching ratio around these nodes and thus provides a certain degree of protection against such attacks.

To summarize the above discussion, it is feasible that molecular networks operating in living cells have organized themselves in an interaction pattern that is both robust and specific. Topologically the specificity of different functional modules is enhanced by limiting interactions between hubs and suppressing the average connectivity of their neighbors. Such correlations are also evident on a more detailed level of local structural motifs such as SIMs and DORs. Each of those network motifs has certain computational properties providing the cell with appropriate responses to environmental and internal changes. On a larger scale there is evidence for interconnections between these modules, although the principles of such global organization of living cells remain unclear from the present day data and analysis tools.

Correlation profiles and statistically significant network motifs allow one to distinguish between different complex networks, even if their connectivity distributions appear identical. Thus, for example, the Internet at the level of Autonomous Systems and physical interactions among yeast proteins are both characterized by power-law connectivity distributions with rather similar exponents. However, correlation profiles of these two networks (Figs. 8.7, 8.11), are qualitatively different from each other. First, in the Internet unlike in molecular networks, connections between the highly connected nodes were not suppressed. In fact any pair of hubs on the Internet was connected to each other by a direct link. Secondly, the protein interaction network in yeast is characterized by an enhancement of connections between nodes with intermediate connectivities , as opposed to the Internet, where such connections were found to be strongly suppressed. Also, unlike protein interaction networks the Internet has a deficit of edges connecting nodes of very low connectivity to each other. This all indicates that the information processing mechanisms relevant to protein interaction networks are qualitatively different from those relevant to the Internet.

The main goal of the present work was to introduce a number of statistical tools necessary for analyzing topological patterns and correlations in networks. These tools allowed us to identify the set of different topological patterns and characteristic building blocks (motifs) present in a broad range of complex networks, which may help to better understand possible

mechanisms for their function and evolution. The advantage of our approach lies also in it iterative nature in which the understanding of more and more complex topological properties of the network gradually builds up on the analysis of its lower level features.

References

[1] For an overview see e.g. other chapters of this book.

[2] V. Batagelj, A. Mrvar, *Pajek - Program for Large Network Analysis*, Connections **21** 2, 47–57 (1998)

[3] P. Erdős, A. Rényi, *On the evolution of random graphs*, Publ. Math. Inst. Hung. Acad. Sci. 5, 1760 (1960).

[4] A.-L. Barabasi, R. Albert, *Emergence of scaling in random networks*, Science **286**, 509–512 (1999).

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).

[6] A. Broder, *et al.*, *Graph Structure in the Web*, Computer Networks **33**, 309-320 (2000).

[7] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A.-L. Barabasi, *The large scale organization of methabolic networks*, Nature **407**, 651–654 (2000).

[8] H. Jeong, S. Mason, A.-L. Barabasi, Z.N. Oltvai, *Centrality and lethality of protein networks*, Nature **411**, 41–42 (2001).

[9] M. E. J. Newman, S. H. Strogatz, and D. J. Watts *Random graphs with arbitrary degree distributions and their applications*, Phys. Rev. E **64**, 026118, 1-17 (2001); See also the chapter by M. E. J. Newman in this book.

[10] D. Watts and S. Strogatz, *Collective Dynamics of Small World Networks*, Nature **293**, 400 (1998).

[11] S. Wasserman and K. Faust, *Social Network Analysis*, Cambridge University Press, 1994.

[12] S. Maslov and K. Sneppen, *Pattern Detection in Complex Networks: Correlation Profile of the Internet*, cond-mat/0205379, (2002).

[13] S. Maslov and K. Sneppen, *Specificity and Stability in Topology of Protein Networks*, Science **296**, 910-913, (2002).

[14] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, *Network motifs in the transcriptional regulation of Escherichia coli*, Nature Genetics **31**(1):64-68 (2002).

[15] Early studies of these algorithms were reported in: D. Gale, *A theorem of flows in networks*, Pacific J. Math. **7**, 1073-1082 (1957); H.J. Ryser, *Matrices of zeros and ones in combinatorial mathematics, in "Recent Advances in Matrix Theory"*, pp. 103-124, Univ. of Wisconsin Press, Madison, (1964). For more recent references see e.g.: R. Kannan, P. Tetali, S. Vempala, *Simple Markov-chain algorithms for generating bipartite graphs and tournaments*, Random Structures and Algorithms **14**, 293-308, (1999).

[16] E.A. Bender and E.R. Canfield, *The asymptotic number of labeled graphs with given degree sequences*, Journal of Combinatorial Theory A **24**, 296–307 (1978).

[17] The set of MATLAB programs can be downloaded at http://cmth.phy.bnl.gov/ maslov/matlab.htm

[18] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995);
M. Molloy and B. Reed, Combinatorics, Probab. Comput. **7**, 295 (1998).

[19] T. Ito, *et al.*, *A comprehensive two-hybrid analysis to explore the yeast protein interactome*, Proc. Natl. Acad. Sci. USA **98**, 4569–4574 (2001).

[20] M. C. Costanzo, *et al.*, *YPD, PombePD, and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information*, Nucleic Acids Research **29**, 75-79 (2001).

[21] A.-C. Gavin, *et al.*, *Functional organization of the yeast proteome by systematic analysis of protein complexes*, Nature **415**, 141 (2002).

[22] P. Uetz, *et al.*, *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae*, Nature **403**, 623–627 (2000).

[23] The supplementary materials to B. Schwikowski, P. Uetz, and S. Fields, *A network of protein protein interactions in yeast*, Nature Biotechnology 18, 1257 - 1261 (2000).

[24] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, *Modeling of protein interaction networks*, cond-mat/0108043 (2001).

[25] Website maintained by the National Laboratory for Applied Network Research (NLANR) Measurement and Network Analysis Group at http://moat.nlanr.net/

[26] The idea of deducing the hierachial level of an AS from its connectivity was proposed in: R. Govindan and A. Reddy, *Analysis of internet inter-domain topology and route stability*, In Proceedings of the IEEE Infocom, pages 851-858, Kobe, Japan, April 1997, avaliable for download at http://citeseer.nj.nec.com/govindan97analysis.html, and later elaborated e.g. in L. Gao, *On inferring autonomous system relationships in the Internet*, in Proc. IEEE Global Internet Symposium, November 2000, avalaible for download at http://citeseer.nj.nec.com/gao00inferring.html.

[27] R. Pastor-Satorras, A. Vazquez, A. Vespignani, *Dynamical and Correlation Properties of the Internet*, Phys. Rev. Lett. **87**, 258701 1-4 (2001).

[28] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **88**, 108701 (2002).

[29] D. Thieffry, A.M. Huerta, E. Pérez-Rueda, & J. Collado-Vides, *From specific gene regulation to global regulatory networks: a characterization of Escherichia coli transcriptional network*, BioEssays 20, 433-440 (1998).

[30] A.M. Huerta, H. Salgado, D. Thieffry & J. Collado-Vides, *RegulonDB: A database on transcriptional regulation in Escherichia coli* Nucleic Acid Res. 26 (1998), 55-59.

[31] F. C. Neidhardt and M. A. Savageau, *Regulation beyond the operon*. In *"Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology,"* Vol. 1 (F. C. Neidhardt, Ed.), pp. 1310-1324, Am. Soc. Microbiol., Washington DC (1996).

[32] S. Kalir, K. McClure, C. Pabaraju, C. Southward, M. Ronen, S. Leibler, M.G. Surette and U. Alon, *Ordering genes in a flagella pathway by analysis expression kinetics from living bacteria*. Science **292**, 2080-2083 (2001).

[33] L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, *From molecular to modular cell biology*, Nature **402** (6761 Suppl), C47–52 (1999).

[34] S.F. Elena & R.E. Lenski, *Test of synergetic interactions among deleterious mutations in bacteria Nature* **390**, 395-398 (1999).

[35] C.K.Stover, *Complete genome sequence of Pseudomonas aerugiosa PA01, an opportunistic pathogen Nature* **406**, 959-398 (2000).

[36] R. Albert, H. Jeong, A.-L. Barabasi, *Error and attack tolerance of complex networks*, Nature **406**, 378-382 (2000).

[37] B. Vogelstein, D. Lane, and A.J. Levine, *Surfing the p53 network*, Nature **408**, 307-310 (2000).

9 Theory of interacting neural networks

Wolfgang Kinzel

9.1 Introduction

Neural networks learn from examples. This concept has extensively been studied using models and methods of statistical physics [1, 2]. In particular the following scenario has been investigated: Feed-forward networks are trained on examples generated by a different network.

Feed-forward networks classify high dimensional data, in the simplest case by a single output bit (1/0, wrong/correct, yes/no). They are adaptive algorithms, their parameters (= synaptic weights) are adapting to a set of training examples, in our case a set of input/output pairs. After the training phase, the networks have achieved some knowledge about the rule which has generated the examples, the network can classify input vectors which it never has seen before, it can generalise.

Several mathematical models studied before use training examples which are generated by a different neural network, called the "teacher". On-line training means that the "student", at each training step, receives a new example from the teacher network. Each example is used only once for training. Hence, in this case training may be considered as dynamics of interacting neural networks: A teacher network is sending signals (= examples) to the student network which is stepwise changing its weights according to the received message.

Mathematical methods have been developed to calculate the properties of the dynamics of interacting networks. In the limit of large networks one can describe the system by a differential equation for a few "order parameters", which determine, for example, the generalisation error as a function of the number of training examples [3].

In this contribution we give an overview over recent work on the theory of interacting neural networks. The model is defined in Section 2. The typical teacher/student scenario is considered in Section 3. A static teacher network is presenting training examples for an adaptive student network. In the case of multilayer networks, the student shows a transition from a symmetric state to specialisation. Neural networks can also generate a time series. Training on time series and predicting it are studied in Section 4. When a network is trained on its own output, it is interacting with itself. Such a scenario has implications on the theory of prediction algorithms, as discussed in Section 5. When a system of networks is trained on its minority decisions, it may be considered as a model for competition in closed markets, see Section 6. In Section 7 we consider two mutually interacting networks. A novel phenomenon is observed: synchronisation by mutual learning. In Section 8 it is shown, how this phenomenon can be applied to cryptography: Generation of a secret key over a public channel.

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

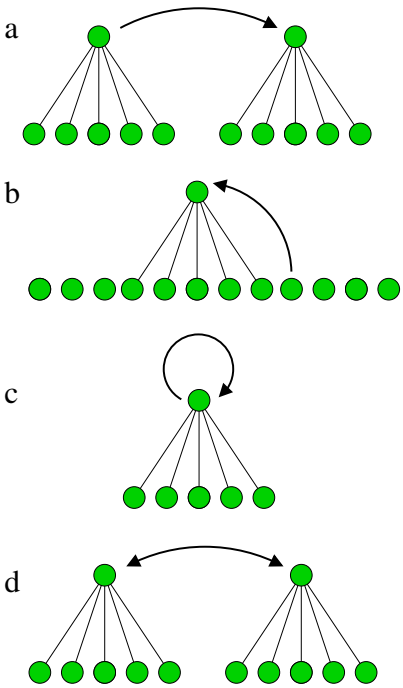


Figure 9.1: Different kinds of interaction discussed in this contribution: (a) static teacher vs. adaptive student, (b) time series generation and prediction, (c) self-interaction, (d) mutual learning

9.2 On-line training

The simplest mathematical neural network is the perceptron. It consists of a single layer of N synaptic weights $\underline{w} = (w_1, ..., w_N)$. For a given input vector $\underline{x}$, the output bit is given by

$$\sigma = \text{sign}\left(\sum_{i=1}^{N} w_i x_i\right) \tag{9.1}$$

The decision surface of the perceptron is just a hyperplane in the N-dimensional input space, $\underline{w} \cdot \underline{x} = 0$. A perceptron may also have a continuous output y as

$$y = \tanh\left(\sum_{i=1}^{N} w_i x_i\right) \tag{9.2}$$

A perceptron may be considered as an elementary unit of a more complex network like an attractor network or a multilayer network. In fact any function can be approximated by a multilayer network if the number of hidden units is large enough [2].

The perceptron can learn from examples. Examples are input/output pairs,

$$(\underline{x}(t), \underline{\sigma}(t)) \qquad t = 1, ..., \alpha N \tag{9.3}$$

On-line training means that at each time step t the weights of the perceptron adapt to a new example, for instance by the rule

$$\underline{w}(t+1) = \underline{w}(t) + \frac{\eta}{N}\sigma(t)\underline{x}(t)\ F(\sigma(t)\underline{x}(t)\cdot\underline{w}(t)) \tag{9.4}$$

$F(z) = 1$ is usually called – after the corresponding biological mechanism – the Hebbian rule, each synapse w_i responds to the activities $\sigma(t)x_i(t)$ at its ends. $F(z) = \Theta(-z)$ is called the Rosenblatt rule: a training step occurs only if the example is misclassified. Finally, the Adatron rule $F(z) = |z|\Theta(-z)$ is important since it gives good results for generalisation, as discussed in the following. For the last two learning rules, in addition to the two neural activities at the synaptic ends, the postsynaptic potential determines the strength of the synaptic adaptation.

9.3 Generalisation

Now we consider two perceptrons. One is called the teacher network which is producing a set of examples. It is receiving a set of random input vectors $\underline{x}(t)$ and generating output bits $\sigma(t)$. The teacher has a fixed weight vector $\underline{w}^T$.

Each time the teacher is producing a new example, the student perceptron is trained on it according to (9.4). As a consequence, the weight vector of the student $\underline{w}^S(t)$ is time dependent. The student tries to approach the teacher, at each training step t its weight vector moves towards the one of the teacher. It is easy to see that for random inputs $\underline{x}$, equation (9.4) gives a kind of random walk in N-dimensional space with a bias towards the teacher vector $\underline{w}^T$.

The distance between student and teacher can be measured by the overlap

$$R = \frac{\underline{w}^T\cdot\underline{w}^S}{|\underline{w}^T||\underline{w}^S|} \tag{9.5}$$

The quantity R determines the angle ϕ between student and teacher weights, $R = \cos\phi$. It turns out that from this overlap the generalisation error e_g can be calculated. The generalisation error is the probability that the student gives an answer to a random input $\underline{x}$ which is different from the one of the teacher. One finds

$$e_g = \frac{\arccos R}{\pi} \tag{9.6}$$

In the limit of infinitely many input units, $N\to\infty$, the dynamics of the overlap $R(t)$ can be calculated analytically. According to (9.4), the size of the training step scales down with $1/N$. Therefore one defines a variable $\alpha = t/N$ which becomes a continuous variable, called time, in the limit of large N.

The time dependence of $R(\alpha)$ is obtained by multiplying (9.4) by $\underline{w}^T$ and $\underline{w}^S$ and by averaging these two equations over the random input vector $\underline{x}$. This can be done since the expressions $\underline{w}\cdot\underline{x}$ are Gaussian variables.

For the Adatron rule one finally obtains the differential equation [4]

$$\frac{dR}{d\alpha} = -\frac{R}{2\pi}\arccos R + \frac{1}{\pi}\Big(1-\frac{R^2}{2}\Big)\sqrt{1-R^2} \tag{9.7}$$

The generalisation error calculated by this equation is shown in Fig. 9.2. If only a finite number of examples has been learned, $\alpha = 0$, the error is 50 %, as bad as random guessing. If the number of training examples is of the order of N, $\alpha > 0$, the student network has obtained some overlap to the teacher. In the limit of large values of α the error decreases to zero, the student has obtained complete knowledge about the parameters of the teacher.

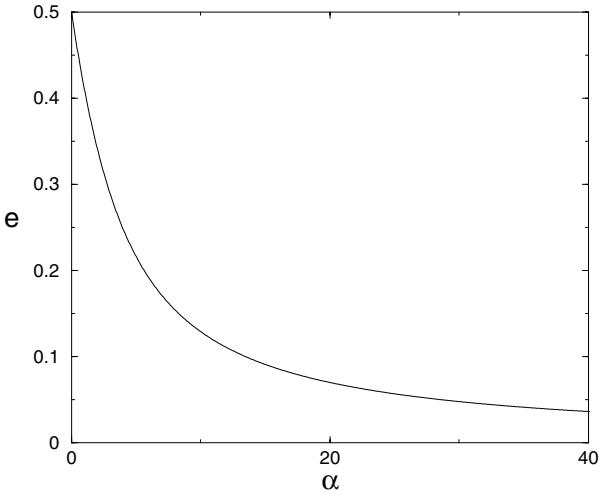


Figure 9.2: Generalisation error as a function of time, from Ref. [4]

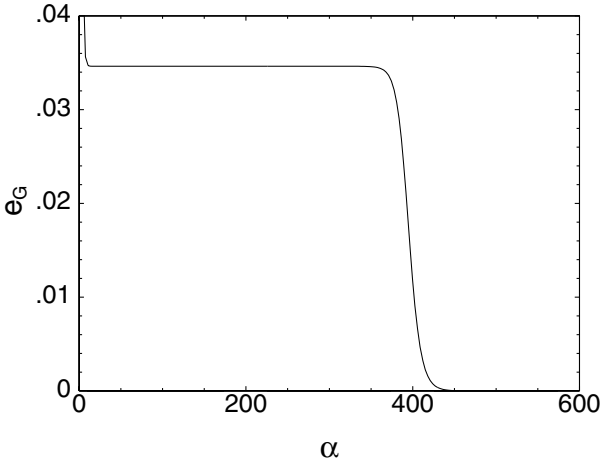


Figure 9.3: Generalisation error as a function of time, for a multilayer network with two hidden units, from Ref. [7]

The asymptotic decay of the generalisation error depends on the learning rule. One finds for the Adatron rule $e_g \propto 3/2\,\alpha^{-1}$. In fact, it has been shown that the error cannot decay faster

than $e_g \propto 0.88\,\alpha^{-1}$ [5]. For the Rosenblatt rule, one finds $e_g \propto \alpha^{-1/3}$, and for the Hebbian rule $e_g \propto \alpha^{-1/2}$. But in all cases the student succeeds to approach the teacher when it makes of the order of N steps. This even holds when the examples are distorted by noise [6].

Learning from examples works for more complex networks, too. Here I would like to mention the work on specialisation of committee machines [7]. Such a network is a multilayer network with several hidden units, similar to Fig. 9.8. The output bits of the continuous hidden units are summed and taken as the output of the network. Teacher and student networks have an identical architecture, and the learning step is just a gradient descent of the training error, the quadratic deviation between teacher and student output.

The corresponding generalisation error is shown in Fig. 9.3. For small number of examples it decreases fast, then it reaches a plateau and only for a huge number of examples it decreases to zero.

The motion of the student network can be expressed by the overlap between the corresponding members of the two machines. The teacher as well as the student consists of two weights vectors, $\underline{w}_1^{T/S}$ and $\underline{w}_2^{T/S}$. A distance between teacher and student can be defined from the overlaps

$$R_{i,j} = \underline{w}_i^T \cdot \underline{w}_j^S \tag{9.8}$$

Initially, all vectors are random, hence up to fluctuations all the overlaps are zero. Then all overlaps increase due to learning. But on the plateau of Fig. 9.3 the overlaps are all identical, $R_{1,1} = R_{2,2} = R_{1,2}$. The student has achieved some knowledge, but it is in a symmetric state. Only if the student receives much more information it can specialise: $R_{1,1} = R_{2,2}$ are much larger than $R_{1,2}$.

In the initial process the two members of the student committee act as being one single perceptron, but later they specialise and follow their partners in the teacher committee until they achieve complete knowledge for $\alpha \to \infty$.

9.4 Time series prediction and generation

Neural networks are successful prediction algorithms [8]. Given a sequence of numbers, a neural network can be trained on this sequence by moving it over the sequence, as shown in Fig.9.4. This sequence can be produced by another network, called the teacher, by generating a new number and using it as an input component in the next step. Hence we have a new kind of interacting networks: The teacher with its static weight vector is a bit or sequence generator [9]. The student is adapting its weights to the sequence generated by the teacher.

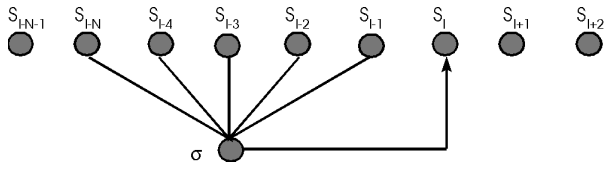


Figure 9.4: Time series generation and prediction by a perceptron

In principle this is the scenario described in the previous section. Here the difference is that the patterns are correlated, the input is not random but generated from the output of the teacher.

If a neural network cannot generate a given sequence of numbers, it cannot predict it with zero error. Hence one has investigated the generation of time series by neural networks [11–14] But this is not the whole story. Even if the sequence has been generated by an (unknown) neural network (the teacher), a different network (the student) can try to learn and to predict this sequence. In this context we are interested in two questions:

1. When a student network with the identical architecture as the teacher's is trained on the sequence, how does the overlap between student and teacher develop with the number of training examples (= windows of the sequence)?

2. After the student network has been trained on a part of the sequence, how well can it predict the sequence several steps ahead?

Recently these questions have been investigated numerically for the simple perceptron, equation (9.1,9.2) [10]. Consider a teacher perceptron with weight vector $\underline{w}^T$ generating the sequence $S_0, S_1, S_2, \ldots, S_t, \ldots$ This sequence follows the equation

$$S_t = f\left(\frac{1}{N}\sum_{j=1}^{N} w_j S_{t-j}\right) \tag{9.9}$$

where $f(x)$ is the transfer function. It has been shown that a perceptron can generate simple as well as complex sequences [11, 14].

If $f(x)$ is monotonic, for instance $f(x) = \tanh(\beta x)$, then in general one obtains quasiperiodic sequences. In fact, the sequence is essentially generated by one Fourier component of the weight vector w_i^T [11]. If the transfer function, however, is not monotonic, for instance $f(x) = \sin(\beta x)$, then the sequence can be chaotic, depending on the model parameters [14]. For both cases, learning and prediction have been investigated [10].

If a quasi periodic sequence is learned on-line, using gradient descent to update the weights,

$$\Delta w_i = \frac{\eta}{N}(S_t - f(h)) \cdot f'(h) \cdot S_{t-i} \;\text{ with }\; h = \beta \sum_{j=1}^{N} w_j S_{t-j} \tag{9.10}$$

then one has found two time scales (time α means the number of training steps divided by N):

1. A short scale on which the overlap $R(\alpha)$ between teacher and student rapidly increases to a value which is still far away from the value $R = 1$, which corresponds to perfect agreement.

2. A long one on which the overlap $R(\alpha)$ increases very slowly. Numerical simulations up to $10^6 N$ training steps yielded an overlap which was close but sell different from the value $R = 1$.

Although there is a mathematical theorem on stochastic optimisation which seems to guarantee convergence to perfect success [15], the on–line algorithm cannot gain much information about the teacher network, at least during reasonable training periods.

This is completely different for a chaotic time series generated by a corresponding teacher network with $f(x) = \sin(x)$. It turns out that the chaotic series appears like a random one: After a number of training steps of the order of N the overlap relaxes exponentially fast to perfect agreement between teacher and student.

Hence, after training the perceptron with a number of examples of the order of N we obtain the two cases: For a quasi periodic sequence the student has not obtained much information about the teacher, while for a chaotic sequence the student's weight vector comes close to the one of the teacher. One important question remains: How well can the student predict the time series?

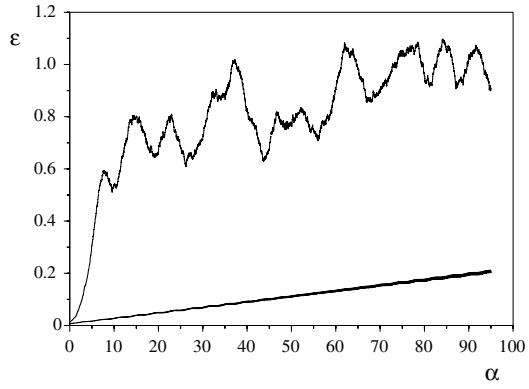


Figure 9.5: Prediction error as a function of time steps ahead, for a quasi periodic (lower) and chaotic (upper) series, from Ref. [10]

Fig.9.5 shows the prediction error as a function of the time interval over which the student makes the predictions. The student network which has been trained on the quasi periodic sequence can predict it very well. The error increases linearly with the size of the interval, even predicting $10N$ steps ahead yields an error of about 10% of the total possible range. On the other side, the student trained on the chaotic sequence cannot make predictions. The prediction error increases exponentially with time; already after a few steps the error corresponds to random guessing, $e_g \simeq 1$. The explanation is that an infinitesimal change in the parameters of a chaotic map has the same effect as a small change in initial conditions, namely, an exponential growth in the distance between original and the disturbed trajectory.

In summary one finds the counterintuitive result:

1. A network trained on a quasiperiodic sequence does not obtain much information about the teacher network which generated the sequence. But the network can predict this sequence over many (of the order of N) steps ahead.

2. A network trained on a chaotic sequence obtains almost complete knowledge about the teacher network. But this network cannot make reasonable long-term predictions on the sequence.

It would be interesting to find out whether this result also holds for other prediction algorithms, such as multi-layer networks.

9.5 Self-interaction

In the previous section the time series was generated by a static teacher network. Now we consider a network which changes its synaptic weights while it is generating a bit sequence. The teacher is interacting with itself. The motivation of this investigation stems from the following problem:

Consider some arbitrary prediction algorithm. It may contain all the knowledge of mankind, many experts may have developed it. Now there is a bit sequence $S_1, S_2, \ldots$ and the algorithm has been trained on the first t bits $S_1, \ldots, S_t$. Can it predict the next bit S_{t+1}? Is the prediction error, averaged over a large t interval, less than 50%?

If the bit sequence is random then every algorithm will give a prediction error of 50%. But if there are some correlations in the sequence then a clever algorithm should be able to reduce this error. In fact, for the most powerful algorithm one is tempted to say that for *any* sequence it should perform better than 50% error. However, this is not true [16]. To see this just generate a sequence $S_1, S_2, S_3, \ldots$ using the following algorithm:

> Define S_{t+i} to be the opposite of the prediction of this algorithm which has been trained on $S_1, \ldots, S_t$.

Now, if the same algorithm is trained on this sequence, it will always predict the following bit with 100% error. Hence there is no general prediction machine; to be successful the algorithm needs some pre-knowledge about the class of problems it is applied to.

The Boolean perceptron is a very simple prediction algorithm for a bit sequence, in particular with the Hebbian on–line training algorithm (9.4). What does the bit sequence look like for which the perceptron fails completely?

Following (9.4), we just have to take the negative value

$$S_t = -\text{sign}\left(\sum_{j=1}^{N} w_j S_{t-j}\right) \tag{9.11}$$

and then train the network on this new bit:

$$\Delta w_j = +\frac{1}{N} S_t \; S_{t-j}. \tag{9.12}$$

The perceptron is trained on the opposite (= negative) of its own prediction. Starting from (say) random initial states $S_1, \ldots, S_N$ and weights $\underline{w}$, this procedure generates a sequence of bits $S_1, S_2, \ldots S_t, \ldots$ and of vectors $\underline{w}, \underline{w}(1), \underline{w}(2), \ldots \underline{w}(t), \ldots$ as well. Given this sequence and the same initial state, the perceptron which is trained on it yields a prediction error of 100%.

It turns out that this simple algorithm produces a rather complex bit sequence which comes close to a random one [17]. After a transient time the weight vector $\underline{w}(t)$ seems to perform

a kind of random walk on an N–dimensional hyper-sphere. The bit sequence runs to a cycle whose average length L scales exponentially with N,

$$L \simeq 2.2^N. \tag{9.13}$$

The autocorrelation function of the sequence shows complex properties: It is close to zero up to N, oscillates between N and $3N$ and it is similar to random noise for larger distances. Its entropy is smaller than the one of a random sequence since the frequency of some patterns is suppressed. Of course, it is not random since the prediction error is 100% instead of 50% for a random bit sequence.

When a second perceptron (=student) with different initial state $\underline{w}^S$ is trained on such a anti-predictable sequence generated by Eq.(9.11) it can perform somewhat better than the teacher: The prediction error goes down to about 78% but it is still larger than 50% for random guessing. Related to this, the student obtains knowledge about the teacher: The angle between the two weight vectors relaxes to about 45 degrees [16, 17]. Hence the complex anti-predictable sequence still contains enough information for the student to follow the time dependent teacher.

9.6 Agents competing in a closed market

We just considered a network interacting with itself. Now we extend this model to a system of many networks interacting with the minority decision of all members. This work was motivated by the following problem of econophysics [18].

Recently a mathematical model of economy receives a lot of attention in the community of statistical physics. It is a simple model of a closed market: There are K agents who have to make a binary decision $\sigma(t) \in \{+1, -1\}$ at each time step. All of the agents who belong to the minority gain one point, the majority has to pay one point (to a cashier which always wins). The global loss is given by

$$G = \left| \sum_{t=1}^{K} \sigma(t) \right| \tag{9.14}$$

If the agents come to an agreement before they make a new decision, it is easy to minimise G : $(K-1)/2$ agents have to choose $+1$, then $G = 1$. However, this is not the rule of the game; the agents are not allowed to make contracts, and communicate only through the global sum of decisions. Each agent knows only the history of the minority decision, $S_1, S_2, S_3, \ldots$, but otherwise he/she has no information. Can the agent find an algorithm to maximise his/her profit?

If each agent makes a random decision, then $\langle G^2 \rangle = K$. It is possible, but not trivial, to find algorithms which perform better than random [19].

Here we use a perceptron for each agent to make a decision based on the past N steps $\underline{S} = (S_{t-N}, ..., S_{t-1})$ of the minority decision. The decision of agent $\underline{w}(t)$ is given by

$$\sigma(t) = \text{sign}(\underline{w}(t)\ \underline{S}). \tag{9.15}$$

After the bit S_t of the minority has been determined, each perceptron is trained on this new example $(\underline{S}, S_t)$,

$$\Delta \underline{w}(t) = \frac{\eta}{N} \; S_t \; \underline{S}. \tag{9.16}$$

This problem could be solved analytically [20]. The average global loss for $\eta \to 0$ is given by

$$\langle G^2 \rangle = (1 - 2/\pi) K \simeq 0.363 \; K. \tag{9.17}$$

Hence, for small enough learning rates the system of interacting neural networks performs better than random decisions. Successful cooperation emerges in a pool of adaptive perceptrons.

9.7 Synchronisation by mutual learning

Before, we have considered a pool of several neural networks interacting through their minority decisions. Now we study the interaction of just two neural networks [20]. Contrary to the teacher/student case, now both of the networks are adaptive, each network is learning the output bit of its partner.

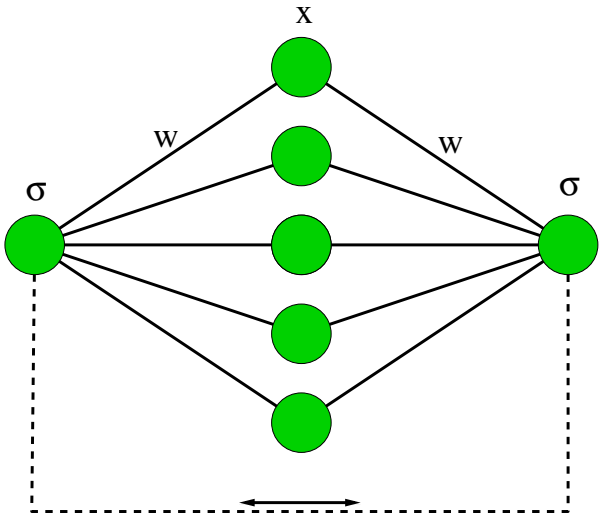


Figure 9.6: Two perceptrons are trained by their mutual output bits, from Ref. [20]

In particular, we consider the model where two perceptrons A and B receive a common random input vector $\underline{x}$ and change their weights $\underline{w}$ according to their mutual bit σ, as sketched in Fig. 9.6. The output bit σ of a single perceptron is given by the equation

$$\sigma = \text{sign}(\underline{w} \cdot \underline{x}) \tag{9.18}$$

$\underline{x}$ is an N-dimensional input vector with components which are drawn from a Gaussian with mean 0 and variance 1. $\underline{w}$ is a N-dimensional weight vector with continuous components which are normalised,

$$\underline{w} \cdot \underline{w} = 1 \tag{9.19}$$

The initial state is a random choice of the components $w_i^{A/B}, i = 1, ..., N$ for the two weight vectors $\underline{w}^A$ and $\underline{w}^B$. At each training step a common random input vector is presented to the two networks which generate two output bits σ^A and σ^B according to (9.18). Now the weight vectors are updated by the Rosenblatt learning rule (9.4):

$$\begin{aligned} \underline{w}^A(t+1) &= \underline{w}^A(t) + \frac{\eta}{N}\underline{x}\, \sigma^B\, \Theta(-\sigma^A\sigma^B) \\ \underline{w}^B(t+1) &= \underline{w}^B(t) + \frac{\eta}{N}\underline{x}\, \sigma^A\, \Theta(-\sigma^A\sigma^B) \end{aligned} \tag{9.20}$$

$\Theta(x)$ is the step function. Hence, only if the two perceptrons disagree a training step is performed with a learning rate η. After each step (9.20), the two weight vectors have to be normalised.

In the limit $N \to \infty$, the overlap

$$R(t) = \underline{w}^A(t)\, \underline{w}^B(t) \tag{9.21}$$

has been calculated analytically [20]. The number of training steps t is scaled as $\alpha = t/N$, and $R(\alpha)$ follows the equation

$$\frac{dR}{d\alpha} = (R+1)\left(\sqrt{\frac{2}{\pi}}\,\eta(1-R) - \eta^2\frac{\phi}{\pi}\right) \tag{9.22}$$

where ϕ is the angle between the two weight vectors $\underline{w}^A$ and $\underline{w}^B$, i.e. $R = \cos\phi$. This equation has fixed points $R = 1, R = -1$, and

$$\frac{\eta}{\sqrt{2\pi}} = \frac{1-\cos\phi}{\phi} \tag{9.23}$$

Fig.9.7 shows the attractive fixed point of (9.22) as a function of the learning rate η. For small values of η the two networks relax to a state of a mutual agreement, $R \to 1$ for $\eta \to 0$. With increasing learning rate η the angle between the two weight vectors increases up to the value $\phi = 133°$ for

$$\eta \to \eta_c \cong 1.816 \tag{9.24}$$

Above the critical rate η_c the networks relax to a state of complete disagreement, $\phi = 180°, R = -1$. The two weight vectors are antiparallel to each other, $\underline{w}^A = -\underline{w}^B$.

As a consequence, the analytic solution shows, well supported by numerical simulations for $N = 100$, that two neural networks can synchronise to each other by mutual learning. Both of the networks are trained to the examples generated by their partner and finally obtain an antiparallel alignment. Even after synchronisation the networks keep moving, the motion is a kind of random walk on an N-dimensional hypersphere producing a rather complex bit sequence of output bits $\sigma^A = -\sigma^B$ [17]. In fact, after synchronisation the system is identical to the single network learning its opposite output bit discussed in section 5.

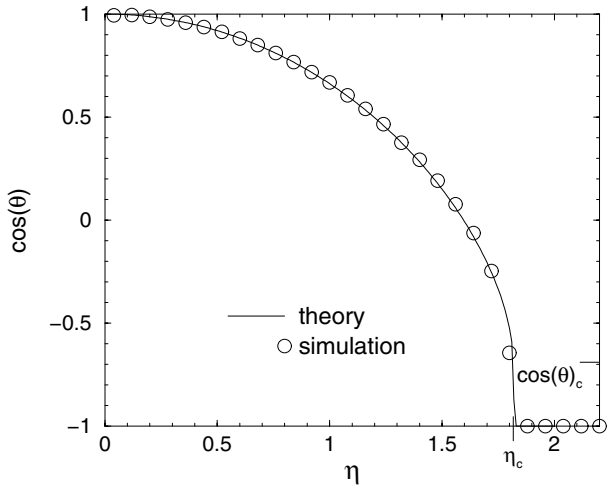


Figure 9.7: Final overlap R between two perceptrons as a function of learning rate η. Above a critical rate η_c the time dependent networks are synchronised. From Ref. [20]

9.8 Cryptography

In the field of cryptography, one is interested in methods to transmit secret messages between two partners A and B. An opponent E who is able to listen to the communication should not be able to recover the secret message.

Before 1976, all cryptographic methods had to rely on secret keys for encryption which were transmitted between A and B over a secret channel not accessible to any opponent. Such a common secret key can be used, for example, as a seed for a random bit generator by which the bit sequence of the message is added (modulo 2).

In 1976, however, Diffie and Hellmann found that a common secret key could be created over a public channel accessible to any opponent. This method is based on number theory: Given limited computer power, it is not possible to calculate the discrete logarithm of sufficiently large numbers [21].

Recently, it has been shown how interacting neural networks can produce a common secret key by exchanging bits over a public channel and by learning from each other [22].

We want to apply synchronisation of neural networks to cryptography. In the previous section we have seen that the weight vectors of two perceptrons learning from each other can synchronise. The new idea is to use the common weights $\underline{w}^A = -\underline{w}^B$ as a key for encryption. But two problems have to be solved yet: (i) Can an external observer, recording the exchange of bits, calculate the final $\underline{w}^A(t)$, (ii) does this phenomenon exist for discrete weights? Point (i) is essential for cryptography, it will be discussed further below. Point (ii) is important for practical solutions since communication is usually based on bit sequences. It will be investigated in the following.

Synchronisation occurs for normalised weights, unnormalised ones do not synchronise [20]. Therefore, for discrete weights, we introduce a restriction in the space of possible vectors

and limit the components $w_i^{A/B}$ to $2L+1$ different values,

$$w_i^{A/B} \in \{-L, -L+1, ..., L-1, L\} \tag{9.25}$$

In order to obtain synchronisation to a parallel – instead of an antiparallel – state $\underline{w}^A = \underline{w}^B$, we modify the learning rule (9.20) to:

$$\begin{aligned} \underline{w}^A(t+1) &= \underline{w}^A(t) - \underline{x}\sigma^A\Theta(\sigma^A\sigma^B) \\ \underline{w}^B(t+1) &= \underline{w}^B(t) - \underline{x}\sigma^B\Theta(\sigma^A\sigma^B) \end{aligned} \tag{9.26}$$

Now the components of the random input vector $\underline{x}$ are binary $x_i \in \{+1, -1\}$. If the two networks produce an identical output bit $\sigma^A = \sigma^B$, then their weights move one step in the direction of $-x_i\sigma^A$. But the weights should remain in the interval (9.25), therefore if any component moves out of this interval, it is set back to the boundary $w_i = \pm L$.

Each component of the weight vectors performs a kind of random walk with reflecting boundary. Two corresponding components w_i^A and w_i^B receive the same random number ± 1. After each hit at the boundary the distance $|w_i^A - w_i^B|$ is reduced until it has reached zero. For two perceptrons with a N-dimensional weight space we have two ensembles of N random walks on the internal $\{-L, ..., L\}$. If we neglect the global signal $\sigma^A = \sigma^B$ as well as the bias σ^A, we expect that after some characteristic time scale $\tau = \mathcal{O}(L^2)$ the probability of two random walks being in different states decreases as

$$P(t) \sim P(0)e^{-t/\tau} \tag{9.27}$$

Hence the total synchronisation time should be given by $N \cdot P(t) \simeq 1$ which gives

$$t_{\text{sync}} \sim \tau \ln N \tag{9.28}$$

In fact, the simulations for $N = 100$ show that two perceptrons with $L = 3$ synchronise in about 100 time steps and the synchronisation time increases logarithmically with N. However, the simulations also showed that an opponent, recording the sequence of $(\sigma^A, \sigma^B, \underline{x})_t$ is able to synchronise, too. Therefore, a single perceptron does not allow a generation of a secret key.

Obviously, a single perceptron transmits too much information. An opponent, who knows the set of input/output pairs, can derive the weights of the two partners after synchronisation. Therefore, one has to hide so much information, that the opponent cannot calculate the weights, but on the other side one has to transmit enough information that the two partners can synchronise.

In fact, it was shown that multilayer networks with hidden units may be candidates for such a task [22]. More precisely, we consider parity machines with three hidden units as shown in Fig.9.8. Each hidden unit is a perceptron (9.1) with discrete weights (9.25). The output bit τ of the total network is the product of the three bits of the hidden units

$$\begin{aligned} \tau^A &= \sigma_1^A\ \sigma_2^A\ \sigma_3^A \\ \tau^B &= \sigma_1^B\ \sigma_2^B\ \sigma_3^B \end{aligned} \tag{9.29}$$

At each training step the two machines A and B receive identical input vectors $\underline{x}_1, \underline{x}_2, \underline{x}_3$. The training algorithm is the following: Only if the two output bits are identical, $\tau^A = \tau^B$, the

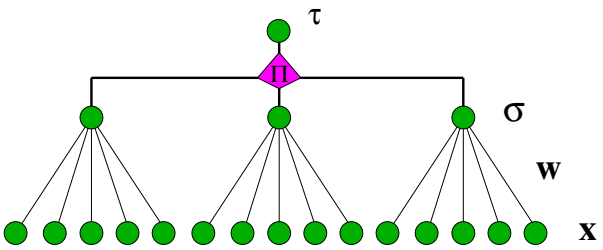


Figure 9.8: Parity machine with three hidden units.

weights can be changed. In this case, only the hidden unit σ_i which is identical to τ changes its weights using the Hebbian rule

$$\underline{w}_i^A(t+1) = \underline{w}_i^A(t) - \underline{x}_i \tau^A \tag{9.30}$$

For example, if $\tau^A = \tau^B = 1$ there are four possible configurations of the hidden units in each network:

$$(+1,+1,+1), (+1,-1,-1), (-1,+1,-1), (-1,-1,+1)$$

In the first case, all three weight vectors $\underline{w}_i, \underline{w}_2, \underline{w}_3$ are changed, in all other three cases only one weight vector is changed. The partner as well as any opponent does not know which one of the weight vectors is updated.

The partners A and B react to their mutual stop and move signals τ^A and τ^B, whereas an opponent can only receive these signals but not influence the partners with its own output bit. This is the essential mechanism which allows synchronisation but prohibits learning. Numerical [22] as well as analytical [23] calculations of the dynamic process show that the partners can synchronise in a short time whereas an opponent needs a much longer time to lock into the partners.

This observation holds for an observer who uses the same algorithm (9.30) as the two partners A and B. Note that the observer knows 1. the algorithm of A and B, 2. the input vectors $\underline{x}_1, \underline{x}_2, \underline{x}_3$ at each time step and 3. the output bits τ^A and τ^B at each time step. Nevertheless, it does not succeed in synchronising with A and B within the communication period.

Since for each run the two partners draw random initial weights and since the input vectors are random, one obtains a distribution of synchronisation times as shown in Fig. 9.9 for $N = 100$ and $L = 3$. The mean value of this distribution is shown as a function of system size N in Fig. 9.10. Even an infinitely large network needs only a finite number of exchanged bits - about 400 in this case - to synchronise.

If the communication continues after synchronisation, an opponent has a chance to lock into the moving weights of A and B. Fig.9.11 shows the distribution of the ratio between the synchronisation time of A and B and the learning time of the opponent. In the simulations for $N = 100$, this ratio never exceeded the value $r = 0.1$, and the average learning time is about 50000 time steps, much larger than the synchronisation time. Hence, the two partners can take their weights $\underline{w}_i^A(t) = \underline{w}_i^B(t)$ at a time step t where synchronisation most probably occurred as a common secret key. Synchronisation of neural networks can be used as a key exchange protocol over a public channel.

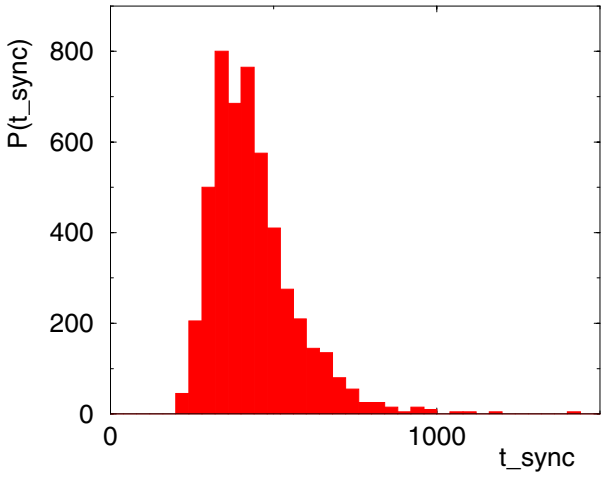


Figure 9.9: Distribution of synchronisation time for $N = 100$, $L = 3$, from Ref. [22]

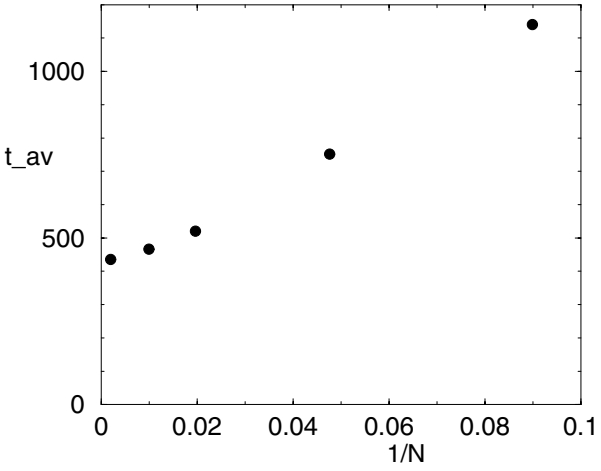


Figure 9.10: Average synchronisation time as a function of inverse system size.

Up to now it is not clear, yet, whether more advanced attacks will finally break this exchange protocol. On the other side, there are several possible extensions of the synchronisation mechanism where tracking seems to be even harder [24].

9.9 Conclusions

The dynamics of interacting neural networks has been studied in the context of a simple model: the perceptron and its extensions. The dynamics of these models can be calculated

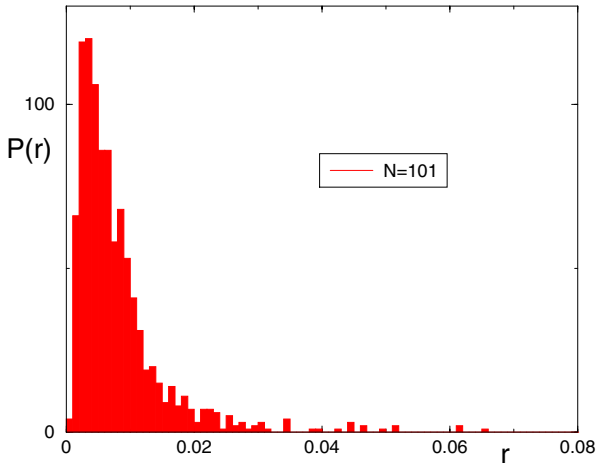


Figure 9.11: Distribution of the ratio of synchronisation time between networks A and B to the learning time of an attacker E.

analytically; macroscopic properties can be described by differential equations for a few order parameters.

Several kinds of interaction processes have been studied. In all cases the networks are trained by a set of examples which are generated by different networks. This is the way networks interact. Some networks are generating pairs of high dimensional input data and a corresponding output signal and transmit this information to other networks. The networks receiving this information are adapting their parameters – their synaptic weights – to each example. One question is to what extend the networks exchanging information are approaching each other in the high dimensional synaptic space.

The teacher/student scenario of a static networks generating the examples in addition to an adaptive network being trained on these examples is the case studied most. The question is: How well does the student learn the rule which is producing the examples, how can it generalise? The analytic solutions for the perceptron show the number of training examples has to be of the number of neurons to achieve generalisation. For the optimal training algorithm the generalisation error decays not faster than the inverse power of training time.

If both teacher and student networks are more complex then new phenomena are observed. For example, for a committee machine specialisation occurs: For short times the student relaxes fast to a configuration for which it has achieved generalisation but it is still in a symmetric state, it behaves like a simple perceptron. Only if the number of examples is increased to a very large value the student network can escape from this configuration and each member of the committee specialises to its corresponding member in the teacher network.

A static teacher network can also generate a time series on which the student network is trained. In addition to the overlap between teacher and student parameters, one is interested in the prediction abilities of the student after the interaction period. It turns out that learning and prediction are not necessarily correlated. One finds perceptrons which learn a chaotic

sequence very well but cannot predict it. On the other side, for quasiperiodic sequences it is difficult to learn the weights of the teacher but it is simple to predict the corresponding sequence.

A very general question about properties of prediction algorithms leads to a perceptron interacting with itself. It produces a rather complex time series which yields 100 % prediction error if the same perceptron is trained on this sequence. But even if a different perceptron is trained on this sequence, it achieves some knowledge about the teacher, hence its prediction error is larger than 50 %.

A community of neural networks can exchange information and learn from each other. It is shown how such a scenario can lead to successful cooperation in the minority game – a model for competing agents in a closed market.

If two networks are exchanging information and learning from each other, they can synchronise. That means, after some training time they relax to a configuration with identical time dependent synaptic weights (up to a common sign). The two networks keep diffusing on a high dimensional hypersphere, but with identical weights. Neural networks can synchronise by mutual learning.

This new phenomenon is applied to cryptography. It is shown how multilayer networks with discrete weights synchronise after a few hundred steps of interactions. However, a third network which is recording the exchange of examples does not synchronise, at least during a short period of time. Learning by mutual learning is fast, but learning by listening is very slow. Hence two partners can agree on a common secret key over a public channel. Any observer who knows all the details of the algorithm and who knows all the training examples cannot calculate the secret key. This is – to my knowledge – the first public key exchange which is not based on number theory. Future research will show how well this new algorithm – based on mutual learning of neural networks – can resist any advanced attacks which still have to be invented.

Synchronisation of neural networks is an active subject of research in neurobiology. Up to now it is – to my knowledge – unclear how synchronisation develops and what is its function. Here the model calculations point to a new direction: Two or several biological networks can achieve a common time dependent state by learning information exchanged between active partners. Any other network receiving the same information without being able to influence the partners cannot lock into this time dependent common state. Hence even in a fully connected network, parts of it can synchronise by mutual learning, at the same time screening their synchronised state from parts which learn only by listening.

In summary, this is the first attempt to develop a theory of interacting neural networks. Several phenomena were discovered from simple models like the perceptron or the parity machines. These phenomena were neither included into the models from the beginning nor are obvious; they are a result of cooperative behaviour of the synaptic weights and can only be understood from the analytical and numerical calculations.

As the different scenarios described in this overview show, the first results of this theory of interacting adaptive systems may be relevant in the fields of cooperative systems, nonlinear dynamics, time series prediction, economic models, biological networks and cryptography.

Acknowledgement This overview is based on enjoyable collaborations with Ido Kanter, Richard Metzler and Michal Rosen-Zvi. I thank Michael Biehl for suggestions on the manuscript. This work has been supported by the German Israel Science Foundation (GIF), the Minerva Center of the Bar Ilan University and the Max-Planck Institute für Physik komplexer Systeme in Dresden.

References

[1] Hertz, J., Krogh, A., and Palmer, R.G.: *Introduction to the Theory of Neural Computation*, (Addison Wesley, Redwood City, 1991)

[2] Engel, A., and Van den Broeck, C.: *Statistical Mechanics of Learning*, Cambridge University Press, 2001)

[3] Biehl, M., and Caticha, N.: Statistical Mechanics of On-line Learning and Generalisation, *The Handbook of Brain Theory and Neural Networks*, ed. by M. A. Arbib (MIT Press, Berlin 2001)

[4] Biehl, M. and Riegler, P.: *On-line learning with a perceptron* Europhys. Lett. **28**, 525 (1994)

[5] Kinouchi, O. and Caticha, N. *Optimal generalisation in perceptrons*, J. Phys. **A 25**, 6243 (1992)

[6] Biehl, M., Riegler, P. and Stechert, M.: *Learning from noisy data: An exactly solvable model*, Phys. Rev. **E 52**, 4624 (1995)

[7] Biehl, M. Riegler, P. and Wöhler, C.: *Transient dynamics of on-line learning in two-layered networks* J. Phys **A 29**,4769 (1996); Saad, D. and Solaa, S.A., *On-line learning in soft committee machines*, Phys. Rev. **E 52**, 4225 (1995)

[8] Weigand, A. and Gershenfeld, N.S.: *Time Series Prediction*, Addison Wesley, Santa Fe (1994)

[9] Eisenstein, E., Kanter, I., Kessler, D.A., and Kinzel, W.: *Generation and Prediction of Time Series by a Neural Network*, Phys. Rev. Letters **74** 1, 6-9 (1995)

[10] Freking, A., Kinzel, W., and Kanter, I., Learning and predicting time series by neural networks, Phys. Rev. E **65**, 050953 (2002).

[11] Kanter, I., Kessler, D.A., Priel, A., and Eisenstein, E.: *Analytical Study of Time Series Generation by Feed-Forward Networks*, Phys. Rev. Lett. **75** 13, 2614-2617 (1995)

[12] Schröder, M. and Kinzel, W.: *Limit cycles of a perceptron*, J. Phys. A **31**, 9131-9147 (1998)

[13] Ein-Dor, L., and Kanter, I.: *Time Series Generation by Multi-layer networks*, Phys. Rev. E **57**, 6564 (1998)

[14] Priel, A., and Kanter, I.: *Robust chaos generation by a perceptron*, Europhys. Lett. **51**, 244-250 (2000)

[15] C. M. Bishop: *Neural Networks for Pattern Recognition* (Oxford University Press, New York 1995)

[16] Zhu, H., and Kinzel, W.: *Anti-Predictable Sequences: Harder to Predict Than A Random Sequence*, Neural Computation **10**, 2219-2230 (1998)

[17] Metzler, R., Kinzel, W., Ein-Dor, L., and Kanter, I.: *Generation of unpredictable time series by a neural network*, Phys. Rev. E **63**, 056126 (2001).

[18] Econophysics homepage: http://www.unifr.ch/econophysics/

[19] Challet, D., Marsili, M., and Zecchina, R.: *Statistical Mechanics of Systems with Heterogeneous Agents: Minority Games*, Phys. Rev. Lett. **84** 8, 1824-1827 (2000)

[20] Metzler, R., Kinzel, W., and Kanter, I.: *Interacting Neural Networks*, Phys. Rev. E **62**, 2555 (2000)

[21] D. R. Stinson, *Cryptography: Theory and Practice* (CRC Press 1995)

[22] I. Kanter, W. Kinzel and E. Kanter, *Secure exchange of information by synchronisation of neural networks* Europhys. Lett. **57**, 141-147 (2002)

[23] M. Rosen-Zvi, I. Kanter and W. Kinzel, cond-mat/0202350 (2002)

[24] Kanter, I. and Kinzel, W.: unpublished

10 Modelling food webs

Barbara Drossel and Alan J. McKane

Abstract

We review theoretical approaches to the understanding of food webs. After an overview of the available food web data, we discuss three different classes of models. The first class comprise static models, which assign links between species according to some simple rule. The second class are dynamical models, which include the population dynamics of several interacting species. We focus on the question of the stability of such webs. The third class are species assembly models and evolutionary models, which build webs starting from a few species by adding new species through a process of "invasion" (assembly models) or "speciation" (evolutionary models). Evolutionary models are found to be capable of building large stable webs.

10.1 Introduction

Ecological systems are extremely complex networks, consisting of many biological species that interact in many different ways, such as mutualism, competition, parasitism and predator-prey relationships. They have been built up over long, evolutionary time scales, and in some cases will contain extremely ancient structures which hold information on the nature of the evolutionary changes which occurred in the distant past. Understanding and modelling such complex networks is one of the major challenges in present-day natural sciences.

Much research focuses on only a small number of species and their interactions, such as hosts and their parasites, or the relationship of a particular species with its prey or predators. Another important direction of research consists in studying larger networks of species by concentrating on their feeding relationships and on competition between predators, neglecting other types of interaction. Such networks, called food webs, are the subject of this review article. We will only be concerned with community food webs, which describe these interactions between species in a particular habitat, and will not discuss sink webs (species identified when tracing interactions down from a particular chosen species) or source webs (species identified when tracing interactions up from a particular chosen species). From now on community food webs will simply be referred to as food webs, or webs.

The early studies of the natural history of a given habitat were descriptive. It was not until the third quarter of the nineteenth century that the idea of listing the basic information on "what eats what" in a particular habitat, and presenting it in the form of a matrix was born. The usual format has the rows representing predators and the columns representing prey. The

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

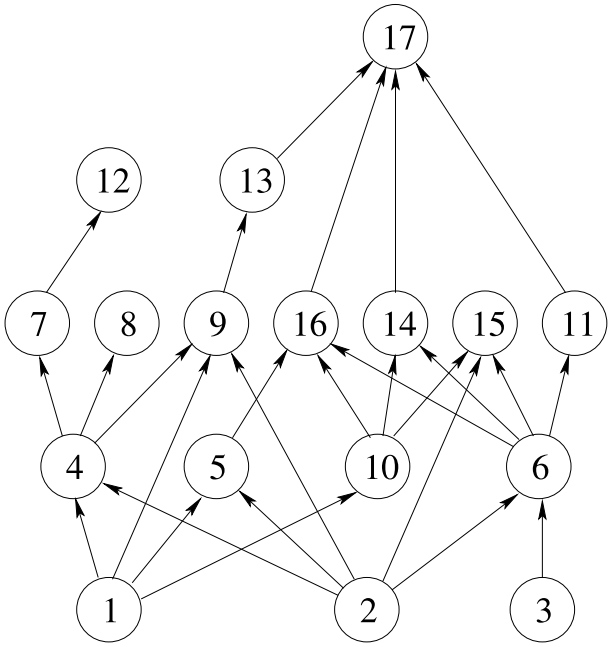


Figure 10.1: Narragansett Bay food web. 1=flagellates, diatoms; 2=particulate detritus; 3=macroalgae, eelgrass; 4=*Acartia*, other copepods; 5=sponges, clams; 6=benthic macrofauna; 7=ctenophores; 8=meroplankton, fish larvae; 9=pacific menhaden; 10=bivalves; 11=crabs, lobsters; 12=butterfish; 13=striped bass, bluefish, mackerel; 14=demersal species; 15=starfish; 16=flounder; 17=man. (After Yodzis 1989. *Introduction to theoretical ecology*, Harper and Row, New York.)

matrix elements might be numbers which specify the amount of food consumed, or, since such detail is very rarely known, simply 0's and 1's specifying the presence or absence of a predator-prey link. The diagrammatic representation of food webs was introduced some time later (see Fig. 10.1 for an example). They consist of vertices representing species in the food web, with a directed link — that is a line with an arrow attached — from vertex A to vertex B, if species A is eaten by species B. Notice that the direction of the arrows signifies the flow of resources. These networks illustrate the bare bones of the predator-prey relationships between the species: they miss much of the fine detail such as temporal variations in diet (daily or seasonal), but allow the longer term picture of predator-prey relationships to emerge. Community food webs cannot include all species in a habitat (such as all the bacteria living within plants and animals), but rather focus on a set of different types of species, which are chosen prior to analyzing their predator-prey relationships. This reduction of the rich, distinctive and complex nature of individual ecosystems tends to be less enthusiastically embraced by field ecologists, than by theoretical ecologists. For many of those in the field, the attraction of the study of natural communities is in their details and unique features. In fact, even when food webs were published by early investigators, they frequently seemed designed to illustrate their complexity, rather than to encourage the theoretical understanding of their

form. It was probably this idea of the distinctiveness and natural complexity of food webs that ensured that few theoretical investigations were carried out during the century after the first food webs were constructed.

The change in emphasis in the study of food webs dates from the late 1970's and early 1980's, when many of the published webs were collected together and various regularities were noticed for the first time. During roughly the same period simple models of food webs were formulated. The first models were dynamic, using simple population dynamics, but without incorporating the evolved nature of the web. Static models were also introduced in which species were simply represented as vertices in a graph and directed links between them were drawn according to some rule. The structure of the resulting webs could then be compared to that of real food webs.

Obtaining data on dynamic properties is much more difficult than the already hard task of collecting information on static webs. In any case, many of the time scales of interest to us will be so long as to be inaccessible by direct methods. One aspect which is in some sense intermediate between the static and dynamic descriptions, and which has attracted a large amount of attention, is the question of the stability of food webs. This is frequently couched in terms of the complexity of the food web (here meaning a larger web or one with greater connectance). In other words: does stability increase with web complexity?

To ecologists working in the 1950's and 1960's the answer was clearly "yes". They pointed to the susceptibility of certain cultivated, or other species poor, communities to large scale invasions by pests, and the relative rarity of such outbreaks in naturally rich ecosystems as evidence [1, 2]. Theoretically it was noted that increases in species number or connectance led rapidly to increases in the available number of food chains or pathways in food webs. It then seemed quite compelling to argue that the disruption or elimination of only a few of these pathways, if they were numerous, was not likely to lead to a complete collapse of the web [3].

This consensus was disrupted in the early 1970's by investigations into the linear stability of model ecosystems having random interactions [4–6]. They showed that these model ecosystems became less stable as species number or connectivity increased. There are two obvious objections to these conclusions. Firstly, real food webs are not random; they are highly evolved structures. Secondly, the criterion of linear stability analysis as applied to population dynamics equations may not be particularly relevant to real ecosystems, not least because they may not be sufficiently close to equilibrium for a local condition such as this to apply.

Over the next few years both objections relating to the validity of the randomness assumption and to the use of linear stability analysis for real webs, were addressed. Dynamical models with Lotka-Volterra type or more complicated population dynamics were introduced, and the stability and dynamical properties of small networks consisting of a few species were investigated, demonstrating a variety of instances where more complex model systems were not less stable than simpler ones. Furthermore, models were introduced that built up a community by a sequence of invasions or speciations. The most recent of these models lead to the formation of large stable webs, thus demonstrating explicitly that large complex networks can be stable.

The different types of models mentioned above will be reviewed in the following sections, with an emphasis on more recent work. The earlier phase of theoretical work has been well documented in a number of books and articles [7–9]. Before concentrating on theoretical developments, however, we will begin (in section 10.2) by introducing the basic concepts

used to describe food webs by reference to real webs, and also briefly allude to the problems in collecting data to construct real webs. In section 10.3, we will briefly discuss static food web models. Section 10.4 describes the various types of dynamical models and their contribution to the complexity-stability debate. In section 10.5, we give an overview of models that build food webs by a sequence of species invasions or speciations. We include simple toy models as well as species assembly models and more sophisticated evolutionary models. Section 10.6 concludes the article with a brief overview and a look to the future.

10.2 Basic properties of food webs

In this section we will characterize food webs by introducing the basic features associated with them, quantifying them as much as possible. This will allow comparison between real webs and model webs. As the concepts are defined, we will also briefly discuss additional aspects such as various assumptions made in the definitions, problems or ambiguities associated with the definitions, typical values found in real webs or difficulties in obtaining accurate measurements in the field.

The most primitive concept is the size of the food web, defined by the number of species in the web, S. In published versions of real webs (see, for instance, [9]), the terms "species" may refer to "trophic species", which is a collective term for all the species having a common set of predators and prey. For this reason general terms such as "ants" or "algae" appear; conversely the same species at different stages in its life-cycle may belong to a different trophic species. The existence of terms such as "detritus" or "dead organic material" in many webs, is also an illustration of the difficulty in deciding what to include and what to omit from the web. Many earlier studies were not very extensive, and frequently the published webs were rather small. For example, the 113 webs listed by Cohen *et al* in their 1990 review [9] vary in size from 5 to 48, with a mean of 17. Since 1990 larger webs have been reported, with several containing more than 100 species [10–13].

The next most important quantity associated with a food web is a measure of the number of the interactions between the species. This is frequently taken to be the ratio L/S, the total number of predator-prey links, L, divided by the total number of species, S. This property, called the linkage density, seems a fairly natural choice, but was also favored by some because analysis of the pre-1990 webs suggested that L/S was independent of S and therefore that this ratio was "scale invariant". The value of the slope for the best fit to L versus S was found to be 1.99 ± 0.07, although there appeared to be a slight tendency for the more recently collected webs in that set to have a higher value [9]. The finding that the more recently collected webs tend to be larger, suggests the possibility that the linkage density is not in fact constant, but slightly increasing with S, and that earlier webs were too small to show this clearly. In fact, this was already noted at the time [9, 14, 15]; the scaling relation $L/S \sim S^{\epsilon}$ with ϵ equal to 0.3 or 0.4 was not ruled out, especially when the sample included larger webs.

An alternative measure of the number of interactions between species, is the connectance, C, of a food web defined as the total number of links in the web divided by the total number of possible links in a web of the same size. Since, excluding links from a species to itself (cannibalism), there are $S(S-1)/2$ pairs of species that can be connected by a link in a web of S species, $C = 2L/S(S-1)$. This quantity was originally introduced by theorists [4–6], since

it is equal to the probability of finding a non-zero entry in the community matrix. Clearly, for large S, the power-law scaling mentioned above gives $C \sim S^{-1+\epsilon}$, with the scale invariance hypothesis leading a hyperbolic S-C relation.

During the 1990's the scale invariance hypothesis became more and more untenable. The values of the linkage density for the larger post-1990 webs ranged from 3.5 to 11.0, compared with the value of 2.0 found for the pre-1990 webs [16]. The scaling form $L/S \sim S^{\epsilon}$, with a value of ϵ close to or equal to 0.3 or 0.4 suggested above was still found to be consistent with data [17], although it was also suggested [18, 19] that ϵ might be as large as 1, leading to the conclusion that the connectance was independent of the web size. There are several reasons why the analysis of more recently collected data might differ from the earlier results [20], but an obvious one involves questions of resolution: the earlier webs might be smaller in part because many species and/or links were omitted due to incomplete or biased recording. We will return to this point later in this section, since it is a criticism which may be applied to other measured quantities. Recent work has confirmed that the scale invariance hypothesis for the linkage density is not correct, but no consensus has emerged to replace it. It may simply be that food webs from diverse communities have different characteristics [21], or that while this may be the case, much of the disagreement between data from early webs and those collected more recently can be explained by the fact that the linkage density is very sensitive to sampling effort [22], or that the observed patterns can be explained, but not with simple conjectures as scale invariance or constant connectance [23].

In addition to quantities which describe the structure of the network, such as S and C, it would be useful to have some characterization of the type of species in the web. In order to be meaningful, this should not be so detailed that generic patterns are not seen, but not so coarse that it contains little information. The simplest and most widely used classification is to divide species in the web up into top, intermediate and basal species. Top species have no predators, basal species have no prey — they obtain all of their resources directly from the environment — and intermediate species have both predators and prey. It is now possible to classify all links in the web into four classes: links between top species and intermediate species, top species and basal species, intermediate species and basal species, and links between intermediate species and other intermediate species. This gives 7 quantities which contain information about the biological aspects of the food-web: the proportions of the species which are top, intermediate and basal (denoted by T, I and B) and the proportion of links between these three types (denoted by, for example, IB, for the proportion of links that go between intermediate and basal species). Only 5 of these are independent, because $T + I + B = 1$ and also the proportion of the links in all four classes add up to unity.

This seems to embody just the right amount of information to allow useful comparison of model webs with real ones. The classification scheme was given further credence when an analysis of the pre-1990 webs suggested that, like L/S, the proportions T, I and B were independent of web size, having values of 29%, 52% and 19% respectively [9]. With L/S and three categories all being independent of web size, it was perhaps not too surprising that the four types of links, TI, TB, IB and II, were also found to be the same in webs of different sizes. The reported values were 35%, 8%, 27% and 30% respectively [9], although, while it is true that the data showed no evidence of increasing or decreasing trends, the scatter of points looked so random that the conceptual jump to the scale-invariance conjecture seemed to be a large one.

As an alternative to the (T, I, B) classification scheme, the proportion of the species which are prey, H, and the proportions which are predators, P, may be used. They are related to the previous set by $H = I + B$ and $P = T + I$. Recall that only two of the set $\{T, I, B\}$ are independent and that $H + P = I + 1$ is greater than unity, since intermediate species are both predators and prey. A frequently quoted statistic for food-webs is the so-called predator-prey ratio (in fact, the prey-predator ratio) H/P. This seems to be the property which shows the least change between the earlier, smaller webs, where it had a value of 0.9, and the more recent, larger webs where it has a mean value only slightly larger than this [16]. There is, however, a considerable spread in actual values for different webs.

The review of food web patterns by Pimm *et. al.* in 1991 [15] was still able to hold on to the belief that many web properties were size-independent, though with high variances and with the possible exception of the linkage density. However, by the time of the next major review in 1993 [16], it was accepted that T, I, B and the links between them did, like L/S or C, vary with the size of the web. This change of view was mainly prompted by two studies [10, 24], which took large food webs and reduced their resolution by lumping more and more of the species together. It was found that many of the quantities discussed above were sensitive to this aggregation. Although the aggregation criteria which were employed were not identical, and thus some of the details of the findings differed, it was clear that generally speaking food-webs properties changed with S. In particular, Martinez [10] found that the highly resolved Little Rock Lake web had larger I, but smaller T and B than the aggregated version. The fact that the latter had similar properties to the pre-1990 webs, led him to speculate that these webs might too be aggregated versions of larger webs.

These studies were followed up by a re-analysis [25] of an earlier attempt at aggregation [26], which had shown little change with aggregation, and by a re-analysis [27] of the pre-1990 webs [9]. The conclusion of these studies was that aggregated webs and the earlier, smaller webs all have lower $L/S, I$ and II and higher T, B and TB than highly resolved or more recently collected, larger webs. The quantities TI and IB do not seem to change in a consistent way with S. These results can be understood to some degree by beginning with two observations. Firstly, it is now believed that species without any predator are very rare, and perhaps non-existent [10, 28], and thus T will be tiny in well resolved webs. Secondly, basal species are already rather coarsely specified, and it would be difficult to aggregate them further. Thus it might be expected that the proportion of basal species would increase as S decreased. If both T and B decrease as S increases, we would also expect TB to decrease, and I and II to increase. Presumably TI and IB do not change in a consistent direction because, unlike TB and II, they link types of species whose proportions change in opposite directions. For a similar reason, the predator-prey ratio $(I+B)/(T+I)$ seems to be extremely robust under aggregation [16].

These ideas can be extrapolated to the limit by considering a food web with only two species on the one hand, and the entire global ecosystem on the other [29]. While this might be of dubious validity, the trends displayed are suggestive. If the two species web consists of a predator and prey, then $T = 50\%, I = 0\%$ and $B = 50\%$. If we assume that the global ecosystem has no, or very few, top species, that animals are intermediate species and plants basal species, and that animals comprise 95% of the species, then $T = 0\%, I = 95\%$ and $B = 5\%$. These asymptotic values, taken together with the previous web results, show a consistent trend of T and B decreasing, and I increasing, with S, and the possibility that food

web properties become scale invariant for S larger than about 1000 [29]. Studies have also been carried out to investigate other types of sampling effects. For example, the threshold for the inclusion of links can be varied [10, 30] and species can also be omitted (as opposed to being aggregated) [31]. Once again, it was found that the poorly sampled versions of these webs were much more similar to the pre-1990 webs, than were the full versions. Conclusions such as these have convinced ecologists of the need to be more systematic and methodological in the collection of food webs [32–34].

What are the other food web attributes which field ecologists should be looking for? In their review, Cohen *et. al.* [9] listed five "laws" of food webs. Three of these dealt with scale-invariance (of the linkage density, of T, I, B, and of the links between these three types). The fourth was that "food chains are short". A chain in a food web is the set of links along a particular path starting from a basal species and ending at a top species. The number of links along this path is the length of that particular food chain. By averaging this over all the chains in a web, a mean chain length may be assigned to each web. For the food webs listed in [9], the mean chain length over all 113 webs is 2.88 and the median of the maximum chain length in each web is 4 links. The observation that food chains are short is not new; it is one of the earliest inherent tendencies noted in the study of food webs [35]. The classic explanation is that energy is transmitted very inefficiently up the chain and after dissipation at more than three or four vertices, is not sufficient to sustain predators at the top of the chain [36]. This should mean that productive ecosystems should have longer food chains, but the evidence for this is mixed [8]. Other hypotheses are discussed by Pimm [7]. As for the properties discussed earlier, the nature of food chains in the more recent, larger webs, differ from those appearing in Cohen *et. al.* [9], being typically much longer. This raises the possibility that mean chain lengths are a function of the size of webs. Certainly, food chain length decreases when webs are aggregated [16].

The fifth "law" was that, excluding cannibalism, cycles are rare [9]. Cycles are sets of links which end with the same species as they started from. Of the 113 webs, only 3 contained cycles, and in each case only one cycle of length 2. By contrast the large Little Rock Lake web [10] contained many cycles. If cycles are at all numerous, the definition of the length of food chains given above has to be modified, since it is clearly ambiguous. Two different algorithms were used to calculate food chain length in the case of the Little Rock Lake web. Even when cycle-forming species are excluded, reducing the mean chain length, the mean of 7 links is still much greater than the 2.88 links for the pre-1990 webs.

A term used frequently in ecology is the "level", or "trophic level" on which a species appears in the food web. This is clearly a useful descriptive term, and when it refers to a single food chain it is obviously unambiguous: it has a value which is one more than the chain length, that is, the number of linkages between it and the basal species in the web [15]. Equally obvious is the fact that it is not a uniquely defined quantity in a web — there will typically be several routes from the species under consideration to basal species. One definition in this case is to list all the possible routes and assign the most common (modal) as the trophic level [7]. Another definition, which seems to us somewhat superior, is to assign the shortest of the possible routes as the trophic level. This choice is based on energy considerations: given the inefficiency of energy transfer along a chain mentioned above, the most important links are likely to be the shortest. This latter definition also has the advantage of being unique; in the former case there may be more than one modal value. While the term "trophic level" is

used extensively in a qualitative way in studies of food webs, little is known quantitatively about the number or other attributes of species on different trophic levels, perhaps because of the absence of a single agreed definition. This is unfortunate, because a quantity such as the fraction of the species on a particular level, is relatively easy to obtain from the data and is another attribute which can be compared to models. It is also slightly less coarse than the T, I, B designations, and also probes the food web hierarchy in a slightly different manner; using the latter definition above, basal species are always on level 1, but intermediate species may also be on level 1, and top species need not be on the highest level.

One important property of food webs which rests on the definition of trophic levels is the degree of omnivory. An omnivorous species is one that feeds on more than one trophic level [7]. Thus, for instance, a species which feeds on its prey's prey is omnivorous. One of the earliest results was that omnivory was less common in some types of real webs, than in randomly generated webs [7]. About 27% of the species in the pre-1990 webs are omnivores, but the overall picture is quite confused, in part because of the different ways that degree of omnivory can be defined [16]. If the assignments of trophic levels has been agreed upon, the most straightforward index of omnivory is simply the proportion of species which are omnivores. Another measure was given by Goldwasser and Roughgarden [37]. They first determined the statistical distribution of the number of links along all pathways from a particular species to the basal species. The mean of this distribution, which they termed the trophic height, gave a generalized trophic level. The standard deviation, on the other hand, gave an indication of the extent to which the species ate on variety of different levels, and was used by them as a second index of onmivory.

In any case, omnivory seems to be less common towards the base of a community web, and therefore the degree to which sampling favors a particular group of predators will have a marked effect on the percentage of species which are omnivores [16]. Some webs have been reported to have a high degree of omnivory (*e.g.* 78% in [28]), so it is again tempting to list omnivory as another attribute which has been underestimated in the older web data, and in fact it has been found to be sensitive to sampling effort [31]. However, Warren [20] points out that connectance may be a key parameter on which other web characteristics depend, and thus increase in omnivory may not be independent of the increase in connectance. Incidently, this type of reasoning may be used to argue that more highly connected webs may have a higher proportion of intermediate species (a species is more likely to have links both to it and from it), more cycles, longer chain length and so on.

The description of food webs given so far in this section has focussed on static, structural properties of webs. In reality, food webs are dynamical systems, and links, population sizes, and species composition change with time. This brings additional difficulties into the quantitative description of web structure. Empirical data are collected over a certain time which may vary. If, for instance, a predator feeds on a certain prey only during harsh seasons when other food is scarce, the link to that prey is present only temporarily, and only when links to other prey species are absent. Large food webs, the data for which have been collected over a long time, may therefore overestimate the number of links that are present at a given moment in time.

There are, of course, a wealth of field observations of the dynamical behavior of food webs, but it has not yet been possible to formulate a quantitative, mathematical description that is generally valid across food webs. There are a variety of different population dynamics

equations containing different interaction terms, which will be discussed in section 10.4. The discussion about which mathematical form is more appropriate is lively and diverse.

This section has not been designed to be an exhaustive review of food webs, but rather a summary of the key ideas, concentrating on those that are the most relevant for the modelling of webs. The rest of the article will be devoted to a discussion of the various models of food webs that have been put forward.

10.3 Static models

This section describes models that build food webs by assigning links between species according to some rule, and then evaluate the properties of the resulting webs. Species are simply represented as points in space or on a line.

The first such models were modified versions of the random graphs introduced by Erdös and Rényi [38], where links are assigned to randomly chosen pairs of points. Cohen [9, 39] suggested several models for randomly generated webs where links have an orientation indicating which of the two species connected by the link is food for the other one. Links have no orientation in conventional random graphs. Many properties of such directed random graphs can be derived analytically, such as the fraction of top and basal species and the numbers of cycles. The agreement with data from real webs is not very good. This is not surprising, since this simple model has many unrealistic features, such as the assumption that every species can in principle be the predator of every other species.

A model that takes into account the fact that some species are higher up in the food chain than others, has become known under the name "cascade model" [9,40]. In this model, species are assigned numbers from 1 to S. Each species can prey only on species that have a lower number, and it preys on any of these species with a probability d/S. Here d, the density of links per species, is a constant which, along with S, is the only parameter of the model which has to be fitted to data. One can easily show that the expected number of links in such a food web is $d(S-1)/2$. Therefore, for not too small values of S, the cascade model predicts that the mean number of species should grow linearly with S: $L \sim dS/2$. As discussed in section 10.2, this is consistent with the pre-1990 webs collected in [9], and with the choice $d = 4$ the mean number of links per species agrees with the then accepted empirical value close to 2. Other properties, such as the fraction of top and basal species, can also be calculated, and they are not far from empirical data for the older collections of webs [9]. For example, the values of T, I and B for large S asymptote to 26%, 48% and 26% respectively when $d = 3.72$. The mean length of the longest chain increases only slowly with the number of species S, and it is around 4 for S between 10^3 and 10^5. However, the cascade model seems less good at predicting chain length statistics, than many of the other measures investigated [9].

In the light of all of the comments made in section 10.2 concerning the difference in trends between data collected in the last decade or so and older data, it is not surprising that the predictions of the cascade model have been found to be in disagreement with more recently collected data [18, 37]. Two of the simplest predictions of the cascade model, that food webs are acyclic and that $L \sim S$, are no longer tenable. In an attempt to generalize the cascade model to avoid these and other predictions which are not borne out, Cohen constructed 13 alternative versions of the model [41]. However, in all but one case these were inferior to

the original cascade model in predicting general web properties. For example, models which assumed that $L/S \sim S^{\epsilon}$ with $\epsilon = 0.35$ made inferior predictions to those models which took $\epsilon = 0$. More recent studies have also pointed to deficiencies in the cascade model, especially when the assumption of the random distribution of links is viewed in terms of aggregated webs [42].

Recently, another static model, called the niche model, was introduced by Williams and Martinez [43]. Just as in the cascade model, the species in this model are put in order: a "niche value" is assigned to them by randomly drawing a number from the interval $[0, 1]$. In contrast with the cascade model, the species are now constrained to consume all prey within a range of values whose randomly chosen center is less than the consumer's niche value. The size of the range is chosen according to a beta distribution with parameters such that the desired mean number of links per species results. In contrast to the cascade model, species with similar niche values often share consumers, and the strict cascade hierarchy is partially relaxed by allowing up to half of the consumer's range to include species with niche values higher than the consumer's value. As in the cascade model there are only two empirical parameters: the number of species and the linkage density (or the connectance). Evaluating 12 different structural properties of the webs generated by the niche model and comparing them to real food web data, the authors found that the agreement between the model web and real webs is in general much better for the niche model than for the cascade model, in particular with respect to features such as cycles and species similarities.

In spite of the apparent success at reproducing properties of real food webs for appropriately chosen parameter values, these static models cannot give a real explanation of the observed web structures. The webs constructed by these models do not result from a dynamical process; links are not assigned according to some biologically inspired rule, and the models do not contain any population dynamics. A good agreement with real data is achieved by capturing some structural features of real webs, but not by incorporating underlying biological properties. In particular, the question of web stability cannot be addressed in these simple models. The question of web stability will be discussed in the next section, where dynamical models are considered, and the question how the structure of webs might follow from evolutionary dynamics combined with biological principles, will be explored in section 10.5.

10.4 Dynamic models

The models in the last section attempt to describe food webs as static objects, which is after all what nearly all of the data collected is concerned with. However, it seems more rational to study the kinds of static structures which emerge from biologically reasonable dynamics, rather than attempt to characterize the currently observed webs in terms of simple properties of graphs. As stressed in the Introduction, more than one time-scale will be relevant in food web dynamics; the long time-scale evolutionary dynamics and the shorter time-scale population dynamics will be both important. Evolutionary dynamics will be discussed in the next section. In this section we will review the population dynamics of predator-prey interactions, with greater emphasis on multispecies communities than is traditional in this subject, and with a focus on the question of under which conditions multispecies communities can be stable.

We will start with a discussion of the two-species model, which is frequently as far as most textbooks go. Then, we will generalize the two-species dynamic equations to an arbitrary number of species. Finally, the stability of such coupled equations for small webs as function of the structural properties of the web, and the types of equations used, will be discussed in subsection 10.4.3.

10.4.1 Two-species models

In a two-species model a predator (or parasite) depends for subsistence on a single species of prey (or host) and cannot turn to an alternative food source. We denote the number of predators at time t by $P(t)$ and the number of prey by $H(t)$ (H can be thought of as an abbreviation for "hosts" or "herbivores"). In most cases, these numbers are understood as individuals per unit area, *i.e.*, the predator and prey densities. Almost all of the models which are formulated in terms of differential equations are a special case of what we will call the standard model [44–46]

$$\begin{aligned} \frac{dH}{dt} &= \phi(H) - g(H,P)\,P\,, \\ \frac{dP}{dt} &= n(H,P)\,P - d_P\,P\,. \end{aligned} \tag{10.1}$$

Here $\phi(H)$ is the growth of the prey in the absence of predators, $g(H,P)$ is the capture rate of prey per predator, $n(H,P)$ is the rate at which each predator converts captured prey into predator births and d_P is the (constant) rate at which predators die in the absence of prey. The function $n(H,P)$, called the numerical response, which describes how the numbers of new predators relate to the captured prey, is not usually very well known. Frequently it is assumed that a constant fraction of the captured prey are used as resource to produce new predators, that is, $n(H,P) = \lambda g(H,P)$, where λ is a constant called the ecological efficiency. Early models assumed that the growth rate of an individual prey in the absence of predators was constant, that is, $\phi(H) = rH$, but most models now include intra-species competition by taking $\phi(H)$ to have the logistic form $\phi(H) = r(1 - H/K)H$, where K is the carrying capacity. With these choices, the type of model is specified solely by the choice of the function $g(H,P)$, called the functional response.

The first model of predator-prey dynamics put forward having the form (10.1) was the Lotka-Volterra model which had exponential (not logistic) growth of the prey ($\phi(H) = rH$) and a linear functional response $g(H,P) = aH$, so that the capture rate for an individual predator increased linearly with the number of prey [45]. This model has the unrealistic feature of neutral stability: it contains a limit cycle with an amplitude which is determined by the initial conditions, rather than by the parameters of the model. Imposing a logistic form for $\phi(H)$ cures this, but only by eliminating limit cycles entirely [46]. During the 1960's the study of the standard model (10.1), with more realistic forms for the functional response began. Rosenzweig and MacArthur [44, 47, 48] developed a graphical method to determine what functional forms for g and ϕ gave rise to stable fixed points and limit cycles, although the analysis was restricted to functions g which only depended on H, and not on P. A broad conclusion was that the most complete range of behaviors were seen in (10.1) if $\phi(H)$ had the logistic form and if $g(H)$ saturated at some constant value for large H (the so-called Type II form).

A specific Type II functional form suggested by Holling [49] is widely used in modelling, partly because of its simplicity, but also because it can be derived in a reasonably convincing way [50]. The essential idea is that the period of searching, T, should be divided into true searching time, T_s, and a "handling time", T_h, which represents the time taken to eat the prey as well as the time taken afterwards to clean, rest and digest the food. Use of T_s, rather than T, in the definition of the functional response, and the assumption of random encounters between predators and prey, gives the Holling form

$$g(H) = \frac{aH}{1 + bH}\,, \tag{10.2}$$

where a and b are constants. Beddington [51] extended this idea by having a second type of "wasted time" in addition to the handling time, namely time wasted when two predators meet. Incorporating this into the definition gave a functional response which depended on the number of predators:

$$g(H, P) = \frac{aH}{1 + bH + cP}\,, \tag{10.3}$$

where c is another constant. Both forms (10.2) and (10.3) are widely accepted as reflecting essential features of predator-prey interactions. However, this acceptance is not universal, and the traditional arguments used to construct them have been criticized [52]. The basis of the criticism is that the function g appearing in the population dynamics equations should be the function calculated on the same time scale as that of the population dynamics, and not that calculated on the same time scale as the behavioral response. When viewed from the slow time scale, prey abundance is assumed to appear as a continuous function. However, when viewed from the fast behavioral time scale, prey production is no longer continuous but appears as successive "bursts". Between these bursts, the predators consume the prey (or the fraction of prey available to predation) by some mechanism (possibly random search). Thus, for a given number of prey, each predator's share is reduced if more predators are present. This suggests that the consumption rate should be a function of prey abundance *per capita*, that is,

$$g(H, P) = \Phi(H/P)\,, \tag{10.4}$$

a ratio-dependent functional response. The form of the function Φ can be deduced by looking at two extreme situations. When the prey is very abundant, predators feed at a constant maximum rate, so that $\Phi \to$ constant, for $H \gg P$. On the other hand, if predators are very abundant they will consume prey at a constant rate, so that $g(H, P)P = a'H$ in the limit $H/P \to 0$. A simple form which has this structure is

$$g(H, P) = \frac{a'(H/P)}{1 + b'(H/P)} = \frac{a'H}{P + b'H}\,. \tag{10.5}$$

Beddington's form and this specific form of the ratio-dependent functional response may be written as

$$g(H, P) = \frac{H}{\alpha + \beta H + \gamma P}\,, \tag{10.6}$$

where α, β and γ are constants. The only difference is that in the ratio-dependent case $\alpha = 0$. Despite these similarities, there has been a vigorous discussion in the literature as to the superiority of one form over the other [53–58]. The essential differences between the two methods of modelling the functional response are discussed in a recent review article written by authors from both camps [59].

10.4.2 Generalized dynamical equations

The generalization of the population dynamics equations (10.1), with realistic growth rates ϕ and functional responses g, to more than two species is straightforward for a food chain [60], or other simple webs, such as two chains with a mobile top predator [61], but less obvious for a general web. For this reason virtually all investigators, starting with May [5], who have studied the population dynamics for a general web have used Lotka-Volterra dynamics. However the well-known unsatisfactory features of these equations [6], together with a desire for greater realism, have resulted in some suggested versions of population dynamics which go beyond the Lotka-Volterra scheme [62–66].

Let us begin with the Lotka-Volterra equations. If N_i is the population size or population density of species i, the Lotka-Volterra equations for a general web may be written as

$$\frac{dN_i(t)}{dt} = N_i \left(b_i + \sum_j a_{ij} N_j \right) , \tag{10.7}$$

where b_i is a positive growth rate for basal species, and a negative death rate for the other species. The b_i and the interaction coefficients a_{ij} are constants, independent of the population sizes. There are three different possible contributions to $\sum_j a_{ij} N_j$: (i) As mentioned in the context of two–species models, many authors include a logistic term for the basal species, implying $a_{ii} < 0$ for basal species, and zero for the other species. (ii) Some authors using Lotka–Volterra models include competition between two predators that share the same prey, *i.e.*, $a_{ij} < 0$ whenever i and j have a prey in common. (iii) The most important contributions are the predator–prey terms. If i is a predator and j is one if its prey species, then $-a_{ji} N_j$ is the functional response g_{ij}, *i.e.*, the number of individuals of species j consumed per unit time by an individual of species i. Often, the identity $-a_{ji} = \lambda a_{ij}$ is used, but some Lotka–Volterra models have independent random (positive) numbers for a_{ij} and $-a_{ji}$, and some models do not even impose opposite signs for a_{ij} and a_{ji}.

We will restrict our discussion of general non-Lotka-Volterra type equations to those that satisfy the balance equations

$$\frac{dN_i(t)}{dt} = \lambda \sum_j N_i(t) g_{ij}(t) - \sum_j N_j(t) g_{ji}(t) - d_i N_i(t) . \tag{10.8}$$

These equations are in many ways the natural generalizations of (10.1), with the first term on the right-hand side representing the growth in numbers of species i due to predation on other species, the second term the decrease in numbers due to predation by other species, and the last term the constant rate of death of individuals of species i, in the absence of interactions with other species. Where there is no predator-prey relationship between species i and species

j, g_{ij} is zero. There are two minor variants on (10.8): the basal species may be treated differently from the other species, and given a positive growth term to represent feeding off the environment, or the environment may be included as a "species 0" and these growth terms represented by functional responses g_{i0}.

Apart from the constant death rates d_i and the ecological efficiency, λ, the model is completely specified once the functional responses have been chosen. Arditi and Michalski [65] have pointed out that these generalized functional responses, if they are to be logically consistent, must leave the balance equations invariant if two identical species are aggregated into a single species. The obvious generalized form of the Holling type functional response, Eq. (10.2), is

$$g_{ij} = \frac{a_{ij} N_j}{1 + \sum_k b_{ik} N_k}\,, \tag{10.9}$$

where the sum in the denominator is taken over all prey k of species i.

Generalizations of more complicated functional responses can only be found in the recent literature. Arditi and Michalski [65] suggest the following generalized Beddington form:

$$g_{ij} = \frac{a_{ij} N_j}{1 + \sum_k b_{ik} N_k + \sum_l c_{il} N_l}\,, \tag{10.10}$$

where the first sum is again taken over all prey k of species i, and the second sum is taken over all those predator species l that share a prey with i.

A possible generalization of the ratio-dependent functional response results from Eq. (10.10) if the 1 in the denominator is cancelled. However, as Arditi and Michalski [65] point out, the idea that predators share the prey, which led to the introduction of ratio-dependent functional responses, is better reflected by the following expression,

$$g_{ij} = \frac{a_{ij} N_j^{r(i)}}{N_i + \sum_{k \in R(i)} b_{ik} N_k^{r(i)}}\,, \tag{10.11}$$

with the self-consistent conditions

$$N_j^{r(i)} = \frac{\beta_{ji} N_i^{C(j)} N_j}{\sum_{k \in C(j)} \beta_{jk} N_k^{C(j)}}, \quad N_k^{C(j)} = \frac{h_{jk} N_j^{r(k)} N_k}{\sum_{l \in R(k)} h_{lk} N_l^{r(k)}}\,.$$

Here β_{ij} is the efficiency of predator i at consuming species j, h_{ij} is the relative preference of predator i for prey j, $R(i)$ are the prey species for predator i, $C(i)$ are the species predating on prey i, $N_j^{r(i)}$ is the part of species j that is currently being accessed as resource by species i and $N_k^{C(j)}$ is the part of species k that is currently acting as consumer of species j. An interesting consequence of this implicit form of the functional response is that not all the links that are in principle possible are realized, by a long way. This is a very realistic feature of the model, since species typically feed on those prey that are most easily available, and resort to other prey only during periods of food shortage. Arditi and Michalski [65] also found that small food webs with this generalized ratio-dependent functional response are far less sensitive to the aggregation of species than webs with prey-dependent functional responses.

A shortcoming of model (10.11) is that the predator preferences h_{ij} are constants that are independent of prey availability. In reality, one can expect that predators assign more effort to those prey from which they obtain more food per unit effort, so that a stationary point is reached only when a predator obtains from each prey the same amount of food per unit effort. This condition is implemented in the generalized ratio-dependent functional response suggested by Drossel, Higgs, and McKane [66], Eq. (10.13), which is discussed in the next section.

10.4.3 The complexity-stability debate

So far in this section we have surveyed the kinds of population dynamics equations which are frequently applied to the modelling of predator-prey systems. As is usual, we have assumed that the parameters of the various models are given, but for a large community these may be hundreds in number. Obviously some way of specifying the parameters is required, and it is at this stage that we move into the question of food-web modelling, since many of these parameters will be related to the underlying web structure. The methods that have been used to go beyond pure population dynamics to incorporate food-web structure fall into three classes (see, for instance, [67], who defines the first two classes). The first class, which is the object of this subsection, studies the stability of small webs as function of their structure, of the choice of dynamic equations, or of the choice of parameter values. The motivation for this type of study is the intuition that real food webs must be stable, Part of this program involves defining exactly what is meant by "stable". The second class, which will be studied in section 10.5.2, assembles communities from a very small original system by bringing in species from a "species pool", and if they can add to the community in a stable way, they are incorporated into the system. In this way, larger ecosystem can, in principle, be built up. Third, in evolutionary models (see section 10.5.3), a community is built up not from a preexisting pool of species, but by modification ("mutation") of existing species.

The first attempt to write down mathematical equations for the dynamics of food webs and to study their stability, is due to May [5]. May performed a linear stability analysis of the population sizes around a supposed equilibrium point:

$$\frac{d}{dt}\left(\delta N_i\right) = \sum_j \alpha_{ij}\, \delta N_j \,, \tag{10.12}$$

where δN_i is the deviation of the population size of species i from its equilibrium value and α_{ij} is the community matrix. In this way he avoided specifying the underlying population dynamics equations, but was constrained to stay near equilibrium. The choice of web structure is equivalent to the choice of the α_{ij}. May chose the diagonal elements of the matrix to be -1. The other elements were taken to be zero with probability $1 - h$. With probability h, they had a random nonzero value chosen from a distribution of width α, so that α is a measure of the average interaction strength. Using results from random matrix theory, he found that ecosystems that are initially stable will become less stable (*i.e.*, the initially negative eigenvalues of the community matrix move towards zero) when $\alpha(SC)^{1/2}$ is increased. Furthermore, Eq. (10.12) will almost certainly be stable if $\alpha(SC)^{1/2} < 1$, and almost certainly be unstable if $\alpha(SC)^{1/2} > 1$. This finding spurred on much of the interest in the relationship between C

and S discussed earlier. The belief that webs with high connectance were unstable supplied a reason why webs with large C were not observed. On the other hand, the result was hard to reconcile with the increasing evidence for the scaling relation $C \sim S^{-1+\epsilon}$, $\epsilon > 0$ unless α was very small, there were complaints from field ecologists that the webs which they had been observing for many years should be unstable according to the May criterion [28], and there were discussions concerning the mathematical basis of the result [68–70].

Although May's work was interesting because it broke new ground, there were obviously several weak points in the analysis. One was the lack of biological realism assigned to the web. It was argued that the web structure should be "plausible", and not just randomly generated [71, 72]. It was suggested, for example, that food webs with "realistic", rather than random, structures had more chance of being stable [72]. These ideas were made more concrete by Yodzis [73], who constructed "plausible community matrices" by using the topologies of real webs, with the correct sign and an estimate of the magnitude of the strength of the links. He then showed that in every case where community matrices were plausible, disrupted forms of these matrices, which no longer represented real communities, were less stable. Other authors [74, 75] started with a large random Lotka-Volterra system (of the order of 50 species) and successively removed those species that were least stable, until a stable smaller food web was obtained, which typically had more positive coefficients than random networks. Still other authors investigated the stability of small Lotka-Volterra food webs (typically 4 to 10 species) as a function of the connectivity pattern and the link strengths. For Lotka-Volterra systems, one simply has $\alpha_{ij} = N_i^* a_{ij}$, where N_i^* is the equilibrium population size of species i. Taking into account differences in body size between predators and their prey, Pimm and Lawton [76] found that webs with more omnivory (*i.e.*, more links) are not always less likely to be stable. The exception occurs in webs where a "predator" is a small parasite i of a large host, j, in which case $|\alpha_{ij}|$ is much larger than $|\alpha_{ji}|$. De Angelis [71] found that small webs are more stable when the ecological efficiency λ is smaller, when species on higher trophic levels have strong self-limitation (*i.e.,* a strong negative a_{ii}), or when the predator population dynamics have little impact on their prey.

Very recent evidence suggests that models with more realistic functional responses tend to be more stable than Lotka-Volterra systems. In those models, the community matrix α_{ij} has no simple relation to the coefficients in the dynamical equations, and its values can therefore be expected to be far from random, even if the parameters in the dynamical equations are chosen in some random way. This was demonstrated explicitly by Pelletier [77], who studied a system of n basal species (prey) and n predator species feeding on these prey, choosing a functional response of the form

$$g_{ij} = a_{ij} N_j \frac{n a_{ij} N_j}{\sum_k a_{ik} N_k} .$$

In this way, a predator can assign more weight to a prey from which it obtains more food. Pelletier found that 85% of these types of food webs (with random values for a_{ij}) are stable, irrespective of the value of n, in contrast to Lotka-Volterra systems, where the percentage of stable webs decreases quickly with n. We will see in the next section that an evolutionary model that uses generalized ratio-dependent functional responses, can build larger food webs than a Lotka-Volterra type model, indicating again that Lotka-Volterra systems are less stable than more realistic ones.

Another weak point of May's analysis is the use of linear stability analysis. Clearly, a model ecosystem need not be at a stable equilibrium point in order to be realistic, but may instead be on a limit cycle or even a chaotic trajectory, as long as the fluctuations are small enough that no species get close to extinction. In such a situation, the question as to whether more complex ecosystems are more stable takes the form "under what conditions have more complex systems smaller fluctuations in population sizes". Intuitive arguments were put forward that the addition of weak links to an existing web with a strong predator-prey coupling should have a dampening effect on the population oscillations of the strongly coupled predator-prey pair. The reason is that the predator can feed on an alternative prey to which it has a weak link when its main prey becomes low in population size, allowing the main prey to recover. Similarly, a weakly linked alternate predator can increase in population size when the main predator decreases, thus preventing a large oscillation in prey population. These arguments are supported by the numerical study of models for small food webs with several weak links. Using a Holling-type generalized functional response, McCann, Hastings, and Huxel [78] found that the weak links have indeed a stabilizing effect on the model dynamics. Polis [79] suggests that the chosen form of the functional response is important for the result, since it makes it impossible for a predator to maintain a high feeding efficiency on many prey at the same time (in contrast to Lotka-Volterra systems). Field data seem to support the hypothesis that stable food webs have many more weak links than strong links [78].

Some authors point out that species rich communities should have less community-level variability (*i.e.*, relative fluctuation of the combined density of all species sharing the same ecological role) than species poor communities, where the oscillations of one species cannot be counterbalanced by different oscillations of another species. This concept of community-level stability is supported by numerical simulation of a Lotka-Volterra type model [80], and is reviewed by Loreau in [81].

An alternative definition of stability, called "species deletion stability", which might have more direct relevance to real webs was introduced by Pimm in 1979 [82]. An ecosystem is defined to be species deletion stable if, when a species is removed from the web, all the remaining species remain at a stable equilibrium involving only positive densities. Species deletion stability decreases with increasing numbers of species and connectance, *i.e.,* decreases with complexity [82], but it also depends crucially on which species are selected for removal [83]. A quantitative measure of the deletion stability of a web is provided by S_d, the fraction of species for which the web is species deletion stable. However, it should be noted that data on experimental species removal show that many real species are not species deletion stable [83]. Recent work [84] showed that the risk of additional species deletions, following the loss of one species in model food webs, decreases with biodiversity. A review of the relation between the complexity and stability of an ecosystem [85] concluded that much of the confusion in the literature to date arose because of the different meanings given to the terms "complexity" and "stability"; many different definitions of perturbations and persistence are possible, and only a few are appropriate for real webs. One of the more fruitful of these has been the idea of "permanence" [86, 87], which will be explored in more detail in the following section, when assembly models are discussed. Recently the diversity-stability debate has been reviewed by McCann [88].

10.5 Assembly models and evolutionary models

This section describes models for food webs which incorporate longer time scales. In contrast to the models presented so far, they allow for the ongoing introduction of new species (due to immigration or speciation) and for species extinctions. As a consequence, the composition and structure of the web changes with time. Studies of assembly models and evolutionary models focus mainly on the features of the food web after a sufficiently long time, when the size of the food web and other properties cease to change in a systematic manner.

While the static models presented in section 10.3 are only concerned with web structure, but cannot address web stability, and while the dynamic models presented in section 10.4 focus on the stability of web subunits but do not deal with the overall web structure, the models presented in this section combine the two aspects of web structure and web stability. Another advantage of assembly models and evolutionary models is that links between species and interaction strengths are shaped by the web's history, instead of being assigned in an ad-hoc manner as in the other two types of models.

Assembly models and evolutionary models can be divided into three classes, which will be presented in the following subsections. The first class comprises toy models that resemble to some extent the static models discussed in section 10.3. They ignore population sizes, and species, and links are added and removed according to simple rules. The structure of the resulting webs is usually different from the structure of real food webs, as described in section 10.2. However, the main focus of these models is on species extinctions rather than on the food web structure, and these extinction events bear some similarity to those seen in the fossil record [89].

The other two classes of models are more realistic, as they take population sizes into account and include such important features as competition for food and link strength, which are not part of the toy models. The second class of models are species assembly models, which, starting from a small initial system, bring in new species from a species pool that are incorporated in the system if they add to the community in a stable way. These species assembly models, which typically lead to an uninvadable, stable system, will be discussed in subsection 10.5.2. The third class of models, reviewed in subsection 10.5.3 are inspired by biological evolution through modification of existing species. Just as the assembly models, they start from a small set of species, and then add new species, which are obtained by modifying existing species. In spite of differences in the population dynamics, the different evolutionary models lead to similar and realistic food web shapes.

10.5.1 Toy models

The purpose of introducing evolutionary toy models was not so much to reproduce realistic web structures, but rather to study the large-scale dynamics of species extinctions. Species are usually characterized by a number which is related to their fitness, and they become extinct when this number falls below a threshold value. The web structure can be a regular lattice, as in the Bak–Sneppen model [90], or a fully connected web, as in the Solé–Manrubia model [91], or new links are added together with a new species according to some rule, as in the Slanina–Kotrla model [92] and the Amaral-Meyer model [93]. Fitness changes are triggered by changes in species linked to a given species, and they also include a stochastic component. An overview of all these models was given by Newman and Palmer [94] and by Drossel [95].

Since the links in these models are in most cases not understood to be feeding relations but interactions of any type, their connection to food webs is only superficial. In the following, we give a description only of the model by Amaral and Meyer, which is the one closest to food webs, since it places the species in trophic layers, with links indicating which species feeds on which other species. The model is defined as follows: Species can occupy niches in a model ecosystem with L levels in the food chain, and N niches in each level. Species from the first level $l = 0$ do not depend on other species for their food, while species on the higher levels l each feed on k or less species in the level $l - 1$. Changes in the system occur due to two processes: (i) Creation of new species with a rate μ for each existing species. The new species becomes located at a randomly chosen niche in the same level or in one of the two neighboring levels of the parent species. If the new species arises in a level $l > 0$, k species are chosen at random from the layer below as prey. A species never changes its prey after this initial choice. (ii) Extinction: At rate p, species in the first level $l = 0$ become extinct. Any species in layer $l = 1$ and subsequently in higher levels, for which all preys have become extinct, also become extinct immediately. This rule leads to avalanches of extinction that may extend through several layers. Amaral and Meyer found from computer simulations that the size distribution of these extinction avalanches is given by a power law $n(s) \sim s^{-\tau}$ with $\tau \simeq 2$. This result $\tau = 2$, which was confirmed by an analytical calculation by Drossel [96], is compatible with the findings of paleontologists that species extinction events of all sizes have occurred in the geological past [89]. A more detailed study which also includes the taxonomy generated by the model is given by Camacho and Solé [97].

Large extinction avalanches are also found in the other toy models mentioned above, and they imply that the internal dynamics of ecosystems place them at the border of stability such that small triggers can have large consequences. However, the more widely accepted view seems to be that ecosystems in themselves are rather stable, but that external events like meteorite impacts or changes in the sea level are to blame for the large extinctions in the geological past. If this is correct, the simple toy models miss important ingredients that are present in real ecosystems. The more realistic models described in the next subsection lead to food webs that are much more stable.

10.5.2 Species assembly models

The more realistic assembly and evolutionary models, which will be discussed in the remainder of this section, include population dynamics. They have two time-scales, which are assumed to be separated. On the faster, ecological time scale, population sizes change until they reach fixed points or stationary orbits. On the slower time scale, new species are introduced by immigration (assembly models) or by modifying one or a few individuals of an existing species (evolutionary models). After introduction of the new species, the population dynamics may either drive this new species to extinction, or the new species becomes established, while possibly one or a few other species become extinct. Even if no species become extinct, the food web may become rearranged, with species abandoning one prey or choosing an additional prey.

Species assembly models take into account that real ecosystems, for instance on an island, are often built up by species immigration. Starting with either one or a few species, species from a "species pool" are added to the system, and they remain in it if the resulting system

is stable. Energetically constrained community assembly was modelled by Yozdis [73, 98]. Starting with N basal species each of which has a "production" P, new species are introduced one by one. The required energy intake e of a new species is chosen from a given probability distribution, and the prey species are chosen one by one with probabilities proportional to their unused production. If a prey has a randomly chosen fraction of its production available, this prey is utilized by the new species, and further links to preys are added until the energy needs of the new species are satisfied. The assembly process ends when the total unutilized production falls below a minimum value. The resulting webs have properties that agree well with real web data.

In all the other models, population dynamics is modelled via Lotka-Volterra equations. The species pool is usually a set of no more than 25 basal species ("plants"), and the same number of "herbivores", "carnivores" and top predators, with interaction coefficients between neighboring layers assigned according to some random rule (sometimes taking the larger body size of consumers into consideration or trying to include "specialists" that feed on only one prey as well as "generalists" that can feed on several prey). The third or fourth trophic level (carnivores and top predators) are missing in some of the models. Although this species pool is usually interpreted as stemming from a large ecosystem, like the mainland, no stability criteria or other criteria inspired from real large webs are applied to it.

After adding a new species with an initially small population size to the system, one of the three following things can happen: (i) The new species increases and coexists with all the other species. (ii) The new species remains in the system, but one or more other species go extinct. (iii) The new species goes extinct. Numerical integration of Lotka-Volterra equations, combined with the criterion of local stability, were used by Post and Pimm [99] and by Drake [100, 101], to construct webs of typically less than 20 species. Since the numerical integration of large Lotka-Volterra systems is very inefficient [102], other authors [103, 104] use the concept of permanence in order to find the composition of the new community after species addition. An ecological community is permanent if all species remain at a positive finite density when the density of each is started at a positive finite value. For Lotka-Volterra systems, the permanence of a system can be quickly tested using only two criteria. Clearly, this criterion of permanence is too strict, since real systems never explore the full space of possible population sizes, and since it seems implausible that for real communities even very unusual combinations of population sizes should not result in species extinctions.

After some time, an invasion-resistant state is achieved, the properties of which can be evaluated as function of the properties of the species pool. The invasion-resistant state may be a single community, or (in a minority of cases) a cyclic sequence of communities. Typically, the size of the resulting community increases with the size of the pool, but saturates when the pool size becomes large [104].

Lockwood *et al* [105] showed that if subsequent invasions follow rapidly (instead of waiting for a stable species configuration after each invasion), before the system can achieve an equilibrium state, communities do not evolve towards an invasion-resistant state, but move through complex cycles of composition, where each species gets its turn. A recent review on community assembly is [106].

To summarize, the species assembly models put forward so far are capable of generating intermediate-size webs with a predetermined number of trophic layers. There are several drawbacks of these models. First, so far only Lotka-Volterra equations have been used. How-

ever, since equations using other functional responses are known to be more realistic and more stable, it would be worthwhile to investigate species-assembly models with other functional responses. Second, the species pool is not usually very large, thus limiting the number of possible modifications of the web. It might well be possible that with a much larger pool, the webs would not evolve towards an invasion-resistant state. Third, the species pool is composed of species which have not co-evolved. Since the assembled web will consist after some time of species that are in some respect adapted to each other, it is very unlikely that a randomly defined additional species could invade the system. In contrast, real species pools contain species that have evolved to be able to survive well in the presence of other species from the pool. A real species pool, even when not large, will therefore contain many more species that can invade the ecosystem under consideration, than do the random species pool used in the models. The evolutionary models presented in the next subsection have no species pool at all, but they introduce new species as modifications of existing ones. New species are therefore much more likely to fit into the existing ecosystem than the randomly generated species in assembly models.

10.5.3 Evolutionary models

Evolutionary food web models introduce new species as variations of existing ones. The first evolutionary food web model that includes population dynamics was introduced by Caldarelli, Higgs and McKane [107]. Species in this model can be characterized as binary strings, with each bit representing a feature that is either present (1) or absent (0) in a species. This representation gives a measure of similarity between species (the number of features they have in common) and allows for "mutations" by randomly swapping a 1-bit and a 0-bit (*i.e.*, by replacing one feature with another). "Scores" between two species are obtained by multiplying the two feature vectors to the right and left of an asymmetric random matrix that is chosen at the beginning of the simulation. Positive scores indicate that the first species can feed on the second species, and negative scores mean that the first species is eaten by the second. The external resources are represented as an additional species of fixed (and large) population size, which does not feed on any species. The population dynamics are simple: at each time step, a fixed percentage of every species that has at least one predator is eaten by the predators of that species. The prey is divided such that all those predators that have a score within a certain narrow range of the maximum score against a prey species obtain a share, the size of which depends linearly on the score. This means that species do not feed on all prey species they can potentially feed on. As the web changes and evolves, the prey species eaten by a given predator can change. Since the dynamical equations are linear, they quickly reach a fixed point. Then, a randomly chosen individual is "mutated", and the new population sizes are calculated. Starting with one species and the external resources, large webs can be built. After some time, a stable species configuration is reached such that no "mutant" can become established. The parameters of the model can be chosen such that the fractions of top and bottom species, the numbers of links per species and other properties of the webs are very similar to those of real food webs.

In a subsequent paper by Drossel, Higgs and McKane [66], a modified version of this model was introduced, which contains more realistic population equations. For all species equations of the form Eq. (10.8) were used. As in the previous model, $\lambda = 0.1$ was chosen, and

the external resources were modelled as an additional species with a large and fixed population size. Apart from the external resources, all species have a death rate $d_i = 1$ and a ratio-dependent functional response of the form

$$g_{ij}(t) = \frac{S_{ij} f_{ij}(t) N_j(t)}{b N_j(t) + \sum_k \alpha_{ki} S_{kj} f_{kj}(t) N_k(t)} \,. \tag{10.13}$$

The S_{ij} are the above–mentioned scores, and f_{ij} is the fraction of its effort (or available searching time) that species i puts into preying on species j. These efforts must satisfy $\sum_j f_{ij} = 1$ for all i, and they are determined self–consistently from the condition

$$f_{ij}(t) = \frac{g_{ij}(t)}{\sum_k g_{ik}(t)}. \tag{10.14}$$

This condition is such that no individual can increase its energy intake by putting more effort into a different prey. The parameters α_{ki} give a measure of the strength of competition between species k and i. They are equal to 1 for $i = k$, and a linear function of the overlap between species i and k (*i.e.*, of the number of features that i and k have in common) for $i \neq k$.

In computer simulations of the model, the population sizes quickly reach a fixed point. As discussed in the previous section, this is generally not the case for Lotka–Volterra type population dynamics. The effects of competition, of predator saturation, the ratio-dependent functional response, and the ability to assign more effort to searching for the better prey species, may all play a crucial role in stabilizing the population dynamics. Using the same evolutionary dynamics as in the previous model, large food webs can again be built that consist of several hundreds of species. Just as with the simpler population dynamics equations of the previous model, the properties of the food webs agree well with the empirical ones. In contrast to the above model, which has simpler population dynamics, no stable species configuration is reached, but there is ongoing species creation and extinction, even after a long time. However, no more than a few species become extinct at the same time, and the size distribution of extinction events has a sharp exponential cutoff. The evolutionary dynamics of the model, combined with the population dynamics, thus create large stable webs, which have ongoing changes due to species overturn, but do not show strong responses to small perturbations. More recent studies of this model can be found in [108, 109]. Figure 10.2 shows an example of a food web generated by this model.

A different dynamical model, which uses Lotka–Volterra type equations, was introduced by Lässig *et al* [110]. In contrast to standard Lotka–Volterra equations, where the quadratic term is only used for predator–prey relationships, these authors also include a predator–predator competition term. The a_{ij} in Eq. (10.7) are equal to a constant γ_+ if j is predator of i, and $-\gamma_-$ (with $0 < \gamma_- < \gamma_+$) if i is predator of j. If i and j have a prey in common, then a competition term is added to a_{ij}, which is -1 for $i = j$ and $-\beta\rho_{ij}$ otherwise, with β being a constant smaller than 1, and ρ_{ij} the link overlap between the two species, which is defined as the geometric mean between the fraction of i's prey species that it shares with j and the fraction of j's prey species that it shares with i. All species have the same death rate. The external resources are represented as a few species with a positive growth rate and no prey.

The "mutations" which generate the evolutionary dynamics in this model consist in a change of a predation link for an individual. Using a mean–field approximation, the authors

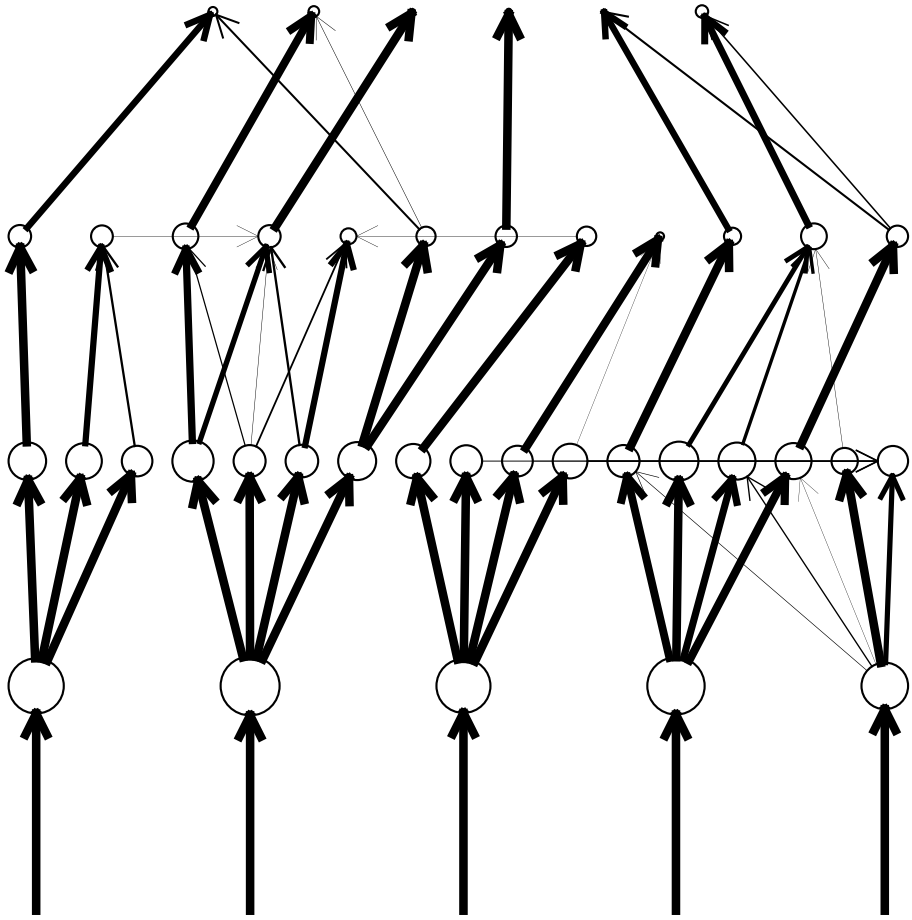


Figure 10.2: Example of a food web generated by the evolutionary model [66]. The radius of the circles is proportional to the logarithm of the population size, and the thickness of a link is a measure of the energy flow along that link.

analyzed this model analytically and obtained web structures which are very similar to realistic food webs. Computer simulations of this model are given in [111]. However, up to now only simulation results for one trophic layer of species have been published. All these species feed on the external resources, which are modelled as permanent species with a constant growth rate. The computer simulations show a constant turnover of species, with considerable fluctuations in the total number of species. This is in strong contrast to model [66], where simulations of one trophic layer always converged to an uninvadable species constellation. This confirms again the finding that Lotka-Volterra systems are less stable than systems with more realistic functional responses. It would be interesting to see how the model [110] behaves if several trophic layers are allowed. We expect that it would be less stable and that the webs would be smaller than for model [66].

The fact that the different evolutionary models give the same overall structure of the food web, indicates that the shape of food webs results from a few general principles, and is not much affected by certain details of the system. All models include competition between predators, and an efficiency of converting eaten prey into predator biomass which is smaller than 1. Furthermore, there are only few different external resources. Qualitatively, one can understand how these ingredients generate the familiar food web shapes: one can argue that very few external resources, combined with competition among the species feeding on these resources and with predation allowing only efficient feeders to survive, leads to only a few basal species that feed on these resources. Since there are more basal species than different external resources, there are more predator species that can coexist feeding on the basal species. The number of species thus increases initially with increasing trophic level, until the total biomass becomes so small that only a few top species can survive at the highest level. This explains the short length of food chains and the small proportion of basal and top species. The small number of links per species must result from the competition between predators, the strength of which can be tuned in all of the models.

10.6 Conclusions

Our aim in this review has been to survey most of the approaches that have been used to model food webs and to discuss real web data in enough detail to provide a background to the model building. A number of clear trends can be seen in both theory and experiment. In the latter case there has been a realization in the last decade or so that data collection needs to become a much more controlled and systematic affair. Much of the older data was collected as a by-product of other projects, and it has become clear that the database of collected webs contains enough hidden biases to raise serious doubts about its usefulness. Recently food web data has been far more painstakingly collected, and has been done as part of projects specifically devoted to the investigation of food webs. Eventually the quality of the ensembles of collected webs will reflect this. The theory of food web structure has also made advances during this time. Much of the early work concerned webs with random, rather than evolved, structures, but there has been a tendency in later work towards building up a web either from a pool of species (assembly models) or, more recently, by creating webs through modification of the existing species (evolutionary models). In parallel with these developments, some suggestions for more realistic population dynamics for whole communities, and not just for a single predator-prey pair, have been put forward. In our opinion these last two developments, taken together, will form the basis for further progress in the theory of food webs.

We have seen that evolved food webs or webs generated through a series of species invasions are generally more stable than randomly assembled food webs, even if realistic food web shape is imposed. Large and complex randomly assembled food webs simply collapse under the population dynamics and are very unlikely to be stable. Evolutionary dynamics can chose the predator–prey links and the competition structure such that stable webs are built step by step, starting from a small initial web. We do not yet know in sufficient detail in what respects the link strengths and link structure of evolved networks differ from ad-hoc compositions.

Furthermore, there is a need to investigate the effects of different functional responses on the structure and stability of food web models, rather than continuing to use the less satisfac-

tory Lotka-Volterra equations. In section 10.4 we drew attention to the existence of theories of population dynamics with realistic functional responses which can be applied to communities with arbitrary structures. The amount of computer time required to construct model webs will not obviously be much greater than using Lotka-Volterra equations, and there will probably be a saving due to the increased stability of the models with more realistic functional responses.

To conclude, we hope that more realistic functional responses will be used to investigate the structure and stability of assembled and evolved webs in greater depth in the future, and that models will be constructed that create webs by both immigration from a species pool and by variation of the species within the community. We believe that these, and other similar studies, will pave the way for a greatly increased understanding of the structure and nature of food webs over the next few years.

Acknowledgements

We thank C. Quince for producing Figure 2. B.D. was supported by the Deutsche Forschungsgemeinschaft (DFG) under Contract No Dr300-2/1.

References

[1] E. P. Odum, *Fundamentals of ecology*, (Saunders, Philadelphia, 1953).

[2] C. S. Elton, *Ecology of Invasions by Animals and Plants*, (Chapman and Hall, London, 1958).

[3] R. H. MacArthur, *Fluctuations of animal populations and a measure of community stability*, Ecology **36** (1955), 533–536.

[4] M. R. Gardner and W. R. Ashby, *Connectance of large dynamic (cybernetic) systems: critical values for stability*, Nature **228** (1970), 784–784.

[5] R. M. May, *Will a large complex system be stable?* Nature **238** (1972), 413–414.

[6] R. M. May, *Stability and complexity in model ecosystems*, (Princeton University Press, Princeton, 1974), Second edition.

[7] S. L. Pimm, *Food webs*, (Chapman and Hall, London, 1982).

[8] J. H. Lawton, *Food webs* in *Ecological Concepts*, J. M. Cherrett, (ed.), (Blackwell, Oxford, 1989), pp 43–78.

[9] J. E. Cohen, F. Briand and C. M. Newman, *Community food webs*, Biomathematics Vol. 20, (Springer-Verlag, Berlin, 1990).

[10] N. D. Martinez, *Artifacts or attributes? Effects of resolution on the Little Rock Lake food web*, Ecol. Monogr. **61** (1991), 367–392.

[11] M. Huxham, S. Beaney and D. Raffaelli, *Do parasites reduce the chances of triangulation in a real food web?* Oikos **76** (1996), 284–300.

[12] D. P. Reagan and R. B. Waide, *The food web of a tropical rain forest*, (U. of Chicago Press, Chicago, 1996).

[13] J. Memmott, N. D. Martinez and J. E. Cohen, *Predators, parasitoids and pathogens: species richness, trophic generality and body sizes in a natural food web*, J. Anim. Ecol. **69** (2001), 1–15.

[14] T. W. Schoener, *Food webs from the small to the large*, Ecology **70** (1989), 1559–1589.

[15] S. L. Pimm, J. H. Lawton and J. E. Cohen, *Food web patterns and their consequences*, Nature **350** (1991), 669–674.

[16] S. J. Hall and D. G. Raffaelli, *Food webs: theory and reality*, Adv. Ecol. Res. **24** (1993), 187–239.

[17] K. Havens, *Scale and structure in natural food webs*, Science **257** (1992), 1107–1109.

[18] N. D. Martinez, *Constant connectance in community food webs*, Am. Nat. **139** (1992), 1208–1218.

[19] N. D. Martinez, *Effect of scale on food web structure*, Science, **260** (1992), 242–243.

[20] P. H. Warren, *Making connections in food webs*, Trends in Ecol. Evol. **9** (1994), 136–141.

[21] P. A. Murtaugh and J. P. Kollath, *Variation of trophic fractions and connectance in food webs*, Ecology **78** (1997), 1382–1387.

[22] L-F. Bersier, P. Dixon and G. Sugihara, *Scale-invariant or scale-dependent behavior of the link density property in food webs: A matter of sampling effort?* Am. Nat. **153** (1999), 676–682.

[23] J. M. Montoya and R. V. Solé, *Topological properties of food webs: from real data to community assembly models*, (Santa Fe Institute working paper 01-11-069).

[24] S. J. Hall and D. G. Raffaelli, *Food web patterns: lessons from a species rich web*, J. Anim. Ecol. **60** (1991), 823–841.

[25] N. D. Martinez, *Effects of resolution on food web structure*, Oikos **66** (1993), 403–412.

[26] G. Sugihara, K. Schoenly and A. Trombla, *Scale invariance in food-web properties*, Science **245** (1989), 48–52.

[27] N. D. Martinez, *Scale-dependent constraints on food-web structure*, Am. Nat. **144** (1994), 935–953.

[28] G. A. Polis, *Complex trophic interactions in deserts: an empirical critique of food-web theory*, Am. Nat. **138** (1991), 123–155.

[29] N. D. Martinez and J. H. Lawton, *Scale and food-web structure — from local to global*, Oikos **73** (1995), 148–154.

[30] K. O. Winemiller, *Spatial and temporal variation in tropical fish trophic networks*, Ecol. Monogr. **60** (1990), 331–367.

[31] L. Goldwasser and J. Roughgarden, *Sampling effects and the estimation of food-web properties*, Ecology **78** (1997), 41–54.

[32] J. E. Cohen *et al*, *Improving food webs*, Ecology **74** (1993), 252–258.

[33] J. H. Lawton, *Webbing and WIWACS*, Oikos **72** (1995), 305–306.

[34] N. D. Martinez, B. A. Hawkins, H. A. Dawah and B. P. Feifarek, *Effects of sampling effort on characterization of food-web structure*, Ecology **80** (1999), 1044–1055.

[35] C. S. Elton, *Animal ecology*, (Sidgwick and Jackson, London, 1927).

[36] G. E. Hutchinson, *Homage to Santa Rosalia or why are there so many kinds of animals?* Am. Nat. **93** (1959), 145–159.

[37] L. Goldwasser and J. Roughgarden, *Construction and analysis of a large Caribbean food web*, Ecology **74** (1993), 1216–1233.

[38] P. Erdös and A. Rényi, *On the evolution of random graphs*, reprinted in J. Spencer (ed) *The art of counting: selected writings*, (MIT Press, Cambridge, 1973).

[39] J. E. Cohen, *Food webs and niche space*, (Princeton University Press, Princeton, NJ, 1978).

[40] J. E. Cohen and C. M. Newman, *A stochastic theory of community food webs. I Models and aggregated data*, Proc. R. Soc. Lond. B **224** (1985), 421–448.

[41] J. E. Cohen, *A stochastic theory of community food webs. VI Heterogeneous alternatives to the cascade model*, Theor. Pop. Biol. **37** (1990), 55–90.

[42] A. R. Solow and A. R. Beet, *On lumping species in food webs*, Ecology **79** (1998), 2013–2018.

[43] R. J. Williams and N. D. Martinez, *Simple rules yield complex food webs*, Nature **404** (2000), 180–183.

[44] J. Maynard Smith, *Models in ecology*, (CUP, Cambridge, 1974).

[45] E. C. Pielou, *Mathematical ecology*, (Wiley, New York, 1977).

[46] J. Roughgarden, *Theory of population genetics and evolutionary ecology: an introduction*, (MacMillan, New York, 1979).

[47] M. L. Rosenzweig and R. H. MacArthur, *Graphical representation and stability conditions of predator-prey interactions*, Am. Nat. **97** (1963), 209–223.

[48] M. L. Rosenzweig, *Why the prey curve has a hump*, Am. Nat. **103** (1969), 81–87.

[49] C. S. Holling, *The functional response of predators to prey density and its role in mimicry and population regulation*, Mem. Ent. Soc. Can. **45** (1965), 1–60.

[50] M. P. Hassell, *The dynamics of arthropod predator-prey systems*, (Princeton University Press, Princeton, 1978).

[51] J. Beddington, *Mutual interference between parasites or predators and its effect on searching efficiency*, J. Anim. Ecol. **51** (1975), 597–624.

[52] R. Arditi and L. R. Ginzburg, *Coupling in predator-prey dynamics: ratio-dependence*, J. Theor. Biol. **139** (1989), 311–326.

[53] I. Hanski, *The functional response of predators: worries about scale*, Trends Ecol. Evol. **6** (1991), 141–142.

[54] R. Arditi, L. R. Ginzburg and N. Perrin, *Scale-invariance is a reasonable approximation in predation models — reply*, Oikos, **65** (1992), 336–337.

[55] P. A. Abrams, *The fallacies of ratio-dependent predation*, Ecology **75** (1994), 1842–1850.

[56] H. R. Akcakaya, R. Arditi and L. R. Ginzburg, *Ratio-dependent predation — an abstraction that works*, Ecology **76** (1995), 995–1004.

[57] A. A. Berryman, A. P. Gutierrez and R. Arditi, *Credible, parsimonious and useful predator-prey models — a reply*, Ecology **76** (1995), 1980–1985.

[58] G. Huisman and R. J. De Boer, *A formal derivation of the Beddington functional response*, J. Theor. Biol. **185** (1997), 389–400.

[59] P. A. Abrams and L. R. Ginzburg, *The nature of predation: prey dependent, ratio-dependent or neither?* Trends Ecol. Evol. **15** (2000), 337–341.

[60] A. Hastings and T. Powell, *Chaos in a three-species food chain*, Ecology **72** (1991), 896–903.

[61] D. M. Post, M. E. Conners and D. S. Goldberg, *Prey preference by a top predator and the stability of linked food chains*, Ecology **81** (2000), 8–14.

[62] W. M. Getz, *Population dynamics: a per capita resource approach*, J. Theor. Biol. **108** (1984), 623–643.

[63] W. M. Getz, *A unified approach to multispecies modeling* Nat. Res. Mod. **5** (1991), 393–421.

[64] A. A. Berryman, J. Michalski, A. P. Gutierrez and R. Arditi, *Logistic theory of food web dynamics*, Ecology **76** (1995), 336–343.

[65] R. Arditi and J. Michalski, *Nonlinear food web models and their responses to increased basal productivity*, in *Food webs: Integration of patterns and dynamics*, (G. A. Polis and K. O. Winemiller, eds), pp. 122–133 (Chapman and Hall, New York, 1996).

[66] B. Drossel, P. G. Higgs and A. J. McKane, *The influence of predator-prey population dynamics on the long-term evolution of food web structure*, J. Theor. Biol. **208** (2001), 91–107.

[67] P. Yodzis, *Introduction to theoretical ecology*, (Harpers and Row, New York, 1989).

[68] H. M. Hastings, *The May-Wigner stability theorem*, J. Theor. Biol. **97** (1982), 155–166.

[69] J. E. Cohen and C. M. Newman, *When will a large complex system be stable?* J. Theor. Biol. **113** (1985), 153–156.

[70] P. Érdi and J. Tóth, *What is and what is not stated in the May-Wigner theorem?* J. Theor. Biol. **145** (1990), 137–140.

[71] D. L. De Angelis, *Stability and connectance in food web models*, Ecology **56** (1975), 238-243.

[72] L. R. Lawlor, *A comment on randomly constructed model ecosystems*, Am. Nat. **112** (1978), 445–447.

[73] P. Yodzis, *The stability of real ecosystems*, Nature **289** (1981), 674–676.

[74] A. Roberts and K. Tregonning, *The robustness of natural systems*, Nature **288** (1981), 265–266.

[75] P. J. Taylor, *The construction and turnover of complex community models having generalized Lotka-Volterra dynamics*, J. Theor. Biol. **135** (1988), 569–588.

[76] S. L. Pimm and J. H. Lawton, *On feeding on more than one trophic level*, Nature **275** (1978), 542–544.

[77] J. D. Pelletier, *Are large complex ecosystems more unstable? A theoretical reassessment with predator switching*, Math. Biosc. **163** (2000), 91–96.

[78] K. McCann, A. Hastings and G. R. Huxel, *Weak trophic interaction and the balance of nature*, Nature **395** (1998), 794–798.

[79] G. A. Polis, *Stability is woven by complex webs*, Nature **395** (1998), 744–745.

[80] A. R. Ives, J. L. Klug, and K. Gross, *Stability and species richness in complex communities*, Ecology Letters **3** (2000), 399–411.

[81] M. Loreau, *Biodiversity and ecosystem functioning: recent theoretical advances*, Oikos **91** (2000), 3–17.

[82] S. L. Pimm, *Complexity and stability: another look at MacArthur's original hypothesis*, Oikos **33** (1979), 351–357.

[83] S. L. Pimm, *Food web design and the effect of species deletion*, Oikos **35** (1980), 139–149.

[84] C. Borrvall, B. Ebenman and T. Jonsson, *Biodiversity lessens the risk of cascading extinction in model food webs*, Ecology Letters **3** (2000), 131–136.

[85] S. L. Pimm, *The complexity and stability of ecosystems*, Nature **307** (1984), 321–326.

[86] W. Jansen, *A permanence theorem for replicator and Lotka-Volterra systems*, J. Math. Biol. **25** (1987), 411–422.

[87] R. Law and J. C. Blackford, *Self-assembling food webs: A global viewpoint of coexistence of species in Lotka-Volterra communities*, Ecology **73** (1992), 567–578.

[88] K. S. McCann, *The diversity-stability debate*, Nature **405** (2000), 228–233.

[89] D. M. Raup, *Biological extinction in earth history*, Science **231** (1986), 1528–1533.

[90] P. Bak, and K. Sneppen, *Punctuated equilibrium and criticality in a simple model of evolution*, Phys. Rev. Lett. **71** (1993), 4083–4086.

[91] R. V. Solé and S. C. Manrubia, *Extinction and self-organized criticality in a model of large-scale evolution*, Phys. Rev. E**54**, (1996) R42–45.

[92] F. Slanina and M. Kotrla, *Extremal dynamics model evolving networks*, Phys. Rev. Lett. **83** (1999), 5587–5590.

[93] L. A. N. Amaral and M. Meyer, *Environmental changes, co-extinction, and patterns in the fossil record*, Phys. Rev. Lett. **82** (1999), 652–655.

[94] M. E. J. Newman and R. G. Palmer, *Modeling Extinction*, Oxford University Press, New York (2002).

[95] B. Drossel, *Biological evolution and statistical physics*, Adv. Phys. **50** (2001) 209–295.

[96] B. Drossel, *Extinction events and species lifetimes in a simple ecological model*, Phys. Rev. Lett. **81** (1998), 5011–5014.

[97] J. Camacho and R. V. Solé, *Extinctions and taxonomy in a trophic model of coevolution*, Phys. Rev. E**62** (2000), 1119–1123.

[98] P. Yodzis, *The structure of assembled communities. II.* J. Theor. Biol. **107** (1984), 115–126.

[99] W. M. Post and S. L. Pimm, *Community assembly and food web stability*, Math. Biosci. **64** (1983), 169–192.

[100] J. A. Drake, *Models of community assembly and the structure of ecological landscapes*, in *Mathematical ecology* (T. Hallam, L. Gross and S. Levin, eds), pp. 584–604, (World Scientific, Singapore, 1988).

[101] J. A. Drake, *The mechanics of community assembly and succession*, J. Theor. Biol. **147** (1990), 213–233.

[102] R. D. Morton, R. Law, S. L. Pimm and J. A. Drake, *On models for assembling ecological communities*, Oikos **75** (1996), 493–499.

[103] R. Law and R. D. Morton, *Permanence and the assembly of ecological communities*, Ecology **77** (1996), 762–775.

[104] R. D. Morton and R. Law, *Regional species pools and the assembly of local ecological communities*, J. Theor. Biol. **187** (1997), 321–331.

[105] J. L. Lockwood, R. D. Powell, P. Nott and S. L. Pimm, *Assembling ecological communities in time and space*, Oikos **80** (1997), 549–553.

[106] R. Law, *Theoretical aspects of community assembly*, in *Advanced ecological theory: principles and applications*, J. McGlade, (ed), (Blackwell, Oxford, 1999), pp 143–171.

[107] G. Caldarelli, P. G. Higgs and A. J. McKane, *Modelling coevolution in multispecies communities*, J. Theor. Biol. **193** (1998), 345–358.

[108] C. Quince, P. G. Higgs and A. J. McKane, *Food web structure and the evolution of ecological communities*, in *Biological Evolution and Statistical Physics* (M. Lässig and A. Valleriani, eds), (Springer-Verlag, Berlin, 2002).

[109] C. Quince, P. G. Higgs and A. J. McKane, *The effects of the removal and addition of species on ecosystem stability*, in *Complexity emerging: a paradigm for ecological thought* (J. A. Drake, C. R. Zimmermann, S. Gavrilets and T. Fukami, eds), (Columbia University Press, New York), to be published.

[110] M. Lässig, U. Bastolla, S. C. Manrubia, and A. Valleriani, *Shape of ecological networks*, Phys. Rev. Lett. **86** (2001), 4418–4421.

[111] U. Bastolla, M. Lässig, S. C. Manrubia, and A. Valleriani, *Dynamics and topology of species networks*, in *Biological Evolution and Statistical Physics* (M. Lässig and A. Valleriani, eds), (Springer-Verlag, Berlin, 2002).

11 Traffic networks

Kai Nagel

Abstract

Transportation systems are complex dynamical systems whose dynamics unfolds on networks as the spatial substrate. Early approaches to the problem have similarities to the computation of equilibrium current flow in electrical networks, with the main difference that in traffic the particles have fixed destinations. These steady state approaches are unrealistic when describing more complex aspects of the dynamics, which is why time-dependent microscopic models are introduced. Such models resemble typical molecular dynamics simulations, except that the spatial substrate is a graph instead of flat space, and particles are "intelligent". Both aspects are discussed in detail, the latter meaning that one has to go far beyond physics and into the area of human behavior and human learning. Another network aspect is the network of interaction between objects in the simulation, where these objects are not only travelers, but also traffic signals, traffic management centers, etc. For fast large scale simulations, one employs distributed computers, and mapping these interactions on the computational system is critical for high computing performance.

11.1 Introduction

Much of this book on networks is about the dynamics *of* networks. The question there is how networks form or change. Examples come from many different areas, from electrical networks to the the blood system, or from the Internet to the networks of socio-economic interaction. This contribution concentrates on another aspect: on dynamics *on* networks. In the particular example of traffic, this means that there is an underlying network, the road network, and the dynamics of the system unfolds on this network. Although this is also true for other networked systems, such as for electrical networks or for biological networks (nerve system, blood transport system), the traffic dynamics on links is relatively complex and thus very interesting. The reason for this is that the particles, or agents, in the traffic simulation are "intelligent", which means that they have strategic, long-term goals.

The context for the work reviewed in this contribution is in transportation planning. This means the prediction of traffic patterns 20 or more years into the future. Let us, as an example, consider the question (relevant for Switzerland) to build a second Gotthard tunnel through the Alps. Initially, such a new tunnel would just relieve congestion (and increase safety). However, on a somewhat slower time scale of days to months, people who previously took a different route because of congestion will switch to the Gotthard route. On an again longer

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

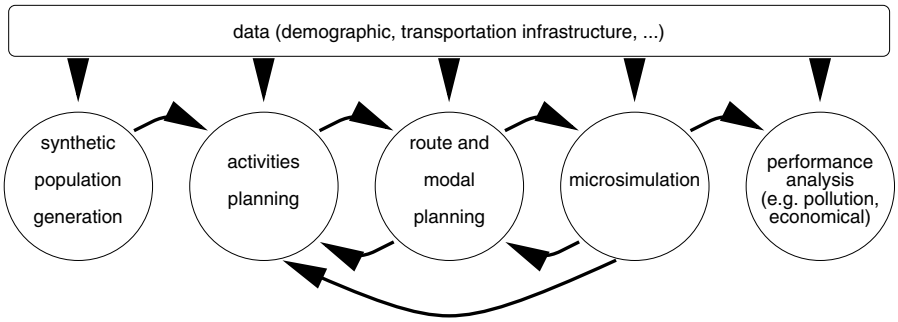


Figure 11.1: TRANSIMS modules

time scale, people will maybe make additional trips which use this. And finally, it is possible that land use changes in reaction to such changes – in this case for example in terms of a leisure park or industry south of the Alps which needs easy access to or from the north. In other cases, people's housing decisions will depend on access to their workplace.

In consequence, there is an emerging consensus that transportation simulations for planning purposes should consist of the following modules (Fig. 11.1):

- **Traffic simulation** – This is where travelers move through the street network by walking, car, bus, train, etc.
- **Modal choice and route generation module** – The travelers in the traffic simulation usually know where they are headed; it is the task of this module to decide which mode they take (walk, bus, car, bicycle, ...) and which route.
- **Activity generation module** – The standard cause why travelers are headed toward a certain destination is that they want to perform a specific activity at that location, for example work, eat, shop, pick someone up, etc. The activity generation module generates synthetic daily plans for the travelers.
- **Life style, housing, land use, freight, etc.** – The above list is not complete; it reflects only the most prominent modules. For example, the whole important issue of freight traffic is completely left out. Also, at the land use/housing level, there will probably be many modules specializing into different aspects.
- **Feedback** – The above modules interact, and the interaction goes in both directions: activities and routes generate congestion, yet (the expectation of) congestion influences activities and routes. This is typically solved via a relaxation method, i.e. modules are run sequentially assuming that the others remain fixed, until the results are consistent.

In addition, there need to be initialization modules, such as the **synthetic population generation module**, which takes census data and generates disaggregated populations of individual people and households. Similarly, it will probably be necessary to generate good default layouts for intersections etc. without always knowing the exact details.

In this review, we will look at this technology specifically from the view of networks. There are four aspects that we will discuss:

1. **Dynamics on networks** (Sec. 11.2) – As pointed out at the beginning of the introduction, the transportation system is a network of roads and other links connected at intersections, train stations, etc. These networks have interesting dynamics, both on the links and at the intersections.

2. **Particles are intelligent** (Sec. 11.3) – Not strictly being a network aspect, it is however still important to note that travelers are "intelligent", when compared to, say, molecules or blood cells. This aspect means that travelers have strategic goals, and they have internal representations of the world around them which they use to reach these goals. This means that two travelers in identical situations will in general make different decisions.

3. **The network of interactions and distributed computing** (Sec. 11.4) – Travelers and other objects in a transportation system interact. For example, congestion is formed by travelers being in each other's way; ride sharing necessitates travelers to meet at a common pick-up location; adaptive traffic lights react to traffic conditions; etc. This network of interactions is generally sparse, and it is also often local although the mechanics of transmission are different from a physical force field – consider e.g. a sensor for an adaptive signal. At the opposite, non-local end, there are global radio broadcasts. There are also increasingly networked services which are both sparse and non-local – for example electronic route guidance systems where only very few travelers communicate with a center. We will review this aspect together with consequences for distributed large scale computation in Sec. 11.4.

4. **Dynamics of networks** (Sec. 11.5) – Finally, people can, via the political process, change the transportation network. Although it appears difficult if not impossible to make any reliable predictions here, it may be worthwhile to explore such models to understand the mechanisms behind it, in particular the effects of self-reinforcing decisions and positive feedback. We will look at this aspect in Sec. 11.5.

The paper will be concluded by a short summary.

11.2 Dynamics on networks

As pointed out above, traffic unfolds its dynamics on a graph, and the dynamics on the links (roads) and nodes (intersections) of this graph are complex and interesting. This section will concentrate on these dynamic aspects. The section will start by a short review of the traditional static assignment method and then proceed to particle-based micro-simulations. As one will see, the static assignment method resembles a steady state current distribution in a fuse network while the micro-simulations resemble molecular dynamics simulation of particles flowing through a graph. Thus, transportation science follows physics on the path to more and more microscopic simulations in the attempt to go beyond steady-state phenomena.

11.2.1 The four step process and static assignment

The traditional method of traffic prediction for transportation planning is based on the four step process:

1. **Trip generation:** This module generates, for each traffic zone, the number of trips starting there and the number of trips ending there. This can be done for arbitrary time slices, but is often done for a typical 24-hour weekday.

2. **Trip distribution:** Trip generation results in sources and sinks, but not how they are connected. This is done in the trip distribution module. The result is an **origin-destination matrix**, which has, at row i and column j, the number of trips going from zone i to zone j.

3. **Modal choice:** In this module, the trips are split between the modes of transportation.

4. **Route assignment:** For each trip, a path is found through the network so that no other path is faster. Congestion is taken into account via the link travel time being a function of the trips using that link.

Route assignment can be formalized in the following way: Let r_{ij} be the number of trips from i to j. Routes from i to j are numbered by k; in consequence, $r_{ij,k} \geq 0$ is the number of trips using the k-th route. Let $\delta_{ij,k,a}$ an indicator if route ij,k uses link a. The number of trips using link a then is

$$x_a = \sum_{ij} \sum_{k} r_{ij,k}\, \delta_{ij,k,a} \; .$$

The link travel time (link cost) is normally defined via a function $t_a(x_a)$. It makes sense to assume that this function is strictly monotonically increasing. The trip time of a route in consequence is

$$T_{ij,k} = \sum_{a} t_a(x_a)\, \delta_{ij,k,a} \; .$$

The problem specification now is that $r_{ij,k}$ need to be found such that

$$\sum_{k} r_{ij,k} = r_{ij}$$

and such that all used routes have the same travel time and no unused route has a faster (= better) travel time.

This is typically solved by making the $r_{ij,k}$ continuous, meaning that also trip generation, trip distribution, and modal choice can be made on real numbers. With real numbers and with the assumption that $t_a(x_a)$ is strictly monotonically increasing, it can be proven that the above problem has a unique solution in terms of the x_a. The problem can in fact be written as a minimization problem, making it amenable to the tools of nonlinear optimization. In consequence, sophisticated algorithms exist which compute numerical approximations to the unique solution [1, 2].

The above problem is very similar to a non-linear static network flow problem in physics, where the link resistance is given via a non-linear $U = R(I)\, I$, and where sources and sinks are given via the result of the trip generation. The only (but important) difference is that in assignment "particles know where they go", meaning that one cannot in general exchange particles as one can, in electrical networks, do with electrons.

Static assignment has many shortcomings. For example, it does not correctly represent dynamic effects such as queue build-up, and it does not have enough microscopic information to do, for example, emission calculations. It also de-couples decisions from individual actors. For example, the only decision available for modal choice is the origin and the destination of the trips; important aspects such as income, car ownership (!), additional trips during the day, etc. are not used. Note, however, that these latter aspects could be overcome by a different software design. What cannot be overcome are the shortcomings in the representation of dynamic effects, which are treated in more detail in the next section.

11.2.2 Simple link dynamics and the queue model

In static assignment, one assumes a function $t_a(x_a)$ for each link. In practice, this function is parametrized by a few numbers, such as the free speed and the capacity of the link. The capacity is the maximum number of vehicles that can traverse a given location on the link per time unit; i.e., it is the maximum throughput or maximum current for that link. From a physics perspective, it is clear that such number must be an average, and that any realization of traffic can deviate from that number, especially for short times. Nevertheless, it should be clear that, if a link has a capacity of C_a, then, in the average no more than C_a vehicles can traverse the link.

Now imagine a scenario as in Fig 11.2, with a road 1 with a capacity of 4000 vehicles per hour connected to a road 2 with a capacity of 2000 vehicles per hour, and a demand zero for $t < 0$ and of 3000 per hour for $t \geq 0$. After one hour, 1000 cars will be queued up at the entry to the bottleneck, and the queue will grow by another 1000 cars in each hour. That is, the steady state solution of a demand in excess of capacity corresponds to an infinitely long queue *upstream of* the overloaded link. Static assignment would represent this via its link travel time functions $t_1(x_1)$ and $t_2(x_2)$. It would show link 2 as overloaded and congested, which is dynamically incorrect, since it is link 1 (and eventually additional upstream links) which bear the consequences, as we just saw.

It would be possible to avoid this situation in static assignment by setting $t_a(x_a) = \infty$ as soon as x_a exceeds capacity C_a, since the route assignment would then avoid to put more than C_a trips on that link. This, however, also does not correspond to reality, since waiting queues at the entry to bottlenecks clearly exist. Alternatively, one could attempt to formulate a mathematical model which includes queues. Although this is in principle feasible, not enough mathematical facts are known about such a model to make it useful in practice (see, e.g., [3]). In addition, with every level of additional complexity the situation looks more hopeless. For example, something like individual preferences, or different link travel speeds (e.g. cars vs. trucks), or the effects of turn pockets/merge lanes, or emissions resulting from acceleration, are difficult to represent in the static assignment framework.

Putting these arguments together, it makes sense to consider microscopic approaches. In microscopic approaches, all individual objects such as vehicles or travelers are represented individually. Here, we will start with a simple microscopic model which is called the queue

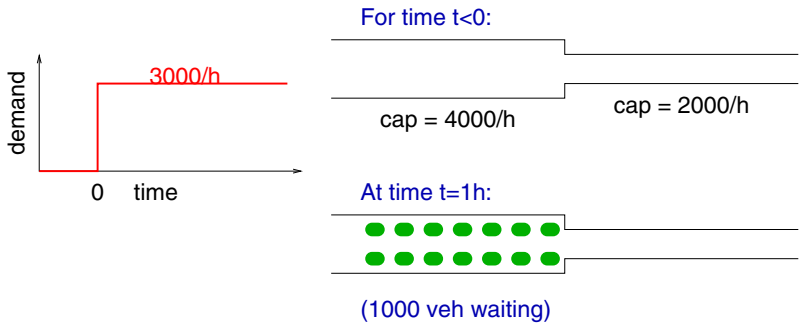


Figure 11.2: Example of failure of static assignment in the representation of dynamics. The scenario is a street with a capacity of 4000 vehs/hour leading into a street with a capacity of 2000 vehs/hour. The demand is zero before $t = 0$ and 3000 vehs/hour afterward. In each hour, an additional 1000 vehs will be waiting in upstream of the bottleneck.

model [64]. In the queue model, links are represented by simple FIFO (first-in first-out) queues. Vehicles entering a link at time t_1 are assumed to travel along the link with free speed V_a, meaning that they cannot leave the link before time $t_2 = t_1 + L_a/V_a$, where L_a is the link length. At the end of the link there is a capacity constraint, meaning that at most C_a vehicles can leave the link per time step. Non-integer C_a are resolved by using probabilities. So far, this is indeed a standard queuing model [4]. An important addition is the introduction of a storage constraint, meaning that there is a limitation of the number of vehicles that one can put on a link. It is the link storage constraint which will eventually make the link "full" and thus cause queue build-up and congestion spill-back.

In spite of their aesthetic appeal and their computational speed, there are a number of problems with queue models. Some are relatively simple geometrical shortcomings, such as the fact that, although for city networks it makes sense to have the capacity constraint at the end, in other situations like for freeways or in Fig. 11.2, they are at the beginning. Others are more severe, such as the fact that intersection design has not been solved satisfactorily with this model. The difficulty stems from the fact that both the capacity and the storage constraint need to be satisfied. Especially at high congestion levels, there are typically at each intersection many vehicles from incoming links competing for the same few slots on outgoing links. In reality, this is solved via explicit prioritization, either based on traffic lights or stop/yield signs, or on explicit legal rules such as "right before left". For transportation planning however, this information is often not available, and it is also subsumed in the capacities. Although this problem seems solvable, some more systematic work will be necessary here. Finally, it is difficult to consistently handle different vehicle speeds, vehicle types, or vehicle classes. An example for vehicle classes, which in this case differ by destination link at an intersection, is depicted in Fig. 11.4.

11.2.3 Virtual reality micro-simulations

An alternative to the queue model, avoiding the problems which come from the reduced geometrical representation, are micro-simulations which run on correct street layouts, including

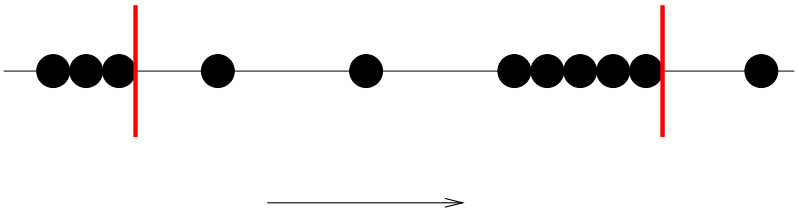


Figure 11.3: Queue model dynamics

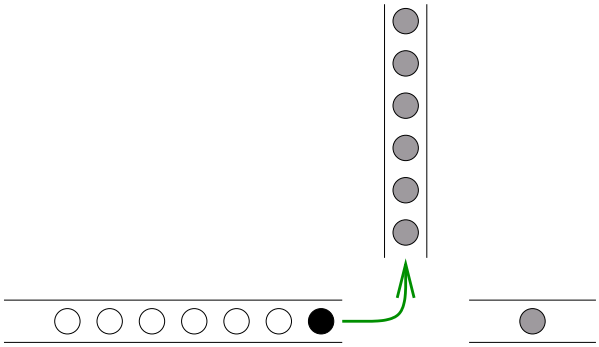


Figure 11.4: Limited geometric representation with queue model: The white vehicles cannot move (to the right) because the black vehicle, which is stuck, is in the way.

merging, turning, and weaving lanes, correct signal schedules etc. In terms of driving rules, such simulations consist of four major elements:

1. Car following. This describes how one vehicle follows another (or many others) on a single lane road.
2. Lane changing. This describes how vehicles change lanes on multi-lane roads. Reasons to change lane can be speed improvement, or anticipated turns in the future (a vehicle that wants to make a left turn needs to get into the left lane). Passing against oncoming traffic also belongs here.
3. Protected turns. This describes how vehicles behave at fully signalized intersections. For transportation planning purposes, it is enough to simply make vehicles stop at red and go at green.
4. Unprotected turns. Often, movements across intersections are not protected by signals, such as a left turn against oncoming traffic, or at a yield sign. Also, there may be special rules such as that the light rail always has priority.

Such microscopic simulations have the advantage that, at least in principle, they can be made arbitrarily realistic by adding more and more rules. In addition, they look very convincing to a non-technical audience (Fig. 11.5), an important aspect since the results of transportation planning simulations are of interest to all citizens.

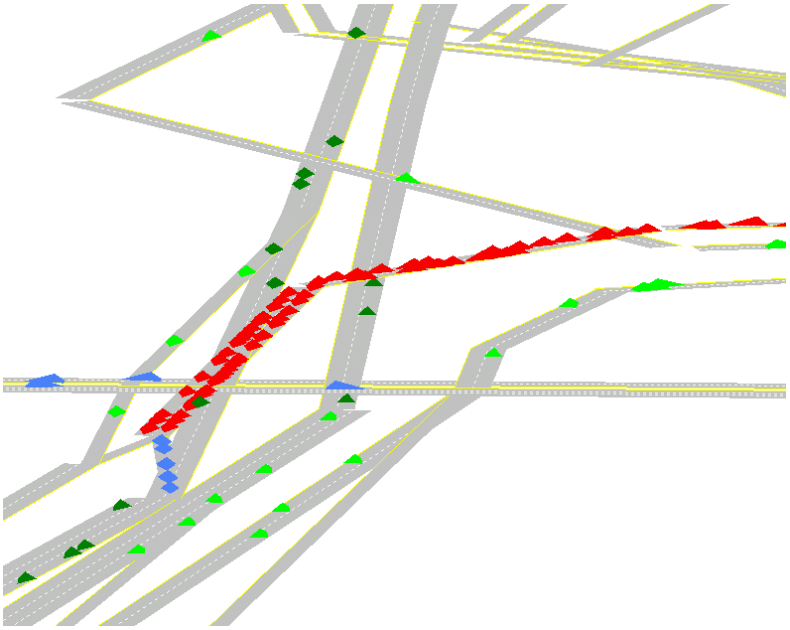


Figure 11.5: Virtual reality representation of simulated traffic in Portland/Oregon.

11.2.4 CA implementations of virtual reality micro-simulations

The maybe simplest approach to implement these rules are cellular automata micro-simulations. In these micro-simulations, roads are segmented into cells, each typically of the size that a single vehicle occupies in a jam (e.g. 7.5 meters). Movement is performed via jumps from one cell to another; for example, a speed of 5 cells per time-step corresponds to a jump of 5 cells. In order to translate those units into the real world, a time step of one second is customary (because of reaction time arguments), meaning that "5 cells per time-step" corresponds to

$$5 \text{ cells/time-step} \times 7.5 \text{ meters/cell} \times 1 \text{ second/time-step}$$

$= 37.5$ meters/second $= 135$ km/h . In such a cell-based system, a simple rule set to achieve the above functionality is as follows.

1. Car following.

   For all cars do in parallel

   - $v' = \min[v(t)+1, gap(t), v_{max}]$, where v' is a temporary variable, $v(t)$ is the current velocity, $gap(t)$ is the number of empty cells ahead, and v_{max} is the maximum speed, for example given by the speed limit.
     gap models the effect that one has to slow down if there is a vehicle ahead; note that this simple formulation assumes infinite braking capability.
   - With probability p_{noise}, $v(t+1) = \max[v' - 1, 0]$, else $v(t+1) = v'$.

After this is done for all vehicles, each vehicle is moved forward according to its speed.

This model [5] has been investigated in much detail in the literature; see Ref. [6] for a review. Its main feature is that at high enough densities, distinctive traffic jams form which would be interpreted as start-stop traffic by an individual driver.

2. Lane changing. Before the speed calculation, do the following:

   For all cars do in parallel

   - Set "reason-to-change-lanes" to true if the vehicle eventually wants to make a corresponding turn. Also set it to true if the other lane is faster than the current lane.
   - Set the safety criterion to true if there is sufficient space on the other lane.

   After this is done for each vehicle, all vehicles for which both criteria are fulfilled change lanes.

   There are more technical details here than with car following. For example, the above criteria need to be filled with quantitative rules, and care needs to be taken that two vehicles do not end up in the same cell during the parallel update. Also, the lane changing in anticipation of turning movements requires care, because vehicles can change lanes too early, meaning they may get stuck in a queue for a different turn, or too late, meaning they may not be able to make the intended turn. For more information see, e.g., Ref. [7, 8].

3. Protected turns. As stated above, this is relatively simple. As long as the modeled vehicles have infinite braking capability, a red light can be modeled by a virtual car of speed zero being inserted and removed at the location of the traffic signal.

4. Unprotected turns. A simple rule is (see Fig. 11.6):

   For each interfering lane (meaning a lane which has priority), check if the gap in front of the interfering vehicle is large enough. If yes, accept the turn, otherwise wait.

   Both with protected and with unprotected turns, there needs to be space available on the outgoing link.

The above is simplified in many respects, in particular with respect to complicated intersection designs, which are here replaced by the assumption that a vehicle makes the complete decision at the waiting position, and once the decision for a movement has been made, it can move freely across the intersection. Another simplification concerns the use of the cellular automata (CA) technique. CA are easy to code for such applications, since most driving rules need spatially-organized access to data, for example to neighboring cells, and the CA technique provides that. It is however also possible to code vehicles as individual particles with continuous position and speed (e.g. [9–11]), similar to a molecular dynamics technique [12]. Such codes are harder to program efficiently since one needs to keep track of spatial neighbors, but when done correctly they are computationally as fast as CA codes. This is helped by the fact that for traffic, taking the limit of $\Delta t \to 0$ where Δt is the computational time step is not useful since human reaction time needs to be modeled.

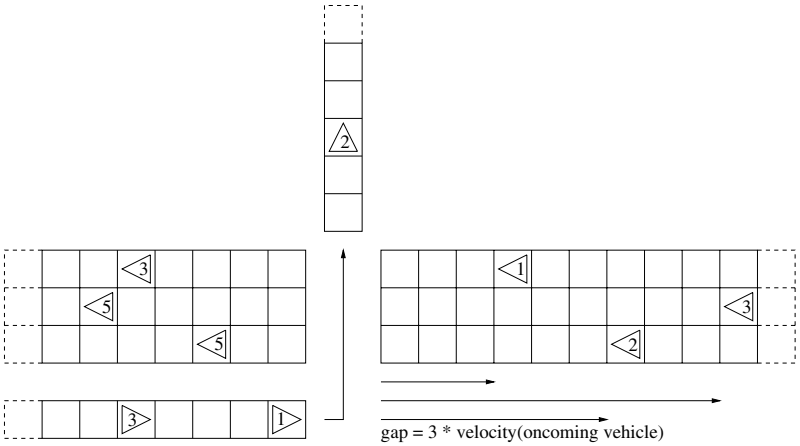


Figure 11.6: Illustration of gap acceptance for a left turn against oncoming traffic. From [8]

The trick with using such microscopic models for transportation planning is to make them computationally fast by making them on the micro-level barely realistic enough to obtain good information on the macro-level. This is consistent with a Statistical Physics approach, where many macroscopic laws can be obtained from much simplified microscopic models. Much progress on this aspect has emerged in recent times. For example, from Monte Carlo runs it has become clear that models as simple as the ones described above are macroscopically reasonable. They generate plausible fundamental diagrams, both on closed links and at intersections (Fig. 11.7, also see [13]), and they can display the emergence of the infamous jam-out-of-nowhere (Fig. 11.8) although there is discussion about its real world importance [14].

In terms of theory, some simple CA models can be provably related to accepted macroscopic theories of traffic: The continuous limit of the Asymmetric Stochastic Exclusion Process (ASEP) is the Burgers equation

$$\partial_t \rho + \partial_x \rho - 2\rho \partial_x \rho = D\, \partial_x^2 \rho \,,$$

which implies a relation of $q = \rho\,(1-\rho)$ between density ρ and flow q (see, e.g., [15]). This, in return, means a speed-density relation of $v = 1 - \rho$, which is known as the Greenshields relation (see [16]) in traffic flow theory. For deterministic continuous microscopic models, it has been shown, in certain cases mathematically and in others by computer simulation, that the mechanism of traffic flow breakdown (i.e. the transition from homogeneous "laminar" traffic to inhomogeneous traffic with stop-and-go waves) is the same as for Navier-Stokes models for traffic flow [11, 17]. Kinetic theory can build an, albeit still fragile, mathematical bridge from microscopic dynamics to fluid-dynamic equations [18, 19]. For an exhaustive review of models for traffic, see [6, 20]. Interestingly, the precise mechanism for traffic breakdown in stochastic models is still under discussion [21, 22]. In particular, under certain circumstances the boundary of jams is weakly fractal [23] (look at Fig. 11.8 for an impression) while under others it is not [10], and this is related to a discussion about a possible phase transition and its order.

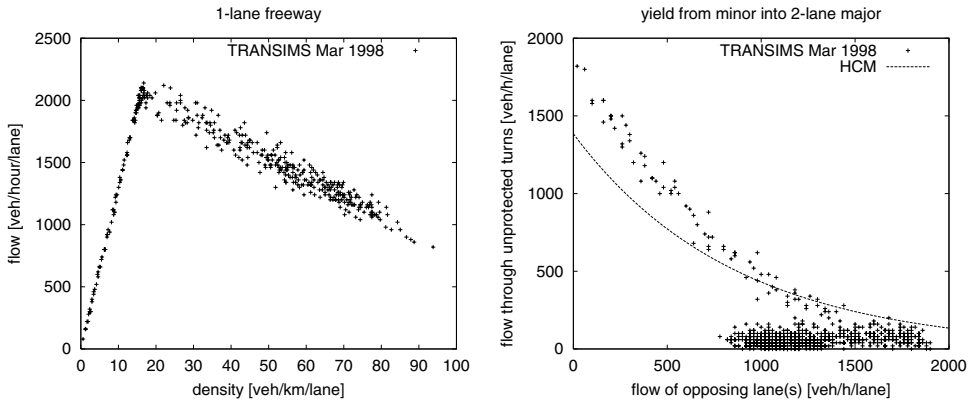


Figure 11.7: Fundamental diagram (flow vs. density; LEFT) and diagram for unprotected yield (unprotected flow vs. major flow; RIGHT).

A disadvantage of a "virtual reality micro-simulation" is that it needs rather a lot of input data. For example, many aspects of the street network are needed, such as merge or turn pockets, signal plans, grades, speed limits, or lane connectivities. Fig. 11.9 shows an example of the last: the arrows denote which incoming lanes are connected to which outgoing lanes, and the micro-simulation needs this information to induce lane changing as described above.

Since these data are usually not readily available, good "default generators" should be developed. They would for example generate plausible intersection designs and signal schedules based on other data, such as link capacities (maximum hourly flows) which are often available. Such synthetic defaults could then be used (with care) until real data became available. Also, the simulation could be used to detect obvious implausibilities, which could then be corrected on a case-by-case basis.

Since macroscopic quantities such as hourly flow are "emergent", there is no method to systematically construct the needed microscopic from the available macroscopic data. In consequence, the only method available is to run systematic tests with many intersection layouts and to record the resulting behavior. From such simulations, lookup tables could be constructed which then generate the microscopic designs from the macroscopic data.

Finally, good care needs to be taken to clearly differentiate between synthetic and field data in the data bases. This is often not done, or it is not done automatically by the system, resulting in questionable data entries necessitating costly manual correction.

11.2.5 Traffic in networks

In the above, we have started from static assignment and pointed out its similarity to equilibrium flow in physical networks. In both systems, when the dynamical aspects become important, the steady-state equilibrium approach is no longer valid. Similar to physical networks, progress can be made by a microscopic approach, which means to represent particles (travelers) individually instead of as part of some steady state rate or steady state flow. We

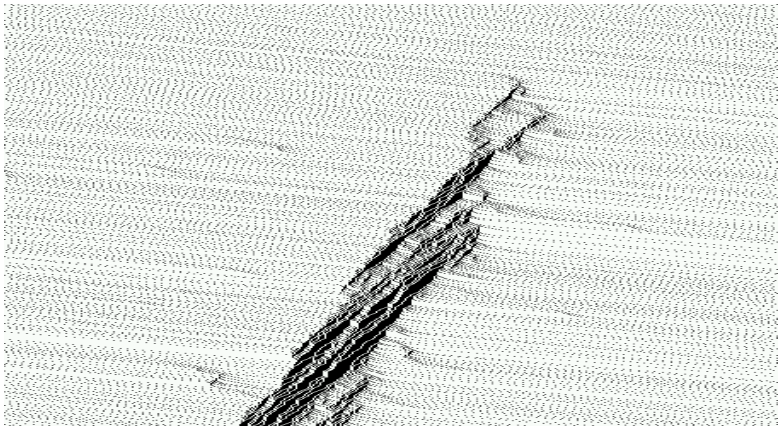


Figure 11.8: Traffic jam out of nowhere. The plot shows a space-time plot, space being horizontal, and time increasing from top to bottom. In consequence, the lines of the plot show consecutive time steps. The jam emerges spontaneously, and it shows a fragmented, weakly fractal structure. Traffic moves from left to right; the jam moves against the traffic direction.

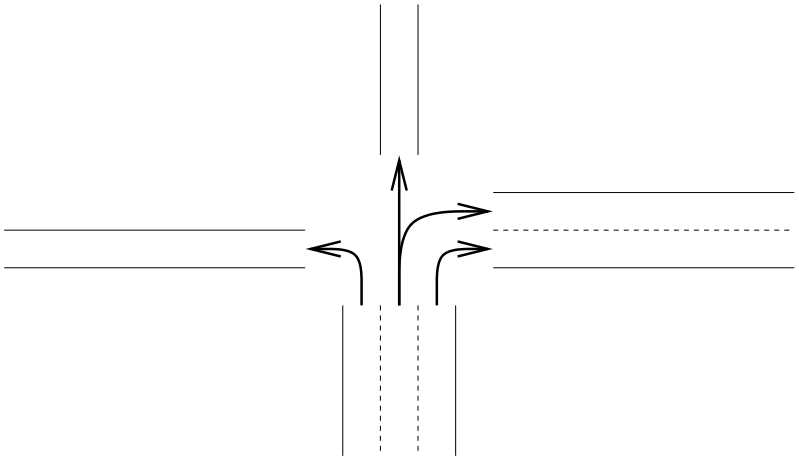


Figure 11.9: Lane connectivities

have then presented two possible micro-simulations, one relatively simple, based on a small but important extension of standard queuing models, and one relatively advanced, with the ability to be extended toward a virtual reality micro-simulation.

When inspecting the two presented systems, one will notice that for the first one (the queue simulation) nearly all the dynamics happens at the intersection, while for the second one there is an equal share of link and intersection dynamics. Many practitioners believe that for traffic in urban networks, the intersection dynamics is much more important than the

link dynamics. This motivates a much simplifying approach to traffic in networks, which is to model traffic on a two-dimensional grid [24–27]. Such simulations show interesting phenomena, such as phase transitions to grid-lock. It is an open question in how far these observations can be translated to more realistic traffic networks, with their more complicated (and more "forgiving") intersection dynamics.

11.3 Particles are intelligent

We have argued in the introduction that transportation planning tools need to include effects ranging from traffic flow via human decision-making up to land-use planning. We have then presented the static assignment approach to transportation planning, where a restricted representation of the traffic dynamics made it impossible or useless to make the other modules more realistic. Following that, we have described two alternatives for the traffic dynamics which "repair" these problems. Both traffic simulations that we describe are dynamic (i.e. time-dependent) and microscopic (i.e. each traveler is individually represented), which are the minimal pre-requisites for the following arguments. We have presented the two approaches to make clear that there is a wide range of models which fulfill these criteria, ranging from relatively simple (such as the queue simulation) to extremely realistic; the only restriction is that computation needs to be fast enough. Clearly, there are other models which fulfill these specifications [28–34].

Given that, it is now possible to improve the other aspects of the four step process. As in the traffic flow simulation, instead of gradual improvement we will focus on a bottom-up approach from first (or microscopic) principles. Again, the microscopic actors of the system are the travelers. As pointed out, the main difference to a simulation of, say, electrons in an electric network is the internal intelligence and adaptability of travelers. In contrast to water molecules, two travelers in exactly the same external situation can make different decisions.

11.3.1 Route generation

Having travelers move around randomly is not enough. For example, if a car approaches an intersection, the driver needs to decide the turning direction. A traditional method is to use turn counts, meaning that there is empirical data with the information about what fraction of the traffic goes into which direction. For any kind of transportation planning question, this is not enough information. The most drastic example is the addition of a new road: There would be no information available of how the traffic at the connecting intersection redistributes when the new road becomes connected. One would also assume that turn counts at other intersections change, since some of the traffic would adapt to use the new road.

This means that for transportation planning simulations it is indispensable to know the destinations, and to have routes for each vehicle. In this way, when a new road or a railway connection is added, every traveler can consider to adapt their routing in order to use this new connection. The route generation module of the transportation planning simulation should be multi-modal (i.e. include other modes besides cars), although some of the mode decision is better done in the demand generation module (see next).

A typical method for route generation is a time-dependent fastest path algorithm. Given a starting time t_0, an origin i and a destination j, and, for each link, information how long it will take to traverse the link when entering at a specific time, this algorithm will compute the fastest path from i to j when starting at time t_0. The time-dependent Dijkstra algorithm which solves this problem is, with a heap implementation, of complexity $M \log N$, where M is the number of links and N is the number of nodes (intersections). This is in fact a very low complexity, and it is difficult to construct a heuristic which is significantly faster [35].

11.3.2 Activity generation

For many questions, having the routes adaptive while the activities remain fixed is not enough. For example, making travel faster usually results in people making more trips. This is called **induced traffic**. Conversely, increasing congestion levels will eventually suppress trips which would otherwise be made, although it is not always clear which trips are suppressed and what congestion level is necessary to have that effect.

In order to deal with these and other effects, one has to make demand generation adaptive to congestion. A recent method for this is activity generation, meaning that, for each individual in the simulation, one generates a list of activities (such as sleeping, eating, working, shopping) plus locations and times (Fig. 11.10). Since in this method each traveler is treated individually, it is possible to use arbitrary decision rules, which means that arbitrary methods can be investigated. The currently best-accepted methods are based on random utility theory and are called discrete choice models [36].

As stated above, activity generation needs to be done in conjunction with mode decisions. For example, having a car clearly changes the list of preferable destinations for a given activity, or may even make other activities more desirable.

11.3.3 Housing, land use, freight, life style, et al

Transportation planning does not stop at activities. For example, making commuting roads faster by increasing capacity usually results in more people moving to the suburbs. That is, housing decisions are closely related to transportation system performance. Similarly, questions of land use (e.g. residential vs. commercial vs. industrial areas) clearly influence and interact with transportation. Freight traffic needs to be considered. Life style choices (e.g. urban life style, often without car ownership, vs. rural life style, usually with car ownership) need to be considered; as already alluded to above, such long-term commitments have strong influence on activity selection and modal/route choice.

11.3.4 Day-to-day learning, feedback, and relaxation

There is strong interaction between the above modules. For example, plans depend on congestion, but congestion depends on plans. A widely accepted method to resolve this is systematic relaxation (e.g. [37]) – that is, make preliminary plans, run the traffic micro-simulation, adjust the plans, run the traffic micro-simulation again, etc., until consistency between modules is reached. The method is somewhat similar to a standard relaxation technique in numerical analysis.

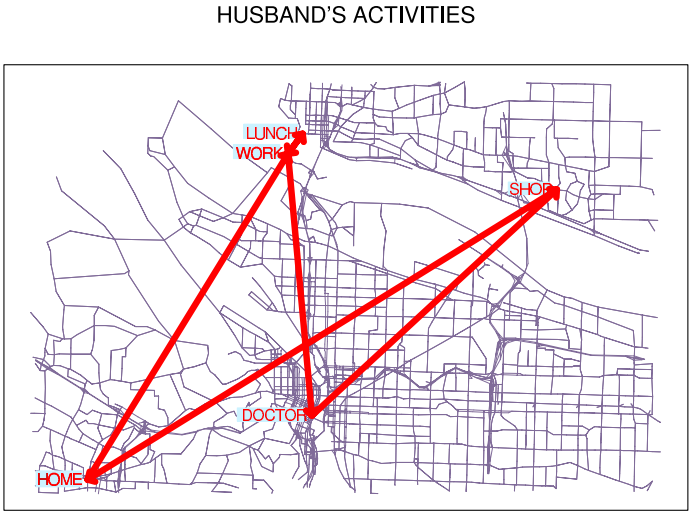


Figure 11.10: Example of a sequence of activities for a person in Portland/Oregon. From R.J. Beckman.

Such iterated simulations can be treated as discrete dynamical systems. A state is the trajectory of the simulation through one day; an iteration is the update from one day (period) to the next (Fig. 11.11). As such, one can search for things like fix points, steady state densities, multiple basins of attraction, strange attractors, etc. Typically, one would first analyze the steady state behavior, and then the transients. Under certain conditions the existence of a unique steady state can be proven [38], although for the computationally feasible number of iterations the possible occurrence of "broken ergodicity" [39] needs to be taken into account. Broken ergodicity is the property of a system to be mathematically ergodic but to remain in parts of the phase space for long periods of time.

11.3.5 Within-day re-planning

All the above lines of thought still assume, in some sense, "dumb" particles. Particles follow routes, but the routes are pre-computed, and once the simulation is started, they cannot be changed, for example being adapted to unexpected congestion and/or a traffic accident. In other words, the intelligence of the agents is external to the micro-simulation. In that sense, such micro-simulations can still be seen as an, albeit much more sophisticated, version of the link cost function $c_a(x_a)$ from static assignment, now extended by influences from other links and made dynamic throughout time. And indeed, many dynamic traffic assignment (DTA) systems work exactly in that way (e.g. [37]), in spite of several problems in particular with quick congestion build-up [40].

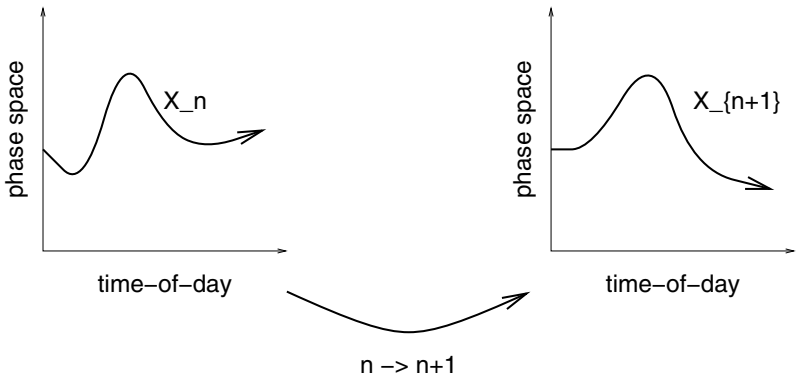


Figure 11.11: Schematic representation of the mapping generated by the feedback iterations. Traffic evolution as a function of time-of-day can be represented as a trajectory in a high dimensional phase space. Iterations can be seen as mappings of this trajectory into a new one.

Another way to look at this is to say that one assumes that the emergent properties of the interaction have a "slowly varying dynamics", meaning that one can, for example, consider congestion as relatively fixed from one day to the next. This is maybe realistic under some conditions, such as commuter traffic, but clearly not for many other conditions, such as accidents, adaptive traffic management, impulsive behavior, stochastic dynamics in general, etc. It is therefore necessary that agents are adaptive (intelligent) also on short time scales not only with respect to lane changing, but also with respect to routes and activities. It is clear that this can be done in principle, and the importance of it for fast relaxation [41,42] and for the realistic modeling of certain aspects of human behavior [43,44] has been pointed out. Nevertheless, we are not aware of operational implementations of this aspect.

11.3.6 Individualization of knowledge

Another aspect of intelligent agents is that their "knowledge" should be private, i.e. each agent should have a different set of knowledge items. For example, people typically only know a relatively small subset of the street network, and they have different knowledge and perception of congestion. This is called "mental maps"; some experimental implementations are Refs. [45–48]. We will come back to computational aspects in Sec. 11.4.2.

11.3.7 State of the art

No simulation package currently integrates all the aspects that are discussed. TRANSIMS [49] comes from the transportation planning side and is maybe the most advanced in terms of using advanced computing methods for large scale scenarios. The TRANSIMS research program is reaching completion in 2002, with a full-scale simulation of a scenario in Portland/Oregon, with a network of 200 000 links and several million travelers. We ourselves are in the process of using TRANSIMS for a full-scale simulation of all of Switzerland [50], see Fig. 11.12.

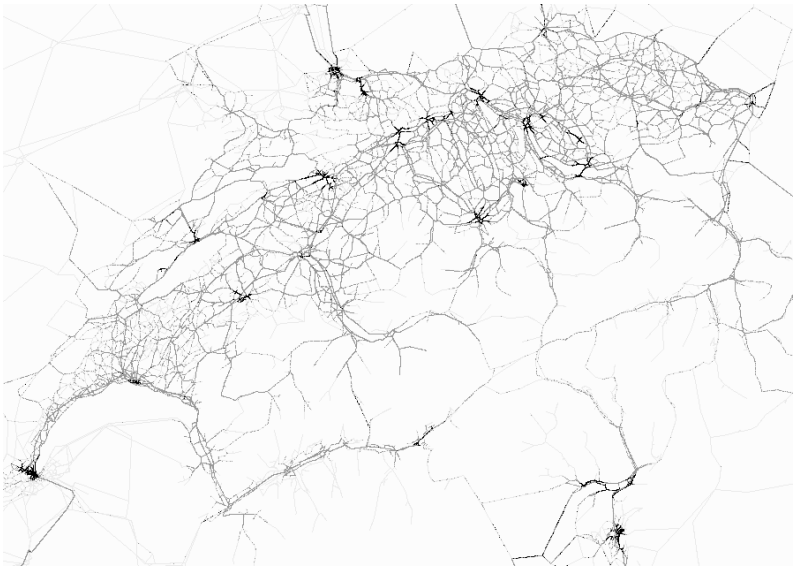


Figure 11.12: Microscopic simulation of all of Switzerland. Preliminary result

DYNAMIT [29] and DYNASMART [30], originally started as transportation simulation tools for the evaluation of ITS (Intelligent Transportation System) Technology, also advance into the area of transportation planning by the addition of the demand generation modules. Some comparison between field data and simulation results, obtained from a queue model micro-simulation and much simplified demand generation, can be found in [51]. METROPOLIS [28] is a package designed to replace static assignment by a simulation-based but very simple dynamic approach. It allows the user the specify arbitrary link-cost functions but in its current version still does not allow the queue build-up as discussed in Sec. 11.2.2. Its strength lies in the self-consistent computation of departure time choice. A more complete overview of regional transportation simulation models can be found in [52].

11.4 Distributed computing and the network of interactions

So far, this text has assumed that agents do their "strategic planning" independently of each other, and interactions occur in the traffic micro-simulation alone. These interactions are entirely local, since cars/drivers react to the situation around them, including other cars, signals, speed limits, etc. In consequence, the network of interactions is similar to, say, a molecular dynamics simulation with short-range interactions, where the network of interactions reflects the spatial dimensionality of the scenario.

We have also seen that the results of interaction, such as congestion, for many reasons cannot assumed to be fixed from one run to the next, and that an on-line (or within-day) replanning capability should be included. From here, it is easy to see that long-range interactions clearly play a role, for example the telephone call to another household member to pick up the child from the kindergarten since one is running late. Other examples for long-range interaction are:

- Interaction of agents with the transportation infrastructure, for example at adaptive traffic lights
- Reaction to radio broadcasts from a traffic management center
- Reaction to reports from friends
- Car sharing
- Coordination (ride sharing, household activities)

The graphs of these interactions can be seen as (meta-)networks. These networks are *sparse* and *dynamic*, meaning that in spite of the long-rangedness only a small number of particles interacts (in contrast to long-range forces in some molecular dynamics simulations), and that the network links re-organize over time.

The network aspects of interaction become particularly clear when one looks at computational aspects. As long as one runs everything on a single CPU, it is in principle possible to write one monolithic software package. In such a software, an agent who wanted to change plans would call a subroutine to compute a new plan, and during this time the computation of the traffic dynamics would be suspended. However, single CPUs are too slow for large scale simulations, and so one uses parallel computers, typically large clusters of connected PCs. In such a cluster, the regional area of the simulation is distributed across the PCs so that each PC only deals with a small part (domain decomposition). Other PCs deal with, say, the computation of routes or of activities. This approach means that interactions between simulation objects often result in interactions between PCs, which need to be explicitly coded, usually via message passing.

11.4.1 Distributed computing of the traffic micro-simulation

The most compute-intensive part of current implementations is usually the traffic micro-simulation. A simple calculation gives an approximate number: Assume a 24-hour simulation ($\sim 10^5$ sec) and a one second time step, 10^7 travelers, and a 1 GHz CPU (10^9 CPU-cycles per sec). Further assume that the computation of one time step for each traveler needs 100 CPU-cycles – remember the driving rules (car following, lane changing, protected turns, unprotected turns) and include overhead for route following etc. The result is that such a simulation takes about

$$\frac{10^5 \times 10^7 \times 10^2}{10^9} = 10^5$$

seconds or approximately 1 day on a single CPU. This is indeed approximately correct for a TRANSIMS simulation of a corresponding Switzerland scenario (5 mio travelers; network with 28 622 links); the queue simulation is 10–100 times faster [53].

The simulations can be accelerated by using parallel computers. This becomes indispensable for large applications when including feedback learning as discussed in Sec. 11.3.4 since this multiplies the computing times by a factor of 50, resulting in 50 days of computing time for the above scenario when using the TRANSIMS micro-simulation. We focus on so-called Beowulf architectures, since they are the most probable ones to be available to prospective users (metropolitan planning organizations; traffic engineering consulting companies; academics). Beowulf clusters consist of regular workstations (such as Pentium PCs running Linux) coupled by regular local area network (such as 100-Mbit Ethernet).

The idea is to divide the simulation area into many pieces, each of which is given to a different CPU. The CPUs communicate e.g. via message passing. In principle, using, say, 100 CPUs should result in a speed-up of 100. In practice, there are many limiting factors coming from the hardware and from the operating system. For traffic micro-simulations, the most important limiting factor is the latency of the Ethernet, which (in an off-the-shelf system without tuning) is of the order of 1 msec [54]. Since each CPU in the average needs to communicate with six other CPUs, this means that each time step needs approx. 6 msec for communication. This limits the speed-up to $1 \text{ sec}/6 \text{ msec} \approx 167$, independent of the number of CPUs that one uses. In practice, "100 times faster than real time" is a good rule of thumb [53, 55]. This domain decomposition approach is similar to a parallel computing approach to "standard" particle dynamics, for example in molecular dynamics [12], with the maybe only distinction that molecular dynamics simulations rarely use a graph instead of regular Cartesian space as spatial substrate.

Unfortunately, in contrast to many other computing aspects, latency does not seem to improve in commodity hardware: is has been virtually unchanged from 10 Mbit Ethernet to 100 Mbit Ethernet to Gbit Ethernet; FDDI is even slower. This has some interesting consequences:

- The above result refers to the speed-up with given system size when using more and more CPUs. Alternatively, one can run larger and larger systems when using more and more CPUs. As is well known, scale-up is much less problematic on parallel computers than speed-up. In consequence, it is possible to run scenarios of virtually arbitrary size 100 times faster than real time.

- Alternatively, one can make the micro-simulations more realistic while still being able to compute 100 times faster than real time.

- It should be noted that parallel supercomputers do not have the same limitation since they employ special purpose hardware for the communication between CPUs. This results in an improvement by a factor of 100 for latency, meaning that for practical scenarios other factors play a more important role.

  While a parallel Beowulf costs of the order of 2000-3000 U.S.-$ per node, a parallel supercomputer is about 20 times more expensive. Since this makes supercomputers irrelevant for the expected users, even when considering the use of a supercomputing center, we have done little research in this direction.

  It is however possible to use more advanced communication hardware for Beowulf clusters, for example Myrinet (www.myri.com). This should improve latency and thus maximum speed-ups by a factor of 10-50.

- Finally, it should be mentioned that, while for 10 Mbit Ethernet the main limiting factor was the hardware, for Gbit Ethernet this is no longer true: Special purpose implementations [56] bring Gbit Ethernet in the range of Myrinet. It is unclear if these improvements will make it into the mainstream.

Alternatively, one can consider other means of speeding up the computation. A possibility is to replace day-to-day replanning by within-day replanning, as discussed in Sec. 11.4. Experiments have shown that this reduces the number of necessary iterations considerably [42]. Possible distributed implementations of this are discussed in Sec. 11.4.2.

11.4.2 Distributed computing of plans generation

Additional complications come in with within-day replanning (Sec. 11.3.5) and with non-local interaction. Two examples:

- Re-planning. On a single CPU, a traveler calling the re-planning subroutine will just suspend the traffic simulation. On a parallel computer, if one traveler on one CPU does this, *all* CPUs have to suspend the traffic simulation since it is not possible (or very difficult) to have simulated time continue asynchronously. A more plausible approach is to have the re-planning module on a different CPU. The traveler then sends out the re-planning request to that CPU, and the traffic simulation keeps going. Eventually, the re-planning will be finished, and it will be sent to the traveler, who picks it up and starts acting on it. An experimental implementation of this using UDP (User Datagram Protocol) for communication shows that it is possible to transmit up to 100 000 requests per second per CPU [47], which is far above any number that is relevant for practical applications. This demonstrates that such a design is feasible and efficient.

  An additional advantage of such a design is that the implementation of a separate "mental map" (Sec. 11.3.6) for each individual traveler does not run into memory or CPU-time problems. Specific route guidance services can be simulated in a similar way.

- Non-local interaction between travelers. So far, everything assumes that travelers make autonomous decisions, and they interact in the micro-simulation only. This is however not always the case, for example for ride sharing, or when family members re-organize the kindergarten pick-up when plans have changed during the day. This will necessitate complicated negotiations between agents, and neither the models nor the computational methods for this are developed.

Some readers may have noticed that, in particular in the first example, success of the re-planning operation is not guaranteed. For example, the new plan may say to make a turn at a specific intersection, and by the time the new plan reaches the traveler, she/he may have driven past that point. Such situations are however not unusual in real life – how often does one recognize that a different decision some time ago would have been beneficial. Thus, in our view the key to success for large scale applications it to not fight asynchronous effects but to use them to advantage. For example, once it is accepted that such messages can arrive late, it is also not a problem to not have them arrive at all, which greatly simplifies message passing.

This design is similar to many robot designs, where the robots are autonomous on short time scales (tactical level) while they are connected via wireless communication to a more powerful computer for more difficult and more long-term time scales (strategic level); see, e.g., Ref. [57] for robot soccer. Also, it seems that the human body is organized along these lines – for example, in ball catching, it seems that the brain does an approximate pre-"computation" of the movements of the hands, while the hands themselves (and autonomously) perform the fine-tuning of the movements as soon as the ball touches them and haptic information is available [58]. This approach is necessitated by the relatively slow message passing time between brain and hands, which is of the order of 1/10 sec, which is much too slow to directly react to haptic information [59].

11.5 Outlook: Dynamics of networks

A further step for transportation simulations could be to make infrastructure changes (such as the addition of roads or train connections) endogeneous to the simulation package. This would mean that one would enable the simulation system to autonomously find out where changes to the transportation systems should be done, and include them into the system. On the simpler level of trail selection, this has been done by the method of active walkers [60]; Yamins et al investigate methods to grow urban roads [61]. For urban planning, one would have to make assumptions about political power distributions and related policies, but based on those it should be possible to run such a model. If it would yield anything useful will remain an open question for the foreseeable future.

11.6 Conclusion

For many systems, the dynamics does not unfold on "flat space", but on a graph or network. Although many concepts of dynamical systems still apply, they need to be adapted for dynamics on a graph. As a non-mathematical example, look at the visualization of dynamics on a graph, which is rather different from visualization of dynamics in two- or three-dimensional space.

Transportation simulation is a prime example of a real-world dynamical system on a graph. It is particularly interesting, since the one-dimensional dynamics on the links interacts with the networks aspect. For example, kinematic waves, as described by the Burgers equation or by an Asymmetric Stochastic Exclusion Process, can travel through an intersection, causing complicated dynamics there [62, 63]. In fact, very little seems to be known of these link-network-interactions, especially for large systems with many links (roads) and vertices (intersections).

In addition, the particles/agents in traffic systems are "intelligent". This means that they have strategic, long-term goals, with the consequence that no two particles are interchangeable, and that different particles, when confronted with the same situation, can make different decisions. In practical terms, for transportation simulations this "intelligence" involves aspects like route choice, mode choice, or activity generation. Moreover, agents adapt or learn, which means that they should be able to remember past behavior and past performance, to construct new plans, and to try them out.

For large scale scenarios, distributed computing is a necessity. The typical starting point is domain decomposition of the traffic micro-simulation, which means that each CPU runs the micro-simulation on a piece of the network. For efficiency reasons, this implies that the "intelligence" modules need to be separate from the traffic simulation itself. Mapping the resulting system well on parallel computer architectures seems to be a necessity for efficient large scale transportation simulations. Finally, one can look at the re-organization of the transportation network as a consequence of a political process. This aspect is touched only very briefly.

In summary, transportation simulations combine elements from many areas, ranging from dynamical systems via networks and graph theory to socio-economic human behavior. The current technology is advanced enough to start helping with policy decisions, yet many aspects remain unsolved and offer challenging problems for years to come.

Acknowledgments

Los Alamos National Laboratory makes the TRANSIMS software available to academic institutions for a small charge.

The Swiss Federal Administration provides the input data for the Switzerland studies.

References

[1] Y. Sheffi. *Urban transportation networks: Equilibrium analysis with mathematical programming methods*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1985.

[2] Michael Patriksson. *The Traffic Assignment Problem: Models and Methods*. Topics in Transportation. VSP, Zeist, The Netherlands, 1994.

[3] Stochastic and dynamic models in transportation, special issue. *Operations Research*, 41(1), 1993.

[4] H.P. Simão and W.B. Powell. Numerical methods for simulating transient, stochastic queueing networks. *Transportation Science*, 26:296, 1992.

[5] K. Nagel and M. Schreckenberg. A cellular automaton model for freeway traffic. *Journal de Physique I France*, 2:2221, 1992.

[6] D. Chowdhury, L. Santen, and A. Schadschneider. Statistical physics of vehicular traffic and some related systems. *Physics Reports*, 329(4–6):199–329, May 2000.

[7] K. Nagel, D.E. Wolf, P. Wagner, and P. M. Simon. Two-lane traffic rules for cellular automata: A systematic approach. *Physical Review E*, 58(2):1425–1437, 1998.

[8] K. Nagel, P. Stretz, M. Pieck, S. Leckey, R. Donnelly, and C. L. Barrett. TRANSIMS traffic flow characteristics. Los Alamos Unclassified Report (LA-UR) 97-3530, Los Alamos National Laboratory, 1997.

[9] R. Wiedemann. Simulation des Straßenverkehrsflusses. Schriftenreihe Heft 8, Institute for Transportation Science, University of Karlsruhe, Germany, 1994.

[10] S. Krauß. *Microscopic modeling of traffic flow: Investigation of collision free vehicle dynamics*. PhD thesis, University of Cologne, Germany, 1997. See www.zpr.uni-koeln.de.

[11] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, and Y. Sugiyama. Dynamical model of traffic congestion and numerical simulation. *Physical Review E*, 51(2):1035, 1995.

[12] D.M. Beazley, P.S. Lomdahl, N. Gronbech-Jensen, R. Giles, and P. Tamayo. Parallel algorithms for short-range molecular dynamics. In D. Stauffer, editor, *Annual reviews of computational physics III*, pages 119–176. World Scientific, 1995.

[13] W. Brilon and N. Wu. Evaluation of cellular automata for traffic flow simulation on freeway and urban streets. In W. Brilon, F. Huber, M. Schreckenberg, and H. Wallentowitz, editors, *Traffic and Mobility: Simulation – Economics – Environment*, pages 163–180. Aachen, Germany, 1999.

[14] C.F. Daganzo. Requiem for second-order fluid approximations of traffic flow. *Transportation Research B*, 29B(4):277, 1995.

[15] H. Spohn. *Large scale dynamics of interacting particles*. Springer, Berlin, 1991.

[16] D. L. Gerlough and M. J. Huber. *Traffic Flow Theory*. Special Report No. 165. Transportation Research Board, National Research Council, Washington, D.C., 1975.

[17] B. S. Kerner and P. Konhäuser. Structure and parameters of clusters in traffic flow. *Physical Review E*, 50(1):54, 1994.

[18] D. Helbing. Gas-kinetic derivation of Navier-Stokes-like traffic equations. *Physical Review E*, 53(3):253–282, 1996.

[19] P. Nelson. Synchronized traffic flow from a modified Lighthill-Whitman model. *Physical Review E*, 62(2):R6052–R6055, 2000.

[20] D. Helbing. Traffic and related self-driven many-particle systems. *Reviews of Modern Physics*, Oct/Nov 2001. Also www.arXiv.org, cond-mat/0012229.

[21] D. Chowdhury et al. Comment on: "Critical behavior of a traffic flow model". *Physical Review E*, 61(3):3270–3271, 2000.

[22] K. Nagel, Chr. Kayatz, and P. Wagner. Breakdown and recovery in traffic flow models. In Y. Sugiyama et al, editor, *Traffic and granular flow '01*. Springer, Heidelberg, in press.

[23] K. Nagel and M. Paczuski. Emergent traffic jams. *Physical Review E*, 51:2909, 1995.

[24] O. Biham, A. Middleton, and D. Levine. Self-organization and a dynamical transition in traffic-flow models. *Physical Review A*, 46:R6124, 1992.

[25] F.C. Martinez, J.A. Cuesta, J.M. Molera, and R. Brito. Random versus deterministic 2-dimensional traffic flow models. *Phys. Rev. E*, 51(2), 1995.

[26] T. Nagatani. Jamming transition in a two-dimensional traffic flow model. *Physical Review E*, 59(5):4857–4864, 1999.

[27] J. Freund and T. Pöschel. A statistical approach to vehicular traffic. *Physica A*, 219(1–2), 1995.

[28] A. de Palma and F. Marchal. Real case applications of the fully dynamic METROPOLIS tool-box: an advocacy for large-scale mesoscopic transportation systems. *Networks and Spatial Economics*, forthcoming.

[29] DYNAMIT, 1999. Massachusetts Institute of Technology, Cambridge, Massachusetts. See its.mit.edu.

[30] H.S. Mahmassani, T. Hu, and R. Jayakrishnan. Dynamic traffic assignment and simulation for advanced network informatics (DYNASMART). In N.H. Gartner and G. Improta, editors, *Urban traffic networks: Dynamic flow modeling and control*. Springer, Berlin/New York, 1995.

[31] T. Schwerdtfeger. *Makroskopisches Simulationsmodell für Schnellstraßennetze mit Berücksichtigung von Einzelfahrzeugen (DYNEMO)*. PhD thesis, University of Karsruhe, Germany, 1987.

[32] A. Dupuis and B. Chopard. Cellular automata simulations of traffic: a model for the city of Geneva. *Networks and Spatial Economics*, forthcoming.

[33] G. D. B. Cameron and C. I. D. Duncan. PARAMICS — Parallel microscopic simulation of road traffic. *J. Supercomputing*, 10(1):25, 1996.

[34] H. A. Rakha and M. W. Van Aerde. Comparison of simulation modules of TRANSYT and INTEGRATION models. *Transportation Research Record*, 1566:1–7, 1996.

[35] R. R. Jacob, M. V. Marathe, and K. Nagel. A computational study of routing algorithms for realistic transportation networks. *ACM Journal of Experimental Algorithms*, 4(1999es, Article No. 6), 1999.

[36] M. Ben-Akiva and S. R. Lerman. *Discrete choice analysis*. The MIT Press, Cambridge, MA, 1985.

[37] J. A. Bottom. *Consistent anticipatory route guidance*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 2000.

[38] E. Cascetta and C. Cantarella. A day-to-day and within day dynamic stochastic assignment model. *Transportation Research A*, 25A(5):277–291, 1991.

[39] R. Palmer. Broken ergodicity. In D. L. Stein, editor, *Lectures in the Sciences of Complexity*, volume I of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 275–300. Addison-Wesley, 1989.

[40] B. Raney. Issues of feedback routing in iterated transportation simulations. *Proceedings of the Swiss Transportation Research Conference*, Monte Verita, Switzerland, March 2002. See www.strc.ch.

[41] J. Esser. *Simulation von Stadtverkehr auf der Basis zellularer Automaten*. PhD thesis, University of Duisburg, Germany, 1998.

[42] M. Rickert. *Traffic simulation on distributed memory computers*. PhD thesis, University of Cologne, Germany, 1998. See www.zpr.uni-koeln.de/~mr/dissertation.

[43] K.W. Axhausen. A simultaneous simulation of activity chains. In P.M. Jones, editor, *New Approaches in Dynamic and Activity-based Approaches to Travel Analysis*, pages 206–225. Avebury, Aldershot, 1990.

[44] S. T. Doherty and K. W. Axhausen. The developement of a unified modelling framework for the household activity-travel scheduling process. In *Verkehr und Mobilität*, number 66 in Stadt Region Land. Institut für Stadtbauwesen, Technical University, Aachen, Germany, 1998.

[45] H. Unger. An approach using neural networks for the control of the behaviour of autonomous individuals. In A. Tentner, editor, *High Performance Computing 1998*, pages 98–103. The Society for Computer Simulation International, 1998.

[46] H. Unger. *Modellierung des Verhaltens autonomer Verkehrsteilnehmer in einer variablen staedtischen Umgebung*. PhD thesis, Technische Universität Berlin, 2002.

[47] Chr. Gloor. Modelling of autonomous agents in a realistic road network (in german). Diplomarbeit, Swiss Federal Institute of Technology ETH, Zürich, Switzerland, 2001.

[48] S. Weinmann. *Simulation of spatial learning mechanisms*. PhD thesis, Swiss Federal Institute of Technology ETH, Zürich, Switzerland, in preparation.

[49] TRANSIMS, TRansportation ANalysis and SIMulation System, since 1992. See transims.tsasa.lanl.gov.

[50] A. Voellmy, M. Vrtic, B. Raney, K. Axhausen, and K. Nagel. Status of a transims implementation for switzerland, forthcoming.

[51] J. Esser and K. Nagel. Iterative demand generation for transportation simulations. In D. Hensher and J. King, editors, *The Leading Edge of Travel Behavior Research*, pages 659–681. Pergamon, 2001.

[52] K. Nagel and P. Wagner (editors). Special issue on regional transportation simulations. *Networks and Spatial Economics*, forthcoming.

[53] N. Cetin. personal communication.

[54] K. Nagel and M. Rickert. Parallel implementation of the TRANSIMS micro-simulation. *Parallel Computing*, 27(12):1611–1639, 2001.

[55] P. Gonnet. A thread-based distributed traffic micro-simulation. Term project, Swiss Federal Institute of Technology ETH, Zürich, Switzerland, 2001.

[56] http://pdswww.rwcp.or.jp/, since 1993.

[57] J.H. Kim (editor). Special issue about the first micro-robot world cup soccer tournament, MIROSOT. *Robotics and Autonomous Systems*, 21(2):137–205, 1997.

[58] D. Sternad. personal communication.

[59] J.D. Rothwell. *Control of Human Voluntary Movement*. Chapman and Hall, 1994.

[60] D. Helbing, F. Schweitzer, J. Keltsch, and P. Molnar. Active walker model for the formation of human and animal trail systems. *Physical Review E*, 56(3):2527–2539, 1997.

[61] D. Yamins, S. Rasmussen, and D. Fogel. Growing urban roads. *Networks and Spatial Economics*, forthcoming.

[62] H.Y. Lee, H.-W. Lee, and D. Kim. Origin of synchronized traffic flow on highways and its dynamic phase transitions. *Physical Review Letters*, 81(5):1130–1133, August 1998.

[63] D. Helbing and M. Treiber. Gas-kinetic-based traffic model explaining observed hysteretic phase transition. *Physical Review Letters*, 81(14):3042–3045, 1998.

[64] C. Gawron. An iterative algorithm to determine the dynamic user equilibrium in a traffic simulation model. *International Journal of Modern Physics C*, 9(3):393–407, 1998.

12 Economic networks

Alan Kirman

12.1 Introduction

Networks and network analysis play a central role in many disciplines but their role in economics is ambiguous. For many economists the study of networks is limited to the analysis of the functioning of physical networks such as the railway, the telephone system or the internet for example. Yet networks, in fact are much more fundamental and pervasive than this. Almost any serious consideration of economic organisation leads to the conclusion that network structures both within and between organisations are important.

Why should this be of interest to economists? The answer is not only that the result in terms of the state of the economy may be different than in the anonymous market situation but there is also something more profound. The relationship between the behaviour of individuals and the behaviour of aggregate variables will be different. Moreover that rationality which we attribute to economic individuals in order to justify and analyse the behaviour of aggregates may have to be modified. Thus what I will argue here is that we should, in fact, be interested in the passage from micro to macro but that this cannot be understood without taking into account the way in which individuals' decisions and actions are influenced by the networks of connections that link them to other agents. In other words, the organisation of the individuals in the economy plays a crucial role in explaining what happens at the aggregate level. Aggregate behaviour cannot be thought of as being like the behaviour of one individual. Furthermore, one will not, in general, be able to represent it as the behaviour of some average or representative individual. Just as neurologists would not think of explaining behaviour by studying the changes in a representative neuron nor should economists try to explain aggregate phenomena in this way.

This does not mean that one should not be interested in what happens at the micro level, but rather, that the passage to the aggregate level is mediated by the network structure in which individuals find themselves. Neurologists will continue to examine what happens at the molecular level but would not argue that there is some simple passage from that level to the aggregate activity of the brain which does not involve the network of interactions between neurons.

Of course, as economists, unlike neurologists, we are not making a descent as far as the level of the neurons of economic agents, but, to make another analogy, we would not expect to be explain how much food is stored by a colony of ants by looking at the behaviour of individual ants in isolation. The organisation of the ants plays an essential role. This example raises an important point. Far from complicating things taking direct account of interaction and the networks which organise it, actually makes life simpler for the economic theorist.

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

This is because the reasoning and calculating capacities we need to attribute to economic agents may be substantially less than it is in standard models. Individuals operating with simple rules in a limited context may, together, generate rather sophisticated behaviour on the aggregate level. In other words, aggregation itself may generate structure. Interaction and the networks through which it operates are important in determining aggregate economic phenomena and that this allows us to start from more plausible models of individuals than those that we normally use. If this is accepted, then we must first understand how networks influence aggregate outcomes. The next step is to understand how these networks form and if, and why, they persist. To many social scientists the interest of the problems examined here must seem evident and it is of some interest to look at the relationship between economic and sociological analysis since the latter tends to take it for granted that network structures are at the heart of the explanation of social phenomena whereas the former has attached rather little importance to them.

12.2 Economics and sociology

The idea that networks of relations at various levels have an important effect on economic activity is familiar in sociology. In that discipline a great deal of emphasis is put on the networks in which individuals are "embedded", (see e.g. Granovetter[1,2] (1985 and 1997) and White[3] (1981)). Lazega (2001), for example puts a lot of weight on the way in which individuals in an organisation are constrained by their relations with those who have come to be regarded as "powerful" figures. Thus, it is suggested that most economic mechanisms are a mixture of formal and informal rules or "conventions". The latter are thought of as emerging as a result of the network structure of the society in which the particular economic mechanism and the individuals who operate within it are situated. Consider the approach adopted by Baker[4] (1984) in his study of a major securities exchange and by Abolafia[5] (1996) in his analysis of the Chicago Commodities Exchange. Baker shows that the volatility of options prices is dependent on the type of network structure and, in particular, the size of the subset within which the agents in a market operate. Abolafia explains how informal rules emerge, are tested, and are consequently modified. The changes in rules after what are considered as abuses are testimony to this. The evolution of rules against "insider trading" and the reaction of the markets to the silver trading episodes of the 1980s are good examples. But what is the role of networks here? An important feature is that the way in which markets are organised into networks allows differential access to information. Whilst models with asymmetric information are widespread in economics little attention is paid to the origins of that asymmetry. Yet the fact that individuals operate within a limited network provides just such an argument. Indeed sociologists defend the idea

[1] Granovetter M, (1985) "Economic action and social structure: the problem of embeddedness", *American Journal of Sociology* 91:481-510

[2] Granovetter J, (1997) (forthcoming) *Society and Economy: The Social Construction of Economic Institutions.* Cambridge: Harvard University Press

[3] White, Harrison (1982), *Identity Control: A Structural Theory of Social Action. Princeton:* Princeton U Press

[4] Baker W.E., (1984), "The Social Structure of National Securities Market", *Amer J of Sociology AJS Vol 89 No 4 pp 775-*

[5] Abolafia M.Y., (1996), *Making Markets: Opportunism and Restraint on Wall Street*, Cambridge Harvard U P

that markets and market outcomes cannot be insulated from social structure because different social relationships will result in informational asymmetries, for example, which will provide some parties with benefits and leave others at a disadvantage.

As economists try to integrate network structures into their considerations they should perhaps be more inclined to take seriously the work of sociologists who have typically regarded networks of various sorts in society as being of great importance in determining how society behaves. However sociologists have, in general, explicitly rejected what has come to be regarded as the "ultra rational" behaviour attributed by economists to the individuals in society. Thus from their point of view the derivation by economists of the characteristics of aggregate situations from the rational behaviour of isolated individuals is at odds with the network approach. Whilst sociologists regard collective behaviour as the result of the interaction between individuals situated in relation to each other in social networks economists, with rare exceptions, have assimilated collective behaviour to that of an individual. Two things then seem to separate the approach of the economists from that of other social scientists. Firstly, there is the lack of any organisational structure in standard models and secondly there is the assumption of complete rationality.

Of course if we accept a more sociological view of the economy, we are faced with a standard but, nonetheless important, difficulty. Which part of the system can be considered as exogenous and which can be thought as endogenous. Social ties also have their "raisons d'être". However, at some point, one is obliged to limit the environment within which one works. But if one limits the scope too much then models become so "partial" that they lose any general relevance. Indeed, as mentioned, it was his dissatisfaction with the scope of economics that motivated Pareto's work on sociology. Ideally, one would like to have an explanation as to how social networks emerge, that is how the whole system organises itself. But even the problem of how economic links emerge is difficult and has received little systematic treatment.

12.3 The economic consequences of networks

Note that two claims have already been made for emphasising the importance of networks in economics. Firstly it is claimed that interaction other than through the market mechanism is important. This may be true even when there is no specific network structure. Indeed there are situations in which interaction is general in that no agent is linked to a particular collection of other agents. Even interaction of this sort, if it is not mediated by a market in which agents are isolated and anonymous can change aggregate outcomes. Consider, for example, the most extreme version of interaction, that studied by game theory Every player takes account of what every other player does and moreover knows that the others do so. The network of links between individuals is complete and, what is more, it is being fully and consciously used. This leads to two problems. Firstly the idea that everyone is taking into account the actions of others means that only limited examples can be fully analysed. Furthermore the assumption that every player is taking account of each others' actions and that he knows that the others are doing so and that they know that he is doing so and so forth leads to basic logical problems, (see, for example Binmore (1990).

There are thus two extreme approaches in economics. On the one hand there is the standard model where individuals are essentially independent, act in isolation and their activities are coordinated by market signals. On the other, there is the full game theoretic model in which individuals are completely interdependent but would have to be endowed with prodigious

powers of reasoning. Models in which agents are linked through a network of contacts lie between the two.

There are many economic situations in which there are links between some but not necessarily all agents. These are frequently referred to as models of local interaction. In order to discuss local interaction we must define what we mean by local and this, in turn, means that we must impose some structure on the space of agents. This will determine the distances between agents, which in turn will determine who is a neighbour of whom. The structure might depend on geographical distance as in locational models (see Gabszewicz and Thisse[6] (1986) for example), closeness in characteristics (see Gilles and Ruys[7] (1990)) or the potential gains from bilateral trade. In particular, we shall consider the interaction structure as being modelled by a graph where the agents are the nodes and two nodes are linked by an edge if the corresponding agents interact. In particular we will restrict attention to the class of undirected graphs, which says that if a interacts with b in some way then b interacts with a in the same way.[8] Within such a framework, it will generally be the case that individuals interact directly with those who are their neighbours, that is those who are near to them. But for this to be meaningful it is necessary to define a notion of "nearness". For example, in a graph, one might ask how many links are there on the shortest path from individual a to individual b, thus defining a "distance" between a and b. In this context, we might, for example, assume that each agent is influenced only by a limited (finite) number of other agents who are within a certain distance of him. Such individuals are usually referred to as the agent's "neighbours". At this point, once the graph is given the "neighbourhood structure" is defined a priori. However, it is important to bear in mind that a basic aim of the study of economic networks is to make this structure endogenous. Nevertheless, before moving to the problem of the formation of links, it is worth analysing the consequences of introducing the idea of local relationships among individuals.

Models with a network structure of interaction iallow us to analyse an important problem which is often alluded to in economics but not often analysed. As agents' interactions are limited to a set of neighbours, changes will not affect all agents simultaneously but rather diffuse across the economy. The way in which such diffusion will occur, and the speed with which it happens, will depend crucially on the nature of the neighbourhood structure. It is the connectivity of the graph of relations that will be essential here. In many cases, a specific structure is imposed and thus the connectivity of the graph will be determined exogenously. For example the most typical example of a graph structure used in economics is that in which agents are thought of as being placed on a lattice[9] and interacting with the individuals nearest to them. Alternative structures of links can be considered, (see Myerson[10] (1977).

---

[6] Gabszewicz J.J. and J.F. Thisse, (1986), "Spatial Competition and the Location of Firms", in *Location Theory, Fundamentals of Pure and Applied Economics*, Harwood Academic Publishers

[7] Gilles R., and P. Ruys, (1990), "Characterisation of Economic Agents in Arbitrary Communication Structures", *Nieuw Archief voor Wiskunde*, vol 8, pp. 325-345

[8] This, of course, excludes a whole class of interest where the link is active and then specifying the nature of the transaction. More difficult to handle is, for example, the transmission of information from a central source which is clearly best handled by considering the links as directed. Evstigneev (1995a and 1995b) has modelled some economic equilibrium problems using directed graphs but, as he indicates, the task is made difficult by the fact that some of the mathematical tools which make the undirected case tractable are not available for directed graphs.

[9] The lattice is usually taken to be one or two dimensional, (see Durlauf (1990), Benabou (1996), Blume (1993) and Ellison (1993)).

[10] Myerson R.B., (1977), "Graphs and Cooperation in Games", *Math Oper Res 2: 225-229*

In the deterministic case it is clear that, once the particular network of communication is established, analysis of many problems becomes straightforward. Suppose that the graph of relationships is given and the "distance" between two agents is defined as the number of links in the shortest path between the agents. The set of individual a's neighbours, $N(a)$, consists of all agents within some distance k of a. Suppose we are interested in the speed with which a signal is propagated through the population. As we have seen, this speed depends on the connectivity of the graph, a convenient measure of which is the "diameter" of the graph, i.e. the maximum of the distances over all the pairs of agents. In the usual lattice type of model, the diameter of the graph becomes very large as the number of agents increases. In the implicit graph associated with a non-cooperative game however, the diameter is one, that is every agent is in contact with every other one regardless of how many agents there are. Thus we can viualise three types of model already.

The classic Walrasian model in economics is one in which individuals are linked to some central figure often referred to as the Walrasian auctioneer who sends price signals and to whom they communicate their supplies and demands. This gives us a star shaped graph as in Figure 1.

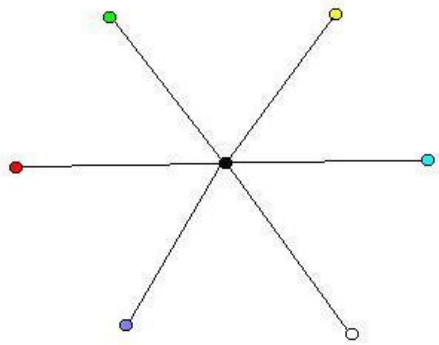


Figure 1

In many of the standard studies of interaction in economics a simple lattice is chosen and the neighbours of an individual are often considered to be the four agents directly linked with that individual. Thus in Figure 2 the blue individual has the red individuals as neighbours.

In the case of game theory all players in an n person game are linked to each other as in Figure 3.

Gilles and Ruys (1990), working within the deterministic approach developed by Myerson (1977) and Kalai, Postlewaite and Roberts (1978), adopt a different approach and use the topological structure of the underlying characteristics space to define the notion of distance. Thus agents nearer in characteristics communicate more directly. Translating this into a network structure means attributing links to those people who are closest in terms of some common characteristic. To an economist interested in how trading relationships are established, this seems a little perverse. In general, the possibility of mutually profitable trade increases with increasing difference in characteristics whether these be differences in tastes, endowments or abilities. Identical agents have no trading possibilities. However, there are other economic problems, the formation of unions and cartels for example, for which this approach may be

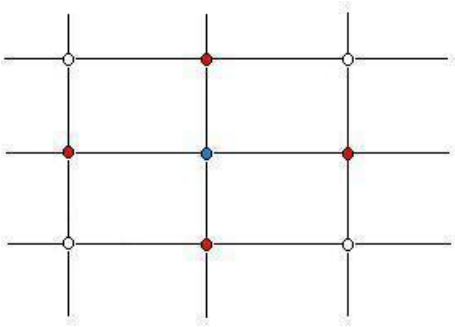


Figure 2

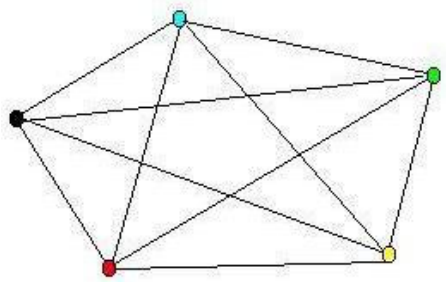


Figure 3

suitable. The important point is that the relational structure is linked to, but not identical with, the topological characteristics structure. Once again the main feature of such models is that the aggregate outcomes that emerge may depend crucially on the underlying communication structure

In the case where the interaction is stochastic, the problem becomes more complicated. To examine the consequences of interaction and to see what aggregate patterns will emerge one has still to specify the nature of the communication graph linking agents. As I have mentioned one might think of a lattice structure. In general, however, one can consider any network and define the "neighbourhoods" accordingly. However the determination of how individuals depend on others is now random. There are two ways of viewing such stochastic interaction.

12.4 Fixed network: stochastic interaction

A first is to consider the graph of communication as given but to assume that the dependence of the state of any individual on his neighbours is stochastic. Thus, once the graph is specified, neighbourhoods are defined and one assumes that the agent a is influenced by his neighbours in $N(a)$ in a probabilistic sense, that is, the probability of the agent a being in a state s is conditioned by the states of his neighbours.

Thus one specifies the probabilistic dependence between individuals and then sees what may happen at the aggregate level. The question now is can we assign probabilities to each state of the whole system in a way which will be consistent with the individual probabilities? Suppose that we can do this. Furthermore suppose that all the individuals are identical in that their local probabilities are the same. Follmer[11] (1974) explains that two disturbing things can happen. There may be more than one aggregate probability law consistent with the individual specifications, and worse some of these may not treat identical agents symmetrically. The latter case in which the conditional aggregate probabilities may vary from agent to agent, is called a symmetry breakdown.

Föllmer's contribution shows that, even with a completely symmetric local structure, non-homogeneities may arise at the aggregate level. Think of an economy with a large number of agents each of whom draws his characteristics, for example his resources and endowments, from the same distribution and each of whom depends on his neighbours in the same way. One might hope that some sort of "law of large numbers" would apply and that with enough individuals the prices that constitute an equilibrium would be independent of the particular draws of the individualowever, in the case in question, even in the limit it may not be possible to specify what the equilibrium prices of such an economy will be. If interaction is strong enough, no heterogeneity is required at the individual level to produce these difficulties. Put another way, if agents are sufficiently dependent on their neighbours, aggregation does not remove uncertainty. Strong local random interaction amongst agents who are a priori identical may prevent the existence of equilibrium. This is very different from models with randomness at the individual level, but where the uncertainty is independent across individuals. Thus even with a highly symmetric underlying micro-structure aggregate behaviour is unpredictable with sufficient interaction and a priori identical agents may not be treated identically. The aggregate patterns which emerge simply do not reflect the symmetry of the microstructure. Thus the very fact that the probabilistic interaction is mediated through a network structure and that dependence is limited to neighbours is enough to change the relation between individual and aggregate outcomes. Moreover it is not the particular structure of the network that is important, even the simplest lattice structure will generate these problems.

Föllmer's random field approach has been extended recently to consider dynamic stochastic processes. Indeed much of the recent work on local interaction and interdependence has examined the evolution of behaviour of interacting agents. The specification of the local interaction plays an important role in characterising the nature of aggregate behaviour that will emerge over time**.** The clearest examples of this are the articles by Blume[12] (1993) and Durlauf (1997) which both illustrate the use of the model adopted by Föllmer. Blume looks at a situation in which individuals play a game against their neighbours. Here the structure of links is fixed and the results of the interaction between players depends on their choice of strategy which players modify over time. The probabilities of playing different strategies are modified according to the difference in performance between a player's own strategies and those played by the opponents. If the parameter governing the revision is very large the process is simply the "best response" process. For smaller values of the revision parameter however, the population may fluctuate between states and may have a limit distribution. The

[11] Föllmer H., (1974), "Random Economies with Many Interacting Agents", *J. Math. Econ.* **1(1)**: 51-62

[12] Blume L., (1993), "The Statistical Mechanics of Strategic Interaction", Games and Economic Behaviour, 5, pp. 387-424

sensitivity of the adjustment of players' strategies to those of their neighbours determines the nature of the aggregate outcomes.

Another example of local interaction with a fixed communication structure is that given by Ellison (1993). He compares a model with global interaction with one in which players are situated on a circle and play against one of their neighbours. Here again the graph structure is very specific and is not itself stochastic. However, there are mutations in individual behaviour which means that from time time to time players randomly make mistakes. In this example the pairs of players play a simple coordination game, which has two equilibria, one of which dominates the other. In both cases the system fluctuates between coordinating on the "good" and "bad" Nash equilibrium. However the switch to the "good" equilibrium is achieved much more easily in the local case.

12.5 Random graphs and networks

An alternative approach is to consider the graph structure itself as random[13]. Suppose that once the communication network is established, a corresponding equilibrium notion is defined. Then one can study the characteristics of the equilibrium that arises from the particular structure one has chosen, since the agents will be constrained in their choices by the network. However, the network, or the links used, will be be a particular realisation of a random drawing and, in consequence, the outcome will, itself, be random. In the basic Markov random field model, interaction is probabilistic but between neighbours in the lattice or more general graph, and the graph structure is given exogenously. Haller (1990) for example, links the deterministic Gilles and Ruys approach to the stochastic one by basing the probability that communicational links exist, directly on the topological structure of the attribute space. However, the stochastic graph approach allows for more complicated neighbourhood structures than this, e.g. by permitting agents, who are not "near" in terms of some underlying characteristics to have a positive probability of communicating.

In the stochastic graph approach, the basic idea is to consider a set of agents A and to attach probabilities to the links between them. Let p_{ab} denote the probability that individual a "communicates" with agent b. In graph terms, this is the probability that an arc exists between nodes a and b. The graph is, as before, taken to be undirected, i.e. the existence of the arc ab implies the existence of ba and thus one way communication is ruled out.[14]

In the case of a finite set A this is easy to envisage. The resulting stochastic graph can be denoted

$$\Gamma(p_{ab}). \tag{12.1}$$

[13] This was introduced by Kirman (1983) and developed by Kirman, Oddou and Weber (1986), Durlauf (1989) and Ioannides (1990)

[14] This, of course, excludes a whole class of interesting economic models in which the direction of communication is important. Much of production involves transactions which are necessarily in one direction from inputs to outputs, for example. This can be taken care of by specifying wheher the link is active and then specifying the nature of the transaction. More difficult to handle is, for example, the transmission of information from a central source which is clearly best handled by considering the links as directed. Evstigneev (1994) has modelled some economic equilibrium problems using directed graphs but, as he indicates, the task is made difficult by the fact that some of the mathematical tools, which make the undirected case tractable, are not available for directed graphs.

If there is no obvious underlying topological structure which is thought of as affecting the probability that agents communicate, then, one could consider $p_{ab} = p$ that is the probability of interaction is the same regardless of "who" or "where" the individuals are. Thus global interaction is a special case of this model.

The graph representing the links through which interaction takes place may, as mentioned previously, become surprisingly highly connected as the number of agents increases, provided that the probability that any two individuals are connected does not go to zero too fast. To understand what is meant by this, consider a result of Bollobas which was used by Kirman, Oddou and Weber (1986) to prove their main result. He shows that if the probability that any two agents know each other in a graph with n nodes, p^n_{ab} is greater than $\frac{1}{\sqrt{n}}$ then as n becomes large it becomes certain that the diameter of the graph, $D\left(\Gamma^n\left(p^n_{ab}\right)\right)$ will be 2. More formally,

$$\underset{n\to\infty}{Lim}\ Prob\left(D\left(\Gamma_n\left(p^n_{ab}\right)\right) \geq 2\right) = 1. \tag{12.2}$$

In other words it is sure that any two individuals will have a "common friend" if the graph is large enough. Thus, as was observed in Kirman et al. (1986), one should say on encountering someone with whom one has a common friend, "it's a large world". This somewhat surprising result suggests that, as sociologists have long observed empirically, relational networks are likely to be much more connected than one might imagine. This is important in economic models, since the degree of connectedness determines how fast information or a "technological shock" diffuses and how quickly an epidemic of opinion or behaviour will occur. Thus the aggregate consequences of a local change will be very much linked to the structure and degree of connectivity of the network through which information flows.

It is important to note that the result just evoked depends crucially on the fact that the actors in the network are linked with uniform probability or, slightly more generally, that the pair of agents with the lowest probability of being linked should still be above the lower bound mentioned. This "small world" problem is rather different to that studied by Watts (2000) but he looked at networks in which the links are changing stochastically and we will come back to this. The dynamic evolution of the state of the individuals linked in a graph like structure is particularly interesting since the stable configurations of states, if there are any, will depend on the graph in question and some of the results from other disciplines (see Weisbuch [1990]) can be evoked in the context of economic models.

In this context it is interesting to examine what happens when, although agents modify their behaviour in the light of their own and their neighbours' experience, the consequences of their behaviour may affect other agents further afield. Weisbuch et al. [1994] for example show how agents may chose polluting or non polluting devices from local experience but their choice may result in pollution which diffuses widely. The consequence of this may be rather sharp division into areas in which all the agents have adopted one type of device while in another area the alternative device will be used.

12.6 Emerging networks

Up to this point economic activity can be considered as being organised with a network structure which is given exogenously. However one of the most interesting challenges in examining

an economy is to include the evolution of the network structures themselves. If one wants to proceed to a theory of endogenous network formation, a first step might be to find out which organisations of individuals are stable. Thus one would look for "rest-points" of a dynamic process of network evolution. Such rest points would be arrangements, or networks, which would not be subject to endogenous pressures to change them. This, as it stands, is not a well formulated concept. More of the rules under which agents operate have to be spelled out. The dynamics of the system will be defined by the way in which links are formed. Two choices in economics have been to consider links as being consciously chosen by individuals while the alternative has been to allow the links to be created as a consequence of the advantage that individuals have accrued from using them. To understand the problems involved, consider first situations in which agents consciously choose their partners

12.7 The strategic formation of networks

A substantial literature has developed in which economic agents deliberately and strategically choose their partners. Thus the strategy of a player will be those players to whom he wishes to be connected. Of course since links are undirected there has to be a rule which determines what happens when the choices are inconsistent. Once such a rule is given one can then see which choices of strategies by individuals constitute an equilibrium. To give a specific example, consider a cooperative game. Think of a partition of the set of players into coalitions and ask whether some other coalitional arrangement might "Pareto improve" the lot of the players. If this were the case, the coalitional structure in question would not be considered as stable. Such problems have been studied extensively since the basic article by Aumann and Dreze (1974). Thus a stability condition is imposed on the set of partitions. We might however, wish to consider the communication structure of the economy as being described by a graph, and then to pose the analagous question as to which graphs would be immune to change? By this, I mean that if agents have the possibility to add or remove links, which graphs would remain unchanged? This is a question which has been studied by Jackson and Wolinsky (1996) and others. They consider models in which links are valuable to players since they connect them directly and indirectly to others. However they may also be costly to form or maintain. They assume that the greater the distance one player is from another, the lower the corresponding utility that their being linked gives. They then examine the characteristics of stable graphs and this provides a reference point against which to set the rest-points, if any, of a model of evolving networks. Here the notion of equilibrium is that of a Nash equilibrium and what one is looking for is a graph structure where no individual would like to change his choice of partners given the choices of the other players. Jackson has contributed a number of papers in this direction as has Goyal. Their aim is to characterise the structures of equilibrium graphs. Thus what one is trying to establish is the characteristics of graphs in which the links are chosen consciously by the players.

To be more specific, consider an example given by Bala and Goyal (2000). The idea is that individuals can profit from the information that others obtain from those to whom they are linked by creating such links themselves. There is a cost to creating a link but one gets the benefit of the knowledge flowing through the network to which one is connected subject to some delay or decay. The architecture of equilibrium networks if knowledge is widely

dispersed is striking. The networks will be of a star form with one agent at the centre and the others linked directly to her. This is, of course precisely the same form as that of the competitive market mentioned earlier. Two characteristics of such graphs are important. They display, by definition, a high degree of centrality and they are very highly connected. The diameter, or longest distance between any two participants is two. In this example only those who create links pay for them and so nobody pays for more than one link and everyone is closely linked with everyone else. A difficulty with such structures and one that has been noted in the context of transportation networks is that they are very vulnerable to the breakdown of any link. Each individual is dependent for all his information on one link to one player.

A second example is due to Goyal and Joshi (2002) and studies an oligopoly model in which firms can collude with others. Firms engage a certain amount of resources in the pairwise arrangements and in return have lower marginal costs. The firms then compete by setting quantities. It turns out that if marginal costs diminish by a constant amount for each link formed then, given the type of competition, there are increasing returns to link formation. The resultant architecture of the graph is interesting. It is formed of a dominant group or clique and a number of isolated players. This configuration is familiar from the work on coalition formation, (see Bloch (1995) and Thoron (2000) for example. It is also characteristic of certain stochastic graphs. What is intriguing is that we can see an asymmetric structure form even though all the players were symmetric to start with. If we return to the example and now allow firms to make side payments to other firms. In this case the star form re-emerges since one player at the centre can benefit from the increasing returns and can make payments to the others.

Jackson and Watts (2000) have also contributed to this literature by examining the equilibrium structures of graphs formed strategically. A number of problems remain in this area since one often assumes that players choose their links according to some pre-established protocol and this protocol is ad hoc. Furthermore, the pay-offs to the players depend on the structure of the graph but the way in which this dependence is determined is often not specified. All that one knows is that each graph structure attributes a payment to each player.

Although, in a certain sense, graphs emerge in this context they are static in the sense that the pay-offs are known and once formed they are fixed. Furthermore, no learning is involved, all the players know everything from the outset and the graphs that are stable are determined by that knowledge.

12.8 Emerging random graphs

This brings me to the most interesting challenge in this area, which is to study the evolution of the communications graph itself. Durlauf [1990] makes a step in this direction when he starts with a given geographical network, but allows agents to choose where to place themselves in the network. This recalls an older model of neighbourhood preferences due to Schelling. However, recent work by Watts (2000) has shown how structure may emerge as links are replaced by other links. What he examines is a situation in which agents have a fixed number of links with others and, for example they are all situated on a circle and are just linked with their immediate neighbours. Then he draws one of the existing links at random and replaces it with a new link. This may be to any agent, in particular one to whome the distance was

great in the original graph. Adding such links drastically increases the connectivity of the graph. This procedure is repeated and what emerges is a typically clustered structure in which closely linked individuals in a small group are linked to other groups through one or two links. In a pure random graph distances are short as we have seen but there is almost no clustering. However, in the sort of small world graphs studied by Watts the situation is different, distances are still short but there is a great deal of clustering. Here the few long links between different groups keep the distance down but most of the interaction is within small groups. This sort of clustering is observed in many empirical situations in economics. Potts (2001) has argued that the small world structure is consistent with the idea that knowledge is concentrated within entities such as firms but is exchanged with other firms through the long links provided by markets. What is lacking for the economist in these models is any explanation of why links are being drawn. If an economic agent is to be linked to another one would like to be able to give some reason for that happening.

As has already been mentioned, an obvious way in which to proceed is to specify models in which the links between agents are reinforced over time by the gain derived from those links. Thus longstanding economic relationships would be derived endogenously from the agents' experience. Instead of thinking of learning only at the individual level the economy as a whole could be considered as learning and the graph representing the communication in the economy as evolving[15].

A paper by Vriend (1994), presents a first step to simulating a model in which either the links themselves or the probability that they will be used over time evolve. He considers a market in which buyers learn where to shop and firms learn, from experience, how much to stock. Thus, the network here consists of the links between a set of buyers and a set of sellers. The interesting question is how will this network evolve over time? In this model firms sell indivisible units of a homogeneous good, the price of which is fixed and agents demand at most one unit. This is the most rudimentary model possible because the only criterion for success that the buyers have is whether they are served, and sellers are only concerned with providing the correct amount to satisfy demand. Nevertheless it is particularly interesting to note the development and persistence of a non degenerate size distribution of firms even though all firms are identical to start with. Furthermore some buyers always return to the same store, whilst others continue to search. There is empirical evidence for this sort of division of activity both on product markets and in financial markets. Thus, in Vriend's model, relationships between traders do evolve over time and a number of stable bilateral arrangements emerge. Vriend adopts what has come to be called the "artificial life" approach, that is, his agents are initially totally ignorant and merely update the probability of performing actions given the rewards that those actions generate[16]. They do this by using conditional rules, consisting of an "if ... then" statement, and then modify the choice of rules in the light of experience.[17]

In an extension to Vriend's original model, Kirman and Vriend (2001) wished to see whether making the behaviour of the buyers and sellers more elaborate would change the sort of network that emerges. They consider individuals who make more than one encounter in a trading day. Sellers now set prices they charge to each of their customers and allow the

[15] For a discussion of the formal problem of evolving networks see Weisbuch [1990].

[16] Tesfatsion (1995) gives a full account of the artificial life approach to economics and its merits and drawbacks.

[17] This is the so-called "classifier" approach introduced by Holland (1992).

latter to choose whether to accept or refuse these prices. Here the number of rules to choose from is vastly greater than in the original model and this poses considerable problems if only for computational reasons. Nevertheless, individual buyers in the model soon learn which prices to accept and which to reject. Furthermore, sellers start to discriminate between buyers and charge their loyal customers different prices than those set to "searchers". Interestingly, some sellers set high prices to buyers and give them priority when there is insufficient stock to serve all the customers. Others do the opposite, giving low prices to loyal customers but serving searchers first at higher prices. Although the former strategy yields higher profits, individuals who adopt the low price strategy for their loyal customers get "locked in" and are unable to learn their way to the alternative strategy. Thus a "dominated" type of behaviour coexists with a superior one.The outcome of the process through which the market organises itself would have been impossible to predict by looking at the individuals in isolation.

In another example of a model in which links evolve, Tesfatsion (1995) considers individuals who have the possibility of accepting trading opportunies with other traders[18]. Trading corresponds to playing a repeated prisoner's dilemma game.However, players can choose whether to play against other players or not on the basis of expected payoffs. These payoffs are based on previous experience with that particular partner, thus the capacity to identify those with whom one interacts is crucial. Positive probability is assigned to being matched with "acceptable" players, while the links between those couples of which one member finds the other unacceptable are assigned probability zero. The criterion for accepting another player is whether the expected payoff from playing against that individual is above some threshold level. The possibility of rejecting partners can lead, as Tesfatsion shows, to the emergence and persistence of groups with different pay-offs.

These results are, like most of the work in this area, obtained from simulations and are thus open to the standard criticisms. What one really wants to obtain are analytic results, providing a benchmark against which to measure the results of such simulations. The basic problem and the approaches to analysing it can all be posed within a simple framework. What we are interested in, is the evolution of the graph representing the interaction beween the different individuals in the economy. The problem is to move on from the previous specification of a random graph to one where that graph itself changes over time. Thus the underlying idea is to look at how the probability that certain links between agents will be used evolves over time. What should become clear is that once again, we are back to the idea that the dynamics of the process we are interested in can be seen as the evolution of probability distributions over time. What is interesting is that, even when, as now, we are interested in the specific links between agents, the same general framework is applicable.

What we wish to do in the present case is to model an evolving graph for a fixed population of size n. One way of doing this is to consider the set of all possible graphs, that is the 2^{n2} $n \times n$ incidence matrices and to define a probability distribution over them. Thus a random directed graph is nothing other than a point in the unit simplex S in $\mathbb{R}^k$ where $k = 2^{n2}$ with the appropriate reduction in dimension for an undirected graph since the matrix is then symmetric[19]. The evolution of the random graph is then described by a mapping from S into

[18] This is a development of the type of model developed by Stanley et al. (1994).

[19] This is more general than the previous way of specifying random graphs which was to draw each of the links a, b with a probability p_{ab}.

S and the dynamics will then be determined by the particular form of learning used to update the probabilities attached to the links. The important point to notice here is that by moving to a specification which involves all the links we greatly increase the size of the space in which we work but we do not fundamentally change the nature of the framework.

To see what is going on observe that a vertex of the simplex corresponds to a deterministic network whilst the barycentre corresponds to the uniform probability model which in turn is the same as the situation in which all agents are linked to each other with probability 1/2. Now the question of how networks evolve can be seen in this general framework. Different models yield different ways of modifying the probability distribution corresponding to the communication network of the economy. The dynamics engendered by the different specifications may yield very different results. In some cases there may be convergence to a deterministic graph, in others, there may be convergence, not to a vertex, but to some other particular point in the simplex. In this case the network will not be deterministic but the probabilities with which the links are used will remain constant over time. It may well also be the case that the dynamics never converge and the probabilities in question change all the time. Careful specification of the updating mechanism will reveal how the resulting network evolves. Thus, depending on the particular specification one chooses one should be able to observe the evolution of trading groups and partnerships in markets and the development of groups playing certain strategies amongst themselves in repeated game models for example. It is, of course, the case that the evolution of purely deterministic graphs is a special case of the model just discussed.

An obvious extension, already mentioned, of this analytic framework is to consider agents as having several different types of functional links. They might be linked with fellow workers in a firm, with trading partners, or with members of a household for example. However the analysis of this sort of multi-layered graph seems to be rather intractable. Therefore, I will briefly examine a number of models in which the links between agents can be seen as evolving within the simple framework just outlined.

To see how the approach just suggested translates into more concrete form, consider a model of a simple market similar to that already mentioned of Kirman and Vriend (2001). Weisbuch et al. (2000) consider a wholesale market in which buyers update their probability of visiting sellers on the basis of the profit that they obtained in the past from those sellers. If we denote by $J_{ij}(t)$ the accumulated profit, that buyer i has obtained from trading with seller j up to period t, then the probability p that i will visit j in that period is given by,

$$p = \frac{e^{\beta J_{ij}}}{\sum e^{\beta J_{ij}}} \quad , \tag{12.3}$$

where β is a reinforcement parameter which describes how sensitive the individual is to past profits.This non-linear updating rule will be familiar as the logit decision, or the quantal response rule. It was developed in statistical physics and has been widely used in economics (see e.g. Blume (1993), Brock and Durlauf (2001a), and Anderson et al. (1992)).

If we consider the graph of probabilistic relations between individuals then what one is doing is seeing how the probabilities attached to each link between a pair of traders evolves over time. What is important is that not all graph structures will be attractors and we can say something quite clear about those that will emerge.

Weisbuch et al. (2000) show that either buyers will become loyal to one seller, (ordered behaviour), or they will shop around with almost uniform probability, (disordered behaviour).

Which of these two situations will develop depends crucially on the parameters β and γ and the profit per transaction. The stronger the reinforcement, the slower the individual forgets, and the higher the profit, the more likely it is that order will emerge. In particular the transition from disorder to order, as β changes, is very sharp.

Given this rule one can envisage the case of 3 sellers, for example as corresponding to the simplex in figure 1 below. Each buyer has certain probabilities of visiting each of the sellers and thus can be thought of as a point in the 3 simplex. If he is equally likely to visit each of the three sellers then he can be represented as a point in the centre of the triangle. If, on the other hand, he visits one of the sellers with probability one then he can be shown as a point at one of the apexes of the triangle. Thus, at any one point in time, the market is

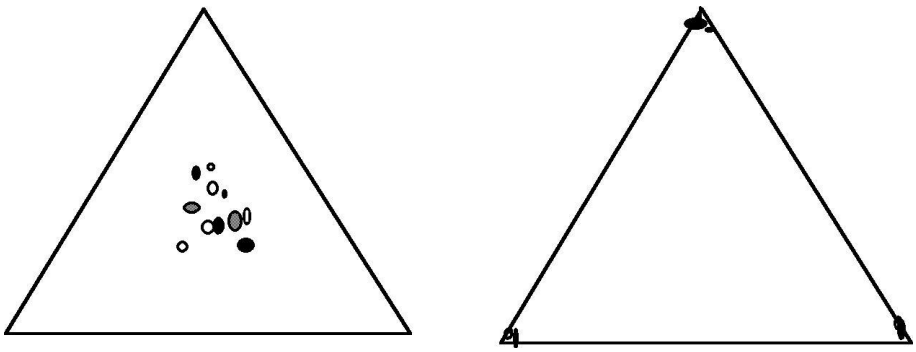


Figure 4

described by a cloud of points in the triangle and the question is how will this cloud evolve? If buyers all become loyal to particular sellers then the result will be be that all the points, corresponding to the buyers will be at the apexes of the triangle as in figure 1a, this might be thought of as a situation in which the market is "ordered". On the other hand, if buyers learn to search randomly amongst the sellers then the result will be a cluster of points at the centre of the triangle,as in figure 4b. In network terms the first case corresponds to a situation in which every buyer has a link with every seller and in the second case each buyer has a link to only one seller. Which of these situations will develop depends crucially on the parameters β, γ and the profit per transaction. The stronger the reinforcement, the slower the individual forgets and the higher the profit the more likely is it that order will emerge. In particular, as mentioned, the transition from disorder to order is very sharp with a change in β.

In the Weisbuch et al. model this sort of "phase transition" is derived using the "mean field" approach. The latter is open to the objection that random variables are replaced by their means and, in consequence, the process derived is only an approximation. The alternative is to consider the fully stochastic process but this is often not tractable, and one can then resort to simulations to see whether the theoretical results from the approximation capture the features of the simulated stochastic process.

Why is it important to understand the nature of the trading relationships that emerge? The answer is that the aggregate efficiency of the market depends on them. A market where agents are predominantly searchers is less efficient than one with many loyal traders. When searching is preeminent, sellers' supply will often not be equal to the demand they face. Some buyers

will be unsatisfied and some sellers will be left with stock on their hands. This is particularly important in markets for perishable goods.

To illustrate the evolution of the networks involved in this model, it is worth looking at figures 5 and 6. As before we are looking at a situation with three sellers and thirty buyers and the evolution of the loyalty of the individuals and the group can be seen as time passes.

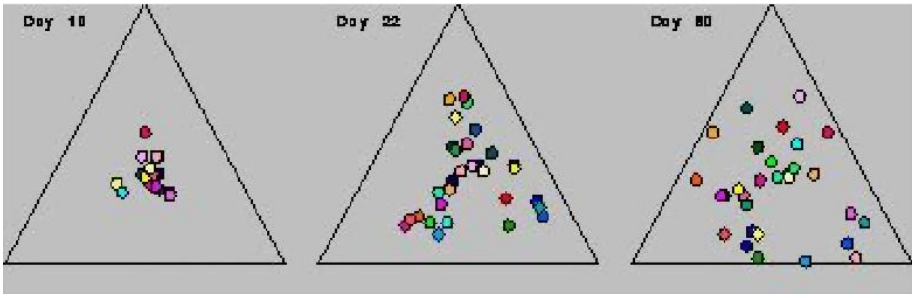


Figure 5

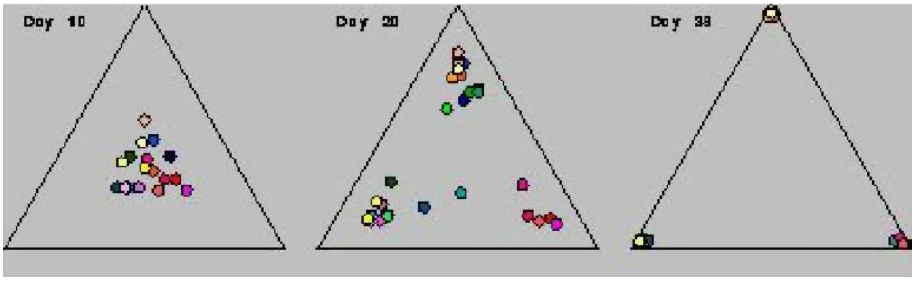


Figure 6

What was done to produce these figures was to simulate the actual stochastic process to see whether the approximate results obtained in the deterministic approximation still held up in the actual process. Thus at each step the market participants decide which rule to use according to its success and then form their offer prices, for the sellers, and reservation prices for the buyers. This, in turn determines the profit the buyers and sellers make in that period and this conditions the buyers' new probability of visiting each seller. They draw a seller according to these probabilities and the process starts again. The only way in which the two figures above differ is in the value of beta. Figure 5 shows the situation in which beta is small and below the critical value. The graph corresponding to this situation is one in which each buyer visits several sellers with a positive probability. Although each agent changes his probability of visiting the sellers they are essentially still in the centre of the simplex.

In Figure 6 however, beta is above the critical value and buyers rapidly become loyal to a single seller. Thus the graph is essentially deterministic and is characterised for each individual by single links with one seller. Thus the graph structure is very different in the two cases.

As has already been observed the deterministic graph is more efficient in terms of resource utilisation. Buyers know where to find what they want and sellers know how much to supply. In the Marseille wholesale fish market where the model was tested those buyers who have high

betas with respect to the critical value are those who come often to the market and who make large profits. The empirical evidence from analysing the data from every single transaction over three years reveals that this is precisely the case, frequent buyers and those who buy large quantities are, indeed, the most loyal.

Yet there is one thing that is unsatisfactory with the way in which the model was simulated to show this. Why do buyers with high betas always find what they want in the simulated model? It may be because there are no random buyers to interfere with their learning process. Yet what we are claiming is that in a market where both types of buyer are present those with high betas develop deterministic links while the low betas continue to shop randomly. Our simulations are of markets with all traders with high betas or all traders with low betas. Might it not be the case that when mixed our results no longer hold?

To test this we ran a simulation with a mixed population. Half of them had betas above the critical value and half below. The result can be seen in Figure 7.

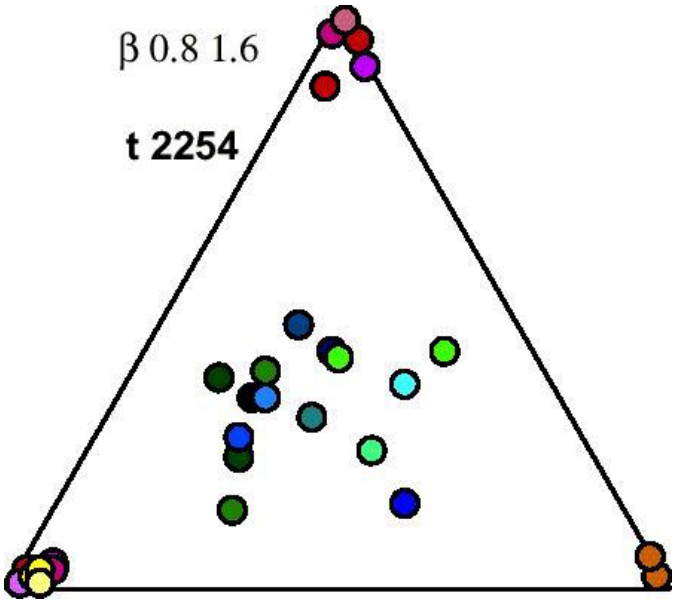


Figure 7

Here the brown, red and purple circles represent high beta buyers while the green and blue individuals have low betas, the critical value being one. There is clearly a separation and if the situation were represented as a graph one would have a mixture of links with probability $p_{ij} = 1$ and others with almost $p_{ij} = \frac{1}{3}$ but none with intermediate values.

This recalls an earlier model of Whittle (1986) where there are two sorts of individuals, farmers and traders. Under certain conditions markets may emerge with well structured relationships between buyers and sellers where previously there were only itinerant traders.

An interesting aspect of these models is the similarity of the underlying formal structure, even if the economic models are rather disparate. A nice example is provided by Paul Krugman's (1991, 1994) seminal work on economic geography. He considers a world in which

there are two activities, manufacturing carried out by mobile workers and agriculture, undertaken by immobile workers. Manufacturing has economies of scale, and incurs transport costs. The latter tend to make for concentration. On the other hand agriculture exerts a centrifugal influence. The evolution of the model involves the dynamics of the allocation of manufacturing workers to different locations. In network terms what one is looking at is the flow of workers from one place to another over time. This means that one can assign weights to the links and if the system converges to a stable distribution the network will be composed of a number of clusters each of a certain size. This requires again, as in the earlier models, examining paths in the n simplex where n is the number of the locations. Thus the problem can be analysed in the same framework as before. It is only the nature of the mapping that has to be redefined. Krugman's models specify some laws of motion in which workers tend to move to places with higher wages, and he studies the resultant dynamics. It is particularly interesting that a great deal of order emerges from an initially random situation. The structure of the distribution of "cities" has quite strong properties and is largely independent of initial conditions. The particular equilibrium distribution may well depend strongly on those conditions, but the general features of its structure may be independent of history.

A contribution which analyses the evolution of historical networks in an explicit way is that by McLean and Padgett (1995). This is related to Weisbuch et al. (2000) and to an early study by Cohen (1966). McLean and Padgett study data on transactions in Renaissance Florence and look at which markets were disordered and which were ordered, in the sense that order corresponds to an essentially deterministic graph in which each buyer is linked with one seller and disorder corresponds to each potential link having the same probability. As the theory developed by Weisbuch et al. predicts, the two markets, those for banking and wool which were the most active and generated the greatest profits were ordered whereas the others were disordered. However, the interpretation by Mclean and Padgett is worth examining. In their vision, disordered markets correspond to perfectly competitive ones since, if every buyer has the same probability of going to any seller then the authors claim that the same price will prevail everywhere. In this case, individuals have no incentive to privilege any particular seller.

Yet, there is another feature of the market which has to be taken into account. This is the inefficiency of the purely random market. As has been mentioned, if agents search at random, there will always be sellers in excess supply or demand, and there will also be unsatisfied buyers. Therefore, even in the case of durable goods, there will either be unnecessary stocks of goods or rationing. Buyers in ordered markets have learned to take into account the possible lack of goods, and their loyalty leads to a more efficient matching of demand and supply. Thus, even with a uniform price, there is pressure for loyalty to emerge. Thus, to describe the random market as reflecting competitive pricing does not seem to be justified. In fact, one might have the opposite view. Switching between suppliers must involve the risk of not being satisfied. Therefore buyers will only "shop around" if there is known to be a non degenerate price distribution. Mclean and Padgett's argument could only be justified if they produced data to show that the "law of one price" held in disordered markets. Yet, as they point out, no data are available on the prices at which the transactions were made in the Florentine markets.

A particularly intriguing, example is given in a paper by Guriev et al. (1995). Their aim is to see how the development of trading structures or graphs of links between the actors in markets influence the allocation of resources in an economy, a phenomenon of particular im-

portance for so-called "transition economies". They have a model with producers who are sellers, consumers who are buyers and traders who may be on either side of the market. One can then look at the trading structures that emerge recalling tht agents can choose whether to be producers or traders. There are costs to trade which depend on the pair of individuals involved and capture difficulties due to poor "infrastructure", whether this be due to communications and other factors, or to spatial distance. Sellers adjust quantities fast but prices slowly, and buyers adjust the probabilities of visiting sellers slowly but the quantities they purchase rapidly. Traders vary the probability of moving to one or the other side of the market in order to maintain a trade balance in the short run and modify prices to maximise profit. The basic adjustments in the model are of parameters of the stochastic processes which govern the shift from one activity to another, e.g. from producing to selling. This in turn governs the flow of basic demand and of supply. Guriev et al. show that, if trading costs are low, the market is close to being competitive whilst, if such costs are high, the traders exhibit monopolistic behaviour and earn high profits. However if trading costs are intermediate, there is a phase transition and complex dynamics are observed with price oscillations, bursts of shortages and long chains of traders. The distances in the graph in the latter case are much higher than in the competitive situation. To quote them,

> "Interactions among the agents at a micro level bring about implicit global phenomena at the macroscopic level, i.e.epiphenomena, for example organization of stable structures in trade networks, or, existence of stable cycles in the dynamical behaviour of the system."

There are many other examples of how graph structures can have an effect on economic outcomes and a number of these are discussed in Casella and Rauchs (2001). In particular, there are examples in which agents generate positive and negative externalities for each other and the activities of individuals will be highly influenced by those around them. Examples range from technological spillovers[20] to pollution, and from imitative behaviour to the contagion effects that can arise in financial markets[21].

12.9 The identification problem

In all of the preceding discussion it was taken for granted that the network structure of economic activity does have an influence on the economic outcomes. In theory this is what is being assumed and many references to observed collective behaviour such as that of fads or fashions or imitation are based on the premise that the influence of networks is important in determining such phenomena. However, as Manski (1995) has pointed out the situation is complicated when it comes to testing such propositions empirically. Manski specifies three different explanations for correlated behaviour in groups or networks.

[20] The sort of model discussed by Arthur (19) and David (19) can be easily extended to examine the diffusion of technologies through a graph structure.

[21] The models of financial markets in which agents are led to change forecasts such as those of Lux and Marchesi (19) and Brock and Hommes (2000) can be interpreted as networks in which individuals are linked to "gurus" who provide them with forecasts and who change their links as a function of the success of the rules, (see Kirman (2000)).

contagious effects, wherein the propensity of an individual to behave in some way varies with the behaviour of the group, *exogenous (contextual) effects*, wherein the propensity of an individual to behave in some way varies with the exogenous characteristics of the group, *correlated effects* wherein individuals in the group have similar individual characteristics or face similar institutional environments"

What we are assuming in many of our economic models is that agents are influenced by their neighbours either by their choices, their expectations or by some other form of direct interaction. How can we distinguish what appears to be a clear case of contagion from a situation in which individuals choose independently but share some characteristics which makes them choose similarly?

The answer is, as pointed out by Brock and Durlauf (2001b), that it is very difficult and that we are in danger of overemphasising network effects. There is therefore a need to be prudent before going down the road of the sociologist Luhmann who implies that individuals do not matter and that activity can be interpreted through the nework alone.

Perhaps what is most lacking in economics is good empirical evidence for the importance of network effects rather than anecdotes. There are isolated examples of such work such as that by Burt (1994), Glaeser et al. (1996) and Granovetter (1985) but this lies at the frontier with sociology and there is relatively little work to examine the results of local interaction mediated through networks. Much of this type of work is surveyed by Brock and Durlauf who use it to give an account of the importance of the identification problem raised by Manski. Until we have more of this sort of work it will be difficult to persuade economists of the importance of these effects.

12.10 Conclusion

This brief survey is intended to give an account of how the direct interaction between individuals who are linked through a network structure can have an important impact on aggregate economic outcomes. The formation of expectations, the trades that are made, people's preferences and the way in which people imitate the behaviour of others all produce collective outcomes which can be very different from those that would occur if individuals were isolated from each other and only interacted through anonymous markets. Throughout this discussion it is the relationship between individuals rather than some sort of mechanical or technical network that is under discussion. The idea here is to rejoin the sociologists who have always argued that understanding the networks in which economic activity is embedded is fundamental to understanding collective behaviour. The notion that economists can safely ignore the way in which individuals interact and the structure that mediates that interaction has had its day. Yet, to move beyond this simple statement and determine the mechanisms through which the network structure exerts its influence on aggregate economic outcomes is a task which is far from being accomplished. The study of how networks emerge is the most challenging. For the time being there are two rather separate strands in the literature, that which is based on the strategic choice of links and that in which the use of links is reinforced by the experience of the users of those links. Whether these two approaches will yield similar results remains to be seen.

References

[Allen B. (1982)] "Some stochastic processes of interdependent demand and technological diffusion of an innovation exhibiting externalities among adopters", International Economic Review, Vol. 23, n. 3, October, pp. 595-608.

[Arthur W.B. (1989)] "Competing technologies, increasing returns and lock-in by historical events", Economic Journal, IC, pp. 116-31.

[Anderson S., A. de Palma, and J.F. Thisse, (1992)] *Discrete Choice Theory of Product Differentiation.* Cambridge, MA: MIT Press

[Aumann R., and J. Dreze., (1974)] "Cooperative Games with Coalition Structures", *International Journal of Game Theory,* vol 3, pp 217-237

[Bala V and S Goyal (2000)] "A Non-Cooperative Model of Network Formation", *Econometrica*, vol. 68, pp. 1181-1230

[Benabou R., (1996)] "Heterogeneity, Stratification and Growth: Macroeconomic Implications of Community Stru cture and School Finance", *American Economic Review*, vol. 86 pp. 584-609

[Bloch F.] "Endogeneous Structures of Association in Oligopolies", *Rand Journal of Economics* 26, (1995), 537-556.

[Blume L. (1993)] "The Statistical Mechanics of Strategic Interaction", *Games and Econ. Behaviour,* 5, pp. 387-424

[Brock W., and S. Durlauf., (2001a)] "Discrete Choice with Social Interactions" Review of Economic Studies, vol.68, pp.235-260

[Brock W., and S. Durlauf., (2001b)] "Interactions-based Models", in J Heckman and E Leamer (eds), *Handbook of Econometrics* Vol.5, Elsevier B V, Amsterdam.

[Cohen J., (1966)] *A Model of Simple Competition*, Harvard University Press, Cambridge, MASS

[Durlauf S., (1990)] "Locally Interacting Systems, Coordination Failure, and the Behaviour of Aggregate Activity", Working Paper, Stanford University, CA

[Durlauf S., (1993)] "Nonergodic Economic Growth", *Review of Econ. Studies* **60(2)**: 349-366

[Durlauf S., (1997)] "Statistical Mechanics Approaches to Socioeconomic Behaviour", in *The Economy as an Evolving Complex System II,* edited by W.B. Arthur, S.N. Durlauf and D.A. Lane, Addison-Wesley, Reading, MASS

[Ellison G., (1993)] "Learning, Local Interaction and Co-ordination", *Econometrica*, vol 61, September, pp. 1047-1072

[Goyal S. and S. Joshi, (2002)] "Networks of Collaboration in Oligopoly", *Games and Economic Behaviour,* forthcoming.

[Guriev S., I. Pospelov, and M. Shakhova., (1995)] "Self-Organization of Trade Networks in an Economy with Imperfect Infrastructure", Mimeo, Computing Center of the Russian Academy of Sciences, Moscow.

[Ioannides Y.M., (1997)] "Evolution of Trading Structures", in *The Economy as an Evolving Complex System II,* eds., W. B. Arthur, S.N. Durlauf, and D.A. Lane, Addison Wesley, pp. 129-167

[Jackson M and A Wolinsky (1996)] "A Strategic Model of Social and Economic Networks", *Journal of Economic Theory,* vol. 71, pp. 44-74

[Kalai E., A. Postlewaite and J. Roberts, (1978)] "Barriers to Trade and Disadvantageous Middlemen: Nonmonotonicity of the Core", *Journal of Econ. Theory*, vol 19, pp. 200-20

[Kirman A.P., C. Oddou and S. Weber, (1986)] "Stochastic communication and coalition formation", *Econometrica*, vol. 54 January, pp. 129-138

[Kirman A.P., and N. Vriend (2001)] "Evolving Market Structure: A Model of Price Dispersion and Loyalty", *Journal of Economic Dynamics and Control*

[Krugman P., (1991)] "Increasing Returns and Economic Geography", *J. of Political Economy,* Vol. 99, pp. 483-499

[Krugman P., (1994)] "Complex Landscapes in Economic Geography", *Amerian Economic Review,* vol. 84 no. 2 pp. 412-416

[Manski C, (1995)] *Identification Problems in the Social Sciences,* Harvard University Press, Cambridge, Mass.

[Mclean P.D., and J.F. Padgett., (1996)] "Was Florence a Perfectly Competitive Market?: Transactional Evidence from the Renaissance", *Theory and Society,* (forthcoming)

[Potts J (2001)] "Knowledge and Markets", *Journal of Evolutionary Economics,* vol. 11, pp. 413-432.

[Tesfatsion L., (1995)] "How Economists Can Get Alife", Mimeo, Dept of Economics, Iowa State University, Ames, Iowa

[Thoron S.] a- "Market organization: Noncooperative Models of Cartel Formation", in *Interaction and Market Structure, Essays on Heterogeneity in Economics*, (D. Delli Gatti, M. Gallegati and A. Kirman, Eds.), Springer-Verlag, 2000, pp. 207-220.

[Vriend N., (1994)] "Self-Organized Markets in a Decentralized Economy", Working Paper 94-03-013, Santa Fe Institute, Santa Fe, NM

[Weisbuch G., (1990)] *Complex System Dynamics*, Addison-Wesley, Redwood City, CA

[Weisbuch G., H. Gutowitz and G. Duchateau-Nguyen, (1994)] "Dynamics of Economic Choices Involving Pollution", Working Paper 94-04-018, Santa Fe Institute, Santa Fe, NM.

[Weisbuch, G., Kirman, A., & Herreiner, D. (2000)] "Market Organisation and Trading Relationships", *Economic Journal*, 110 pp.411-436

[Whittle P. (1986)] *Systems in Stochastic Equilibrium*, John Wiley and sons, New York.

13 Local search in unstructured networks

Lada A. Adamic, Rajan M. Lukose, and Bernardo A. Huberman

13.1 Introduction

Recently, studies of networks in a wide variety of fields, from biology to social science to computer science, have revealed some commonalities [4]. It has become clear that the simplest classical model of random networks, the Erdős-Rényi model [8], is inadequate for describing the topology of many naturally occurring networks. These diverse networks are more accurately described by power-law or scale-free link distributions. In these highly skewed distributions, the probability that a node has k links is approximately proportional to $1/k^{\tau}$. The link graph of the World Wide Web [1], the Internet router backbone [10], certain representations of biological pathways [14], and some social networks [3, 6, 19, 23], each have approximately power-law distributions, in contrast to the Poisson distribution consistent with the Erdős-Rényi random graph model.

In addition to the characterization of the topological structure of these networks, other important questions concerning the growth, robustness, and dynamics on such networks have been addressed. For example, the question of what dynamical models of graph growth tend to generate power-law networks has been investigated [6, 13, 18], as well as their robustness with respect to error and attack [5].

Another important dynamical question is the behavior of local search strategies on networks. Much of the recent work on networks has been motivated by the "small world" phenomenon, in which even very large networks (possibly possessing local clustering or structure) have very short diameters. Here the diameter is defined as the average shortest path length between the nodes in the network. The existence of this phenomenon has been demonstrated in different kinds of networks [22], and the property of short paths is obviously important for dynamic models such as disease spreading [24] and message passing between arbitrary nodes in a network.

The classic social experiment of Milgram [21] found that people could find a short chain of acquaintances in order to pass a message to each other, a phenomenon often referred to as "six degrees of separation". This result was surprising given that most people's interactions tend to be tied to their local communities, with relatively few longer range connections. Watts and Strogatz [26] revitalized interest in the small-world problem by showing that even in highly structured and clustered graphs, a few long range connections dramatically reduce the average shortest path length between nodes.

It is however another question how exactly participants in a Milgram-style experiment might find these short paths, since they do not have global knowledge of the whole graph. That is, even if short paths exist, how can one (approximately) find them using local information?

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

Kleinberg [16, 17] considered this question for a lattice topology with distance dependent shortcuts and found an elegant characterization of the conditions under which it is possible to pass messages efficiently. Kleinberg assumed a very structured topology and considered algorithms which use the target's position on a regular 2-D lattice to direct the search.

While the question of local search in real social networks is an intriguing one, it also relates in an interesting way to recent developments in information technology. The internet and the World Wide Web have certainly had an impact on the way that millions of people all over the world communicate, affecting the structure and dynamics of what we think of as traditional social networks [27]. These ever more ubiquitous technologies, wired and wireless, tend to make geography and distance less relevant for communication between people.

But the relationship is also bi-directional. A social network is also a metaphor that is relevant for understanding popular internet technologies such as peer-to-peer (p2p) file-sharing networks. These networks share some of the topological features of social networks. The Gnutella system connects users computers directly with others to share files, without a central point of coordination. In such networks, the name of the target file may be known, but due to the network's ad hoc nature, until a real-time search is performed the node holding the file is not known. In order to find files on the system, peers pass messages along to the other peers that they know of. In contrast to the scenario considered by Kleinberg, there is no global information about the position of the target, and hence it is not possible to determine whether a step is a move towards or away from the target.

These networks, while not centrally planned in structure, grow according to a simple self-organizing process. Recent measurements of Gnutella [7] and simulated Freenet networks [12] show that they have power-law degree distributions. The resulting highly unstructured networks need efficient search algorithms in order to function well. These algorithms should rely on local information in order to avoid a dependence on a central point of failure, and to accommodate their dynamic nature.

In this chapter, we will discuss a number of message-passing algorithms that can be efficiently used to search through power-law networks. We will discuss relevant work from both the statistical physics community as well as the computer science community. Most of these algorithms are meant to be improvements for peer-to-peer file sharing systems, and some may also shed some light on how unstructured social networks with certain topologies might function relatively efficiently with local information. Like the networks that they are designed for, these algorithms are completely decentralized, and they exploit the power-law link distribution in the node degree. The algorithms use local information such as the identities and connectedness of their neighbors, and their neighbors' neighbors, but not the target's global position. We demonstrate that some of these search algorithms can work well on real Gnutella networks, scale sub-linearly with the number of nodes, and may help reduce the network search traffic that tends to cripple such networks.

The chapter is organized as follows. Sections 13.2, 13.3, 13.4, and 13.5 review results from Adamic et al. [2] regarding localized search. Sections 13.2 and 13.3 present analytical and simulation results, section 13.4 compares search in Poisson random graphs and section 13.5 describes the application of the algorithms to Gnutella. Section 13.6 examines the length of the paths found in search, section 13.7 looks into shortest paths in power-law graphs, while section 13.8 examines search strategies based on information learned about the network, and section 13.9 concludes.

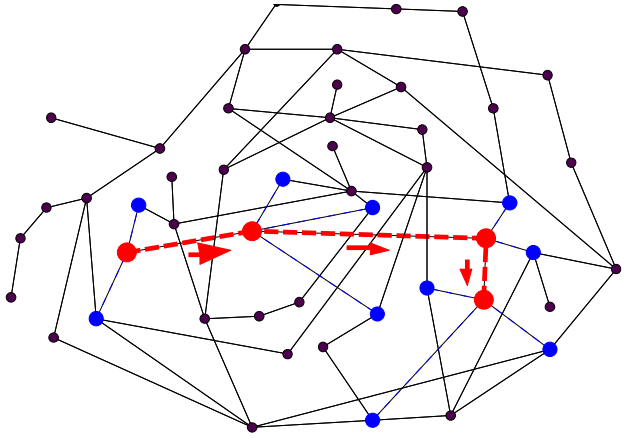


Figure 13.1: An example of a search on a 50 nodes Poisson graph. Starting at a node having three neighbors, the search finds 14 nodes in four steps.

13.2 Search in power-law random graphs

13.2.1 Intuition

The local search strategies we will be discussing use the intuition that connections tend to be disproportionately distributed among nodes and that the well-connected nodes should provide access to a greater portion of the network. In figures 13.1 and 13.2 we compare a sample walk on a standard random graph with a Poisson degree distribution and a power-law graph with the same number of nodes and edges. We plot the number of nodes accessible as a message is passed through two graphs, starting at a random node and proceeding toward the next most highly connected neighbor. Since each node has knowledge of its neighbors, we count reaching a node as finding all of its previously undiscovered neighbors.

The search on the power-law graph finds 30 nodes in 4 steps, while the same approach on the Poisson graph finds only 14 nodes in spite of the initial node having higher degree. Even though the two graphs have the same total number of edges, the distribution of edges allows one to search the power-law graph more rapidly, using only local information.

In following section we will follow up on this intuition and use the generating function formalism introduced by Newman et al. [22] for graphs with arbitrary degree distributions to analytically characterize search-cost scaling in graphs.

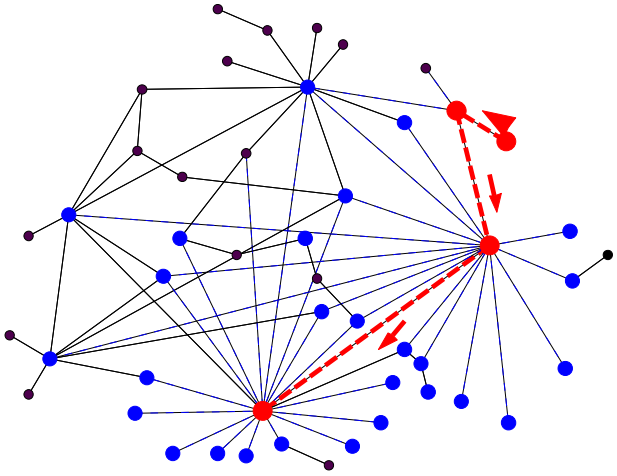


Figure 13.2: An illustration of a search on a power-law graph. Starting at a node having a single neighbor, the search finds 31 nodes in four steps.

13.2.2 Random walk search

First we examine the number of nodes encountered in a random walk on the graph. Let $G_0(x)$ be the generating function for the distribution of the vertex degree k. Then

$$G_0(x) = \sum_0^\infty p_k x^k \tag{13.1}$$

where p_k is the probability that a randomly chosen vertex on the graph has degree k.

For a graph with a power-law distribution with exponent τ, minimum degree $k = 1$ and an abrupt cutoff at $m = k_{max}$, the generating function is given by

$$G_0(x) = c \sum_1^m k^{-\tau} x^k \tag{13.2}$$

with c a normalization constant which depends on m and τ to satisfy the normalization requirement

$$G_0(1) = c \sum_1^m k^{-\tau} = 1 \tag{13.3}$$

The average degree of a randomly chosen vertex is given by

$$z_1 =< k >= \sum_1^m k p_k = G_0'(1) \tag{13.4}$$

Note that the average degree of a vertex chosen at random and one arrived at by following a random edge are different. A random edge arrives at a vertex with probability proportional to the degree of the vertex, i.e. $p'(k) \sim kp_k$. The correctly normalized distribution is given by

$$\frac{\sum_k k p_k x^k}{\sum_k k p_k} = x \frac{G_0'(x)}{G_0'(1)} \tag{13.5}$$

If we want to count the number of outgoing edges from the vertex we arrived at, but not include the edge we just came on, we need to divide by one power of x. Hence the number of new neighbors encountered on each step of a random walk is given by the generating function

$$G_1(x) = \frac{G_0'(x)}{G_0'(1)} \tag{13.6}$$

where $G_0'(1)$ is the average degree of a randomly chosen vertex as mentioned previously.

Since we are concerned with local search algorithms, we make the reasonable assumption that nodes may have at least some knowledge of their neighboring nodes' neighbors. Hence, we now compute the distribution of second neighbors. The probability that any of the 2nd neighbors connects to any of the first neighbors or to one another goes as N^{-1} and can be ignored in the limit of large N. Therefore, the distribution of the second neighbors of the original randomly chosen vertex is determined by

$$\sum_k p_k [G_1(x)]^k = G_0(G_1(x)) \tag{13.7}$$

It follows that the average number of second neighbors is given by

$$z_{2A} = \left[\frac{\partial}{\partial x} G_0(G_1(x)) \right]_{x=1} = G_0'(1) G_1'(1) \tag{13.8}$$

Similarly, if the original vertex was not chosen at random, but arrived at by following a random edge, then the number of second neighbors would be given by

$$z_{2B} = \left[\frac{\partial}{\partial x} G_1(G_1(x)) \right]_{x=1} = [G_1'(1)]^2 \tag{13.9}$$

In both Equation 13.8 and Equation 13.9, the fact that $G_1(1) = 1$ was used. Both these expressions depend on the values $G_0'(1)$ and $G_1'(1)$ so we calculate those for given τ and m. For simplicity and relevance to most real-world networks of interest we assume $2 < \tau < 3$.

$$G_0'(1) = \sum_1^m ck^{1-\tau} \sim \int_1^m x^{\tau-1}dx = \frac{1}{\tau-2}(1-m^{2-\tau}) \tag{13.10}$$

$$G_1'(1) = \frac{1}{G_0'(1)}\frac{\partial}{\partial x}\sum_1^m ck^{1-\tau}x^{k-1} \tag{13.11}$$

$$= \frac{1}{G_0'(1)}\sum_2^m ck^{1-\tau}(k-1)x^{k-2} \tag{13.12}$$

$$\sim \frac{1}{G_0'(1)}\frac{m^{3-\tau}(\tau-2)-2^{2-\tau}(\tau-1)+m^{2-\tau}(3-\tau)}{(\tau-2)(3-\tau)} \tag{13.13}$$

for large cutoff values m. Now we impose the cutoff of Aiello et al. [3] at $m \sim N^{1/\tau}$. The cutoff is chosen so that in an non-truncated distribution the expected number of nodes among N having exactly the cutoff degree is 1. No nodes of degree higher than the cutoff are present in the graph. In real world graphs one does frequently observe nodes of degree higher than this imposed cutoff, so that our calculations describe a worse case scenario. Since m scales with the size of the graph N and for $2 < \tau < 3$ the exponent $2-\tau$ is negative, we can neglect terms constant in m. This leaves

$$G_1'(1) = \frac{1}{G_0'(1)}\frac{m^{3-\tau}}{(3-\tau)} \tag{13.14}$$

Substituting into Equation 13.8 (the starting node is chosen at random) we obtain

$$z_{2A} = G_0'(1)G_1'(1) \sim m^{3-\tau} \tag{13.15}$$

We can also derive z_{2B}, the number of 2nd neighbors encountered as one is doing a random walk on the graph.

$$z_{2B} = [G_1'(1)]^2 = \left[\frac{\tau-2}{1-m^{2-\tau}}\frac{m^{3-\tau}}{3-\tau}\right]^2 \tag{13.16}$$

Letting $m \sim N^{1/\tau}$ as above, we obtain

$$z_{2B} \sim N^{2(\frac{3}{\tau}-1)} \tag{13.17}$$

Thus, as the random walk along edges proceeds node to node, each node reveals more of the graph since it has information not only about itself, but also of its neighborhood. The search cost s is defined as the number of steps until approximately the whole graph is revealed so that $s \sim N/z_{2B}$, or

$$s \sim N^{3(1-2/\tau)} \tag{13.18}$$

In the limit $\tau \rightarrow 2$, equation 13.16 becomes

$$z_{2B} \sim \frac{N}{ln^2(N)} \tag{13.19}$$

and the scaling of the number of steps required is

$$s \sim ln^2(N) \tag{13.20}$$

13.2.3 Search utilizing high degree nodes

Random walks in power-law networks naturally gravitate towards the high degree nodes, but an even better scaling is achieved by intentionally selecting high degree nodes. For τ sufficiently close to 2 one can approximately walk down the degree sequence, visiting the node with the highest degree, followed by a node of the next highest degree, etc. Let $m - a$ be the degree of the last node we need to visit in order to scan a certain fraction of the graph. We make the self-consistent assumption that $a \ll m$, i.e. the degree of the node has not dropped too much by the time we have scanned a fraction of the graph. Then the number of first neighbors scanned is given by

$$z_{1D} = \int_{m-a}^{m} Nk^{1-\tau} dk \sim Nam^{1-\tau} \tag{13.21}$$

The number of nodes having degree between $m - a$ and m, or equivalently, the number of steps taken is given by $\int_{m-a}^{m} k^{-\tau} \sim a$. The number of second neighbors when one follows the degree sequence is given by:

$$z_{1D} * G_1'(1) \sim Nam^{2(2-\tau)} \tag{13.22}$$

which gives the number of steps required as

$$s \sim m^{2(\tau-2)} \sim N^{2-\frac{4}{\tau}} \tag{13.23}$$

We now consider when and why it is possible to go down the degree sequence. We start with the fact that the original degree distribution is a power-law:

$$p(x) = \left(\sum_{1}^{m} x^{-\tau} \right)^{-1} x^{-\tau} \tag{13.24}$$

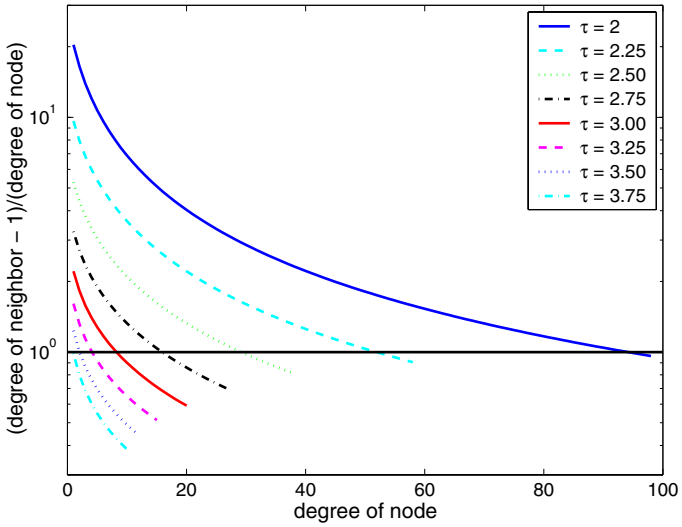


Figure 13.3: Ratio r (the expected degree of the richest neighbor of a node whose degree is n divided by n) vs. n for τ (top to bottom) = 2.0, 2.25, 2.5, 2.75, 3.00, 3.25, 3.50, and 3.75. Each curve extends to the cutoff imposed for a 10,000 node graph with the particular exponent.

where $m = N^{1/\tau}$ is the maximum degree. A node chosen by following a random link in the graph will have its remaining outgoing edges distributed according to

$$p'(x) = \left[\sum_{0}^{m-1} (x+1)^{(1-\tau)}\right]^{-1} (x+1)^{(1-\tau)} \tag{13.25}$$

At each step one can choose the highest degree node among the n neighbors. The expected number of the outgoing edges of that node can be computed as follows. In general, the cumulative distribution (CDF) $P_{max}(x,n)$ of the maximum of n random variables can be expressed in terms of the CDF $P(x) = \int_0^x p(x\prime)\,\mathrm{d}x\prime$ of those random variables: $P_{max}(x,n) = P(x)^n$. This yields

$$p'_{max}(x,n) = n(1+x)^{1-\tau}(\tau-2)\left[1-(x+1)^{2-\tau}\right]^{n-1}(1-N^{2/\tau-1})^{-n} \tag{13.26}$$

for the distribution of the number of links the richest neighbor among n neighbors has.

Finally, the expected degree of the richest node among n is given by

$$E[x_{max}(n)] = \sum_{0}^{m-1} x p'_{max}(x,n) \tag{13.27}$$

Numerically integrating the above equation yields the ratio between the degree of a node and the expected degree of its richest neighbor, plotted in Figure 13.3. For a range of exponents

and node degrees, the expected degree of the richest neighbor is higher than the degree of the node itself. However, as one moves to nodes of higher and higher degree, the probability of finding a neighbor with an even higher degree starts falling (the precise point depends strongly on the power-law exponent).

What this means is that one can approximately follow the degree sequence across the entire graph for a sufficiently small graph or one with a power-law exponent close to 2 ($2.0 < \tau < 2.3$). At each step one chooses a node with degree higher than the current node, quickly finding the one with the highest degree. Once the highest degree node has been visited, it will be avoided, and a node of approximately second highest degree will be chosen. Effectively, after a short initial climb, one goes down the degree sequence. This is the most efficient way to do this kind of sequential search.

13.3 Simulation

We used simulations on a random network with a $\tau = 2.1$ power-law link distribution and a simple cutoff at $m \sim N^{1/\tau}$ to validate our analytical results. The graph is generated by assigning links at random between nodes of pre-assigned degree drawn from the power-law distribution. For $2 < \tau < 3.48$, a graph contains a giant connected component (GCC), the largest group of nodes such that any node can be reached from any other node following links [3]. All our measurements were performed on the GCC which contained the majority of the nodes of the original graph and most of the links as well. The link distribution of the GCC is nearly identical to that of the original graph with a slightly smaller number of nodes of degree 1 and 2.

Next we apply our message passing algorithm to the network. Two nodes, the source and the target, are selected at random. At each time step the node which has the message passes it on to one of its neighbors. The process ends when the message is passed on to a neighbor of the target, that, knowing the identity of its neighbors, passes the message to the target directly. The process is analogous to performing a random walk on a graph, where each node is 'visited' as it receives the message.

There are several variants of the algorithm, depending on the strategy and the amount of local information available.

1. The node passes the message on to one of its neighbors at random, optionally avoiding a node which has already seen the message.

2. The node knows the degrees of its neighboring nodes and chooses to pass the message onto the neighbor with the most neighbors.

3. The node knows who its neighbors' neighbors are and passes the message onto a neighbor of the target if possible.

In order to avoid passing the message to a node that has already seen the message, the message itself must be signed by the nodes as they receive the message. Further, if a node has passed the message, and finds that all of its neighbors are already on the list, it puts a special mark next to its name, which means that it is unable to pass the message onto any new node. This is equivalent to marking nodes as follows:

white Node has not been visited.

gray Node has been visited, but all its neighbors have not.

black Node and all its neighbors have been visited already.

Here we compare two strategies. The first performs a random walk, where only retracing the last step is disallowed. In the message passing scenario, this means that if Bob just received a message from Jane, he wouldn't return the message to Jane if he could pass it to someone else. The second strategy is a self avoiding walk which avoids passing the message to previously visited nodes and prefers high degree nodes to low degree ones. In both strategies the first and second neighbors are scanned at each step.

Figure 13.4(a) shows the scaling of the average search time with the size of the graph for the two strategies. The scaling (exponent 0.79 for the random walk and 0.70 for the high degree strategy) is not as favorable as in the analytic results derived above (0.14 for the random walk and 0.1 for the high degree strategy when $\tau = 2.1$) .

Consider, on the other hand, the number of steps it takes to cover half the graph. That is, instead of asking how long it would take on average to find any node in the graph, we ask how long it would take to find the first 50% of the nodes. Such a measure is reasonable in a network where more than one node is likely to be able to satisfy a request. In a social context, there might be more than one person who has a particular item or can share expertise on a subject. In the context of a file sharing network, there might be more than one node having the requested file.

For this measure we observe a scaling which is much closer to the ideal. As shown in Figure 13.4(b), the cover time scales as $N^{0.37}$ for the random walk strategy vs. $N^{0.15}$ from Equation 13.18. Similarly, the high degree strategy cover time scales as $N^{0.24}$ vs. $N^{0.1}$ in Equation 13.23.

The difference in the value of the scaling exponents of the cover time and average search time implies that a majority of nodes can be found fairly efficiently, but others demand high search costs. As Figure 13.4(c) shows, a large portion of the 10,000 node graph is covered within the first few steps, but some nodes take as many steps or more to find as there are nodes in total. For example, the high degree seeking strategy finds about 50% of the nodes within the first 10 steps (meaning that it would take about $10 + 2 = 12$ hops to reach 50% of the graph). However, the skewness of the search time distribution bring the average number of steps needed to 217.

Some nodes take a long time to find because the random walk, after a brief initial period of exploring fresh nodes, tends to revisit nodes. It is a well-known result that the stationary distribution of a random walk on an undirected graph is simply proportional to the distribution of links emanating from a node. Thus, nodes with high-degree are often revisited in a walk.

A high-degree seeking self-avoiding walk is an improvement over the random walk taking 13 times fewer steps, but still cannot avoid retracing its steps. Figure 13.4(d) shows the color of nodes visited on such a walk for a $N = 1000$ node power-law graph with exponent 2.1 and an abrupt cutoff at $N^{1/2.1}$. The number of nodes of each color encountered in 50 step segments is recorded in the bar for that time period, showing that some grey and black nodes were encountered before the all of the nodes were found.

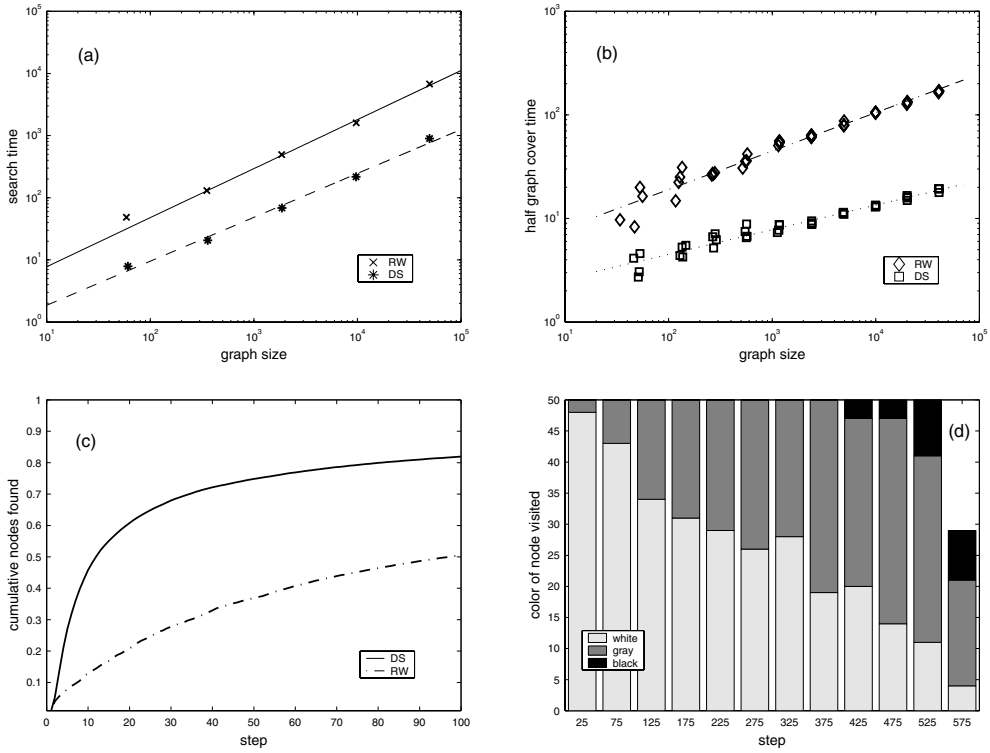


Figure 13.4: (a) Scaling of the average node-to-node search cost in a random power-law graph with exponent 2.1, for random walk (RW) and high-degree seeking (DS) strategies. The solid line is a fitted scaling exponent of 0.79 for the RW strategy and the dashed is an exponent of 0.70 for the DS strategy. (b) The observed and fitted scaling for half graph cover times for the RW and DS strategies. The fits are to scaling exponents of 0.37 and 0.24, respectively. (c) Cumulative distribution of nodes seen vs the number of steps taken for the RW and DS strategies on a 10 000 node graph. (d) Bar graph of the color of nodes visited in DS search of a random 1000 node power-law graph with exponent 2.1. White represents a fresh node, gray represents a previously visited node that has some unvisited neighbors, and black represents nodes for which all neighbors have been previously visited.

Although revisiting nodes slows down search, it is the form of the link distribution that is responsible for changes in the search cost scaling. In a graph with a uniform link distribution the number of new nodes discovered at every step would be proportional to the number of unexplored nodes in the graph. The factor by which the search is slowed down through revisits would be independent of the size of the graph.

In contrast, in a power-law graph, a large number of links point to only a small subset of high degree nodes. When a new node is visited, its links do not let us uniformly sample the graph, they preferentially lead to high degree nodes, which have likely been seen or visited in a previous step. Ironically, the presence of high degree nodes, so useful to our search strategies, also worsens the search cost scaling from the ideal scaling found in sections 13.2.2 and 13.2.3.

This would not be true of a Poisson graph, where all the links are randomly distributed and hence all nodes have approximately the same degree. We will explore and contrast the search algorithm on a Poisson graph in the following section.

13.4 Comparison with Poisson distributed graphs

In a Poisson random graph with N nodes and z edges, the probability $p = z/N$ of an edge between any two nodes is the same for all nodes. The generating function $G_0(x)$ is given by [22]:

$$G_0(x) = e^{z(x-1)} \tag{13.28}$$

In this special case $G_0(x) = G_1(x)$, so that the distribution of outgoing edges of a node is the same whether one arrives at the vertex by following a link or picks the node at random. This makes the analysis of search in a Poisson random graph particularly simple. The expected number of new links encountered at each step is a constant p, so that the number of steps needed to cover a fraction c of the graph is $s = cN/p$. If p remains constant as the size of the graph increases, the cover time scales linearly with the size of the graph. This has been verified via simulation of the random walk search as shown in Figure 13.5.

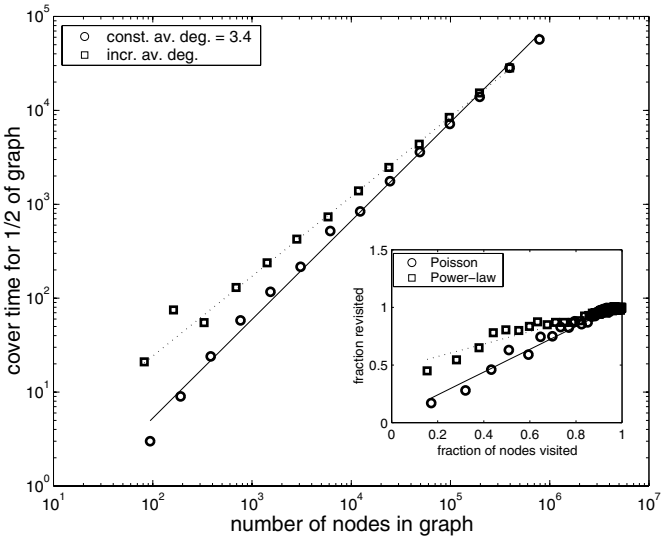


Figure 13.5: Squares are scaling of cover time for 1/2 of the graph for a Poisson graph with a constant average degree/node (with fit to a scaling exponent of 1.0). Circles are the scaling for Poisson graphs with the same average degree/node as a power-law graph with exponent 2.1 (with fit to a scaling exponent of 0.85). The inset compares revisitation between search on Poisson versus power-law graphs, as discussed in the text.

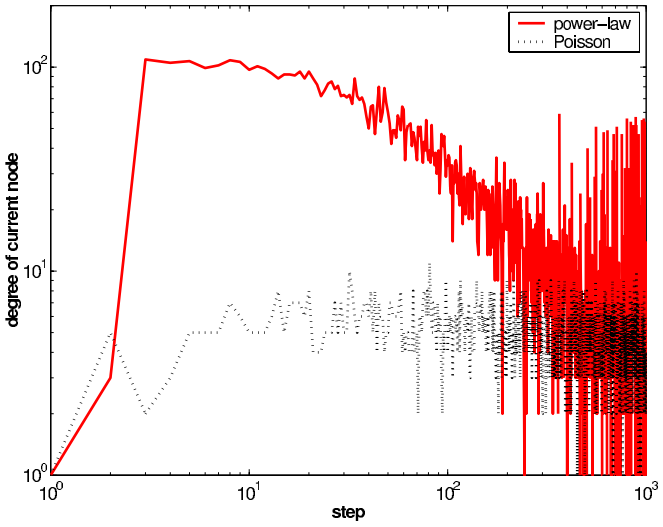


Figure 13.6: Degrees of nodes visited in a single search for power-law and Poisson graphs of 10,000 nodes.

In our simulations, in order to keep the total number of edges equal between power-law and Poisson graphs of the same size, the probability p increases with the size. It grows slowly towards its asymptotic value because of the particular choice of cutoff at $m \sim N^{(1/\tau)}$ for the power-law link distribution. We generated Poisson graphs with the same number of nodes and links for comparison. Within this range of graph sizes, growth in the average number of links per node appears as $N^{0.6}$, making the average number of 2nd neighbors scale as $N^{0.15}$. This means that the scaling of the cover time scales as $N^{0.85}$, as shown in Figure 13.5.

Note how well the simulation results match the analytical expression. This is because nodes can be sampled in an approximately even fashion by following links as is illustrated in Figure 13.5(inset). If links are evenly distributed among the nodes, then when the search has covered 50% of the graph, one would expect to revisit previously seen nodes about 50% of the time. This is indeed the case for the Poisson graph.

However, for the power-law graph, when 50% of the graph has been visited, nodes are revisited about 80% of the time, which implies that the same high degree nodes are being revisited before new low-degree ones. This bias introduces a discrepancy between the analytic scaling and the simulated results in the power-law case. However, even the simulated $N^{0.35}$ scaling for a random, minimally self-avoiding strategy on the power-law graph out-performs the ideal $N^{0.85}$ scaling for the Poisson graph. It's also important to note that the the high degree node seeking strategy has a much greater success in the power-law graph because it relies heavily on the fact that the number of links per node varies considerably from node to node. In the Poisson graph, the variance in the number of links is much smaller, making the high degree node seeking strategy comparatively ineffective as shown in.

Figure 13.6 shows an illustration of this point. We repeat the experiment in Figures 13.1 and 13.2 on larger power-law and Poisson graphs with $N = 10,000$. In the power-law graph we start from a randomly chosen node. In this case the starting node has only one link, but two steps later we find ourselves at a node with the highest degree. From there, one approximately follows the degree sequence, that is, the node richest in links, followed by the second richest node, etc. The strategy has allowed us to scan the maximum number of nodes in the minimum number of steps. In comparison, the maximum degree node of the exponential graph is 11, and it is reached only on the 81st step. Even though the two graphs have a comparable number of nodes and edges, the exponential graph does not lend itself to quick search.

13.5 Gnutella

Gnutella is a peer-to-peer filesharing system which treats all client nodes as functionally equivalent and lacks a central server which can store file location information. This is advantageous because it presents no central point of failure. The obvious disadvantage is that the location of files is unknown. When a user wants to download a file, she sends a query to all the nodes within a neighborhood of size ttl, the time to live assigned to the query. Every node passes on

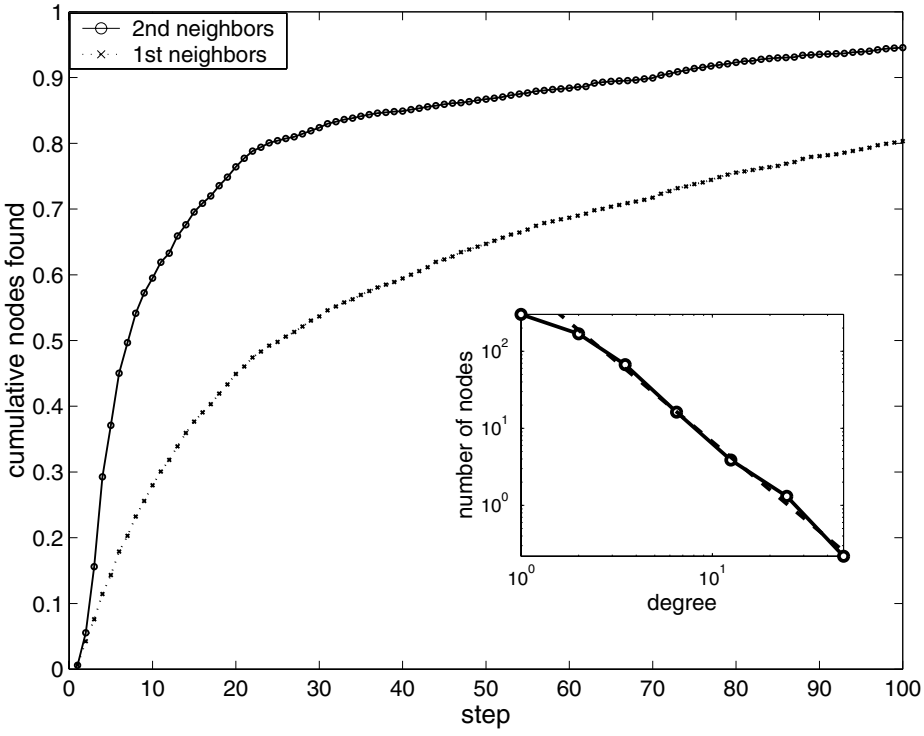


Figure 13.7: Cumulative number of nodes found at each step in the Gnutella network.

the query to all of its neighbors and decrements the ttl by one. In this way, all nodes within a given radius of the requesting node will be queried for the file, and those who have matching files will send back positive answers.

This broadcast method will find the target file quickly, given that it is located within a radius of ttl. However, broadcasting is extremely costly in terms of bandwidth. Every node must process queries of all the nodes within a given ttl radius. In essence, if one wants to query a constant fraction of the network, say 50%, as the network grows, each node and network edge will be handling query traffic which is proportional to the total number of nodes in the network.

Such a search strategy does not scale well. As query traffic increases linearly with the size of Gnutella graph, nodes become overloaded as was shown in a study by Clip2 [7]. 56k modems are unable to handle more than 20 queries a second, a threshold easily exceeded by a network of about 1,000 nodes. With the 56k nodes failing, the network becomes fragmented, allowing users to query only small section of the network.

The search algorithms described in the previous sections may help ameliorate this problem. Instead of broadcasting a query to a large fraction of the network, a query is only passed onto one node at each step. The search algorithms are likely to be effective because the Gnutella network has a power-law connectivity distribution as shown in the inset of Figure 13.7.

Typically, a Gnutella client wishing to join the network must find the IP address of an initial node to connect to. Currently, ad hoc lists of "good" Gnutella clients exist [7]. It is reasonable to suppose that this ad hoc method of growth would bias new nodes to connect preferentially to nodes which are already fairly well-connected, since these nodes are more likely to be "well-known". Based on models of graph growth [6, 13] where the "rich get richer", the power-law connectivity of ad hoc peer-to-peer networks may be a fairly general topological feature.

By passing the query to every single node in the network, the Gnutella algorithm fails to take advantage of its connectivity distribution. To implement our algorithm the Gnutella clients must be modified to keep lists of the files stored by their first and second degree neighbors have. This information must be passed at least once when a new node joins the network, and it may be necessary to periodically update the information depending on the typical lifetime of nodes in the network. The importance of localized indexing to scalability has been illustrated by the growth of the FastTrack [25] network whose size has reached hundreds of thousands of nodes. FastTrack is a network similar to Gnutella, with no central server, but using local indexing. A fraction of FastTrack clients with high bandwidth and reliability are selected to be supernodes. Supernodes index the files of other nodes and route queries on their behalf. We note that unlike FastTrack, our algorithm requires each node to store a local index.

Keeping track of the filenames of its neighbors' files places an additional cost on every node. Since network connections saturated by query traffic are a major weakness in Gnutella, and since computational and storage resources are likely to remain much less expensive than bandwidth, such a tradeoff is readily made. However, now instead of every node having to handle every query, queries are routed only through high connectivity nodes, a situation similar to that of supernodes in the FastTrack network. Since nodes can select the number of connections that they allow, high degree nodes are presumably high bandwidth nodes that can handle the query traffic. The network has in effect created local directories valid within a two

link radius. It is resilient to attack because of the lack of a central server. As for power-law networks in general [5], the network is more resilient than random graphs to random node failure, but less resilient to attacks on the high degree nodes.

Further adjustments to the present Gnutella clients to implement our algorithm involve switching from broadcasting queries to passing them only to the highest degree nodes. To execute a self-avoiding search, nodes need to append their IDs to the query as they process it.

Figure 13.7 shows the success of the high degree seeking algorithm on the Gnutella network. We simulated the search algorithm on a crawl by Clip2 of the actual Gnutella network of approximately 700 nodes. Assuming that every file is stored on only one node, 50% of the files can be found in 8 steps or less. Furthermore, if the file one is seeking is present on multiple nodes, the search will be even faster.

To summarize, we have argued that truly peer-to-peer networks like Gnutella are likely to have a power-law structure, and that the local search algorithms we have described can be effective. As the number of nodes increases, the (already small) number of nodes that will need to be queried increases sub-linearly. As long as the high degree nodes are able to carry the traffic, the Gnutella network's performance and scalability may improve by using these search strategies.

13.6 Path finding

So far we have only discussed the amount of time it takes to locate a node a single time. But in the process of searching for a node, one is also mapping out a path which could be used to contact that node in the future. Removing loops and backtracking steps from the search path leaves a route to the desired node. This route could be reused should one desire to communicate with the node again.

Kim et al. [15] have shown that following a high-degree seeking strategy on power-law graphs produces paths which scale on average as the logarithm of the size of the network. While the paths found are not always the shortest paths themselves, they share in the logarithmic scaling of the average shortest path. In contrast, random walker strategies, or strategies on non-power-law graphs such as Poisson random graphs or small world graphs defined by Watts and Strogatz [26], produce paths which whose scaling is power-law.

Following the methods of Kim et al, we constructed a scale-free network of Barabasi and Albert (BA) [6] type. Starting with a small number ($m_0 = 2$) of vertices, a new vertex with $m = 2$ edges is added at each time step such that the probability of an edge connecting to a vertex is proportional to the degree of the vertex. This method yields a power-law network which has an exponent $\tau = 3$ but is not truly random because correlations between node degrees do exist. Although $\tau = 3$ lies outside the regime favorable to the previously discussed search strategies, requiring many steps to locate a node, the paths obtained with the loops removed scale logarithmically with the size of the network.

Figure 13.8 shows a comparison between the actual shortest paths and the shortest paths found using various search strategies on BA power-law graphs and Poisson graphs with the same total number of vertices and edges. Figure 13.8a shows the that shortest paths scale logarithmically in the size of the graph in both power-law and Poisson graphs, but the average shortest path in a power-law graph grows more slowly as the number of nodes increases. In

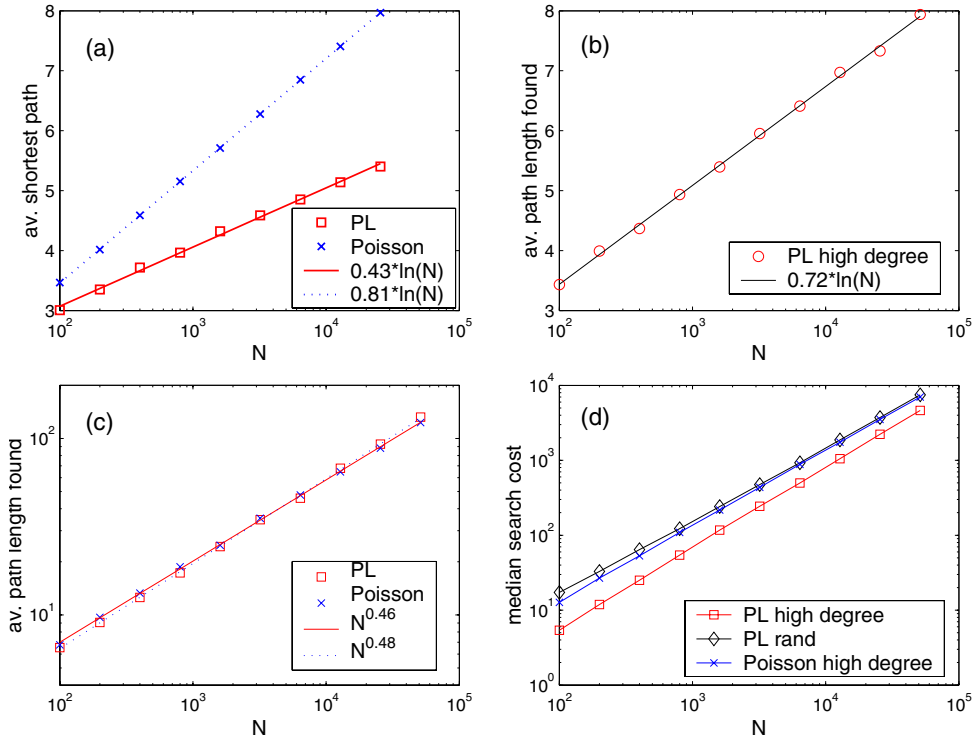


Figure 13.8: Scaling of path finding strategies: (a) average shortest path found using breadth first search for Poisson and power-law graphs, (b) average path length found using a high degree strategy on a power-law graph, (c) average path length found using a random strategy on a power-law graph and a high degree strategy on a Poisson graph, (d) median number of steps required to find a path between two nodes

effect, high degree nodes are drawing the graph closer together. This will be discussed further in section 13.7.

In order to find the exact shortest path, a broadcast method equivalent to a breadth first search (BFS) must be used. As mentioned in our discussion of Gnutella, broadcasting can overwhelm the bandwidth resources of the network. Kim et al. propose instead a search similar to the strategies discussed in sections 13.2.2 and 13.2.3. The message is passed from only one node at each step, either randomly, or to the highest degree neighbor, using knowledge only of the first degree neighbors and their degree. When following the high-degree strategy, a node passes the message to the highest degree node it personally has not passed the message to previously.

This strategy is not truly self-avoiding in the sense that a node does not try to avoid passing the message onto a node that others have already contacted. Curiously, we find that a truly self-avoiding strategy, while locating nodes more quickly, does not yield short paths in the end.

Kim et al. also note that if the strategy chooses nodes probabilistically, with the probability of a node being chosen proportional to its degree, the logarithmic scaling is lost. It is possible that both the self-avoiding and probabilistic methods fare worse because they return to the higher degree nodes less frequently. Because the majority of paths pass through high-degree nodes, the deterministic strategy which routinely revisits high degree nodes before moving forward is more likely to find a shorter path.

For comparison, we also plot in Figure 13.8c the length of the shortest paths found in the BA graph by choosing nodes at random rather than based on their connectivity. The paths found have a much less favorable power-law scaling of approximately $N^{0.5}$, compared to the logarithmic scaling of the shortest path. A similar result is obtained when using a high degree strategy on an equivalent Poisson graph, where extremely high degree nodes are absent.

Even in the case where short paths can be found using a high degree strategy on a power-law graph, the approach may be too costly. While the length of the average path found grows slowly as the size of the network increases, the average cost in the amount of time necessary to find the path, shown in Figure 13.8d, scales nearly linearly. The median number of steps required to find a node grows into the thousands while targets remain less than 10 steps away.

Although the above discussion of path finding strategies demonstrates how nodes could in principle find shortest paths between each other, the extremely high cost of this procedure suggests that additional clues as to the location of the target or knowledge of second degree neighbors would be necessary to make such an approach worthwhile.

13.7 Shortening the shortest path

The previous sections described the role high degree nodes play in locating nodes and constructing a short path to a target. A further twist however, is the fact that the presence of high degree nodes shortens the shortest paths themselves. The average shortest path grows more slowly as the size of the network increases in a power-law graph than in an equivalent Poisson random graph.

Figure 13.9 shows the number of neighbors who are r steps away from a randomly chosen vertex given by the formula of Newman et al. [22]:

$$z_r = [\frac{z_{2A}}{z_1}]^{r-1} z_1 \tag{13.29}$$

with z_1 and z_{2A} given by Equations 13.4 and 13.8. By choice, the Poisson graph has the same average number z_1 of first degree neighbors. Using the above result that the expected number of outgoing edges following a link is equal to the average vertex degree, the number of second degree neighbors is simply z_1^2.

The number of nodes at distance r scales as $\exp(\alpha r)$, $\alpha = z_{2A}/z_1$. The actual number of neighbors of course cannot continue to increase exponentially with r due to the finite size of the graph. For a Poisson graph $\alpha = z_1$, where z_1 is determined by the degree distribution of the power-law graph and asymptotes as the size of the graph increases. The ratio of the number of second degree to first degree neighbors grows more rapidly for a power-law graph as shown in the inset of Figure 13.9.

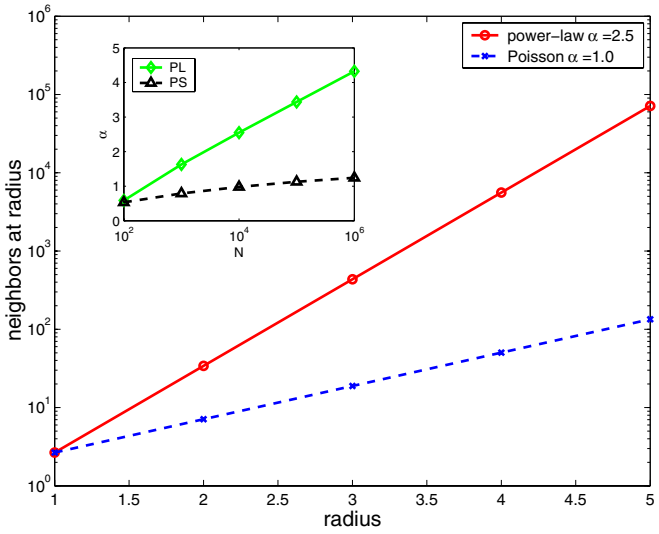


Figure 13.9: The expected number of nodes at a given distance from a node is plotted for $\tau = 2.1$ power-law and Poisson graphs with 10,000 nodes and the same number of edges. The number of neighbors as a function of the radius r is $\exp(\alpha * r)$. $\alpha = 2.5$ for the power-law graph, while $\alpha = 1.0$ for the Poisson one. The actual number of neighbors at higher radii is limited by the finite size of the graphs. The inset shows the variation of α with the size of the network.

13.7.1 Iterative deepening

The fact that the number of hops between nodes is shorter in a power-law graph implies that the broadcasting method of locating nodes and resources will return results more quickly. This is because, as shown in Figure 13.9, compared to other graph topologies, many more nodes are available at the same radius. Yang and Garcia-Molina [28] and Lv et al. [20] have experimented with a local search method that benefits from this fact in order to improve upon the standard fixed-radius broadcast that the default Gnutella protocol uses.

Yang and Garcia-Molina's method, which they call iterative deepening, begins by sending out standard Gnutella queries in a sequence. Queries in the sequence differ only in that they have increasing ttl settings. For example, the first query may be a broadcast that is 2 levels deep. Then the sending client might wait a pre-specified time for a response, and if no results are returned, may send out another query with a ttl of 3. The method is therefore parameterized by a sequence of ttl values and a waiting value.

The method is an improvement over the default protocol when the queries can be satisfied by nodes closer than the maximum radius defined by the ttl of the default. In that case, bandwidth and processing cost are saved. Their experiments on a live Gnutella client showed very good improvements. The bandwidth used and processing cost was 19% and 41% of the default policy, and they argue that the entire network's performance would increase significantly

if each client adopted the iterative deepening policy. Some similar results of simulations on different graph topologies are reported in [20].

13.8 Adaptive search

The above sections have examined strategies for finding a node on a network knowing nothing other than the identities of one's first and second neighbors. However, a node can learn about the network over time and adapt its search strategies. Yang and Garcia-Molina [28] performed experiments on the Gnutella network in which a modified Gnutella node selectively passed a query onto one of its neighbors. The neighbors thereafter would follow the standard Gnutella protocol and broadcast the query to all of their neighbors. To make the experiment realistic, the queries were sampled from a collection gathered by passively listening in on Gnutella traffic.

Yang et al. found that selecting the node which had previously delivered a specified number of results in the least amount of time outperformed a strategy which selects a random or a high-degree neighbor in the first step. The result showed that adapting the search algorithm to incorporate information learned about the network can deliver results comparable to BFS (broadcast) search while using considerably less processing power and bandwidth.

While nodes can adapt their search strategies based on the changing performance of nodes in the network, the network itself can grow and restructure in order to facilitate search. Freenet [12] is an example of a network which dynamically changes connections and distributes data files as a result of queries passing through it. Although decentralized, the Freenet network allows for nodes to specialize in locating subsets of files and for nodes to direct queries to nodes most likely to be able to route or satisfy the query.

Each node stores a routing table of files identified by a unique key and the node which is storing the file. When a node receives a request for a file listed in its routing table, it forwards the request to the node listed as having the file. If the there is no file matching the key, it will forward the request to the location of a file with the 'closest' key to the key requested. If the query is eventually satisfied, the file will be passed back along the same route as the query, and the node will mark the node's location. In this way nodes learn of the locations of files with keys similar to the ones already listed in their routing tables and can specialize in a particular region of the key space, expediting the search further.

Nodes that reliably answer queries will be added to more routing tables and hence will be contacted more often than nodes that do not. In simulations of the network this leads to high degree nodes acquiring even more connections, and, unsurprisingly, to a power-law distribution.

Figure 13.10 shows the number of hops required to satisfy a request as a simulated Freenet network grows from 20 to 200,000 nodes. The median path length scales as $N^{0.28}$, and is a mere 8 hops for a network of 10,000 nodes. The result shows that using a focused search in combination with an adaptive network can improve scalability of a p2p network.

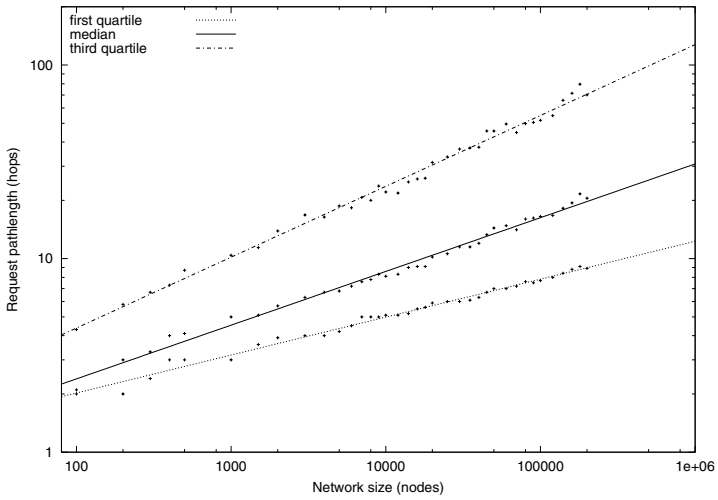


Figure 13.10: Request path length versus Freenet network size. The median path length in the network scales as $N^{0.28}$. Source: Theodore Hong [9]

13.9 Conclusion

In this chapter we have shown that local search strategies in power-law graphs have search costs which scale sub-linearly with the size of the graph, a fact that makes them very appealing when dealing with large networks. The most favorable scaling was obtained by using strategies which preferentially utilize the high connectivity nodes in these power-law networks. We also established the utility of these strategies for searching on the Gnutella peer-to-peer network. Furthermore, we reviewed the effectiveness of other improvements to simple broadcast on Gnutella such as iterative deepening and adaptive search.

Our results on high-degree seeking local search strategies may extend to social networks. However, in social networks, it is clear that people have a wide variety of additional cues to help them find who and what they need. Nevertheless, our results suggest that even strategies that neglect those cues may perform reasonably well on large power-law networks when they take advantage of the connectedness of nodes. These strategies have intuitive appeal, since people naturally ask those they perceive to be well-connected when trying to locate others in a social network.

It may not be coincidental that several large networks are structured in a way that naturally facilitates search. For example, large social networks, such as the AT&T call graph [3] and the collaboration graph of film actors, have exponents in the range ($\tau = 2.1 - 2.3$) which according to our analysis makes them especially suitable for searching using simple, local algorithms. Being able to reach remote nodes by following intermediate links allows communication systems and people to get to the resources they need and distribute information within these informal networks. At the social level, our analysis supports the hypothesis that highly connected individuals do a great deal to improve the effectiveness of social networks in terms of access to relevant resources [11].

Furthermore, it has been shown that the Internet backbone has a power-law distribution with exponent values between 2.15 and 2.2 [10], and web page hyperlinks have an exponent of 2.1 [6]. While in the Internet search is more structured, using routing tables for directing packets and search engines for finding web pages, high degree nodes still play a very significant role. Packets are usually routed through high degree hubs, and people searching for information on the Web turn to highly connected nodes, such as directories and search engines, which can bring them to their desired destinations. On the other hand, a system such as the power grid of the western United States, which does not serve as a message passing network, has an exponential degree distribution.

Networks for which locating and distributing information play a vital role, even without perfect global information, tend to be power-law with exponents favorable to local search. Actually, we find it likely that these networks could have evolved so as to facilitate search and information distribution.

Acknowledgements

We would like to thank Clip2 for the use of their Gnutella crawl data.

References

[1] Lada A. Adamic. The small world web. In *Proceedings of the 3rd European Conf. on Digital Libraries*, volume 1696 of *Lecture notes in Computer Science*, pages 443–452. Springer, 1999.

[2] Lada A. Adamic, Rajan M. Lukose, Amit R. Puniyani, and Bernardo A. Huberman. Search in power-law networks. *Phys. Rev. E*, 64:046135, 2001.

[3] William Aiello, Fan Chung, and Linyuan Lu. A random graph model for massive graphs. In *STOC '00, Proceedings of the thirty-second annual acm symposium on Theory of computing*, pages 171–180, 2000.

[4] R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Review of Modern Physics*, 74:47–94, 2002.

[5] R. Albert, H. Jeong, and A.-L. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406:37, 2000.

[6] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[7] Clip2. Gnutella: To the bandwidth barrier and beyond. *http://www.clip2.com/gnutella.html*, 2000.

[8] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.

[9] Ian Clarke et al. Protecting free expression online with freenet. *IEEE Internet Computing*, 6(1):40–49, 2002.

[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM*, pages 251–262, 1999.

[11] Malcolm Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Little Brown & Company, New York, NY, 2000.

[12] Theodore Hong. Performance. In Andy Oram, editor, *Peer-to-Peer: Harnessing the Benefits of a Disruptive Technology*, chapter 14, pages 203–241. O'Reilley, 2001.

[13] B.A. Huberman and L.A. Adamic. Growth dynamics of the world wide web. *Nature*, 401:131, 1999.

[14] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[15] Beom Jun Kim, Chang No Yoon, Seung Kee Han, and Hawoong Jeong. Path finding strategies in scale-free networks. *Phys. Rev. E*, 65:027103, 2002.

[16] Jon M. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.

[17] Jon M. Kleinberg. Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems (NIPS)*, page 14, 2001.

[18] P. L. Krapivsky, G. J. Rodgers, and S. Redner. Degree distributions of growing random networks. *Phys. Rev. Lett.*, 86:5401–5404, 2001.

[19] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg. The web of human sexual contacts. *Nature*, 411:907–908, 2001.

[20] C. Lv, P. Cao, E. Cohen, E. Felten, X. Li, and S. Shenker. Search and replication in unstructured peer-to-peer networks. In *Proc. 2002 ACM SIGMETRICS*, 2002.

[21] Stanley Milgram. The small-world problem. *Psychology Today*, 1:62–67, 1967.

[22] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distribution and their applications. *Phys. Rev. E*, 64:026118, 2001.

[23] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proc. Natl. Acad. Sci.*, 99:2566–2572, 2002.

[24] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–03, 2001.

[25] Kelly Truelove and Andrew Chasin. Morpheus out of the underworld. http://www.openp2p.com/pub/a/p2p/2001/07/02/morpheus.html, 2001.

[26] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[27] Barry Wellman. Computer networks as social networks. *Science*, 293:2031–34, 2001.

[28] Beverly Yang and Hector Garcia-Molina. Improving search in peer-to-peer networks. *ICDCS*, 2002.

14 Accelerated growth of networks

Sergei N. Dorogovtsev and Jose F. F. Mendes

Abstract

In many real growing networks, the mean number of connections per vertex increases with time. The Internet, the World Wide Web, collaborations networks, and many others display this behavior. Such growth can be called *accelerated*. We show that this acceleration influences the distribution of connections and may determine the structure of a network. We discuss general consequences of the acceleration and demonstrate its features by use of simple illustrative examples. In particular, we show that accelerated growth explains fairly well the structure of the Word Web (the network of interacting words of human language). Also, we use models of the accelerated growth of networks to describe a wealth condensation transition in evolving societies.

14.1 Acceleration

The great majority of models of evolving networks contain a very important assumption. These models suppose that the total number of edges in a growing network is a linear function of its size, that is, of the total number of vertices. This linear growth does not change the average degree of the network [1–3]. (Here, following standard terminology from graph theory, we call the total number of connections of a vertex its "degree". Physicists often call this quantity "connectivity". The number of incoming edges of a vertex in directed networks is called "in-degree"; the number of outgoing edges is "out-degree".)

The first model for the growth of networks under the mechanism of preferential linking, namely, the Barabási-Albert model [4] (see also Ref. [5]), is only one example of a linearly growing network from a very long list [6–13]. Thus, a linear type of growth is usually supposed to be a natural feature of growing networks. But let us ask ourselves whether this very particular case, that is, linear growth, is so widespread in real networks. To answer this question, we must look at existing empirical data. Let us start from the most well known nets.

(i) *The World Wide Web:*

Recall that the World Wide Web (WWW) is the array of its documents (pages) plus hyperlinks, namely, mutual references in these documents. The WWW is a directed network. Although hyperlinks are directed, pairs of counterlinks, in principle, may produce undirected connections. Links inside pages (self-references) are usually not considered as edges of the WWW, so this network does not contain "tadpoles" (closed one-edge loops).

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

According to Ref. [14], in May 1999, from the point of view of Altavista, the WWW consisted of 203×10^6 vertices (URLs, i.e., pages) and 1466×10^6 hyperlinks. The average in- and out-degrees were $\overline{k}_i = \overline{k}_o = 7.22$. The average in- and out-degrees are equal to each other, since all the connections are inside the WWW. (Notice that "physical" time is unimportant for us, so that, in principle, we might not mention any date.) In October 1999 there were already 271×10^6 and $2130\, times 10^6$ hyperlinks. The average in- and out-degrees had become $\overline{k}_i = \overline{k}_o = 7.85$. Thus, the average degree of the WWW is increasing.

(ii) *The Internet:*

Very roughly speaking, the Internet is a set of vertices, which are interconnected by wires. The vertices of the Internet are hosts (computers of users), servers (computers or programs providing a network service that also may be hosts), and routers that arrange traffic across the Internet. Connections are naturally undirected (an undirected network), and traffic (including its direction) changes all the time. Web documents are accessible through the Internet (wires and hardware), and this determines the relation between the Internet and the WWW. Routers are united in domains; however, this notion is not well defined for the Internet. In January 2001, the Internet already contained about 100 million hosts. One should emphasize that it is not the hosts that determine the structure of the Internet, but rather the routers and domains. In July 2000, there were about 150 000 routers in the Internet [15]. Since then, the number has risen to 228 265 (data from Ref. [16]). Thus, one can consider the topology of the Internet on a router level or inter-domain topology [17]. In the latter case, it is actually a small network.

According to the data of Ref. [17] for the inter-domain level of the Internet, in November 1997 it consisted of 3015 vertices and 5156 edges, so that the average degree was $\overline{k} = 3.42$. In April 1998 there were 3530 vertices and 6432 edges, and the average degree was $\overline{k} = 3.65$. In December 1998 there were 4389 and 8256 edges, so the average degree was already equal to 3.76. Since then, the average degree of the Internet on the inter-domain level has been increasing.

We have noted that domains in the Internet are poorly defined. Also, the last data of Ref. [17] are for December 1998. However, one may use more recent data on "autonomous systems". Extensive data on the connections of operating "autonomous systems" (AS) in the Internet are being collected by the National Laboratory for Applied Network Research (NLANR). For nearly each day, starting from November 1997, NLANR has a map of connections of AS. These maps are closely related to the Internet graph on the inter-domain level. Statistical analysis of these data was made in Refs. [18, 19]. The data were averaged, and for 1997 the average degree $\overline{k} = 3.47$ was obtained; in 1998, the average degree was 3.62, and in 1999, it was 3.82. Again we see that the average degree of the Internet on the inter-domain level (more rigorously speaking, on the AS level) is increasing. One should add that the growth of the average degree of the net of AS was also indicated in Ref. [20].

Unfortunately, there are no reliable empirical data on the router level of the Internet to arrive at precise conclusions. In 1995, the Internet included 3888 routers with 5012 interconnections [17], that is, $\overline{k} \sim 2.6$. In 2000, there were $\sim 150\,000$ routers and $\sim 200\,000$ interconnections between them, so that $\overline{k} \sim 2.7$ [15]. These data are taken from different sources; they are not precise and cannot be compared.

(iii) *Networks of citations in scientific literature:*

The vertices of citation networks are scientific papers, and the directed edges are citations. One cannot update the list of references in a published paper, so that new edges do not emerge between old papers. The direction of an edge between two papers is rigorously determined by their ages, so that one may forget about the directedness of citation networks. Such citation graphs (see Fig. 14.1) are actually very simple growing networks, and most demonstration models of growing networks belong to this class. Note that in electronic archives one can update old papers and lists of references in them. This produces new links between old papers, so that the networks of citations of electronic archives are not quite classical citation graphs.

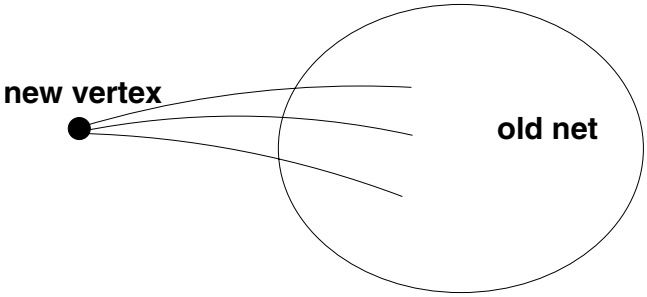


Figure 14.1: Scheme of the growth of a citation graph. New connections emerge only between a new vertex and old ones. New connections between old vertices are impossible.

The statistics of citations in scientific journals was studied in Ref. [21] (see the earlier empirical study of the issue in Ref. [22]). These data were collected for a number of journals (about 10) in the period 1991-1999. In all the journals that were studied in Ref. [21], the average number of references in papers was found to increase.

(iv) *Collaboration networks:*

In the simplest version of a collaboration network, vertices are collaborators. A pair of vertices is connected together by an undirected edge if there was at least one act of collaboration between them [4, 23]. For example, in scientific collaboration networks (networks of coauthorships), vertices are authors, and edges are coauthorships [24]. Such networks are projections of more complex and informative bipartite graphs, which contain two types of vertices: collaborators, and acts of collaboration. Each collaborator is connected to all the acts of collaboration in which he was involved. Empirical data are mostly collected for simple one-mode collaboration networks.

Empirical data of Refs. [25, 26] for large scientific collaboration networks indicate the linear growth of their average degree with increasing number of their vertices. This means that the total number of edges in a network increases as the square of the total number of vertices.

Thus we see that the accelerated growth of networks is not an exception but rather the rule. On the contrary, linear growth is a simple but very particular case.

14.2 Reasons for acceleration

Why is accelerated growth widespread? As an example, consider the growth of the WWW. Let us discuss how new pages appear in the WWW (see Fig. 14.2) [3]. Discussion of the growth of the WWW may be found in Refs. [27,28]. Suppose that you want to create your own personal home page. You prepare it, put references to some pages of the WWW (usually, there are several such references, but in principle the references may be absent), etc. But this is only the first step. You must make your page accessible in the WWW. Your system administrator puts a reference to it (usually one reference) in the home page of your institution, and your page in the WWW.

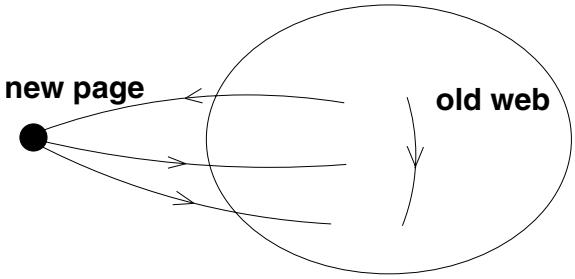


Figure 14.2: Scheme of the growth of the WWW. A new document of the WWW must have at least one incoming hyperlink to become accessible. It may contain any number of references to other pages of the WWW, but usually there are several such outgoing hyperlinks. Also, new hyperlinks emerge between old pages of the WWW.

However, you proceed to work with your page. From time to time, you add new references to it. Of course, you may remove some old references, but usually the total number of references in a page grows. Then the average degree of the WWW increases, that is, the growth of the WWW is naturally accelerated.

14.3 Degree distributions of networks

14.3.1 Types of degree distribution

In this paper we restrict ourselves to degree distributions of networks. Most empirical results are obtained for this simple basic characteristic. Unfortunately, a degree distribution (in-, out-degree distribution) is a restricted characteristic of networks. Indeed, degree is a one-vertex quantity, so that, in general, degree distribution does not yield information about the global topology of a network.

In most cases, for example, for growing networks, in which correlations between degrees of vertices are strong [11, 18, 19], a degree distribution is only the tip of the iceberg (see Fig. 14.3, a). Of course, if degree-degree correlations in a network are absent, then, knowing the degree distribution of a network, one can completely characterize the net (see Fig. 14.3, b). We face this situation in many equilibrium networks.

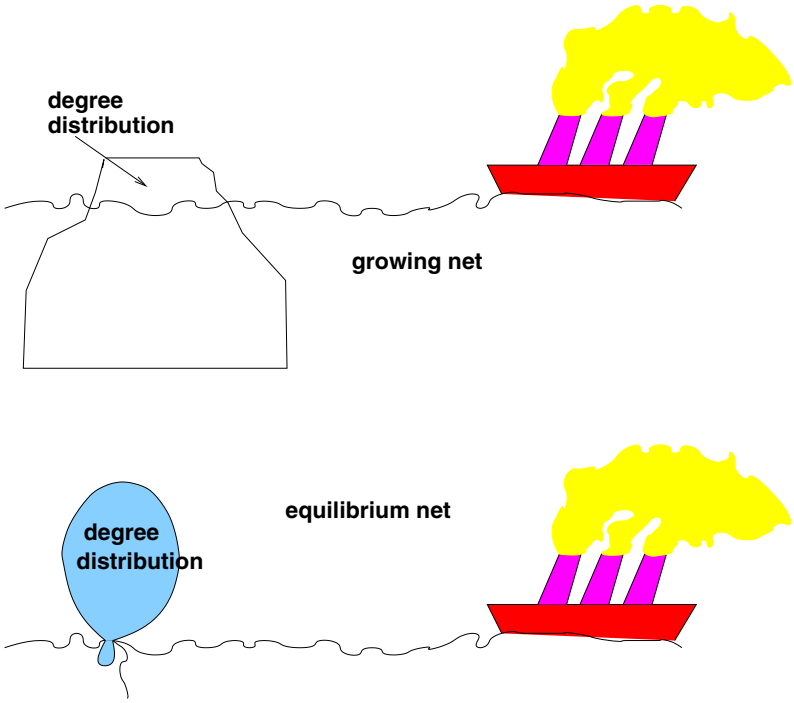


Figure 14.3: Degree-degree correlations, which are necessarily present in growing networks, make a degree distribution a far less informative characteristic (a). The degree distribution of the equilibrium uncorrelated network contains complete information about its structure (b).

Furthermore, analytical results on percolation on networks [29, 30], disease spread within them [31, 32], etc., have been obtained just for a simple construction without degree-degree correlations. This construction is a standard model of a maximally random graph with an arbitrary degree distribution taken from mathematical graph theory ("random graphs with restricted degree sequences") [33]. Luckily, it seems that the main percolation and disease spread results that were obtained for equilibrium networks are still valid for non-equilibrium nets.

What kinds of degree distributions are realized in networks? Here we list the main types with some simple examples of the corresponding networks.

(a) Poisson degree distribution, $P(k) = e^{-\overline{k}}\overline{k}^k/k!$ (see Fig. 14.4, a).

The Poisson distribution is realized in a classical random equilibrium graph of Erdös and Rényi [34, 35] in the limit of an infinite network, that is, when the total number of vertices N is infinite. Pairs of randomly chosen vertices are connected by edges. Multiple edges ("melons") are forbidden. One may create L edges in the graph, or connect pairs of vertices with the probability $L/[N(N-1)/2]$. In both these cases, the resulting graph is the same in the limit $N \to \infty$.

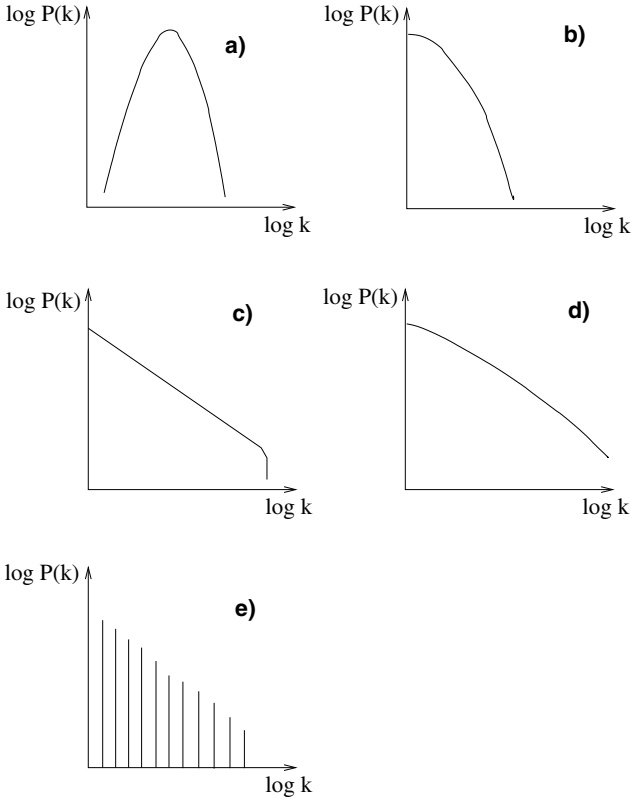


Figure 14.4: "Zoology" of degree distributions in networks. Main types of a degree distribution in log-log plots. Poisson (a), exponential (b), power-law (c), multifractal (d), and discrete (e) distributions.

(b) Exponential degree distribution, $P(k) \sim \exp(-k/\text{const})$ (see Fig. 14.4, b).

A citation graph (see Fig. 14.1) with attachment of new vertices to randomly chosen old ones produces the exponential distribution, but this is only one possible example. (Let each new vertex have the same number of connections, that is, the growth is linear.)

Also, the exponential degree distribution is rather common for many equilibrium networks that are constructed by mechanism of preferential linking.

(c) Power-law degree distribution, $P(k) \sim k^{-\gamma}$ (see Fig. 14.4, c).

Here the standard example is the Barabási-Albert model [4] (see also Ref. [5]). This growing network is a linearly growing citation graph in that new vertices are attached to preferentially chosen old ones. "Popular" old vertices attract more new connections than "failures": "*popularity is attractive*". This is a quite general principle, and is, for example, the one incorporated in the Simon model [36,37]. In the Barabási-Albert model, the probability that an edge becomes attached to some vertex is proportional to the degree

k of this vertex. This yields $\gamma = 3$. If the probability is proportional to $k + \text{const}$ (a linear preference function), γ takes values between 2 and ∞ as the constant changes from -1 to ∞ [7].

Power-law distributions are usually called scale-free or fractal.

(d) Multifractal degree distributions (see Fig. 14.4, d).

This distribution has a continuum spectrum of power laws with different weights. The growth of a network may produce such a degree distribution if new vertices partially copy degrees of old ones [38]. In particular, multifractal degree distributions emerge in some models of networks of protein-protein interactions [39]. Multifractal distributions are a more general case of fat-tailed distributions than power-law distributions. Numerous empirical data have been fitted by a power-law dependence. However, there have been no attempts to check the possibility that at least some empirical degree distributions are multifractal.

(e) Discrete degree distributions (see Fig. 14.4, e).

Deterministic growing graphs have a discrete spectrum of degrees. Recently, it was demonstrated that some simple rules of deterministic growth may produce discrete degree distributions with a power-law decay [40]. Moreover, deterministic graphs from Refs. [3, 41, 42] have an average shortest-path length, which is proportional to the logarithm of their size. Figure 14.5 shows a simple deterministic graph [3,41] with the discrete degree distribution that is characterized by exponent $\gamma = 1 + \ln 3/\ln 2$.

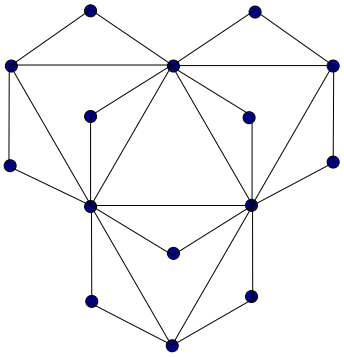


Figure 14.5: A simple deterministic graph [3,41] with a power-law discrete degree distribution. The growth starts from a single edge between two vertices. At each time step, each edge of the graph generates a new vertex, which becomes attached to both the end vertices of the mother edge. The average shortest-path length of this graph grows logarithmically with the total number of vertices.

14.3.2 Power-law degree distribution

Power-law (that is, "scale-free") degree distributions are a prominent particular case of fat-tailed degree distributions, which are widespread in real networks (both natural and artificial) [4, 5, 23]. Let us discuss briefly the general features of power-law distributions.

One may ask, what are the possible values for γ? The first natural restriction follows from the normalization condition $\int dk\, P(k) = 1$ (in this discussion we change the corresponding sum to the integral). We need not be worried about the low-degree region, since the degree distribution is certainly restricted below some characteristic degree k_0. Only the large-degree behavior of the degree distribution is of interest to us. Therefore, the strong restriction is $\gamma > 1$, otherwise the integral is divergent.

If a network grows linearly, so that the first moment of the distribution, that is, the average degree $\overline{k}$, is independent of time, then we have the second restriction $\int dk\, kP(k) < \infty$. Therefore, $\gamma > 2$ for linearly growing networks.

A finite size effect cuts the power-law part of the degree distribution at large degrees. This produces size-dependent degree distributions. One may easily estimate the position of the cutoff k_{cut} in the situation where $\gamma > 2$. Let the total number of vertices in the net be t, and k_0 be some characteristic degree, below which the distribution is, for example, constant or even zero. Then, using the normalization $\int dk\, P(k) = 1$ gives the power-law part of the degree distribution of the form $P(k) \sim [(\gamma - 1)k_0^{\gamma-1}]k^{-\gamma}$ for $k_0 < k < k_{cut}$.

When one measures the degree distribution of a network using only one realization of the growth process, strong fluctuations are observed at degree $k_f(t)$ that is determined by the condition $tP(k_f(t)) \sim 1$. This means that only one vertex in the network has such degree. (More rigorously speaking, the number of such vertices is of the order of one.) This is the first natural scale of the degree distribution.

One may improve the statistics by measuring many realizations of the growth process, or, for example, by passing to the cumulative distribution $P_{cum} \equiv \int_k^\infty dk\, P(k)$. Both these tricks allow us to reduce the above fluctuations. However, we still cannot surpass the next threshold that originates from the second natural scale, k_{cut}: $tP_{cum}(k_{cut}(t)) \sim 1$. This means that only one vertex in the network is of degree greater than k_{cut}. (Again, more rigorously, the number of such vertices is of the order of one.) Using the above expression for $P(k)$ gives

$$k_{cut} \sim k_0\, t^{1/(\gamma-1)}\,. \tag{14.1}$$

Notice that the only reason for this estimate for the cutoff is the natural scale of the problem. Hence more convincing arguments are necessary. The estimate was checked for some specific models. A growing network [13] was solved exactly, and the exact position of the cutoff coincided with Eq. (14.1). The degree distribution of this network has a typical form (see Fig. 14.6). Notice a hump near k_{cut} in Fig. 14.6. This is a trace of initial conditions. Simulation of a scale-free equilibrium network [43] also yielded the cutoff at this point. However, the introduction of the death of vertices in the network may change the estimate (14.1). This factor also removes the hump from the degree distribution. Here we do not consider such situations.

The cutoff (14.1) hinders measurements of power-law dependences in networks [13]. From Eq. (14.1) one sees that measurements of large enough γ are actually impossible. Indeed, in this case k_{cut} is small even for very large networks, and there is no room $\ln k_0 < \ln k < \ln k_{cut}$ for fitting.

What is the nature of power laws in networks? One may directly relate them to self-organized criticality. While growing under the mechanism of preferential linking, networks self-organize into scale-free structures, that is, are in a critical state. This critical state is realized for a wide range of parameters of preferential linking, namely for any linear preference

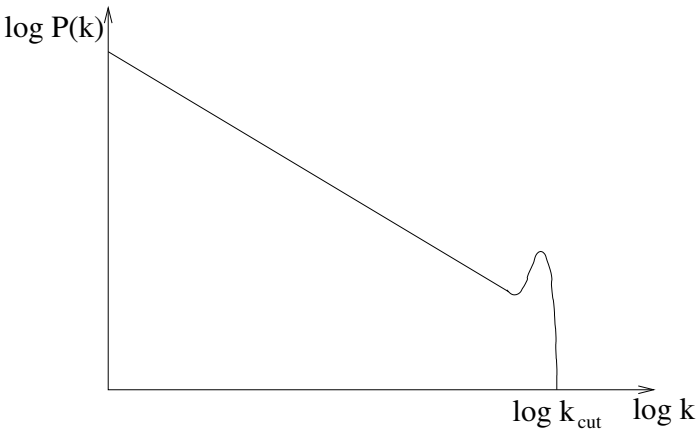


Figure 14.6: The typical form of a power-law degree distribution of finite growing networks. The finite size cutoff is given by Eq. (14.1). The hump near the cutoff depends on initial conditions (we do not account for the factor of mortality).

function (more rigorously, for any preference function that is asymptotically linear at large k [6]). The linear growth of networks may produce scale-free structures. Then, one may ask: What degree distributions does accelerated growth produce?

14.4 General relations for accelerated growth

Let us start with general considerations and not restrict ourselves to some specific model. Let the average degree grow as a power of t, $\overline{k} \propto t^a$, that is, the total number of edges $L(t) \propto t^{a+1}$. Here $a > 0$ is the growth exponent. This consideration is valid not only for degree, but also for in- and out-degrees, so we use the same notation k for all of them. We have chosen the power-law type of acceleration since one may hope that it provides scale-free networks. We suppose from the very beginning that this is the case and then check our assumption.

For accelerated growth, the degree distribution may be non-stationary. It is natural to choose its power-law part in the form

$$P(k,t) \sim t^z k^{-\gamma} . \tag{14.2}$$

Here we have introduced the new exponent $z > 0$ [3, 44, 45] (recall that we consider only $a > 0$). This form is valid only in the range $k_0(t) < k < k_{cut}(t)$. Using the normalization condition $\int_{k_0(t)}^{\infty} dk\, t^z k^{-\gamma} \sim 1$ gives

$$k_0(t) \sim t^{z/(\gamma-1)} . \tag{14.3}$$

This estimate is valid for any $\gamma > 1$.

The cutoff $k_{cut}(t)$ is estimated from the condition $t \int_{k_{cut}(t)}^{\infty} dk\, t^z k^{-\gamma} \sim 1$. Therefore,

$$k_{cut}(t) \sim t^{(z+1)/(\gamma-1)} \tag{14.4}$$

(compare with Eq. (14.1) for linear growth.) Equation (14.4) holds for any $\gamma > 1$.

We will consider two cases (see Fig. 14.7), $1 < \gamma < 2$ and $\gamma > 2$. Recall that we do not account for the mortality of vertices.

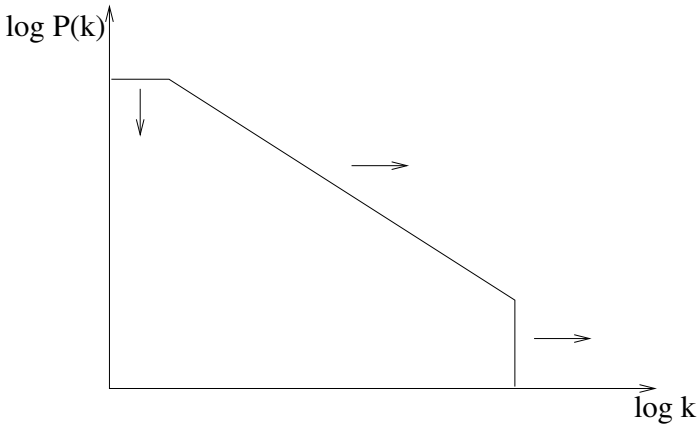


Figure 14.7: Schematic plot of a time-dependent degree distribution of networks that grow in the accelerated mode. Arrows show how the degree distribution changes with time.

(i) $1 < \gamma < 2$.

Recall that the average degree distribution $\overline{k}(t) \sim t^a$. Then

$$t^a \sim \int^{t^{(z+1)/(\gamma-1)}} dk\, kt^z k^{-\gamma} \sim t^{-1+(z+1)/(\gamma-1)}\,.$$

Here the value of the integral is determined by its upper limit. Therefore, $(z+1)/(\gamma-1) = a+1$, and the cutoff is of the order of the total number of edges in the network,

$$k_{cut}(t) \sim t^{a+1} \sim L(t)\,. \tag{14.5}$$

But this is the maximum possible degree in the problem. In this sense, any cutoff of a degree distribution is absent if $\gamma < 2$. From the last relation, we obtain the exponent γ in such a situation,

$$\gamma = 1 + \frac{z+1}{a+1}\,. \tag{14.6}$$

Here, for $\gamma < 2$, one assumes that $z < a$. The lower boundary for γ, namely $\gamma = 1 + 1/(a+1)$, is approached when $z = 0$, that is, when the distribution is stationary.

(ii) $\gamma > 2$.

The integral for the average degree is determined by its lower limit

$$t^a \sim \int_{tz/(\gamma-1)} dk\, kt^z k^{-\gamma} \sim t^{z-z(\gamma-2)/(\gamma-1)}\,.$$

Hence

$$\gamma = 1 + \frac{z}{a} \tag{14.7}$$

and $z > a$ to keep $\gamma > 2$. Notice that this relation is not valid for $a = 0$. One sees that, in this case, the degree distribution cannot be stationary: $z > a > 0$.

14.5 Scaling relations for accelerated growth

For simple scale-free networks that grow in a linear mode, simple scaling relations can be written [7,9]. Let us briefly describe the corresponding scaling relations for accelerated growth. If vertices in a growing network do not die, one can label them by their "birth date" $0 < s < t$. We denote by $p(k, s, t)$ the probability that the vertex s is of degree q. The average degree of a vertex s at time t is $\overline{k}(s, t) \equiv \int dk\, kp(k, s, t)$.

For the networks that we consider, $\overline{k}(s, t)$ is

$$\overline{k}(s, t) \propto t^{\delta} \left(\frac{s}{t}\right)^{-\beta} , \tag{14.8}$$

where β and γ are scaling exponents. One can show [45] that $p(k, s, t) = [1/\overline{k}(s, t)]g[k/\overline{k}(s, t)]$, where $g[\,]$ is some scaling function; therefore

$$p(k, s, t) = t^{-\delta} \left(\frac{s}{t}\right)^{\beta} g\left[kt^{-\delta} \left(\frac{s}{t}\right)^{\beta}\right] . \tag{14.9}$$

Using the relation $P(k, t) = t^{-1} \int_0^t ds\, p(k, s, t)$ yields

$$\int_0^{\infty} dx\, t^{-\delta} x^{\beta} g[kt^{-\delta} x^{\beta}] \propto t^{\delta/\beta} k^{-1-1/\beta} \propto t^{z} k^{-\gamma} , \tag{14.10}$$

whence we obtain relations for the scaling exponents:

$$\gamma = 1 + 1/\beta \tag{14.11}$$

and

$$z = \delta/\beta . \tag{14.12}$$

Taking account of these relations gives the scaling form:

$$p(k, s, t) = \frac{s^{1/(\gamma-1)}}{t^{(z+1)/(\gamma-1)}} g\left[k \frac{s^{1/(\gamma-1)}}{t^{(z+1)/(\gamma-1)}}\right] . \tag{14.13}$$

Similarly, one can find the scaling form for the degree distribution:

$$P(k, t) = t^{z} k^{-\gamma} G(kt^{-(1+z)\beta}) = t^{z} k^{-\gamma} G(kt^{-(1+z)/(\gamma-1)}) , \tag{14.14}$$

where $G(\,)$ is a scaling function. When $z = 0$, Eqs. (14.13) and (14.14) coincide with the scaling relations [7,9] for linearly growing networks.

Notice that it is sufficient to know a and only one exponent of γ, β, z, δ, or x to find all the others.

14.6 Degree distributions produced by acceleration

Let us discuss several illustrative examples. To begin with, we consider a network growing under the mechanism of preferential linking, in which the number of new connections increases as a power law in time. At this point we do not discuss the origin of this power-law dependence. Let it be equal to $c_0 t^a$, where c_0 is some positive constant. Here it is convenient to study the in-degree distribution, so that k will be the in-degree. In such an event we are interested only in incoming connections, so that the outgoing ends of new edges may be attached to any vertices of the network or even be outside of the net.

Let the probability that a new edge becomes attached to a vertex of in-degree k be proportional to $k + A(t)$, where $A(t)$ is some additional attractiveness of vertices. Two particular cases of this linear preferential linking are considered below in the framework of a simple continuum approach [5, 9, 45].

14.6.1 Model for $\gamma < 2$

If the additional attractiveness is constant, $A = \text{const}$, the continuum equation for the average in-degree $\overline{k}(s,t)$ of individual vertices that are born at time s and are observed at time t is of the form

$$\frac{\partial \overline{k}(s,t)}{\partial t} = c_0 t^a \frac{\overline{k}(s,t) + A}{\int_0^t du\, [\overline{k}(u,t) + A]}\,, \tag{14.15}$$

with additional starting and boundary conditions $\overline{k}(0,0) = 0$ and $\overline{k}(t,t) = 0$. Here we have supposed that new vertices have no incoming edges. We use this assumption only for brevity. Naturally, the total in-degree of the network is $\int_0^t du\, \overline{k}(u,t) = c_0 t^{a+1}/(a+1)$. This can also be seen by integrating both sides of Eq. (14.15) over s. Taking account of the last equality yields the solution of Eq. (14.15):

$$\overline{k}(s,t) = A \left(\frac{s}{t}\right)^{-(a+1)}\,. \tag{14.16}$$

Therefore, the β exponent equals $a + 1 > 1$, so that using scaling relation (14.11) gives

$$\gamma = 1 + \frac{1}{a+1} < 2\,. \tag{14.17}$$

One may also apply the following simple relation of the continuum approach:

$$P(k,t) = \frac{1}{t}\int_0^t ds\, \delta(k - \overline{k}(s,t)) = -\frac{1}{t}\left(\frac{\partial \overline{k}(s,t)}{\partial s}\right)^{-1}\Bigg|_{s=\overline{k}(s,t)}\,. \tag{14.18}$$

This equality follows from the fact that the solution of the master equation for the probability $p(k,s,t)$ in the continuum approximation is the δ-function. From Eqs. (14.16) and (14.17) we obtain the in-degree distribution

$$P(k,t) = \frac{A^{1/(a+1)}}{a+1} k^{-[1+1/(a+1)]}\,, \tag{14.19}$$

which is stationary. We have shown in Sec. 14.4 that when $\gamma = 1 + 1/(a+1)$, the (in-)degree distribution must be stationary, and the exponent z is zero. This is the case for the network under consideration.

14.6.2 Model for $\gamma > 2$

Now we choose a different rule of attachment of new edges to vertices. Let the additional attractiveness be time-dependent. Furthermore, let it be proportional to the average in-degree of the network, $c_0 t^a/(a+1)$, at the birth of an edge, $A(t) = Bc_0 t^a/(a+1)$. Here $B > 0$ is some constant. Analogously to the above we obtain the non-stationary in-degree distribution

$$P(k,t) \sim t^{a(1+B)/(1-Ba)} k^{-[1+(1+B)/(1-Ba)]} \tag{14.20}$$

for $k \gg t^a$. Hence the γ exponent is

$$\gamma = 1 + \frac{1+B}{1-Ba} > 2\,. \tag{14.21}$$

The scaling regime is realized when $Ba < 1$.

14.6.3 Dynamically induced accelerated growth

We have shown above that the power-law growth of the total number of edges in a network (or its average degree) produces fat-tailed distributions. Now we discuss the reasons for the power-law growth.

Consider an undirected citation graph, in which each new vertex becomes attached to a randomly chosen old one and to some of its nearest neighbors, to each one of which with probability p (see Fig. 14.8). For the total number of edges $L(t)$ one can write

$$L(t+1) - L(t) = 1 + p\overline{k}(t)\,. \tag{14.22}$$

Here we use the continuum approximation, $\overline{k}(t) = 2L(t)/t$, and therefore

$$\frac{1}{2}\frac{d}{dt}[t\overline{k}(t)] = 1 + p\overline{k}(t)\,. \tag{14.23}$$

For $p < 1/2$, the solution of this equation approaches the stationary limit $\overline{k} = 2/(1-2p)$ as $t \to \infty$. In this case the degree distribution is stationary, and the γ exponent is $\gamma = 1 + 1/p > 3$.

The situation for $p > 1/2$ is quite different, the average degree of network growth being a power law, $\overline{k}(t) \sim t^{2p-1}$ for large networks. This produces a non-stationary distribution $P(k) \propto t^z k^{-\gamma}$ with $\gamma = 1 + 1/(1-p) > 3$ and $z = 1/(1-p) - 2$. Of course, other mechanisms for accelerated growth are also possible.

14.6.4 Partial copying of edges and multifractality

From Eq. (14.1) for the cutoff of a power-law (or, which is the same, fractal) distribution, one sees that the size dependence of the moments $M_m(t) \equiv \int dk\, k^m P(k,t)$ of this distribution is

$$M_m(t) \propto t^{\tau(m)}\,, \tag{14.24}$$

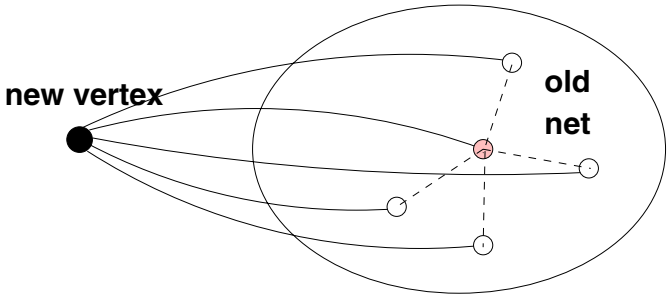


Figure 14.8: One of the possibilities to obtain the acceleration. In this citation graph, each new vertex becomes attached to a randomly chosen old one as well as to some of its nearest neighbors.

where the exponent $\tau(m)$ is a linear function of the order m of the moment,

$$\tau(m) = (m-1)/(\gamma-1) - (\gamma-2)/(\gamma-1).$$

Just the linearity of $\tau(m)$ defines a fractal distribution. The size dependence of the moments of a multifractal distribution also has the form (14.24), but its $\tau(m)$ exponent is a nonlinear function of m.

Multifractal distributions are a more general case of fat-tailed distributions than a power-law (fractal, scale-free) dependence. In Sec. 14.6.3 we have shown how accelerated growth may generate fractal distributions. However, this is only one particular possibility. Partial copying (partial inheritance) of degrees of old vertices by newborn ones together with the preferential attachment of some extra new edges usually provide networks that grow in a nonlinear way and have multifractal degree distributions.

A simple consideration of this problem can be found in Ref. [38]. Note that the acceleration and the multifractality of the degree distribution were obtained in a similar model [39] for protein-protein interaction networks. In this model, duplication of vertices with edges attached to them and breaking of some connections of parent vertices were used instead of partial copying in Ref. [38].

14.7 Evolution of the Word Web

The weak point of network science is the absence of a convincing comparison of the numerous schematic models with real networks. Most models of growing networks only demonstrate intriguing effects but, in fact, are very far from reality. Available empirical data usually can be explained by applying various models with fitting parameters. As a rule, only the exponent of the empirical degree distribution is used for comparison.

Here we consider an exceptional situation, where a reasonable comparison of the model of a growing network with empirical data is possible *without any fitting*. Moreover, it is the idea of accelerated growth that yields an excellent agreement.

The problem of human language is a matter of immense interest in various sciences. How did language begin? How does language evolve? What is its structure? Quite recently, a novel

approach to language was proposed [46]. Human language was considered as a complex network of interacting words. Vertices in this "Word Web" are distinct words of language, and undirected edges are connections between interacting words.

Words interact when they meet in sentences. Different reasonable definitions yield very similar structures of the Word Web. For example, we can connect the nearest neighbors in sentences. This means that an edge between two words of language exists if these words are the nearest neighbors in at least one sentence in the bank of language. One sees that multiple connections are absent. Of course, this is a rather naive definition, but it is also possible to account for other types of correlations between words in a sentence [46]. The resulting network gives the image of language, which is available for statistical analysis.

The empirical degree distribution [46] of the Word Web is very complex (see Fig. 14.9). Therefore, a perfect description of these data without fitting would be convincing. Indeed, it is hardly possible to describe such a complex form of the distribution completely by coincidence. We show below that a minimal model of the evolving Word Web [47], with only known parameters of this network, provides such a perfect description.

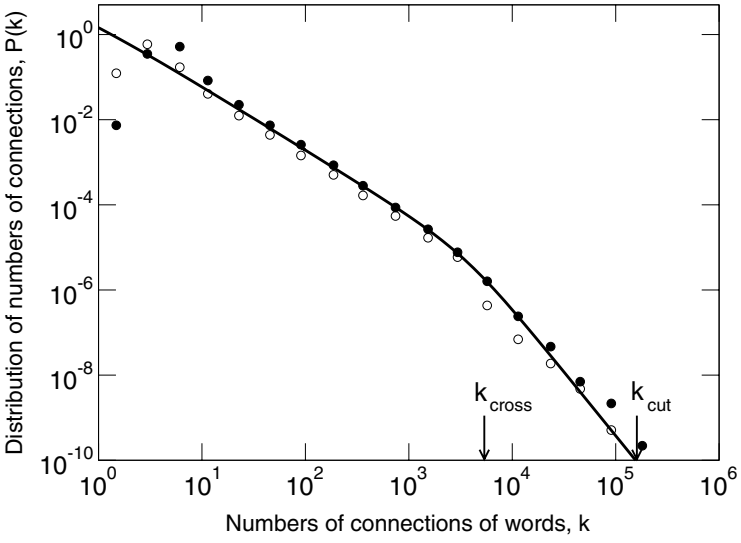


Figure 14.9: Empirical degree distribution of the Word Web (points) [46]. Empty and filled circles correspond to different definitions of the interactions between words in sentences. The solid line [47] shows the result of our calculations using the known parameters of the Word Web, namely the size $t \approx 470\,000$ and the average number of connections, $\overline{k}(t) \approx 72$. The arrows indicate the theoretically obtained point of crossover, k_{cross}, between the regions with exponents $3/2$ and 3, and the cutoff k_{cut} of the power-law dependence due to a finite size effect.

In Ref. [46], the Word Web was constructed after processing $3/4$ million words of the British National Corpus. The British Corpus is a collection of text samples of both spoken and written modern British English. The resulting network contains about $470\,000$ vertices. The average degree is $\overline{k} \approx 72$. These are the only parameters of the network we know and can use in the model.

Notice that the quality of the empirical data is [46] is high: the range of degrees is five decades. The empirical degree distribution has two power-law regions with exponents 1.5 and about 3 (the latter value is less precise, since the statistics in this region is worse). The crossover point and the cutoff due to the finite size effect can be easily indicated (see Fig. 14.9).

We treat language as a growing network of interacting words. At its birth, a new word already interacts with several old ones. New interactions between old words emerge from time to time, and new edges emerge. All the time a word lives, it enters into new "collaborations". Therefore the number of connections grows more rapidly than the number of words: the growth of the Word Web is accelerated.

How do words find their collaborators in language? Here we again use the idea of preferential linking [4]; again the principle "*popularity is attractive*" works.

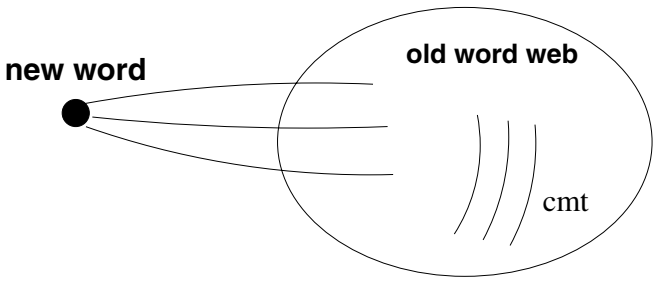


Figure 14.10: Scheme of the Word Web growth. At each time step, a new word emerges, so that t is the total number of words. It connects to $m \sim 1$ preferentially chosen old words. Simultaneously cmt new edges emerge between pairs of preferentially chosen old words. We use the simplest rule for preferential attachment when a node is chosen with probability proportional to the number of its connections.

We use the following rules for network growth (see Fig. 14.10) [47].

(1) At each time step, a new vertex (word) is added to the network, and the total number of words is t.

(2) At its birth, a new word connects to several old ones. On average, let this number be m, so that this number is not necessarily an integer. We use the simplest natural version of preferential linking: a new word becomes connected with some old one i with a probability proportional to its degree k_i, like in the Barabási-Albert model [4].

(3) In addition, cmt new edges emerge between old words, where c is a constant coefficient that characterizes a particular network. If each vertex makes new connections at a constant rate, this linear dependence on time naturally arises. These new edges emerge between old words i and j with a probability proportional to the product of their degrees $k_i k_j$ [10].

These simple rules define the minimal model that can be solved exactly. Here we discuss only the results of the continuum approach. In this case, the approach gives an excellent description of the degree distribution and the proper values of exponents.

In the model that we discuss, words are actually considered as collaborators in language. In our approach, the essence of the evolution of language is the evolution of collaborations between words. Therefore the situation for the Word Web should be rather similar to that for networks of collaborations. The equivalent model was applied to scientific collaboration nets [25], but the more complex nature of these networks makes the comparison impossible.

As above, in the continuum approximation, we can write the equation for the average degree at time t of the word that emerged at time s:

$$\frac{\partial \overline{k}(s,t)}{\partial t} = (m + 2cmt)\frac{\overline{k}(s,t)}{\int_0^t du\, \overline{k}(u,t)}\,, \tag{14.25}$$

where the initial condition is $\overline{k}(0,0) = 0$ and the boundary condition is $\overline{k}(t,t) = m$.

One can see that the total degree of the network is $\int_0^t du\, \overline{k}(u,t) = 2mt + cmt^2$, so that its average degree at time t is equal to $\overline{k}(t) = 2m + cmt$.

The solution of Eq. (14.25) is of a singular form

$$\overline{k}(s,t) = m\left(\frac{cmt}{cms}\right)^{1/2}\left(\frac{2m + cmt}{2m + cms}\right)^{3/2}\,. \tag{14.26}$$

The form of this equation indicates the presence of two distinct regimes in this problem. Using Eqs. (14.18) and (14.26) readily yields the non-stationary degree distribution

$$P(k,t) = \frac{1}{ct}\frac{cs(2+cs)}{1+cs}\frac{1}{k}\,, \tag{14.27}$$

where $s = s(k,t)$ is the solution of Eq. (14.26). Notice that, formally speaking, the number m is absent in Eq. (14.27). This is the consequence of our definition of the coefficient cm (see above).

From Eqs. (14.26) and (14.27), one sees that the non-stationary degree distribution has two regions with different behaviors separated by the crossover point

$$k_{cross} \approx m\sqrt{ct}(2+ct)^{3/2}\,. \tag{14.28}$$

The crossover moves in the direction of large degrees as the network grows. Below this point, the degree distribution is stationary,

$$P(k) \cong \frac{\sqrt{m}}{2}k^{-3/2}\,. \tag{14.29}$$

Above the crossover point, we obtain the behavior

$$P(k,t) \cong \frac{(2m+cmt)^3}{4}k^{-3}\,. \tag{14.30}$$

so that the degree distribution is non-stationary in this region. Thus, we have obtained two distinct values for the degree distribution exponent, namely, $3/2$ and 3.

The model that we consider has two limiting cases. When $c = 0$, it turns out to be the Barabási-Albert model, where $\gamma = 3$. When m is small but cm is large, we come to the network from Sec. 14.6.1 which has $\gamma = 3/2$ and a stationary degree distribution. Thus these two values of γ are not surprising. The important point is that the crossover is observable even though $cmt \gg m$.

The degree distribution has one more important point, the cutoff produced by a finite size effect. We estimate its position from the condition $t \int_{k_{cut}}^{\infty} dk\, P(k,t) \sim 1$ (see Secs. 14.3 and 14.4). This yields

$$k_{cut} \sim \sqrt{\frac{t}{8}}\,(2m + cmt)^{3/2}\,. \tag{14.31}$$

Using Eqs. (14.28) and (14.30) one can estimate the number of words above the crossover:

$$N_c \approx t \int_{k_{cross}}^{\infty} dk\, P(k,t) \sim \frac{m}{8c}\,. \tag{14.32}$$

We know only two parameters of the Word Web that was constructed in Ref. [46], namely $t = 0.470 \times 10^6$ and $\overline{k}(t) = 72 = 2m + cmt \approx cmt$. About m we know only that it is of the order of 1. From the above relations, one sees that the dependence on m is actually weak and is not noticeable on log-log scale plots. In fact, m is an inessential parameter of the model. Hence we can set its value to 1.

In Fig. 14.9, we plot the degree distribution of the model (the solid line). To obtain the theoretical curve, we used Eqs. (14.26) and (14.27) with $m = 1$ and $c \approx \overline{k}(t)/t$. The rather inessential deviations from the continuum approximation are accounted for in the small-degree region ($k \sim 10$). One sees that the agreement with the empirical data [46] is fairly good. Note that we do not use any fitting. However, for a better comparison, in Fig. 14.9, the theoretical curve is displaced upwards. Actually, this is not a fitting, since we have to exclude two empirical points with the smallest degrees. These points are dependent on the method of construction of the Word Web, on specific grammar, so that any comparison in this region is meaningless in principle.

From Eqs. (14.28) and (14.31), we find the characteristic values for the crossover and cutoff, $k_{cross} \approx 5.1 \times 10^3$, that is, $\log_{10} k_{cross} \approx 3.7$, and $\log_{10} k_{cut} \approx 5.2$. From Fig. 14.9 we see that these values coincide with the experimental ones. We should emphasize that the extent of agreement is truly surprising. The minimal model does not account for numerous, at first sight, important factors, e.g., the death of words, the variations of words during the evolution of language, etc.

The agreement is convincing since it is approached over the whole range of values of k, that is, over five decades. In fact, the Word Web turns out to be very convenient in this respect, since the total number of edges in it is extremely high, about 3.4×10^7 edges, and the value of the cutoff degree is large.

Note that there are few words in the region above the crossover point $k_{cross} \approx 5.1 \times 10^3$. These words have a different structure of connections than words from the rest of the language. With the growth of language, k_{cross} increases rapidly but, as follows from Eq. (14.32), the total number N_c of words of degree greater than k_{cross} does not change. It is a constant of the order of $m^2/(8cm) \sim 1/(8c) \approx t/(8\overline{k}) \sim 10^3$, that is, of the order of the size of

the small set of words forming the kernel lexicon of British English, which was estimated as 5000 words [48] and is the most important core part of the language. Therefore, our concept suggests that the number of words in this part of the language does not depend essentially on the size of the language. Formally speaking, the size of this core is determined by the value of the average rate c at which words find new partners in the language.

If our simple theory of the evolution of language is reasonable, then the sizes of the cores of primitive languages are close to those for modern "developed" languages.

14.8 Wealth distribution in evolving societies

Ideas from network science can be applied to various problems. Here we show how the idea of nonlinear growth works in econophysics.

One of the basic problems of econophysics is wealth distribution. Usually, wealth distribution is treated by using so-called stochastic multiplicative models. The standard description of these stochastic multiplicative processes is provided by the generalized Lotka-Volterra equation [49,50]. The preferential linking mechanism, that is, the general "*popularity is attractive*" principle, provides the stochastic multiplicative dynamics of networks [3]. Therefore, results that were obtained for networks may be easily interpreted in terms of wealth distribution.

Let us discuss briefly wealth distribution in stable (stagnating), developing, and degrading (dying) societies. For simplicity, in our very schematic consideration, we do not account for mortality, redistribution and loss of money, inflation, and many other important factors. Let there be one birth per time step. Therefore, there are t members of society at time t. Thus we consider growing (non-equilibrium) societies.

In stable societies, wealth per member (average capital, average amount of money) does not change with time, and the input flow of capital is constant. In developing societies, the average wealth and the input flow of capital grow with time. In degrading societies, these quantities decrease.

One introduces the distribution function of wealth, $P(k,t)$. If this distribution is a power law, $P(k) \sim k^{-\gamma}$, and $\gamma < 2$, the society is *"unfair"*: few persons keep a finite fraction of the total wealth. If $\gamma > 2$, the society is *"fair"*. The wealth condensation transition [51] occurs when $P(k)$ passes over the k^{-2} dependence. When $P(k)$ decreases more rapidly than a power law, e.g., the function is exponential, the society is *"superfair"*.

To study wealth distribution in various societies, we consider the simplest demonstration case of a power-law input flow of capital t^{α}. Growth exponent α indicates the type of society: $\alpha = 0$ corresponds to stable societies; positive and negative α exponents provide evolving and degrading societies, respectively.

Let us discuss the simplest situation. We assume that money attracts money. While trying to diminish inequality, society permanently distributes some fraction of wealth "fairly" (equally) between its members. Another way to make life better for all is to provide everybody with a starting capital. Society also provides its members with educational, etc., "capital", which can also attract money. Such a factor, additional attractiveness, A, is also proportional to the average wealth. It may be provided only once, at birth, $A(s)$ ($s < t$ is the birth time of an individual), or it may increase equally for all persons, $A(t)$ (t is the age of a

society), but in both cases the effect is qualitatively similar to starting capital. We consider the first possibility, that is, providing some starting capital at birth as the simplest case.

We again apply the continuum approach. Then $\overline{k}(s,t)$ is the average wealth of the person born at time $s < t$, t being the present time.

14.8.1 Stable (stagnating) societies

Let m_s be the starting capital and let m extra wealth be distributed at each time step. A is a constant additional attractiveness. The total input flow of wealth is equal to $m+m_s$. A fraction p of the flow m is distributed among members of the society randomly, that is, "fairly"; and the flow $(1-p)m$ is distributed preferentially with probability proportional to your wealth. The continuum approach equation for the average individual wealth $\overline{k}(s,t)$ is of the form

$$\frac{\partial \overline{k}(s,t)}{\partial t} = \frac{pm}{t} + (1-p)m \frac{\overline{k}(s,t)+A}{\int_0^t du\,[\overline{k}(u,t)+A]}\,, \tag{14.33}$$

with the initial condition $\overline{k}(0,0) = 0$ and the boundary condition $\overline{k}(t,t) = m_s$. Integrating (by parts) both sides of Eq. (14.33) over s yields naturally $\int_0^t ds\,\overline{k}(s,t) = (m+m_s)t$. Similarly to the calculations of Sec. 14.6.1, we obtain the power-law wealth distribution with exponent γ:

$$\gamma = 2 + \frac{pm + m_s + A}{(1-p)m} > 2\,. \tag{14.34}$$

Thus, in stable societies, $\gamma > 2$, so that a stagnating society is fair.

14.8.2 Developing and degrading societies

Here we discuss a natural case: let your starting capital be proportional to the average wealth in the society at your birth, $m_s(t) = dmt^\alpha$, where d is a positive constant. In addition, wealth mt^α is distributed among the members of the society at each increment of time. The wealth pmt^α is distributed equally, and the wealth $(1-p)mt^\alpha$ is distributed preferentially (money comes to money). For brevity, we set $A(s,t) = 0$. Then we have

$$\frac{\partial \overline{k}(s,t)(s,t)}{\partial t} = mt^\alpha \frac{p}{t} + (1-p)mt^\alpha \frac{\overline{k}(s,t)}{\int_0^t du\,(u,t)}\,. \tag{14.35}$$

The initial and boundary conditions are $\overline{k}(0,0) = 0$ and $\overline{k}(t,t) = dmt^\alpha$, respectively. From Eq. (14.35), one sees that $\int_0^t ds\,\overline{k}(s,t) = m(1+d)t^{\alpha+1}/(\alpha+1)$.

From Eq. (14.35) we obtain the wealth distribution for various values of the parameters of the problem, p, d, and α.

(i) When $\alpha > (1-p)/(p+d)$, or, in other words, $p > (1-\alpha d)/(1+\alpha)$, the wealth distribution is exponential (the "superfair society").

(ii) For $\alpha < (1-p)/(p+d)$, we obtain the power-law wealth distribution with exponent

$$\gamma = 2 + \frac{(1+\alpha)(p+d)}{1-p-\alpha(p+d)}\,. \tag{14.36}$$

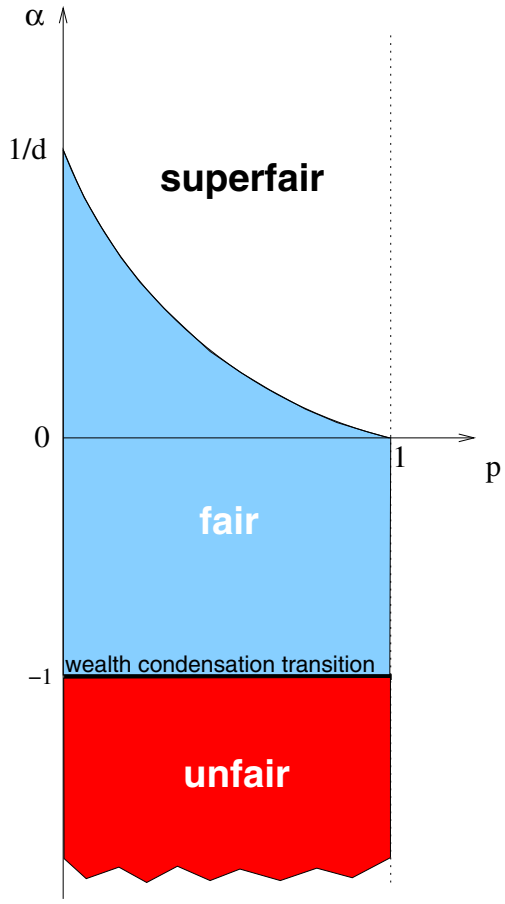


Figure 14.11: Phase diagram of evolving societies. At each time step, the wealth pmt^{α} is distributed equally between all the members of the society, and the wealth $(1-p)mt^{\alpha}$ is distributed preferentially (money comes to money). The individual starting capital is dmt^{α}. In our schematic model, the wealth distribution of the "superfair" society is exponential. The wealth distribution of the fair society is a power law with exponent $\gamma > 2$. For the unfair society $\gamma < 2$.

One sees that $\gamma = 2$ at $\alpha = -1$. This corresponds to the "wealth condensation transition" from the "fair" society ($\gamma > 2$ for $\alpha > -1$) to the "unfair" one ($\gamma < 2$ for $\alpha < -1$). The resulting phase diagram is shown in Fig. 14.11. Note that the position of the wealth condensation transition does not depend on the particular values of p and d. Therefore, even if a significant part of new wealth is distributed equally, rapidly degrading societies are necessarily unfair!

The general picture of wealth distribution in our minimal approach is quite natural. Extremely degrading societies are unfair. It is impossible to approach any "fairness" by the "fair" distribution of any part of new wealth in such a situation. "Fair" societies are possible only if there is some progress or the degradation is rather modest. Only in fair societies does "fair" distribution of new wealth produce visible results.

Conclusions

The nonlinear growth of networks is a more general situation than linear growth. In real evolving networks, nonlinear growth, in particular, accelerated growth, is widespread and is the rule and not the exception. In many cases, it is impossible to understand the nature of an evolving network without accounting for this acceleration.

The complicating circumstance is that existing empirical data clearly indicate the presence of acceleration but usually fail to yield its quantitative description. Theoreticians may easily choose any functional form for the nonlinear growth, but do these beautiful dependences have any relation to reality?

Acknowledgments

S.N.D. thanks PRAXIS XXI (Portugal) for a research grant PRAXIS XXI/BCC/16418/98. S.N.D. and J.F.F.M. were partially supported by the project POCTI/99/FIS/33141. We also thank A.V. Goltsev and A.N. Samukhin for many useful discussions.

References

[1] S.H. Strogatz, Exploring complex networks, Nature **410**, 268 (2001).

[2] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, cond-mat/0106096, Rev. Mod. Phys., **74**, 47-97 (2002).

[3] S.N. Dorogovtsev and J.F.F. Mendes, Evolution of networks, cond-mat/0106144 v. 2, Adv. Phys., **51**, 1079 (2002).

[4] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).

[5] A.-L. Barabási, R. Albert, and H. Jeong, Mean-field theory for scale-free random networks, Physica A **272**, 173 (1999).

[6] P.L. Krapivsky, S. Redner, and F. Leyvraz, Connectivity of growing random network, Phys. Rev. Lett. **85**, 4629 (2000).

[7] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin, Structure of growing networks with preferential linking, Phys. Rev. Lett. **85**, 4633 (2000).

[8] R. Albert and A.-L. Barabási, Topology of evolving networks: Local events and universality, Phys. Rev. Lett. **85**, 5234 (2000).

[9] S.N. Dorogovtsev and J.F.F. Mendes. Evolution of networks with aging of sites, Phys. Rev. E **62**, 1842 (2000).

[10] S.N. Dorogovtsev and J.F.F. Mendes, Scaling behaviour of developing and decaying networks, Europhys. Lett. **52**, 33 (2000).

[11] P.L. Krapivsky and S. Redner, Organization of growing random networks, Phys. Rev. E **63**, 066123 (2001).

[12] P.L.Krapivsky, G.J. Rodgers, and S. Redner, Degree distributions of growing networks, Phys. Rev. Lett. **86**, 5401 (2001).

[13] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Size-dependent degree distribution of a scale-free growing network, Phys. Rev. E **63**, 062101 (2001).

[14] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Graph structure of the web, Proceedings of the 9th WWW Conference, Amsterdam, 15-19 May, 2000, 309.

[15] R. Govindan, and H. Tangmunarunkit, Heuristics for Internet map discovery, Proceedings of the 2000 IEEE INFOCOM Conference, Tel Aviv, Israel, March, 2000, 1371; http://citeseer.nj.nec.com/govindan00heuristics.html.

[16] S.-H. Yook, H. Jeong, and A.-L. Barabási, Modeling the Internet's large-scale topology, cond-mat/0107417.

[17] M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationships of the Internet topology, Comput. Commun. Rev. **29**, 251 (1999).

[18] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Dynamical and correlation properties of the Internet, Phys. Rev. Lett. **87**, 258701 (2001).

[19] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, Large-scale topological and dynamical properties of Internet, Phys. Rev. E **65**, 066130 (2002).

[20] K.-I. Goh, B. Kahng, and D. Kim, Fluctuation-driven dynamics of the Internet topology, Phys. Rev. Lett. **88**, 108701 (2002).

[21] A. Vázquez, Statistics of citation networks, cond-mat/0105031.

[22] S. Redner, How popular is your paper? An empirical study of citation distribution, Eur. Phys. J. B **4**, 131 (1998).

[23] L.A.N. Amaral, A. Scala, M. Barthelemy, and H.E. Stanley, Classes of small-world networks, Proc. Nat. Acad. Sci. USA **97**, 11149 (2000).

[24] M.E.J. Newman, The structure of scientific collaboration networks, Proc. Nat. Acad. Sci. USA **98**, 404 (2001).

[25] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, Evolution of the social network of scientific collaborations, cond-mat/0104162.

[26] H. Jeong, Z. Néda, and A.-L. Barabási, Measuring preferential attachment for evolving networks, cond-mat/0104131.

[27] S. Bornholdt and H. Ebel, World Wide Web scaling exponent from Simon's 1955 model, Phys. Rev. E **64**, 035104 (2001).

[28] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin, WWW and Internet models from 1955 till our days and the "*popularity is attractive*" principle, cond-mat/0009090.

[29] M. Molloy and B. Reed, A critical point for random graphs with a given degree sequence, Random Struct. Algorithms **6**, 161 (1955).

[30] M.E.J. Newman, S.H. Strogatz, and D.J. Watts, Random graphs with arbitrary degree distribution and their applications, Phys. Rev. E **64**, 026118 (2001).

[31] R. Pastor-Satorras and A. Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett. **86**, 3200 (2001).

[32] R. Pastor-Satorras and A. Vespignani, Epidemic dynamics and endemic states in complex networks, Phys. Rev. E **63**, 066117 (2001).

[33] B. Bollobás, A probabilistic proof of an asymptotic formula for the number of labelled random graphs, Eur. J. Comb. **1**, 311 (1980).

[34] P. Erdős and A. Rényi, On random graphs, Publ. Math. **6**, 290 (1959).

[35] P. Erdős and A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).

[36] H.A. Simon, On a class of skew distribution functions, Biometrica **42**, 425 (1955).

[37] H.A. Simon, *Models of Man* (Wiley, New York, 1957).

[38] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin, Multifractal properties of growing networks, cond-mat/0011077; Europhys. Lett. **57**, 334 (2002).

[39] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, Modeling of protein interaction networks, cond-mat/0108043.

[40] A.-L. Barabási, E. Ravasz and T. Vicsek, Deterministic scale-free networks, Physica A **299**, 559 (2001).

[41] S.N. Dorogovtsev, A.V. Goltsev, and J.F.F. Mendes, Pseudofractal scale-free web, Phys. Rev. E **65**, 066122 (2002).

[42] S. Jung, S. Kim, and B. Kahng, A geometric fractal growth model for scale free networks, cond-mat/0112361.

[43] Z. Burda, J.D. Correia, and A. Krzywicki, Statistical ensemble of scale-free random graphs, Phys. Rev. E **64**, 046118 (2001).

[44] S.N. Dorogovtsev and J.F.F. Mendes, Effect of the accelerating growth of communications networks on their structure, Phys. Rev. E 63, 025101 (R) (2001).

[45] S.N. Dorogovtsev and J.F.F. Mendes, Scaling properties of scale-free evolving networks: Continuous approach. Phys. Rev. E **63**, 056125 (2001).

[46] R. Ferrer and R.V. Solé, The small-world of human language, Proc. Roy. Soc. London B **268**, 2261 (2001).

[47] S.N. Dorogovtsev and J.F.F. Mendes, Language as an evolving Word Web, Proc. Roy. Soc. London B **268**, 2603 (2001).

[48] R. Ferrer and R. Sole, Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revised, J. Quant. Linguistics **8**, 165–173 (2001); Working Papers of Santa Fe Institute, 00-12-068 (2000) http://www.santafe.edu/sfi/publications/Abstracts/00-12-068abs.html.

[49] S. Solomon and M. Levy, Spontaneous scaling emergence in generic stochastic systems, Int. J. Phys. C **7**, 745 (1996).

[50] D. Sornette and R. Cont, Convergent multiplicative processes repelled from zero: power laws and truncated power laws, J. Phys. I (France) **7**, 431 (1997).

[51] J.P. Bouchaud and M. Mézard, Wealth condensation in a simple model of economy, Physica A **282**, 536 (2000).

15 Social percolators and self organized criticality

Gérard Weisbuch and Sorin Solomon

Abstract

We discuss the influence of information contagion on the dynamics of choices in social networks of heterogeneous buyers. In the case of non-adaptive agents, the dynamics results in either the contagion process being stuck and very few agents actually buying (flops) or in a 'hit' where most agents a priori interested in getting the product actually buy it. We also show that when buyers and sellers try to adjust bids and asks the tatonement process does not converge to equilibrium at some intermediate market share and that large amplitude swings are actually observed across the percolation threshold.

15.1 Introduction

A number of recent studies (see e.g. Albert and Barabasi 2002) concern the structure of social networks although the subject itself has more than half century a history. Computers and computer networks have made data gathering and processing easier, hence the availability of more comprehensive data and new models of network structures. The study of what dynamics occur on a network and of its functional role, in term of a social institution the network can provide has not proceeded as fast, which is not uncommon in science: characterising dynamics is often more difficult than taking statistics. In the case of social networks, the problem is made harder because of the difficulty of simply modeling the local interactions occurring along the edges connecting agents.

A very simple process of data and information exchange among economic or political agents in a social context, is the idea that connected agents exchange pieces of information. Agents never possess a complete information about the world and can only take decision based upon incomplete information (and non-perfect information processing abilities). But they try to gather complementary pieces of information from their "neighbours" (agents to which they are connected).

One of the most studied dynamics on network is then the dynamics of epidemics, the "pathogen" being either a true pathogen as in epidemiology, or a rumor, an opinion, some piece of information etc.

Economists base the theory of social choice upon the idea that economic decisions are taken under the influence of other agents. They use the simplest mode of interaction among agents imitation. A number of economists characterized the dynamic outcome of such processes for identical agents: under a wide range of hypothesis, one observes homogeneity of

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

agent choices. These dynamics are often described as herding behavior and empirical phenomena such as bubbles in financial markets are interpreted along these lines.

When one tries to take into account the fact that economic agents are not identical, the situation gets more intricate and uniformity of choices is not the standard macro-behaviour as we can observe in real life. Since macroscopic behaviour of disordered systems can often display phase transitions, as it is the case for percolation, it is tempting to suggest that some surprising "stylised facts" observed in socio-economic systems could be explained by the same features. The purpose of this contribution is to shed some light on empirically observed phenomena with the help of two concepts proposed in natural sciences, percolation (Flory 1941, Broadbent and Hammersley 1957, Stauffer and Aharony 1994) and self organized criticality (Bak and Tang 1989).

A stylised fact observed in a number of economic and social phenomena could be denoted "hits or flops". In situation such as:

- markets such as toys and gadgets (Farrell 1998);
- the movie industries;
- the adoption of technological changes, or political and economical measures,
- the political arena where voters choose parties and political options.

one may observe a sharp contrast between between hits which market share is a large fraction of the market and flops with a nearly negligible success. One possible interpretation is the occurrence of percolation across a social network which we discuss in this contribution.

The next section concerns passive actors with fixed preferences for buyers (or adopters or voters) facing passive sellers (or decision makers or parties) with fixed offers: we then consider a "one shot" possible purchase or adoption situation.

But in the case of a series of purchases on different occasions, one would expect changes in the hits and flops situation in the presence of adaptive sellers which would want to avoid flops (!) and would be happier if they could obtain sales hits at a lower production cost. Buyers could also adjust their preference according to whether they were able to get what they wanted in the past. One of the question we raise in the later sections is whether a "tatonement" processes would restore equilibrium and balance between supply and demand. In fact, we will show by computer simulation that the resulting dynamics is characteristic of self organized criticality.

15.2 Social percolation

15.2.1 Simple models

Although the present discussion could apply to a number of equivalent situations, let use the case of a market with buyers and sellers to be more specific. We first start from the idea that agents lack full information about a product before buying it. But agents are not isolated. They are members of social networks: any of them is related to a number of other agents, called in our framework neighbours. An agent can get lacking information about a product from those of her neighbours who bought the product since they were able to use

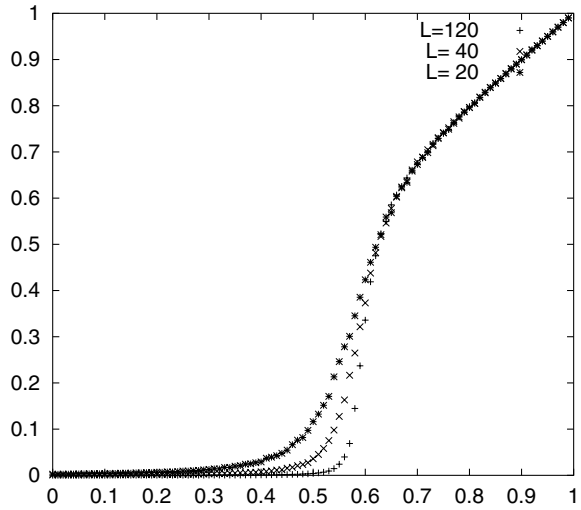


Figure 15.1: Fraction of purchasers as a function of the percentage of agents with preferences smaller than the quality of the product for square lattices of linear dimension 20, 40 and 120.

it. A simple hypothesis is that information is reducible to a scalar quantity that we call the quality q of the product, transferable from any agent who purchased to all her neighbours. On receiving the information, an agent can then take a decision about buying herself. In fact all agents might not want to buy a product of quality q: information is only one necessary condition to take a decision; if it happens that the information from their neighbours is "the quality of the product is less than what you demand", the agents take a negative decision. A simple way to implement such a decision process is to suppose that agents i each have individual preferences p_i. Heterogeneous agents of course have different preferences. The two conditions for purchase can be written as:

An agent will buy a product (or adopt a new technology, or vote for a party) when:

- she obtains information about the quality q from one of her neighbours who made a positive choice;
- if $q > p_i$.

As presently posed the problem becomes equivalent to what physicists call site percolation (Stauffer and Aharony 1994). Let us consider a network, lattice or with random connections, with a random uniform distribution of p_i's. If one starts from initial conditions with a large majority of undecided agents (by lack of information) and a small fraction of "early birds" who have already taken a positive decision, contagion proceeds across the net from newly decided agents to their neighbours. The early birds could be individuals who got the product for free or who happened to get direct relevant information in the case of markets, promoters of a new technology in case of the adoption of new technologies or party members in politics. Waves of purchase (or adoption) propagate across the network, new purchasers carrying the information at each time step to their neighbours. We can predict a priori that none of the

agents with preferences larger than the quality of the product will participate to the above described contagion process across the social network; they can then be removed as totally inactive. In fact for random distribution of p_i's the outcome of the process is:

- surprisingly predictable: the fraction of actual buyers (or adopters) only depends upon the fraction f of agents with
$$p_i < q$$
and not upon the initial distribution of early birds[1];
- surprisingly non-linear: the fraction of actual buyers is nearly a step function of fraction f with a threshold p_c.

Figure 1 summarizes simulations on square lattices of linear size 20, 40 and 120. f is varied from 0 to 1 by steps of 0.01. Each point correspond to an average over 1000 random samplings. Periodic boundary conditions were used. The initial condition is one early bird in the center of the lattice. The steepness of the percolation transition depends upon L: the slope would be vertical for infinite L.

The comparison of f to the percolation threshold p_c on the social network allows to predict the success or failure of the product with a very high probability:

- When $f > p_c$ most sites with preferences smaller than q are invaded, which results in a large success for the seller (or the promoters of the new technology, or the political party presenting the platform). The fraction of actual buyers is nearly equal to f in this regime as seen in figure 1.
- By contrast when $f < p_c$, contagion soon stops and the product is a commercial failure with very small market share.

Let us recall some well known figures (Stauffer and Aharony 1994) in the case of uniform random distributions of p_i's: for a square lattice with four neighbours per site $p_c = 0.593$, with eight neighbours $p_c = 0.407$, and for a random network with large connectivity k, $p_c = 1/(k-1)$ (the connectivity of a network is the number of neighbours per site). The above contrasted dynamic behaviours resulting from the necessity of contagion for adoption of a new product are very different from the prediction of "perfectly rational economics" which would yield a market share exactly equal to f.

This contrasted behaviour between the two dynamical regimes is a generic property independent of the details of the network connectivity. Furthermore it is not restricted to the hypothesis of unique criterion for agent's decision. In management science, a standard approach to decision making is multi-criteria analysis. Several criteria can influence agent decisions: they might take into account not only immediate or delayed profits but also time spent, risks, or even non-economic considerations such as moral, aesthetic or social values. The qualitative behavior that we above described, hits or flops, is also observed for multi-criteria analysis whatever procedure is used to process the criteria, as long as we suppose that all the information necessary for decision is transmitted at the same time by the neighborhood.

[1] Still prediction is difficult in the transition region, with outcome fluctuating largely for different realizations of the sampling of preferences. Furthermore what we mean by independence of initial distribution, refer to genericity: the property is true with a probability converging to one for neworks of infinite size and for random samplings.

15.3 Adjustment meta-dynamics

Among the situations that we try to model some are occurring rarely such as major technological changes, while some others are recurrent as in the movie industry: people often visit movie theaters (or restaurants) in their home town, and producers are producing new movies for the market at a frequency high enough for both parties to adjust supply and demand according to their previous experience. In fact, one should distinguish among two regimes according to whether the percolation process is faster or slower than the tatonnement process.

15.3.1 Slow adjustment

In the case of the movie industry, a director is producing movies at a slow rate with respect to the time it takes for the movies to eventually percolate across the social network of moviegoers. Readjustment is slower than percolation. The phenomenon can be formalised (Salomon et al. 2000) as:

- after opportunities during which they went to the movie, the agents will be more demanding and typically increase their expectations (here the preferences p_i); on the opposite, those who did not go, lower their preferences.
- after hits (resp. flops) the movie producers will decrease (resp. increase) the quality q of the produced movie(s), in their effort to remain above the threshold while minimizing expenses.

We have made computer simulations based on a fast contagion process leading eventually to percolation, embedded in a slower adjustment process as described above: we iterated a series of steps each one composed of a contagion process which was left evolving until percolation or its absence was checked, and of the resulting adjustment process. The early birds were agents with $p_i < q$ aligned along one side of the square lattice. Percolation (resp. its absence) was considered to be achieved when one site of the purchasing cluster reached the other side (resp. when the contagion process stopped earlier). All details concerning the algorithms that were used, including the Leath algorithm, are described in Solomon et al (2000).

After a transient adjustment period, we observed an alternation of hits and flops. Such a dynamics is often described as self organised criticality: the adjustment process brings and maintains the system parameters p_i's and q in the neighborhood of the percolation threshold.

The simulations were done on square lattices. We restricted ourselves to the simplest dynamics: the quality of the movie increased by δq if no cluster spanned from top to bottom, while it decreased by δq otherwise. The viewer's preference p_i, initially distributed randomly between 0 and 1, changed by $\pm\delta p$ depending on whether i went to the movie or not.

The results can be summarized as :

- Adaptive movie quality: for fixed agent preferences, $\delta p = 0$, and for adaptive quality $\delta q > 0$, we observed that the quality q moves to the usual percolation threshold; in this limit, the dynamics of our system is reminiscent of self-organization mechanisms arising in thermal critical phenomena [11], where a suitable feedback mechanism may push the temperature towards the critical temperature.

- Adaptive customer preferences: $\delta p > 0$ but fixed quality $\delta q = 0$, the p_i distribution drifts towards a single peak centered on the fixed q value, taken equal to 0.5 (no percolation) or 0.593 (some percolating clusters).

- When both δp and δq are positive, p_i and q drift towards $p_c = 0.593$, even if the initial q was 0.5. Our dynamic percolator thus drifts towards the standard percolation threshold of 0.593, exhibiting self-organized criticality (for longer times the p_i and q may drift away together).

We thus observe large-amplitude variations of the fraction of movie goers which is reminiscent of the fat tails observed in the distribution of return in financial markets. The above analysis is consistent with the interpretation of fat tails as due to strong cooperative effects, (here the constraints imposed by information contagion on the buy/sell decisions of the agents), in the neighbourhood of a transition[2]. Variants of this model are discussed in Ahmed and Abdusalam (2000), Goldenberg et al. (2000), Huang (2000), Das Gupta (2000), Weisbuch and Stauffer (2000), and Proykova and Stauffer (2002).

15.3.2 Fast adjustment

In the present case, we suppose that the time necessary for the information to propagate across the social net is larger than the time between successive purchases by the same buyer. Although purchases are repetitive, we further suppose that after purchasing a good, a buyer is not active during some kind of a refractory period m in analogy with the phenomena observed in nerve tissues or chemical reactions. The rationale for this hypothesis in economics is the existence of some consumption time for a commodity or some decay time for an investment between successive purchases. The regime of fast adjustment is observed when the refractory period is smaller than the time it takes for the purchase front to sweep across the lattice. In the iterated percolation model three conditions are then necessary for purchase by an agent:

- The agent has not taken any decision (to purchase or not) for a period at least equal to the refractory period m;

- he gets some private information from one of his neighbours;

- the quality q of the product is higher than his preference/ expectations p_i.

We further add preferences adjustment dynamics to the above decision process. We suppose that agents which purchased before the refractory period increase their preference coefficient for the next period by choosing p_i through a random sampling between p_i and 1. Alternatively, we suppose that agents who did not purchase before the refractory period decrease their preference coefficient for the next period by choosing p_i through a random sampling between 0 and q. Those agents which did not receive the information because they might have been screened by refusers don't readjust their preference.

[2] We do not imply here that the General Equilibrium theory is wrong, but simply that some of its assumptions about convexity don't apply to situations where the decision process gives such large variations as observed around the percolation transition

In the case of repetitive purchases, waiting until one of your neighbours purchases still makes sense for the interpretation of contagion based on positive externalities. For the interpretation based on information propagation, waiting for private information from a neighbour after the first purchase might seem unnecessary, since the agent should already know about q. The idea here is that q might change through some adjustment process of the producer, hence the necessity for a prospective buyer to wait for up-to-date private information before any decision. Since only the sign of the quantity $q - p_i$ matters for the decision process, readjusting randomly p_i should be equivalent to readjusting randomly both q and p_i. We ran most simulations with fixed q and tested the hypothesis on equivalence by a few runs with adaptive q.

Simulations set-up

The model defined above was run on square lattices with four neighbours and periodic boundary conditions. We ran many numerical simulations (see Weisbuch and Solomon 2000). We used the Leath algorithm to follow the time evolution of purchases. In the description of the results, one time step corresponds to the propagation of information and of the purchasing process from all purchasers at the current time step to all the immediate neighbours (at the next time step). The following quantities were monitored:

- Time evolution of purchases at each time step (figure 2).
- Patterns of purchase in the network at given time steps (figure 3).
- Average fraction of actual purchases at any given time step. (These average fractions were taken after an initial growth process of 100 time steps).
- The purchase process eventually comes to an end if no purchase is made during a time step. We have measured the average final time whenever the purchase process actually stopped, and the frequency of such occurrences. Averages were most often made over 5 000 runs. (The averages for the fractions of purchase and of agents willing to purchase are only taken for the set of non-stopping runs).
- For the set of non-stopping runs we averaged the power spectrum of purchases (figure 4).

The simulations were made on square lattices of size L equal to 40, 60, 80 and 160. We used periodic boundary conditions. The refractory period m was varied from 4 to L. For m larger than L the purchase process stops after one sweep of the network by the purchase front (see further) by lack of potential purchasers. Preferences p_i were initially randomly distributed (uniform distribution). One single run normally goes on for 4096 time steps: this is more than necessary to take fast Fourier transform since coherence is lost after roughly $2L$ time steps.

Typical results for a representative set of parameters

Let us analyse a first experimental condition with $L = 40, m = 10$, runtime of 4096 steps and averages taken on 500 runs. The figure 2 displays a typical time plot of sales. One can notice large fluctuations with short (around 25 steps) time scale correlations, very different from white noise.

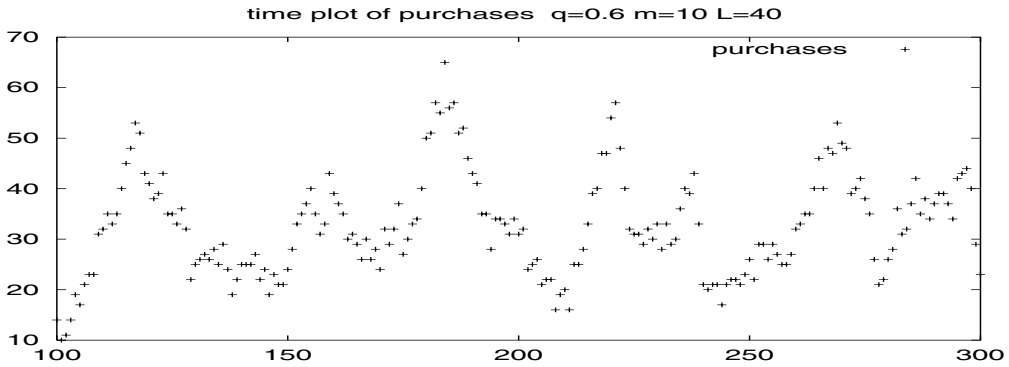


Figure 15.2: Time plot of purchases after 100 time steps for a 40×40 square lattice, with a refractory period of 10 and quality 0.6 slightly above the percolation threshold.

A typical pattern displaying the states of the agents is shown in figure 3. White cells correspond to "awake" agents that are ready to make a choice when triggered by eventual active neighbours. Their previous choices occurred more than m time steps ago. Other cells made their choice more recently: black cells correspond to agents who purchased and grey cells to agents who refused to purchase. On average (500 runs and 4096 time steps) 2 percent of the agents (O(1/L)) purchase at any time step. They are located on the black positions at the edges of the dark clusters, forming a disconnected purchasing front.

On-line monitoring of the patterns dynamics shows the displacement of the purchasing fronts across the lattice. The typical time scale, around 25, observed in figure 2 for the low frequency of the oscillation is similar to the time it takes to a front to move across the lattice and it scales with L.

We observed an average fraction of 0.593 potential purchasers, very close to the percolation threshold, which explains why quite often (19 percent of observed runs) the runs stop at an average time of 358, i.e. after a few sweeps across the lattice.

Finally, figure 4 displays the power spectrum of purchases averaged on those 81 percent of the 5000 runs which carried across 4096 time steps without stop. We observe a power law with a -2 exponent between periods 4 and 40. The flat spectrum at larger periods corresponds to the loss of coherence for times larger than the sweep time of fronts across the lattice.

Influence of simulation parameters and model variants

Systematic tests of simulation parameters and model variants were made.

The quality parameter q

When q is less than the percolation threshold, purchases stop very early since there is no percolation across the lattice. When q is higher than the percolation threshold, the dynamics soon re-shuffle the distribution of individual preferences p_i such that the fraction of agents with preferences below q gets closer to the percolation threshold.

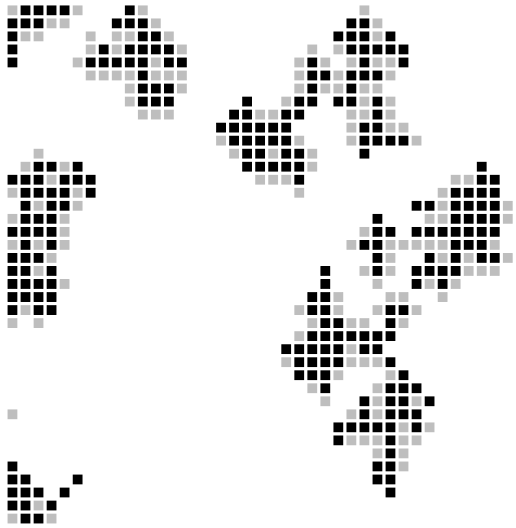


Figure 15.3: Pattern of purchases on a 40×40 square lattice. White cells correspond to "awake" agents that are ready to make a choice when triggered by their active neighbours. Black cells correspond to agents during their refractory period who purchased and grey cells to agents who refused to purchase.

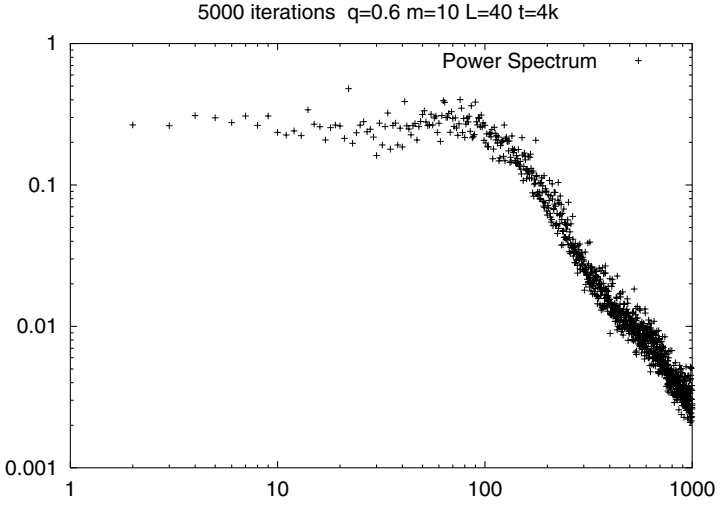


Figure 15.4: Log-log plot of the power spectrum of purchases computed on a time span of 4096 time steps and averaged over 3300 runs. 1000 on the x axis correspond to 4 periods, 100 to 40 periods. The high frequency (above 100) spectrum obeys a power law with a -2 exponent.

The refractory period m

At low values of m, the sweeping process of the purchase front across the lattice is less dominant and a lot of re-purchasing can occur locally. Purchase rate is doubled at $m = 4$ with respect to $m = 10$ and nearly all runs persist for long time, except 9 percent which stop very early after 7 time steps on average. The power spectrum does not display a well characterized power law.

Variants of the initial model

As mentioned in the introduction, we also tested the equivalence between the model with fixed q and a model with adaptive q to test the equivalence of dynamic properties of both model. We supposed that after each time step, a producer changes the quality of the proposed product by a "tatonnement" process: when a sale is made (resp. refused), she decreases q by δq (resp. increases) through equation:

$$q(t+1) = q(t) \pm \delta q. \tag{15.1}$$

We tried $\delta q = 0.0001$ and $\delta q = 0.00001$ corresponding to adjustments of roughly 0.1 and 0.01 per lattice sweep for q. The obtained Fourier spectra display the same scaling behaviour as for fixed q.

In the small world (Watts and Strogatz 1998) variant a small fraction (typically a few per thousand) of the connections were randomly assigned across the whole lattice without taking into account any notion of neighbourhood. The results were qualitatively analogous to those obtained when all connections are periodic.

We also tried to change the readjustment process of those agents which refused to purchase. Rather than randomly redistribute them in the segment $[0, q]$, we redistributed them in the segment $[0, p_i]$. The same dynamics were observed.

Finally another change was a slower adjustment of preferences: purchasers preferences being redistributed in $[p_i, p_i + \alpha(1 - p_i)]$ and non purchasers' preferences in $[p_i, (1 - \alpha)p_i]$ with $0 < \alpha < 1$. Once more we did not observe big changes.

15.4 Conclusions

In conclusion, we have seen that information contagion processes do not average but rather amplify at the "macro" level the consequences of the heterogeneity of individual agents.

In the absence of any readjustment dynamics of prices and preferences, contrasted behavior of either hits or flops are observed on either side of the percolation threshold. The imitation or polling processes, such as Polya urns, often described in the economics literature, also result in contrasted behavior depending upon initial conditions. On the opposite, for the contagion dynamics described here, no sensitivity to initial conditions is observed except in the neighborhood of the threshold.

Rather than driving the economic systems towards equilibrium with small eventual fluctuations, readjustment dynamics brings it close to the percolation transition resulting in large non-Gaussian fluctuations. In the case of slow readjustment, the system alternates from a flop regime with no contagion to the hit regime where the purchasing cluster percolate across the network.

Let us stress that although most simulations reported here were done on square lattices, we have no reason to doubt from the validity of the qualitative results for all sort of networks such as random nets or small world. The percolation transition is a generic property observable on any kind of networks.

One might also ask what are the assumptions that are critical for the observed dynamics. Randomness of the initial distribution of preferences is important. We also suppose that early birds are a minimal fraction of all possible buyers, and that their distribution is random, with no correlation with the distribution of preferences.

The percolation model is not the only instance of disorder in dynamical systems. A more general approach is that of INCA (Inhomogeneous Cellular Automata, Weisbuch 1990). Rather than a contact process such as one buying neighbour being a condition for purchase, in voters (or polling) models, an agents polls his neighbours and after comparing the number of buyers to a threshold, decide to purchase when this number is larger than the threshold. Inhomogeneity is introduced by having specific thresholds for each agents. Neural nets and zero temperature spin glasses are variants of this approach. The percolation contact process would correspond to a threshold for adoption being either one adopting neighbour or a threshold higher than the number of neighbors.

The INCA model gives rise to crossover transitions between growth and no growth of a buying process according to the parameters characterising the distribution of thresholds (Weisbuch and Boudjema 1999). We might then expect that the adjustment mechanism that we used for the percolation model would also display non-gaussian noise. We performed preliminary simulations based on the following adjustment process:

- increase the threshold for purchase after purchase;
- decrease the threshold for purchase after refusal.

Non gaussian fluctuations of the fraction of purchasing agents are indeed observed, but no clear scaling law appear.

Finally, along the same ideas, an alternative mechanism was proposed by Plouraboué et al. (1998) and Steyer and Zimmermann (2000) to explain long term correlations is also based on information diffusion on a social network but implies a Hebbian learning mechanism: connections among neighbours are of varying intensity, and are reinforced when agents take similar decisions. The decision process of individual agents is based on polling their neighbours rather than simply being triggered by any active neighbour. Steyer and Zimmermann report in the presence of Hebbian learning an even more correlated behaviour with a $1/f$ power spectrum instead of our $1/f^2$ result.

Our conjecture based on the several sets of simulations and on reasonable arguments is that long term correlations and non-exponential distribution of fluctuation sizes should be expected whenever decision processes are based on information contagion and seller/customer adaptation. Lux and Ausloos (2001) express the same idea by referring more generally to interactions among economic agents as the generating mechanism. Reciprocally, when long term correlations are observed, an information contagion process is a possible explanation for the effect.

These observations about dynamics can be used not only to describe empirical phenomena, but also to build “selling” strategies for sellers or policy implementers.

Acknowledgments

We thank ZhiFeng Huang, Abhijit Kar Gupta, Jean-Pierre Nadal, Dietrich Stauffer, Antonio Turiel and Jean Vannimenus for collaborations and helpul discussions.

Parts of this work were supported by CNRS, the Arc-en-Ciel program of French embassy in Tel-Aviv, SFB 341 and by Cray-T3E time of the Jülich supercomputer center. Part of it was achieved during visits of D.S and S.S to Ecole Normale which we thank for its hospitality.

References

[1] Ahmed E. and H.A. Abdusalam (2000) *Eur. Phys. J. B 16, 569*.

[2] Albert R. and Barabasi A. L. (2002) Statistical mechanics of complex networks, *Reviews of Modern Physics 74, 47*, cond-mat/0106096.

[3] Bak, P. and Tang, C. (1989) Earthquakes as an SOC phenomenon, *J. Geophys. Res.* 94, 15635-15637.

[4] Broadbent, S.K. and J.M. Hammersley (1957) Percolation processes I. Crystals and mazes, *Proc. Camb. Phil. Soc. 53, 629-641*.

[5] Farrell W. (1998), *How hits happen*, HarperCollins, New York.

[6] Flory, P.J. (1941) Thermodynamics of high polymer solutions, Journal of the American Chemical Society 63, 3083-3100.

[7] Goldenberg J., B. Libai, S. Solomon, N. Jan, D. Stauffer, (2000), Marketing percolation *Physica A* 284, 335-347.

[8] Huang, Z.F. (2000) *Int. J. Mod. Phys.* C 11, 287, and *Eur. J. Phys. B*, 16, 379.

[9] Kar Gupta, A. and Stauffer, D. (2000), *Int. J. Mod. Phys. C* 11, 695.

[10] Levy H., Levy M., and Solomon S. (2000) *Microscopic Simulation of Financial Markets*, Academic Press, New York.

[11] Lux T. and Ausloos M. (2001), Market Fluctuations I: Scaling, Multi-Scaling and Their Possible Origins in A. Bunde and H.-J. Schellnhuber (Hg.): *Facets of Universality in Complex Systems: Climate, Biodynamics and Stock Markets*, Berlin.

[12] Plouraboué F., Steyer A. and Zimmermann J.B., *Economics of Innovation and New Technology* 6, 73 (1998).

[13] Proykova, A. and Stauffer D. (2002) Social percolation and the influence of mass media, *Physica A* 312, p. 300.

[14] Solomon S., Weisbuch G., de Arcangelis L., Jan N., and Stauffer D. (2000) *Physica A* 277, 239.

[15] Stauffer D. and Aharony A. (1994) *Introduction to Percolation Theory*, Taylor and Francis, London.

[16] Steyer A. and Zimmermann J.B. (2000) Self Organised Criticality in Economic and Social Networks: The case of innovation diffusion in *Economics with Heterogeneous Interacting*, A. Kirman and Zimmermann ed. Springer, Berlin.

[17] Watts D. J. and S. H. Strogatz (1998) *Nature* 393, 440.

[18] Weisbuch G. (1990) *Complex Systems Dynamics* Santa-Fe Institute Studies in the Sciences of Complexity, Addison-Wesley, Redwood City, CA, USA.

[19] G. Weisbuch and G. Boudjema (1999), Dynamical aspects in the adoption of agri-environmental measures, *Advances in Complex Systems* 2, pp. 11-36.

[20] G. Weisbuch and D. Stauffer Hits and Flops Dynamics *Physica A* 287, 3-4, 563-576, (2000).

[21] G. Weisbuch and S. Solomon Self Organized Percolation and Critical Sales Fluctuations *Int. Jour. Mod. Phys. C*, Vol 11, No. 6, 1263-1272, (2000).

16 Graph theory and the evolution of autocatalytic networks

Sanjay Jain and Sandeep Krishna

Abstract

We give a self-contained introduction to the theory of directed graphs, leading up to the relationship between the Perron-Frobenius eigenvectors of a graph and its autocatalytic sets. Then we discuss a particular dynamical system on a fixed but arbitrary graph, that describes the population dynamics of species whose interactions are determined by the graph. The attractors of this dynamical system are described as a function of graph topology. Finally we consider a dynamical system in which the graph of interactions of the species coevolves with the populations of the species. We show that this system exhibits complex dynamics including self-organization of the network by autocatalytic sets, growth of complexity and structure, and collapse of the network followed by recoveries. We argue that a graph theoretic classification of perturbations of the network is helpful in predicting the future impact of a perturbation over short and medium time scales.

16.1 Introduction

Studies of networks are useful at several different levels (for recent reviews see [1–4]). At one level one is interested in describing the structure of natural and man-made networks such as food webs in ecosystems, biochemical and neural networks in organisms, networks of social interaction among agents in societies, and technological networks like the internet, etc. A useful representation of a network is a graph (and its generalizations) where the components of the network (which could be species, neurons, agents, etc.) are represented by nodes, and their mutual interactions by the links of the graph. Graph theory provides important tools to capture various aspects of the network structure.

At a second level one wants to know how the network structure of the system influences what happens in the system. E.g., the food-web structure of an ecosystem affects the dynamics of populations of the species, the network of human contacts influences the spread of a contagious disease, etc. At this level of discussion the network is typically taken to be static on the time scales of interest; the prime concern is the dynamics of other variables on a network with some particular type of (fixed) structure. Here dynamical systems theory is a major tool, and network variables (like the adjacency matrix elements of the underlying graph) appear as fixed parameters in the dynamics of other system variables like population, etc.

Handbook of Graphs and Networks: From the Genome to the Internet. Edited by S. Bornholdt, H. G. Schuster

ISBN: 3-527-40336-1

At a third level one is interested in how networks themselves change with time. Biochemical, neural, ecological, social and technological networks are not static, but are products of evolution. Moreover this evolution is quite complex in real systems. Networks sometimes self-organize and grow in size and complexity, and sometimes disintegrate. Their evolution is usually intertwined with other system variables, e.g., a food-web influences populations of species, and if a species goes extinct, the food-web changes. Understanding the processes and mechanisms involved in the evolution of complex networks is a major intellectual challenge.

A problem that illustrates all these levels is the problem of the origin of life on earth. The simplest living structure that we know — a bacterial cell — is a complex collection of several thousand types of molecules interacting with each other in a complex network of chemical interactions. The network may be described by a graph in which the nodes represent the molecular types or molecular species, and links connecting nodes represent chemical interactions between the molecular species. By participating in specific chemical reactions each molecular species or node plays a rather definite functional role in the organization of the cell: it permits or creates certain specific processes or spatial structures. Note that the complex chemical network of a cell is needed to produce the processes and structures that exist in it, and conversely, the same processes and structures are essential for maintaining the network and allowing it to evolve. If we assume that life originated on earth about 3.5 to 3.8 billion years ago as suggested by the microfossil evidence, then about 4 billion years back there was neither such a complex network of interactions nor such processes and structures existing anywhere on the earth. One of the puzzles of the origin of life on earth is: how did the network and the processes and spatial structures bootstrap themselves into existence when none was present — how did a chemical 'organization' emerge with individual molecular species playing definite roles in it?

A second puzzle concerns the highly 'structured' nature of the organization. The molecules appearing in cells are very special (a small subset in a very large space of possible molecules) and so is the graph that describes their interactions (a special kind of graph in the very large space of graphs). The probability of such structures arising by pure chance is astronomically small. If we assume that it was not an unlikely chance event that created life, we are led to the question: what then are the mechanisms that can create highly structured or 'ordered' organizations? A similar question is relevant for economic and social networks.

In order to address such questions in a mathematical model, one is naturally led to dynamical systems in which the graph describing the network is also a dynamical variable, whose dynamics is coupled to that of other variables such as the population of the molecular species. Here we present a model with such a structure, which has been inspired by the work in refs. [5–10]. The analysis of such dynamical systems is facilitated by the development of some new tools in graph theory. Another purpose of this article is to discuss some of these new tools. Together, the model and these tools address the above two questions about the origin of life, and provide partial answers. The model exhibits a mechanism by which a chemical organization can emerge where none existed through the formation of small *autocatalytic sets* of molecular species. In the model we also observe a *self-organizing process* which results in the growth of the initial autocatalytic set into a complex and highly structured chemical organization in a short time.

In addition, the model also captures, in an analytically tractable form, several phenomena that one associates with the evolution of other biological and social systems. These include

emergence of cooperation and interdependence in the system; crashes and recoveries of the system as a whole; 'core-shifts'; appearance of 'keystone species'; etc. We also argue that the juxtaposition of graph theory and dynamical systems provides the possibility of formulating more precisely notions that are important and useful in everyday language but otherwise difficult to pin down. In particular we attempt to formulate the notion of 'innovation' in this dynamical system, and classify innovations into categories according to their graph theoretic structure. It turns out that different categories of innovation have different short and longer term impact on the dynamics of the system.

This article is organized roughly according to the three kinds of network studies indicated above. In section 2 we discuss aspects of graph theory in a self-contained manner, reviewing older results as well as recent work. Among other things we describe a relation between topological properties of a graph (namely its autocatalytic sets) and its algebraic properties (the structure of the eigenvectors of its adjacency matrix). In section 3 we discuss a simple dynamical system describing molecular population dynamics on a fixed interaction graph. Here we show how structure of the graph influences the dynamics of the system; in particular relating the nature of its attractors to graph topology. Section 4 describes a model of graph evolution, motivated by the origin of life problem. In section 5 we show that the dynamics of this model exhibits self-organization and growth of cooperation and structure in the network, with analytical estimates of the time scales involved. Section 6 discusses the phenomena of crashes and recoveries exhibited by the model. In this section we also formulate a definition of innovation that seems appropriate for this model, and discuss a hierarchy of different categories of innovation and the roles they play in the ups and downs of the system. Finally, section 7 contains a discussion of some limitations of the model, speculations regarding the origin of life problem and possible future directions.

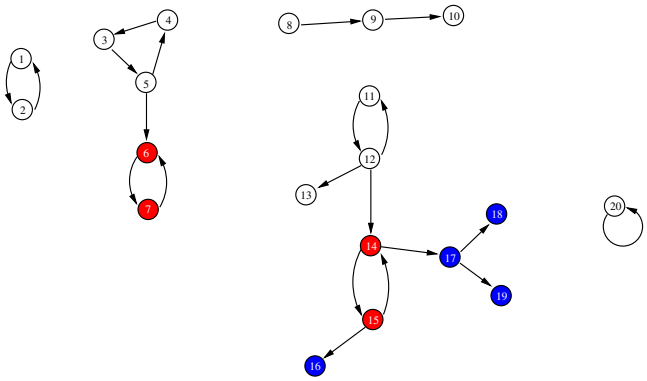
16.2 Graph theory and autocatalytic sets

16.2.1 Directed graphs and their adjacency matrices

A *directed graph* $G = G(S, L)$, often referred to in the sequel as simply a *graph*, is defined by a set S of 'nodes' and a set L of 'links' (or 'arcs'), where each link is an ordered pair of nodes [11, 12]. It is convenient to label the set of nodes by integers, $S = \{1, 2, \ldots, s\}$ for a graph of s nodes. An example of a graph is given in Figure 16.1a where each node is represented by a small labeled circle, and a link (j, i) is represented by an arrow pointing from node j to node i. A graph with s nodes is completely specified by an $s \times s$ matrix, $C = (c_{ij})$, called the *adjacency matrix* of the graph, and vice versa. The matrix element in the i^{th} row and j^{th} column of C, c_{ij}, equals unity if L contains a directed link (j, i) (arrow pointing from node j to node i), and zero otherwise. (This convention differs from the usual one where

$c_{ij} = 1$ if and only if there is a link from node i to node j; our adjacency matrix is the transpose of the usual one. We have chosen this convention because it is more natural in the context of the dynamical system to be discussed in subsequent sections.) Figure 16.1b shows the adjacency matrix corresponding to the graph in Figure 16.1a. We will use the terms 'graph' and 'adjacency matrix' interchangeably: the phrase 'a graph with adjacency matrix C' will often be abbreviated to 'a graph C'. Undirected graphs are special cases of directed

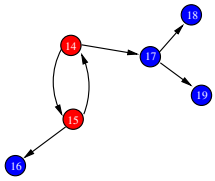
a)



b)

$$C = \begin{pmatrix}
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1
\end{pmatrix}$$

c)



d)

$e_1 = (1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0) / 2$

$e_2 = (0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0) / 2$

$e_3 = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 0) / 6$

$e_4 = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1)$

e)

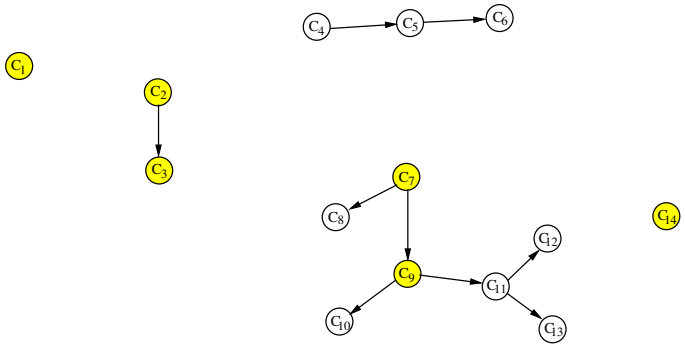


Figure 16.1: a. A directed graph with 20 nodes. **b.** The adjacency matrix of the graph in Figure 16.1a. **c.** A subgraph of the graph in Figure 16.1a. The adjacency matrix of the subgraph is the shaded portion of the matrix in Figure 16.1b. **d.** Four Perron-Frobenius eigenvectors (PFEs) of the graph in Figure 16.1a. The first three vectors have been divided by factors of 2, 2 and 6 respectively to normalize them. **e.** The irreducible decomposition of the graph in Figure 16.1a into subgraphs C_α, with $\alpha = 1, 2, \ldots, 14$. Each of the 14 nodes of the graph in Figure 16.1e represents either an irreducible subgraph of the graph in Figure 16.1a, or a single node that is not part of any irreducible graph. The basic subgraphs of the graph in Figure 16.1a are represented by yellow nodes. The dotted lines in Figure 16.1b demarcate the adjacency matrices corresponding to the subgraphs C_α. Colours identify the attractor of the dynamics discussed in section 3, except in Figure 16.1e. In all graphs in the article (except Figure 16.1e), white nodes have zero relative population in the attractor, $X_i = 0$, while blue and red nodes have $X_i > 0$. In graphs that have an autocatalytic set, red nodes belong to the core of the dominant autocatalytic set of the graph, blue nodes to its periphery, and white nodes are outside the dominant autocatalytic set.

graphs whose adjacency matrices are symmetric. A single (undirected) link of an undirected graph between, say, nodes j and i, can be viewed as two directed links of a directed graph, one from j to i and the other from i to j.

A graph $G' = G'(S', L')$ is called a *subgraph* of $G(S, L)$ if $S' \subset S$ and $L' \subset L$. We will use the term 'subgraph' if G' satisfies a stronger property: every link in L with both endpoints in S' also belongs to L'. That is, for us, a subgraph will be a subset of nodes together with *all* their mutual links. (This is often called an 'induced subgraph' in the literature [12].) The graph in Figure 16.1c (comprising nodes 14, 15, 16, 17, 18 and 19 and all their mutual links) is thus a subgraph of the graph in Figure 16.1a. For a subgraph we will often find it more convenient to label the nodes not by integers starting from 1, but by the same labels the corresponding nodes had in the parent graph. The adjacency matrix of a subgraph can be obtained by deleting all the rows and columns from the full adjacency matrix that correspond to the nodes outside the subgraph. The highlighted portion of the matrix in Figure 16.1b is the adjacency matrix of the subgraph in Figure 16.1c.

A *walk* of length n (from node i_1 to node i_{n+1}) is an alternating sequence of nodes and links $i_1 l_1 i_2 l_2 \ldots i_n l_n i_{n+1}$ such that link l_1 points from node i_1 to node i_2 (or $l_1 = (i_1, i_2)$), l_2 points from i_2 to i_3 and so on. A walk with all nodes distinct (except possibly the first and last nodes) will be called a *path*. If the first and last nodes i_1 and i_{n+1} of a walk or path are the same, it will be referred to as a *closed* walk or path. The existence of even one closed walk in the graph implies the existence of an infinite number of distinct walks in the graph. In the graph of Figure 16.1a, there is an infinite number of walks from node 11 to node 17 (e.g., $11 \to 12 \to 14 \to 17$, $11 \to 12 \to 11 \to 12 \to 14 \to 17$, ...) but no walks from node 11 to node 10. An undirected graph trivially has closed walks if it has any undirected links at all.

In the graph theory literature, what we have defined above to be a 'closed path' is usually referred to as a 'cycle'. However, for later convenience, we define a cycle somewhat differently. We define an *n-cycle* to be a subgraph with $n \geq 1$ nodes which contains exactly n links and also contains a closed path that covers all n nodes. E.g., the subgraph formed by node 20 and its self link is a 1-cycle, that formed by nodes 1 and 2 is a 2-cycle and by nodes 3,4 and 5 a 3-cycle. The subgraph formed by nodes 1,2,3,4 and 5 is not a 5-cycle because it does not have a closed path covering all the five nodes. The word 'cycle' will be used generically for an n-cycle of unspecified length. Given a directed graph C, its *associated undirected graph* (or 'symmetrized version') $C^{(s)}$ can be obtained by adding additional links as follows: for every link (j, i) in L, add another link (i, j) if the latter is not already in L. Two nodes of a directed graph C will be said to be *connected* if there exists a path between them in the associated undirected graph $C^{(s)}$, and disconnected otherwise. Thus any directed graph can be decomposed into 'connected components' which are maximal sets of connected nodes (e.g., the graph of Figure 16.1a has five connected components that are disconnected from each other). In a directed graph C, we refer to a node i as being 'downstream' from a node j if there is a path in C leading from j to i, and no path from i to j. Similarly i is 'upstream' from j if there is a path in C leading from i to j, and no path from j to i. Thus in Figure 16.1a, node 17 is downstream from node 11, or equivalently node 11 is upstream from node 17. Node 10 is neither upstream nor downstream from node 11 since they are not connected, and node 12 is neither upstream nor downstream from 11 because each can be reached from the other along some directed path.

If C is the adjacency matrix of a graph then it is easy to see that $(C^n)_{ij}$ equals the number of distinct walks of length n from node j to node i. E.g., $C^2_{ij} = \sum_{k=1}^{s} C_{ik}C_{kj}$; each term in the sum is unity if and only if there exists a link from j to k and from k to i; hence the sum counts the number of walks from j to i of length 2.

Perron-Frobenius eigenvalues and eigenvectors (PFEs)

A vector $\mathbf{x} = (x_1, x_2, \ldots, x_s)$ is said to be an eigenvector of an $s \times s$ matrix C with an eigenvalue λ if for each i, $\sum_{j=1}^{s} c_{ij}x_j = \lambda x_i$. The eigenvalues of a matrix C are roots of the *characteristic equation* of the matrix: $|C - \lambda I| = 0$ where I is the identity matrix of the same dimensionality as C and $|A|$ is the determinant of the matrix A. In general a matrix will have complex eigenvalues and eigenvectors, but an adjacency matrix of a graph has special properties, because it is a 'non-negative' matrix, i.e., it has no negative entries.

For any non-negative matrix, the Perron-Frobenius theorem [13, 14] guarantees that there exists an eigenvalue which is real and larger than or equal to all other eigenvalues in magnitude. This largest eigenvalue is often called the Perron-Frobenius eigenvalue of the matrix, which we will denote by $\lambda_1(C)$ for a graph C. Further the theorem also states that there exists an eigenvector of C corresponding to $\lambda_1(C)$ (which we will refer to as a Perron-Frobenius Eigenvector, PFE) all of whose components are real and non-negative. The Perron-Frobenius eigenvalue of the graph in Figure 16.1a is 1. Four PFEs of the graph in Figure 16.1a are displayed in Figure 16.1d.

The presence or absence of closed paths in a graph can be determined from the Perron-Frobenius eigenvalue of its adjacency matrix (see ref. [16] for a simple proof):

Proposition 1. *If a graph, C,*
(i) *has no closed walk then* $\lambda_1(C) = 0$,
(ii) *has a closed walk then* $\lambda_1(C) \geq 1$,
(iii) *has a closed walk and all closed walks only occur in subgraphs that are cycles then* $\lambda_1(C) = 1$.

Note that λ_1 cannot take values between zero and one because of the discreteness of the entries of C which are either zero or one. (Thus, for an undirected graph, if it has even one undirected link, $\lambda_1(C) \geq 1$.) Several results pertaining to the relationship of the graph structure to the structure of its PFEs can be found in ref. [15].

Irreducible graphs and matrices

A subgraph of a directed graph is termed *irreducible* if there is a path within the subgraph from each node in the subgraph to every other node in the subgraph. The simplest irreducible subgraph is a 1-cycle. In Figure 16.1a the subgraph comprising nodes 3,4 and 5 is irreducible, as is the subgraph of nodes 6 and 7, but the subgraph of nodes 3,4,5,6 and 7 is not irreducible since there is, for example, no path from node 6 to node 5.

If a graph or subgraph is irreducible then the corresponding adjacency matrix is also termed *irreducible*. Thus a matrix C is *irreducible* if for every ordered pair of nodes i and j there exists a positive integer k such that $(C^k)_{ij} > 0$. Refs. [13, 14] describes further properties of irreducible matrices.

The nodes of any graph can be grouped into a unique set of irreducible subgraphs as follows:

(1) Pick any node, say i. Find all the nodes which have paths leading to them starting at i. Denote this set by S_1; it may include i itself. Similarly find all the nodes which have paths leading to i. Denote this set by S_2. Denote the subgraph formed by the set of nodes $\{i\} \cup (S_1 \cap S_2)$ and all their mutual links as C_1. If $S_1 \cap S_2 \neq \phi$, then C_1 is an irreducible graph because every node of C_1 has a path within C_1 to every other node in it. If $S_1 \cap S_2 = \phi$, then i does not belong to any irreducible subgraph and C_1 consists of just the node i and no links.

(2) Pick another node which is not in C_1 and repeat the procedure with that node to get another subgraph, C_2. The sets of nodes comprising the two subgraphs will be disjoint.

(3) Repeat this process until all nodes have been placed in some C_α, $\alpha = 1, 2, \ldots, M$. Each C_α is either an irreducible subgraph or consists of a single node with no links.

Irrespective of which nodes are picked and in which order, this procedure will produce for any graph a unique set of disjoint subgraphs (upto labelling of the C_α) encompassing all the nodes of the graph. The graph in Figure 16.1a will decompose into 14 such subgraphs (see Figure 16.1e).

We say there is a path from an irreducible subgraph C_1 to another irreducible subgraph C_2 if there is a path in C from any node of C_1 to any node of C_2. The terms 'downstream' and 'upstream' can thus be used unambiguously for the C_α.

Structure of a general graph

A general adjacency matrix can be rewritten in a useful form by renumbering the nodes by the following procedure [13, 14]:
Determine all the subgraphs $C_1, C_2, \ldots, C_M$ of the graph as described above. Construct a new graph of M nodes, one node for each C_α, $\alpha = 1, \ldots, M$. The new graph has a directed link from C_β to C_α if, in the original graph, any node of C_β has a link to any node of C_α. Figure 16.1e illustrates what this new graph looks like for the graph of Figure 16.1a.

Clearly the resulting graph cannot have any closed paths. For if it were to have a closed path then the C_α subgraphs comprising the closed path would together have formed a larger irreducible subgraph in the first place. Therefore we can renumber the C_α such that if $\alpha > \beta$, C_β is never downstream from C_α. Now we can renumber the nodes of the original graph such that nodes belonging to a given C_α occupy contiguous node numbers, and whenever a pair of nodes i and j belong to different subgraphs C_α and C_β respectively, then $\alpha > \beta$ implies $i > j$. Such a renumbering is in general not unique, but with any such renumbering the adjacency matrix takes the following canonical form:

$$C = \begin{pmatrix} C_1 & & & 0 \\ & C_2 & & \\ & & \ddots & \\ & & & \ddots \\ & & & \ddots \\ R & & & C_M \end{pmatrix}$$

where 0 indicates that the upper block triangular part of the matrix contains only zeroes while the lower block triangular part, R, is not equal to zero in general. It can be seen that the graph

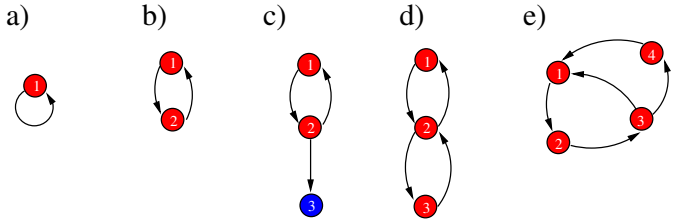


Figure 16.2: Various autocatalytic sets (ACSs). **a.** A 1-cycle, the simplest ACS. **b.** A 2-cycle. **c.** An ACS which is not an irreducible graph. **d,e** Examples of ACSs which are irreducible graphs but not cycles.

in Figure 16.1a is already in this canonical form. In Figure 16.1b, the dotted lines demarcate the block diagonal portions which correspond to the C_α.

From the above form of C it follows that

$$|C - \lambda I| = |C_1 - \lambda I| \times |C_2 - \lambda I| \times \ldots \times |C_M - \lambda I|$$

Therefore the set of eigenvalues of C is the union of the sets of eigenvalues of $C_1, \ldots, C_M$. $\lambda_1(C) = \max_\alpha\{\lambda_1(C_\alpha)\}$.

Therefore if a given graph C has a Perron-Frobenius eigenvalue $\lambda_1 > 0$ then it contains at least one irreducible subgraph with Perron-Frobenius eigenvalue λ_1. When $\lambda_1 > 0$, all irreducible subgraphs of C with Perron-Frobenius eigenvalue equal to λ_1 are referred to as *basic* subgraphs. The yellow nodes in Figure 16.1e correspond to the basic subgraphs of Figure 16.1a.

16.2.2 Autocatalytic sets

The concept of an autocatalytic set (ACS) was first introduced in the context of a set of catalytically interacting molecules. There it was defined to be a set of molecular species which contains a catalyst for each of its member species [17–19]. Such a set of molecular species can collectively self-replicate under certain circumstances even if none of its component molecular species can individually self-replicate. This property is considered important in understanding the origin of life. If we imagine a node in a directed graph to represent a molecular species and a link from j to i as signifying that j is a catalyst for i, this motivates the following graph-theoretic definition of an ACS in any directed graph: An *autocatalytic set* (ACS) is a subgraph, each of whose nodes has at least one incoming link from a node belonging to the same subgraph.

Figure 16.2 shows various ACSs. The simplest ACS is a 1-cycle; Figure 16.2a. There is the following hierarchical relationship between cycles, irreducible subgraphs and ACSs: all cycles are irreducible subgraphs and all irreducible subgraphs are ACSs, but not all ACSs are irreducible subgraphs and not all irreducible subgraphs are cycles. Figures 16.2a and 16.2b are graphs that are irreducible as well as cycles, 16.2c is an ACS that is not an irreducible

subgraph and hence not a cycle, while 16.2d and 16.2e are examples of irreducible graphs that are not cycles. It is not difficult to see the following [16]:

Proposition 2.
(i) *An ACS must contain a closed path*. Consequently,
(ii) *If a graph C has no ACS then $\lambda_1(C) = 0$.*
(iii) *If a graph C has an ACS then $\lambda_1(C) \geq 1$.*

Relationship between autocatalytic sets and Perron-Frobenius eigenvectors

The ACS is a useful graph-theoretic construct in part because of its connection with the PFE. Let $\mathbf{x}$ be a PFE of a graph. Consider the set of all nodes i for which x_i is non-zero. We will call the subgraph of all these nodes and their mutual links the 'subgraph of the PFE $\mathbf{x}$'. If all the components of the PFE are non-zero then the subgraph of the PFE is the entire graph. For example the subgraph of the PFE $\mathbf{e}_3$ mentioned in Figure 16.1d is is the graph shown in Figure 16.1c. One can show that [16]

Proposition 3
If $\lambda_1(C) > 0$, then the subgraph of any PFE of C is an ACS.

For the PFEs of Figure 16.1d this is immediately verified by inspection. Note that this result relates an algebraic property of a graph, its PFE, to a topological structure, an ACS. Further, this result is not true if we considered irreducible graphs intead of ACSs. E.g., the subgraph of $\mathbf{e}_3$, shown in Figure 16.1c, is not an irreducible graph.

Note also that the converse of the above statement is not true, i.e., there need not exist a PFE for every ACS in a given graph. Thus in Figure 16.1a, nodes 3,4,5,6 and 7 form an ACS but there is no eigenvector with eigenvalue λ_1 for which all these and only these components are non-zero.

Let $\mathbf{x}$ be a PFE of a graph C, and let C' denote the adjacency matrix of the subgraph of $\mathbf{x}$. Let $\lambda_1(C')$ denote the Perron-Frobenius eigenvalue of C'. It is not difficult to see that $\lambda_1(C') = \lambda_1(C)$. Figure 16.3 illustrates this point.

For the graph in Figure 16.3a $\lambda_1 = 1$. Figure 16.3b shows a PFE of the graph and how it satisfies the eigenvalue equations. For this PFE, nodes 1, 5 and 6 have $x_i = 0$. Removing these nodes produces the PFE subgraph shown in Figure 16.3c. Its adjacency matrix, C', is obtained by removing rows 1, 5, 6 and columns 1, 5, 6 from the original matrix. Figure 16.3d illustrates that the vector constructed by removing the zero components of the PFE is an eigenvector of C' with eigenvalue 1. The logic of this example is easily extended to a general proof that $\lambda_1(C') = \lambda_1(C)$.

We can now perform a graph decomposition of C' into irreducible subgraphs as before; since $\lambda_1(C') = \lambda_1(C)$, it follows that C' must contain at least one of the basic subgraphs of C. If C' contains only one of the basic subgraphs of C we will refer to $\mathbf{x}$ as a *simple* PFE, and to C' as a *simple ACS*. The graph in Figure 16.1a has only four simple PFEs which are displayed in Figure 16.1d. All PFEs of C are linear combinations of its simple PFEs.

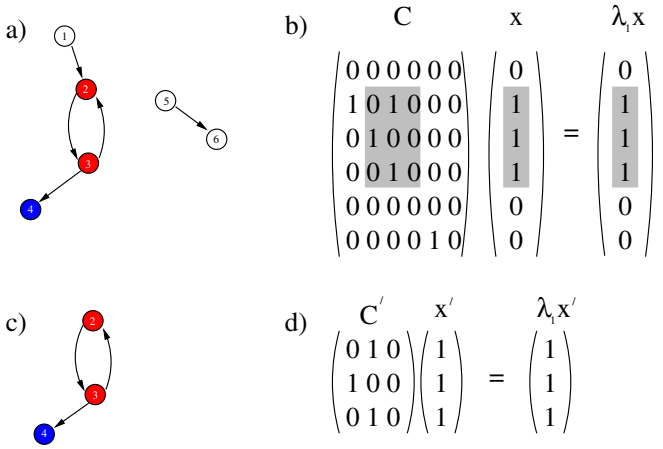


Figure 16.3: Example showing that the λ_1 of a PFE subgraph equals the λ_1 of the whole graph. **a.** A directed graph with 6 nodes. **b.** $\mathbf{x}$ is an eigenvector of its adjacency matrix C with eigenvalue $\lambda_1 = 1$, which is the Perron-Frobenius eigenvalue of the graph. The non zero components of $\mathbf{x}$ and the corresponding rows and columns of C are highlighted. **c.** The subgraph of the PFE $\mathbf{x}$. **d.** The vector $\mathbf{x}$' constructed by removing the zero components of $\mathbf{x}$ is an eigenvector of the adjacency matrix, C', of the PFE subgraph. Its corresponding eigenvalue is unity, which is also the Perron-Frobenius eigenvalue of the PFE subgraph.

Core and periphery of a simple PFE

If C' is the subgraph of a simple PFE, the basic subgraph of C contained in C' will be called the *core* of C' (or equivalently, the 'core of the simple PFE'), and denoted Q'. The set of the remaining nodes and links of C' that are not in its core will together be said to constitute the *periphery* of C'. For example, for the PFE in Figure 16.1c the core is the 2-cycle comprising nodes 14 and 15. Note that the periphery is not a subgraph in the sense we are using the word 'subgraph', since it contains links not just between periphery nodes but also from nodes outside the periphery (like the link from node 15 to 16 in Figure 16.1c).

The core and periphery can be shown to have the following topological property (which justifies the nomenclature):

Proposition 4. *From every node in the core of (the subgraph of) a simple PFE there exists a path leading to every other node of the PFE subgraph. From no periphery node is there any path leading to any core node.*
Thus all periphery nodes are downstream from all core nodes. Starting from the core one can reach the periphery but not vice versa.

It follows from the Perron-Frobenius theorem for irreducible graphs [13] that $\lambda_1(Q')$ will necessarily increase if any link is added to the core. Similarly removing any link will decrease $\lambda_1(Q')$. Thus λ_1 measures the multiplicity of internal pathways in the core. Figure 16.4 illustrates this point.

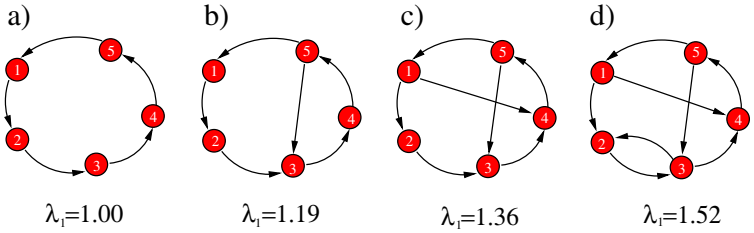


Figure 16.4: λ_1 is a measure of the multiplicity of internal pathways in the core of simple PFE. Four irreducible graphs are shown. An irreducible graph always has a unique PFE that is simple and whose core is the entire graph. The Perron-Frobenius theorem ensures that adding a link to the core of a simple PFE necessarily increases its Perron-Frobenius eigenvalue λ_1. The figure also illustrates the concept of keystone nodes (see section 3).

Core and periphery of a non-simple PFE

Since any PFE of a graph can be written as a linear combination of a set of simple PFEs (this set is unique for any graph), the definitions of core and periphery can be readily extended to any PFE as follows:
The *core of a PFE*, denoted Q, is the union of the cores of those simple PFEs whose linear combination forms the given PFE. The rest of the nodes and links of the PFE subgraph constitute its periphery. It follows from the above discussion that $\lambda_1(Q) = \lambda_1(C)$. When the core is a union of disjoint cycles then $\lambda_1(Q) = 1$, and vice versa.

The structure of PFEs when there is no ACS

The above discussion about the structure of PFEs was for graphs C with $\lambda_1(C) > 0$. If $\lambda_1(C) = 0$, the graph has no ACS. Then the structure of PFEs is as follows: there exists a PFE for every connected component of the graph. Since there are no closed walks in the graph, all walks have finite lengths. Consider the longest paths in a given connected component. Identify the nodes that are the endpoints of these longest paths. The PFE corresponding to the given connected component will have $x_i > 0$ for each of the latter nodes and $x_i = 0$ for all other nodes in the graph. Again a general PFE is a linear combination of all such PFEs, one for each connected component of the graph. In this case since there is no closed path there is no core (or periphery) for any PFE of the graph. The core of all PFEs of such a graph may be defined to be the null set, $Q = \phi$.

16.3 A dynamical system on a fixed graph

In the previous section we have discussed the properties of graphs and their associated adjacency matrices, eigenvalues and eigenvectors. In this section we discuss the dynamical significance of the same constructs. In particular, we present an example of a dynamical system on

a fixed graph described by a set of coupled ordinary differential equations, whose attractors are precisely the PFEs discussed above. This dynamical system arises as an idealization of population dynamics of a set of chemicals.

Consider the simplex of normalized non-negative vectors in s dimensions: $J = \{\mathbf{x} \equiv (x_1, x_2, \ldots, x_s) \in \mathbf{R}^s | 0 \leq x_i \leq 1, \sum_{i=1}^s x_i = 1\}$. For a fixed graph $C = (c_{ij})$ with s nodes, consider the set of coupled ordinary differential equations [20]

$$\dot{x}_i = \sum_{i=1}^{s} c_{ij} x_j - x_i \sum_{j,k=1}^{s} c_{kj} x_j. \tag{16.1}$$

This will be the dynamical system of interest to us in this section.

Note that the dynamics preserves the normalization of $\mathbf{x}$, $\sum_{i=1}^s \dot{x}_i = 0$. For non-negative C it leaves the simplex J invariant. (For negative c_{ij}, additional conditions have to be added (see [21]) but we do not discuss that case here.)

The links of the graph represent the interactions between the variables x_i that live on the nodes. x_i could represent, for example, the relative population of the i^{th} species in a population of s species, or the probability of the i^{th} strategy among a group of s strategies in an evolutionary game, or the market share of the i^{th} company among a set of competing companies, etc. It is useful to see how equation (16.1) arises in a population dynamic context.

Let $i \in \{1, \ldots, s\}$ denote a chemical (or molecular) species in a chemical reactor. Molecules can react with each other in various ways; we focus on only one aspect of their interactions: catalysis. The catalytic interactions can be described by a directed graph with s nodes. The nodes represent the s species and the existence of a link from node j to node i means that species j is a catalyst for the production of species i. In terms of the adjacency matrix, $C = \{c_{ij}\}$ of this graph, c_{ij} is set to unity if j is a catalyst of i and is set to zero otherwise. The operational meaning of catalysis is as follows:

Each species i will have an associated non-negative population y_i in the pond which changes with time. In a certain approximation (discussed below) the population dynamics for a fixed set of chemical species whose interactions are given by C, will be given by

$$\dot{y}_i = \sum_{j=1}^{s} c_{ij} y_j - \phi y_i, \tag{16.2}$$

where $\phi(t)$ is some function of time. To see how such an equation might arise, assume that species j catalyses the ligation of reactants A and B to form the species i, $A + B \xrightarrow{j} i$. Then the rate of growth of the population y_i of species i in a well stirred reactor will be given by $\dot{y}_i = k(1 + \nu y_j) n_A n_B - \phi y_i$, where n_A, n_B are reactant concentrations, k is the rate constant for the spontaneous reaction, ν is the catalytic efficiency, and ϕ represents a common death rate or dilution flux in the reactor. Assuming the catalysed reaction is much faster than the spontaneous reaction, and that the concentrations of the reactants are large and fixed, the rate equation becomes $\dot{y}_i = K y_j - \phi y_i$, where K is a constant. In general since species i can have multiple catalysts, we get $\dot{y}_i = \sum_{j=1}^s K_{ij} y_j - \phi y_i$, with $K_{ij} \sim c_{ij}$. We make the further idealization $K_{ij} = c_{ij}$ giving equation (16.2).

The relative population of species i is by definition $x_i \equiv y_i / \sum_{j=1}^s y_j$. Therefore $\mathbf{x} \equiv (x_1, \ldots, x_s) \in J$, since $0 \leq x_i \leq 1, \sum_{i=1}^s x_i = 1$. Taking the time derivative of x_i and

using (16.2) it is easy to see that $\dot{x}_i$ is given by (16.1). Note that the ϕ term, present in (16.2), cancels out and is absent in (16.1).

We remark that the quasispecies equation [17] has the same form as equation (16.2), albeit with a different interpretation and a special structure of the C matrix that arises from that interpretation.

16.3.1 Attractors of equation (16.1)

The rest of this section consists of examples and arguments to justify the
Proposition 5. *For any graph C,*
(i) *Every eigenvector of C that belongs to J is a fixed point of (16.1), and vice versa.*
(ii) *Starting from any initial condition in the simplex J, the trajectory converges to some fixed point (generically denoted $\mathbf{X}$) in J.*
(iii) *For generic initial conditions in J, $\mathbf{X}$ is a Perron-Frobenius eigenvector (PFE) of C.* (For special initial conditions, forming a space of measure zero in J, $\mathbf{X}$ could be some other eigenvector of C. Henceforth we ignore such special initial conditions.)
(iv) *If C has a unique PFE, $\mathbf{X}$ is the unique stable attractor of (16.1).*
(v) *If C has more than one linearly independent PFE, then $\mathbf{X}$ can depend upon the initial conditions. The set of allowed $\mathbf{X}$ is a convex linear combination of a subset of the PFEs.* The interior of this convex set in J may then be said to be the 'attractor' of (16.1), in the sense that for generic initial conditions all trajectories converge to a point in this set.
(vi) *For every $\mathbf{X}$ belonging to the attractor set, the set of nodes i for which $X_i > 0$ is the same and is uniquely determined by C. The subgraph formed by this set of nodes will be called the 'subgraph of the attractor' of (16.1) for the graph C.* Physically, this set consists of nodes that always end up with a nonzero relative population when the dynamics (16.1) is allowed to run its course, starting from generic initial conditions.
(vii) *If $\lambda_1(C) > 0$, the subgraph of the attractor of (16.1) is an ACS.* This ACS will be called the *dominant ACS* of the graph. The dominant ACS is independent of (generic) initial conditions and depends only on C.

For example for the graph of Figure 16.1a, $\mathbf{X}$ is a convex linear combination of $\mathbf{e}_2$ and $\mathbf{e}_3$, $\mathbf{X} = a\mathbf{e}_2 + (1-a)\mathbf{e}_3$, with $0 \le a \le 1$. a depends upon initial conditions; generically $0 < a < 1$. The subgraph of the attractor contains eight nodes, 6,7,14-19. Starting with generic initial conditions where all the x_i are nonzero, the trajectory will converge to a point $\mathbf{X}$ where these eight nodes have nonzero X_i and each of the other twelve nodes have $X_i = 0$. The eight populated nodes form an ACS, the dominant ACS of the graph.

To see (i), let $\mathbf{x}^\lambda \in J$ be an eigenvector of C, $\sum_j c_{ij}x_j = \lambda x_i$. Substituting this on the r.h.s. of (16.1), one gets zero. Conversely, if the r.h.s. of (16.1) is zero, one finds $\mathbf{x} = \mathbf{x}^\lambda$, with $\lambda = \sum_{k,j} c_{kj}x_j$.

To motivate (ii) and (iii) it is most convenient to consider the underlying dynamics (16.2) from which (16.1) is derived: Since (16.1) is independent of ϕ, we can set $\phi = 0$ in (16.2) without any loss of generality. With $\phi = 0$ the general solution of (16.2), which is a linear system, can be schematically written as:

$$\mathbf{y}(t) = e^{Ct}\mathbf{y}(0),$$

where $\mathbf{y}(0)$ and $\mathbf{y}(t)$ are viewed as column vectors. Suppose $\mathbf{y}(0)$ is a right eigenvector of C

with eigenvalue λ, denoted $\mathbf{y}^\lambda$. Then

$$\mathbf{y}(t) = e^{\lambda t}\mathbf{y}^\lambda.$$

Since this time dependence is merely a rescaling of the eigenvector, this is an alternative way of seeing that $\mathbf{x}^\lambda = \mathbf{y}^\lambda / \sum_{j=1}^s y_j^\lambda$ is a fixed point of (16.1). If the eigenvectors of C form a basis in R^s, $\mathbf{y}(0)$ is a linear combination: $\mathbf{y}(0) = \sum_\lambda a_\lambda \mathbf{y}^\lambda$. In that case, for large t it is clear that the term with the largest value of λ will win out, hence

$$\mathbf{y}(t) \overset{t\to\infty}{\sim} e^{\lambda_1 t}\mathbf{y}^{\lambda_1}$$

where λ_1 is the eigenvalue of C with the largest real part (which we know is the same as its Perron-Frobenius eigenvalue) and $\mathbf{y}^{\lambda_1}$ an associated eigenvector. Therefore, for generic initial conditions the trajectory of (16.1) will converge to $\mathbf{X} = \mathbf{x}^{\lambda_1}$, a PFE of C. If the eigenvectors of C do not form a basis in R^s, the above result is still true (as we will see in examples).

Note that λ_1 can be interpreted as the 'population growth rate' at large t, since $\dot{\mathbf{y}}(t) \overset{t\to\infty}{\sim} \lambda_1\mathbf{y}$. In the previous section we had mentioned that λ_1 measures a topological property of the graph, namely, the multiplicity of internal pathways in the core of the graph. Thus in the present model, λ_1 has both a topological and dynamical significance, which relates two distinct properties of the system, one structural (multiplicity of pathways in the core of the graph), and the other dynamical (population growth rate). The higher the multiplicity of pathways in the core, the greater is the population growth rate of the dominant ACS.

Part (iv) follows from the above. We will give examples as illustrations of (v) and (vi). Further, from Proposition 3, previous section, we know that the subgraph of a PFE has to be an ACS, whenever $\lambda_1 > 0$. That explains (vii). It is instructive to consider examples of graphs and see how the trajectory converges to a PFE.

Example 1. A simple chain, Figure 16.5a:
The adjacency matrix of this graph has all eigenvalues (including λ_1) zero. There is only one (normalized) eigenvector corresponding to this eigenvalue, namely $\mathbf{e} = (0, 0, 1)$ and this is the unique PFE of the graph. (This is an example where the eigenvectors of C do not form a basis in R^s.) Since node 1 has no catalyst, its rate equation is (henceforth taking $\phi = 0$) $\dot{y_1} = 0$. Therefore $y_1(t) = y_1(0)$, a constant. The rate equation for node 2 is $\dot{y_2} = y_1 = y_1(0)$. Thus $y_2(t) = y_2(0) + y_1(0)t$. Similarly $\dot{y_3} = y_2$ implies that $y_3(t) = (1/2)y_1(0)t^2 + y_2(0)t + y_3(0)$. At large t, $y_1 = \text{constant}$, $y_2 \sim t$, $y_3 \sim t^2$; hence y_3 dominates. Therefore, $X_i = \lim_{t\to\infty} x_i(t)$ is given by $X_1 = 0, X_2 = 0, X_3 = 1$. Thus we find that $\mathbf{X}$ equals the unique PFE $\mathbf{e}$, independent of initial conditions.

Example 2. A 1-cycle, Figure 16.5b:
This graph has two eigenvalues, $\lambda_1 = 1$, $\lambda_2 = 0$. The unique PFE is $\mathbf{e} = (1, 0)$. The rate equations are $\dot{y_1} = y_1, \dot{y_2} = 0$, with the solutions $y_1(t) = y_1(0)e^t, y_2(t) = y_2(0)$. At large t node 1 dominates, hence $\mathbf{X} = (1, 0) = \mathbf{e}$. The exponentially growing population of 1 is a consequence of the fact that 1 is a self-replicator, as embodied in the equation $\dot{y_1} = y_1$.

Example 3. A 2-cycle, Figure 16.5c:
The corresponding adjacency matrix has eigenvalues $\lambda_1 = 1, \lambda_2 = -1$. The unique normal-

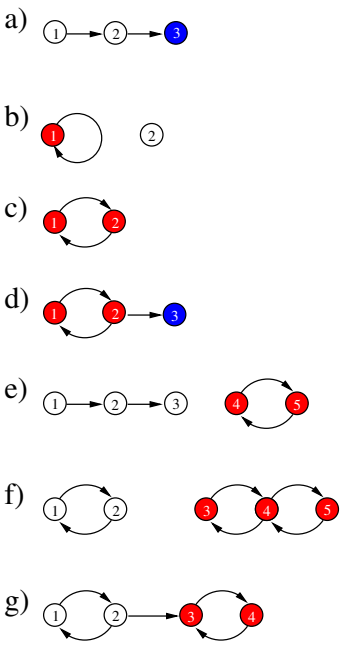


Figure 16.5: Examples of graphs with a unique PFE. The subgraph of the PFE coincides with the nodes that are populated in the attractor.

ized PFE is $\mathbf{e} = (1/2, 1/2)$. The population dynamics equations are $\dot{y_1} = y_2, \dot{y_2} = y_1$. The general solution to these is (note $\ddot{y_1} = y_1$)

$$y_1(t) = Ae^t + Be^{-t}, \qquad y_2(t) = Ae^t - Be^{-t}.$$

Therefore at large t, $y_1 \to Ae^t, y_2 \to Ae^t$, hence $\mathbf{X} = (1,1)/2 = \mathbf{e}$. Neither 1 nor 2 is individually a self-replicating species, but collectively they function as a self-replicating entity. This is true of all ACSs.

Example 4. A 2-cycle with a periphery, Figure 16.5d:
This graph has $\lambda_1 = 1$ and a unique normalized PFE $\mathbf{e} = (1,1,1)/3$. The population equations for y_1 and y_2 and consequently their general solutions are the same as Example 3, but now in addition $\dot{y_3} = y_2$, yielding $y_3(t) = Ae^t + Be^{-t} + \text{constant}$. Again for large t, y_1, y_2, y_3 grow as $\sim Ae^t$, hence $\mathbf{X} = (1,1,1)/3 = \mathbf{e}$. The dominant ACS includes all the three nodes.

This example shows how a parasitic periphery (which does not feed back into the core) is supported by an autocatalytic core. This is also an example of the following general result: when a subgraph C', with largest eigenvalue λ_1', is *downstream* from another subgraph C'' with largest eigenvalue $\lambda_1'' > \lambda_1'$, then the population of the former also increases at the rate λ_1''. Therefore if C'' is populated in the attractor, so is C'. In this example C' is the single node 3 with $\lambda_1' = 0$ and C'' is the 2-cycle of nodes 1 and 2 with $\lambda_1'' = 1$.

Example 5. A 2-cycle and a chain, Figure 16.5e:
The graph in Figure 16.5e combines the graphs of Figures 16.5a and c. Following the analysis of those two examples it is evident that for large t, $y_1 \sim t^0, y_2 \sim t^1, y_3 \sim t^2, y_4 \sim e^t, y_5 \sim e^t$. Because the populations of the 2-cycle are growing exponentially they will eventually completely overshadow the populations of the chain which are growing only as powers of t. Therefore the attractor will be $\mathbf{X} = (0, 0, 0, 1, 1)/2$ which, it can be checked, is a PFE of the graph (it is an eigenvector with eigenvalue 1).

In general when a graph consists of one or more ACSs and other nodes that are not part of any ACS, the populations of the ACS nodes grow exponentially while the populations of the latter nodes grow at best as powers of t. Hence ACSs always outperform non-ACS structures in the population dynamics (see also Example 2). This is a consequence of the infinite walks provided by the positive feedback inherent in the ACS structure, while non-ACS structures have no feedbacks and only finite walks.

Example 6. A 2-cycle and another irreducible graph disconnected from it, Figure 16.5f:
One can ask, when there is more than one ACS in the graph, which is the dominant ACS? Figure 16.5f shows a graph containing two ACSs. The 2-cycle subgraph has a Perron-Frobenius eigenvalue 1, while the other irreducible subgraph has a Perron-Frobenius eigenvalue $\sqrt{2}$. The unique PFE of the entire graph is $\mathbf{e} = (0, 0, 1, \sqrt{2}, 1)/(2 + \sqrt{2})$ with eigenvalue $\sqrt{2}$. The population dynamics equations are $\dot{y_1} = y_2, \dot{y_2} = y_1, \dot{y_3} = y_4, \dot{y_4} = y_3 + y_5, \dot{y_5} = y_4$. The first two equations are completely decoupled from the last three and the solutions for y_1 and y_2 are the same as for Example 3. For the other irreducible graph the solution is (since $\ddot{y_4} = \dot{y_3} + \dot{y_5} = 2y_4$)

$$y_4(t) = Ae^{\sqrt{2}t} + Be^{-\sqrt{2}t}, \qquad y_3(t) = \frac{1}{\sqrt{2}}(Ae^{\sqrt{2}t} + Be^{-\sqrt{2}t}) + C,$$

$$y_5(t) = \frac{1}{\sqrt{2}}(Ae^{\sqrt{2}t} + Be^{-\sqrt{2}t}) - C.$$

Thus, the populations of nodes 3,4 and 5 also grow exponentially but at a faster rate, reflecting the higher Perron-Frobenius eigenvalue of the subgraph comprising those nodes. Therefore this structure eventually overshadows the 2-cycle, and the attractor is $\mathbf{X} = \mathbf{e}$. The dominant ACS in this case is the irreducible subgraph formed by nodes 3,4 and 5.

More generally, when a graph consists of several disconnected ACSs with different individual λ_1, only the ACSs whose λ_1 is the largest (and equal to $\lambda_1(C)$) end up with non-zero relative populations in the attractor.

Example 7. A 2-cycle downstream from another 2-cycle, Figure 16.5g:
What happens when the graph contains two ACSs whose individual λ_1 equals $\lambda_1(C)$, and one of those ACSs is downstream of another? In Figure 16.5g nodes 3 and 4 form a 2-cycle which is downstream from another 2-cycle comprising nodes 1 and 2. The unique PFE of this graph, with $\lambda_1 = 1$, is $\mathbf{e} = (0, 0, 1, 1)/2$. The population dynamics equations are $\dot{y_1} = y_2, \dot{y_2} = y_1, \dot{y_3} = y_4 + y_2, \dot{y_4} = y_3$. Their general solution is:

$$y_1(t) = Ae^t + Be^{-t}, \qquad y_2(t) = Ae^t - Be^{-t},$$

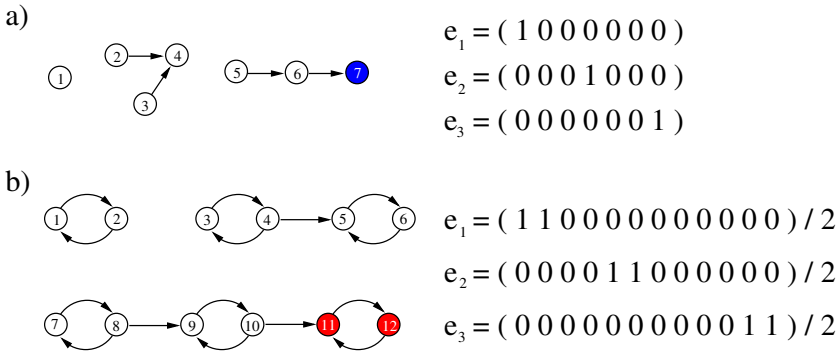


Figure 16.6: Examples of graphs with multiple PFEs. (a) $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ are all eigenvectors with eigenvalue $\lambda_1 = 0$. Only $\mathbf{e}_3$ is the attractor. Thus for generic initial conditions, only node 7, which sits at the end point of the longest chain of nodes is populated in the attractor. (b) $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ are all eigenvectors with eigenvalue $\lambda_1 = 1$, but only $\mathbf{e}_3$ is the attractor. Only the 2-cycle of nodes 11 and 12, which sits at the end of the longest chain of cycles, is populated in the attractor.

$$y_3(t) = \frac{t}{2}(Ae^t - Be^{-t}) + Ce^t + De^{-t},$$

$$y_4(t) = \frac{t}{2}(Ae^t + Be^{-t}) + (C - \frac{A}{2})e^t + (\frac{B}{2} - D)e^{-t}.$$

It is clear that for large t, $y_1 \sim e^t, y_2 \sim e^t, y_3 \sim te^t, y_4 \sim te^t$. While all four grow exponentially with the same rate λ_1, as $t \to \infty$ y_3 and y_4 will overshadow y_1 and y_2. The attractor will be therefore be $\mathbf{X} = (0, 0, 1, 1)/2 = \mathbf{e}$. Here the dominant ACS is the 2-cycle of nodes 3 and 4. This result generalizes to other kinds of ACSs: if one irreducible subgraph is downstream of another with the same Perron-Frobenius eigenvalue, the latter will have zero relative population in the attractor.

The above examples displayed graphs with a unique PFE, and illustrated Proposition 5 (iv). The stability of the global attractor follows from the fact that the constants A, B, C, D, etc., in the above examples, which can be traded for the initial conditions of the populations, appear nowhere in the attractor configuration $\mathbf{X}$. Now we consider examples where the PFE is not unique.

Example 8. Graph with $\lambda_1 = 0$ and three disconnected components, Figure 16.6a:
As mentioned in section 2 this graph has three independent PFEs, displayed in Figure 16.6a. The attractor is $\mathbf{X} = \mathbf{e}_3$. This is an immediate generalization of Example 1 above. Using the same argument as for Example 1, we can see that $y_i \sim t^k$ if the longest path ending at node i is of length k. Therefore the attractor will have nonzero components only for nodes at the ends of the longest paths. Thus the populations of nodes 1,2,3 and 5 are constant, those of 4 and 6 increase $\sim t$ for large t, and of 7 as $\sim t^2$, explaining the result.

Example 9. Several connected components containing 2-cycles, Figure 16.6b:
Here again there are three PFEs, one for each connected component. The population of nodes in 2-cycles which are not downstream of other 2-cycles (nodes 1,2,3,4,7 and 8) will grow as e^t. As in Example 7, Figure 16.5g, the nodes of 2-cycles which are downstream of one 2-cycle (nodes 5,6,9 and 10) will grow as te^t. It can be verified that the populations of nodes in 2-cycles downstream from two other 2-cycles (nodes 11 and 12) will grow as t^2e^t. The pattern is clear: in the attractor only the 2-cycles at the ends of the longest chains of 2-cycles will have non-zero relative populations, explaining the result.

Example 10. Figure 16.1a:
From previous examples it is evident how the populations will change with time for Figure 16.1a. Here we list the result:

$$y_8 \sim t^0, \qquad y_9 \sim t^1, \qquad y_{10} \sim t^2,$$

$$y_1, y_2, y_3, y_4, y_5, y_{11}, y_{12}, y_{13}, y_{20} \sim e^t,$$

$$y_6, y_7, y_{14}, y_{15}, y_{16}, y_{17}, y_{18}, y_{19} \sim te^t.$$

Thus, starting from a generic initial population, only the eight nodes, 6,7,14-19, will be populated in the attractor. This explains the comments just after the statment of Proposition 5.

Note the structure of the dominant ACS in the above examples when $\lambda_1 > 0$. If there is a unique PFE in the graph, the dominant ACS is the subgraph of the PFE. If there are several PFEs only a subset of those may be counted as illustrated in Examples 9 and 10, Figures 16.6b and 16.1a, respectively. A general construction of the dominant ACS for an arbitrary graph will be described elsewhere.

How long does it take to reach the attractor?

The timescale over which the system reaches its attractor depends on the structure of the graph C. For instance in Example 2, the attractor is approached as the population of node 1, y_1, overwhelms the population y_2. Since y_1 grows exponentially as e^t, the attractor is reached on a timescale $\lambda_1^{-1} = 1$. In contrast, in Example 1, the attractor is approached as y_3 overwhelms y_1 and y_2. Because in this case all the populations are growing as powers of t, the timescale for reaching the attractor is infinite. In general, this timescale depends on the difference in growth rate between the fastest growing population and the next fastest growing population.

For graphs which have no basic subgraphs, i.e., graphs with $\lambda_1 = 0$ like those in Example 1 and 8, all populations grow as powers of t, hence the timescale for reaching the attractor is infinite.

For graphs which have one or more basic subgraphs (i.e., $\lambda_1 \geq 1$) but all the basic subgraphs are in different connected components, such as Examples 2-6, the timescale for reaching the attractor is given by $(\lambda_1 - \mathrm{Re}\lambda_2)^{-1}$, where λ_2 is the eigenvalue of C with the next largest real part, compared to λ_1.

For graphs having one or more basic subgraphs with at least one basic subgraph downstream from another basic subgraph, the ratio of the fastest growing population to the next fastest growing will always be a power of t (as in Examples 7, 9 and 10) therefore the timescale for reaching the attractor is again infinite.

Core and periphery of a graph

Since the dominant ACS is given by a PFE, we will define the core of the dominant ACS to be the core of the corresponding PFE. If the PFE is simple, the core of the dominant ACS consists of just one basic subgraph. If the PFE is non-simple the core of the dominant ACS will be a union of some basic subgraphs. Further, the dominant ACS is uniquely determined by the graph. This motivates the definition of the core and periphery of a graph: The *core* of a graph C, denoted $Q(C)$, is the core of the dominant ACS of C. The *periphery* of C is the periphery of the dominant ACS of C. This definition applies when $\lambda_1(C) > 0$. When $\lambda_1(C) = 0$, the graph has no ACS and by definition $Q(C) = \phi$. In all cases $\lambda_1(Q(C)) = \lambda_1(C)$. For all the graphs depicted in this paper, except the one in Figure 16.1e, the red nodes constitute the core of the graph, the blue nodes its periphery, and the white nodes are neither core nor periphery – they are nodes that are not in any of the PFE subgraphs. [1]

Core overlap of two graphs

Given any two graphs C and C' whose nodes are labeled, the *core overlap* between them, denoted $Ov(C, C')$, is the number of common links in the cores of C and C', i.e., the number of ordered pairs (j, i) for which Q_{ij} and Q'_{ij} are both non-zero [22]. If either of C or C' does not have a core, $Ov(C, C')$ is identically zero.

Keystone nodes

In ecology certain species are referred to as keystone species – those whose extinction or removal would seriously disturb the balance of the ecosystem [24–27]. One might similarly ask for the notion of a keystone node in a directed graph that captures some important organizational role played by a node. Consider the impact of the hypothetical removal of any node i from a graph C. One can, for example, ask for the core of the graph $C - i$ that would result if node i (along with all its links) were removed from C. We will refer to a node i as a *keystone node* if C has a non-vanishing core and $Ov(C, C - i) = 0$ [23]. Thus a keystone node is one whose removal modifies the organizational structure of the graph (as represented by its core) drastically. In each of Figures 16.4a-d, for example, the core is the entire graph. In Figure 16.4a, all the nodes are keystone, since the removal of any one of them would leave the graph without an ACS (and hence without a core). In general when the core of a graph is a single n-cycle, for any n, all the core nodes are keystone. In Figure 16.4b, nodes 3, 4 and 5 are keystone but the other nodes are not, and in Figure 16.4c only nodes 4 and 5 are keystone. In Figure 16.4d, there are no keystone nodes. These examples show that the more internal pathways a core has (generally, this implies a higher value of λ_1), the less likely it is to have keystone species, and hence the more robust its structure is to removal of nodes.

Figure 16.7 illustrates another type of graph structure which has a keystone node. The graph in Figure 16.7a consists of a 2-cycle (nodes 4 and 5) downstream from an irreducible subgraph consisting of nodes 1,2 and 3. The core of this graph is the latter irreducible subgraph. Figure 16.7b shows the graph that results if node 3 is removed with all its links. This

[1] The definition of the core of a graph given in refs. [22, 23] is a special case of this definition, holding only for graphs where each connected component of the dominant ACS has no more than one basic subgraph.

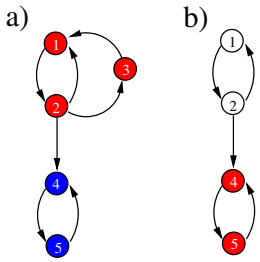


Figure 16.7: Example illustrating the notion of keystone species and the phenomenon of a core-shift. Node number 3 is keystone node of the graph in part **a** because its removal produces the graph in **b** which has a zero core overlap with the graph in **a**. The core nodes of both graph are coloured red. An event in which the core before the event and after the event have zero overlap is called a 'core-shift'.

consists of one 2-cycle downstream from another. Though both 2-cycles are basic subgraphs of the graph, as discussed in Example 7, Figure 16.5g, this graph has a unique (upto constant multiples) PFE, whose subgraph consists of the downstream cycle (nodes 4 and 5) only. Thus the 2-cycle 4-5 is the core of the graph in Figure 16.7b. Clearly $Ov(C, C-3) = 0$ therefore node 3 in Figure 16.7a is a keystone node.

We remark that the above purely graph theoretic definition of a keystone node turns out to be useful in the dynamical system discussed in this and the following sections. For other dynamical systems, other definitions of keystone might be more useful.

16.4 Graph dynamics

So far we have discussed the algebraic properties of a fixed graph, and the attractors of a particular dynamical system on arbitrary, but fixed graphs. However one of the most interesting properties of complex systems is that the graph of interactions among their components evolves with time, resulting in many interesting adaptive phenomena. We now turn to such an example, where the graph itself is a dynamical variable, and display how phenomena such as self-organization, catastrophes, innovation, etc, can arise. We shall see that the above discussion of (static) graph theory will be crucial in understanding these phenomena.

We consider a process which alters a graph in discrete steps. The series of graphs produced by such a process can be denoted $C_n, n = 1, 2, \ldots$. A graph update event will be one step of the process, taking a graph from C_{n-1} to C_n. In fact the process we consider is a specific example of a Markov process on the space of graphs. At time $n-1$, the graph C_{n-1} determines the transition probability to all other graphs. The stochastic process picks the new graph C_n using this probability distribution and the trajectory moves forward in graph space. In the example we consider, the transition probability is not specified explicitly. It arises implicitly as a consequence of the dynamics (16.1) that takes place on a fast time scale for the fixed graph C_{n-1}.

The graph dynamics is implemented as follows [20]:
Initially the graph is random: for every ordered pair (i, j) with $i \neq j$, c_{ij} is independently chosen to be unity with a probability p and zero with a probability $1 - p$. c_{ii} is set to zero for all i. Each x_i is chosen randomly in $[0, 1]$ and all x_i are rescaled so that $\sum_{i=1}^{s} x_i = 1$.

Step 1. With C fixed, $\mathbf{x}$ is evolved according to (16.1) until it converges to a fixed point, denoted $\mathbf{X}$. The set $\mathcal{L}$ of nodes with the least X_i is determined, i.e, $\mathcal{L} = \{i \in S | X_i = \min_{j \in S} X_j\}$.

Step 2. A node, say node k, is picked randomly from $\mathcal{L}$ and is removed from the graph along with all its links.

Step 3. A new node (also denoted k) is added to the graph. Links to and from k to other nodes are assigned randomly according to the same rule, i.e, for every $i \neq k$ c_{ik} and c_{ki} are independently reassigned to unity with probability p and zero with probability $1 - p$, irrespective of their earlier values, and c_{kk} is set to zero. All other matrix elements of C remain unchanged. x_k is set to a small constant x_0, all other x_i are perturbed by a small amount from their existing value X_i, and all x_i are rescaled so that $\sum_{i=1}^{s} x_i = 1$.

This process, from step 1 onwards, is iterated many times.

Notice that the population dynamics and the graph dynamics are coupled: the evolution of the x_i depends on the graph C in step 1, and the evolution of C in turn depends on the x_i through the choice of which node to remove in step 2. There are two timescales in the dynamics, a short timescale over which the graph is fixed while the x_i evolve, and a longer timescale over which the graph is changed.

This dynamics is motivated by the origin of life problem, in particular the puzzle of how a complex chemical organization might have emerged from an initial 'random soup' of chemicals, as discussed in section 1. Let us consider a pond on the prebiotic earth containing s molecular species which interact catalytically as discussed in the previous section, and let us allow the chemical organization to evolve with time due to various natural process which remove species from the pond and bring new species into the pond. Thus over short timescales we let the populations of the species evolve according to (16.1). Over longer timescales we imagine the prebiotic pond to be subject to periodic perturbations from storms, tides or floods. These perturbations remove existing species from the pond and introduce new species into it. The species most likely to be completely removed from the pond are those that have the least number of molecules. The new species could have entirely different catalytic properties from those removed or those existing in the pond. The above rules make the idealization that the perturbation eliminates exactly one existing species (that has the least relative population) and brings in one new species. The behaviour of the system does not depend crucially on this assumption [23].

While in previous sections we have considered graphs with 1-cycles, the requirement $c_{ii} = 0$ in the present section forbids 1-cycles in the graph. The motivation is the following: 1-cycles represent self-replicating species (see previous section, Example 2). Such species, e.g., RNA molecules, are difficult to produce and maintain in a prebiotic scenario and it is generally believed that it requires a complex self supporting molecular organization to be in place *before*

an RNA world, for example, can take off [28, 29]. Thus, we wish to address the question: can we get complex molecular organizations without putting in self-replicating species by hand in the model? As we shall see below, this does indeed happen, since even though self-replicating individual species are disallowed, collectively self-replicating autocatalytic sets can still arise by chance on a certain time scale, and when they do, they trigger a wave of self-organization in the system.

The rules for changing the graph implement *selection* and *novelty*, two important features of natural evolution. Selection is implemented by removing the species which is 'performing the worst', with 'performance' in this case being equated to a species' relative population (step 2). Adding a new species introduces novelty into the system. Note that although the actual connections of a new node with other nodes are created randomly, the new node has the same average connectivity as the initial set of nodes. Thus the new species is not biased in any way towards increasing the complexity of the chemical organization. Step 2 and step 3 represent the interaction of the system with the external environment. The third feature of the model is dynamics of the system that depends upon the interaction among its components (step 1). The phenomena to be described in the following sections are all consequences of the interplay between these three elements – selection, novelty and an internal dynamics.

16.5 Self Organization

We now discuss the results of graph evolution. Figure 16.8 shows the total number of links in the graph versus time (n, the number of graph updates). Three runs of the model described in the previous section, each with $s = 100$ and different values of p are exhibited. Also exhibited is a run where there was *no selection* (in which step 2 is modified: instead of picking one of the nodes of $\mathcal{L}$, any one of the s nodes is picked randomly and removed from the graph along with all its links. The rest of the procedure remains the same). Figure 16.9 shows the time evolution of two more quantities for the same three runs with selection displayed in Figure 16.8.

The quantities plotted are the number of nodes with $X_i > 0$, s_1, and the Perron-Frobenius eigenvalue of the graph, λ_1. The values of the parameters p and s for the displayed runs were chosen to lie in the regime $ps < 1$. Much of the analytical work described below, such as estimation of various timescales, assumes that $ps \ll 1$. Figure 16.10 shows snapshots of the graph at various times in the run shown in Figure 16.9b, which has $p = 0.0025$. It is clear that without selection each graph update replaces a randomly chosen node with another which has on average the same connectivity. Therefore the graph remains random like the starting graph and the number of links fluctuates about its random graph value $\approx ps^2$. As soon as selection is turned on the behaviour becomes more interesting. Three regimes can be observed. First, the 'random phase' where the number of links fluctuates around ps^2 and s_1 is small. Second, the 'growth phase' where l and s_1 show a clear rising tendency. Finally, the 'organized phase' where l again hovers (with large fluctuations) about a value much higher than the initial random graph value, and s_1 fluctuates (again with large fluctuations) about its maximum value s. The time spent in each phase clearly depends on p, and we find it also depends on s. This behaviour can be understood by taking a look at the structure of the graph in each of these phases, especially the ACS structure, and using the results of sections 2 and 3.

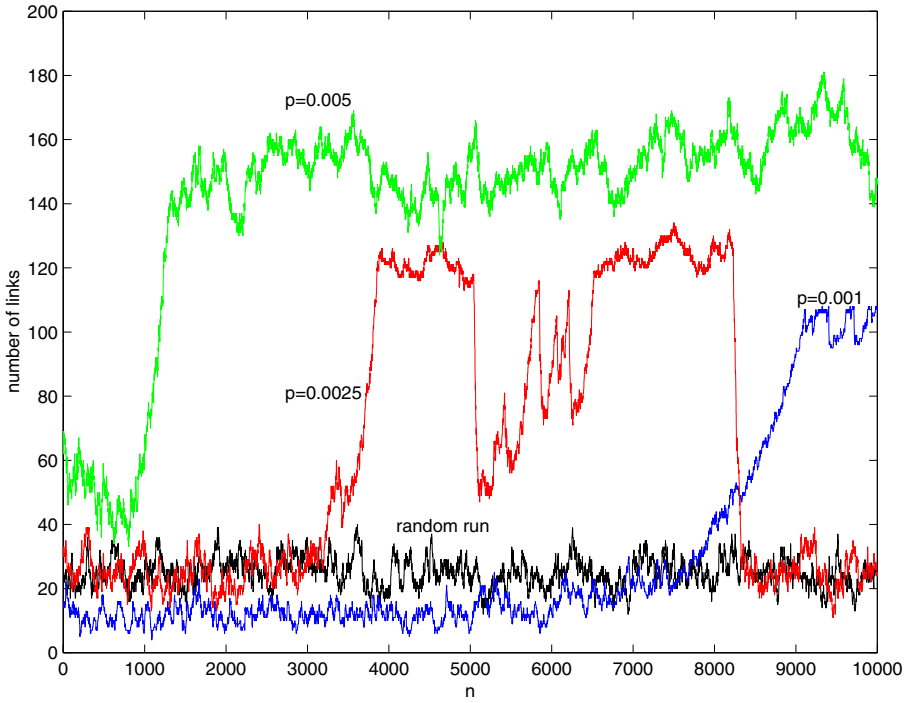


Figure 16.8: The number of links versus time (n) for various runs. Each run had $s = 100$. The black curve is a run with selection turned off; a random node is picked for removal at each graph update. The other curves show runs with selection turned on and with different p values: Blue $p = 0.001$, Red $p = 0.0025$, Green $p = 0.005$.

16.5.1 The random phase

Initially, the random graph contains no cycles, and hence no ACSs, and its Perron-Frobenius eigenvalue is $\lambda_1 = 0$. We have seen in section 3 that for such a graph the attractor will have nonzero components for all nodes which are at the ends of the longest paths of nodes, and zero for every other node. (In Figure 16.10a, there are two paths of length 4, which are the longest paths in the graph. Both end at node 13, which is therefore the only populated node in the attractor for this graph.) These nodes, then, are the only nodes protected from elimination during the graph update. However, these nodes have high relative populations *because they are supported by other nodes*, while the latter (supporting) nodes do not have high relative populations. Inevitably within a few graph updates a supporting node will be removed from the graph. When that happens a node which presently has nonzero X_i will no longer be at the end of the longest path and hence will get $X_i = 0$. For example node 34, which belongs to $\mathcal{L}$, is expected to be picked for replacement within $\approx O(s)$ graph update time steps. In fact it is replaced in the 8th time step. After that node 13 becomes a singleton and joins the set $\mathcal{L}$. Thus no structure is stable when there is no ACS. Eventually, all nodes are removed and replaced, and the graph remains random.

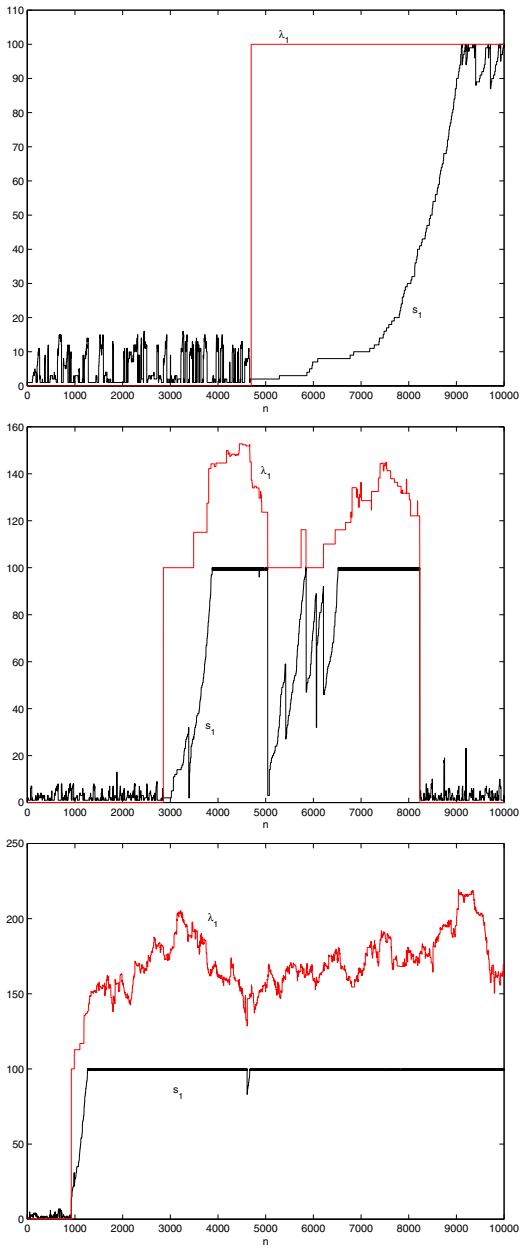


Figure 16.9: Number of populated nodes, s_1, (black curve) and the Perron-Frobenius eigenvalue of the graph, λ_1, (red curve) versus time, n, for the same three runs shown in Figure 16.8. Each run has $s = 100$ and $p = 0.001, 0.0025$ and 0.05 respectively. The λ_1 values shown are 100 times the actual value.

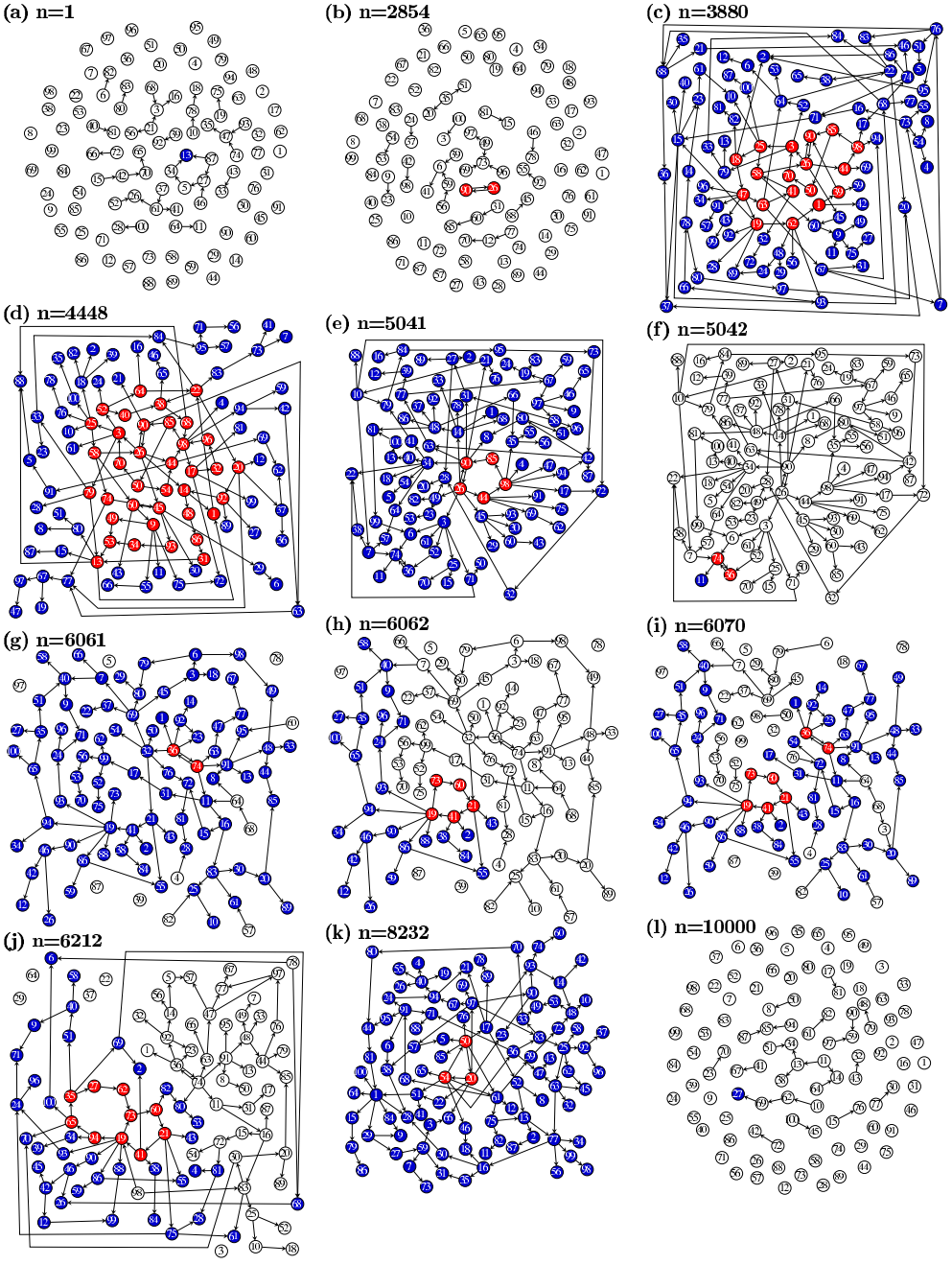


Figure 16.10: Snapshots of the graph at various times for the run shown in Figure 16.9b with $s = 100$ and $p = 0.0025$. See text for a description of the major events. In all graphs, white nodes are those with $X_i = 0$. All coloured nodes have $X_i > 0$. In graphs which have an ACS, the red nodes are core nodes and the blue nodes are periphery nodes.

Note that the inital random graph is likely to contain no cycles when p is small ($ps \ll 1$). If larger values of p are chosen, it becomes more likely that the initial graph will contain a cycle. If it does, there is no random phase; the system is then in the growth phase, discussed below, right from the initial time step.

16.5.2 The growth phase

At some graph update an ACS is formed by pure chance. The probability of this happening can be closely approximated by the probability of a 2-cycle (the simplest ACS) forming by chance, which is $p^2 s$ (= the probability that in the row and column corresponding to the replaced node in C, any matrix element and its transpose both turn out to be unity). Thus the average time of appearance of an ACS is $1/p^2 s$. In the run whose snapshots are displayed in Figure 16.10, a 2-cycle between nodes 26 and 90 formed at $n = 2854$. This is a graph which consists of a 2-cycle and several other chains and trees. For such a graph we have shown in Example 3 in section 3 that the attractor has non-zero X_i for nodes 26 and 90 and zero for all other nodes. The dominant ACS consists of nodes 26 and 90. Therefore these nodes cannot be picked for removal at the graph update and hence a graph update cannot destroy the links that make the dominant ACS. *The autocatalytic property is guaranteed to be preserved until the dominant ACS spans the whole graph.*

When a new node is added to the graph at a graph update, one of three things will happen:

1. The new node will not have any links from the dominant ACS and will not form a new ACS. In this case the dominant ACS will remain unchanged, the new node will have zero relative population and will be part of the least fit set. For small p this is the most likely possibility.

2. The new node gets an incoming link from the dominant ACS and hence becomes a part of it. In this case the dominant ACS grows to include the new node. For small p, this is less likely than the first possibility, but such events do happen and in fact are the ones responsible for the growth of complexity and structure in the graph.

3. The new node forms another ACS. This new ACS competes with the existing dominant ACS. Whether it now becomes dominant, overshadowing the previous dominant ACS or it gets overshadowed, or both ACSs coexist depends on the Perron Frobenius eigenvalues of their respective subgraphs and whether (and which) ACS is downstream of the other. It can be shown that this is a rare event compared with possibilities 1 and 2.

Typically the dominant ACS keeps growing by accreting new nodes, usually one at a time, until the entire graph is an ACS. At this point the growth phase stops and the organized phase begins. As a consequence it follows that λ_1 *is a nondecreasing function of* n *as long as* $s_1 < s$ [16].

Time scale for growth of the dominant ACS.

If we assume that possibility 3 above is rare enough to neglect, and that the dominant ACS grows by adding a single node at a time, we can estimate the time required for it to span the entire graph. Let the dominant ACS consist of $s_1(n)$ nodes at time n. The probability that the new node gets an incoming link from the dominant ACS and hence joins it is ps_1. Thus in Δn graph updates, the dominant ACS will grow, on average, by $\Delta s_1 = ps_1 \Delta n$ nodes. Therefore

$s_1(n) = s_1(n_a)exp((n - n_a)/\tau_g)$, where $\tau_g = 1/p$, n_a is the time of arrival of the first ACS and $s_1(n_a)$ is the size of the first ACS (=2 for the run shown in Figure 16.10). Thus s_1 is expected to grow exponentially with a characteristic timescale $\tau_g = 1/p$. The time taken from the arrival of the ACS to its spanning is $\tau_g \ln(s/s_1(n_a))$. This analytical result is confirmed by simulations (see Figure 16.11).

In the displayed run, after the first ACS (a 2-cycle) is formed at $n = 2854$, it takes 1026 time steps, until $n = 3880$ for the dominant ACS to span the entire graph (Figure 16.10c). This explains how a autocatalytic network structure and the positive feedback processes inherent in it can bootstrap themselves into existence from a small seed. The small seed, in turn, is more or less guaranteed to appear on a certain time scale ($1/p^2 s$ in the present model) just by random processes.

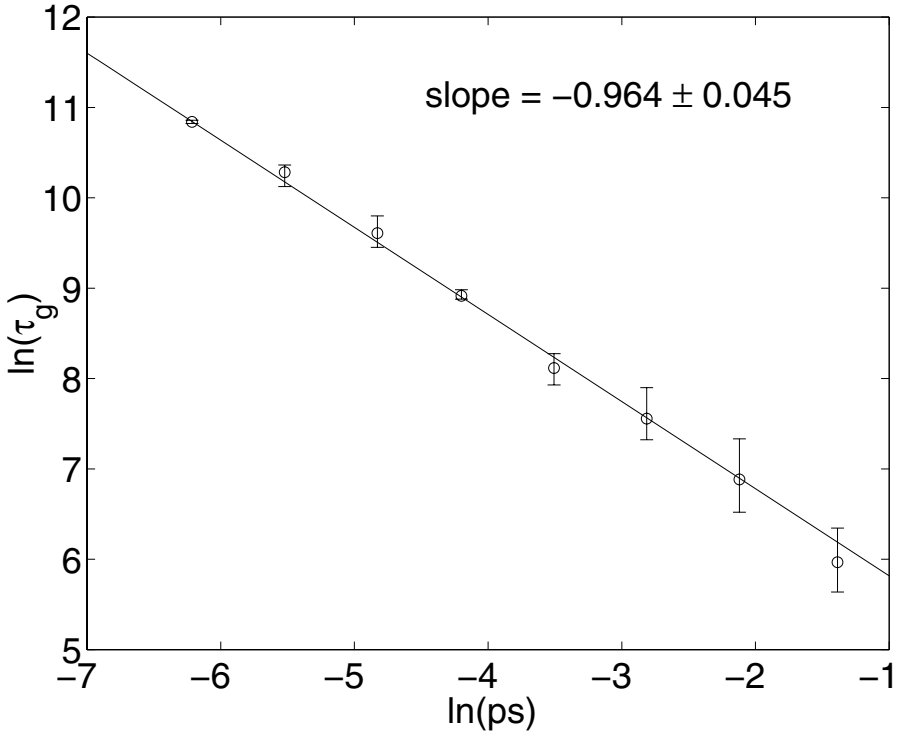


Figure 16.11: Each data point shows the average of τ_g (the growth timescale for an ACS) over 5 different runs with $s = 100$ and the given p value. The error bars correspond to one standard deviation. The solid line is the best linear fit to the data points on a log-log plot. Its slope is consistent with the analytically predicted slope -1 (see the discussion of the growth phase in section 5.)

A measure of the 'structure' of the evolved graph.

A fully autocatalytic graph is a highly improbable structure. Consider a graph of s nodes and let the probability of a positive link existing between any pair of nodes be p^*. Such a graph has on average $m^* = p^*(s-1)$ incoming or outgoing positive links per node since links from a node to itself are disallowed. For the entire graph to be an ACS, each node must have at least one incoming link, i.e. each row of the matrix C must contain at least one positive element. Hence the probability, P, for the entire graph to be an ACS is

$$\begin{aligned} P &= \text{probability that every row has at least one positive entry} \\ &= [\text{probability that a row has at least one positive entry}]^s \\ &= [1 - (\text{probability that every entry of a row is zero})]^s \\ &= [1 - (1-p^*)^{s-1}]^s \\ &= [1 - (1 - m^*/(s-1))^{s-1}]^s \end{aligned}$$

Note from Figure 16.8 that at spanning the number of links is $O(s)$. Thus the average degree m^* at spanning is $O(1)$. We have found this to be true in all the runs we have done where the initial average degree (at $n = 1$) was $O(1)$ or less.

For large s and $m^* \sim O(1)$, $P \approx (1 - e^{-m^*})^s \sim e^{-\alpha s}$, where α is positive, and $O(1)$. Thus a fully autocatalytic graph is exponentially unlikely to form if it were being assembled randomly. In the present model nodes are being added completely randomly but the underlying population dynamics and the selection imposed at each graph update result in the inevitable arrival of an ACS (in, on average, $\tau_a = 1/p^2 s$ time steps) and its inevitable growth into a fully autocatalytic graph in (on average) an additional $\sim \tau_g \ln s$ time steps.

It is a noteworthy feature of self-organization in the present model that an organization whose a priori probability to arise is exponentially small, $\sim e^{-\alpha s}$, arises inevitably in a rather short time, $\sim \frac{1}{p} \ln s$ (for large s). Why does that happen? First a small ACS of size $s_1(n_a) \sim O(1)$ forms by pure chance. The probability of this happening is not exponentially small; it is in fact quite substantial. Once this has formed, it is a cooperative structure and is therefore stable. Its appearance ushers in an exponential growth of structure with a time scale $\tau_g = 1/p$. Hence a graph whose 'structuredness' (measured by the reciprocal of the probability of its arising by pure chance) $= e^{\alpha s}$ arises in only $\frac{1}{p} \ln s$ steps.

As mentioned in the introduction, one of the major puzzles in the origin of life is the emergence of very special chemical organizations in a relatively short time. We hope that the mechanism described above, or its analogue in a sufficiently realistic model, will help in addressing this puzzle. The relevance of this mechanism for the origin of life is discussed in ref. [21]. We remark that other models of self-organization (e.g. the well-stirred hypercycle) do not seem to be able to produce complex structured organizations from a simple starting network (see ref. [23]).

Another graph theoretic measure of the structure of the evolved graph is 'interdependency' among the nodes, discussed in [16, 21]. Like the links and s_1, the interdependency is low in the random phase, then rises in the growth phase to a value that is about an order of magnitude higher.

16.5.3 The organized phase

Once an ACS spans the entire graph the effective dynamics again changes although the microscopic dynamical rules are unchanged. At spanning, for the first time since the formation of the initial ACS, a member of the dominant ACS will be picked for removal. This is because at spanning all nodes by definition belong to the dominant ACS and have non zero relative populations; one node nevertheless has to be picked for removal. Most of the time the removal of the node with the least X_i will result in minimal damage to the ACS. The rest of the ACS will remain with high populations, and the new node will keep getting repeatedly removed and replaced until it once again joins the ACS. Thus s_1 will fluctuate between s and $s-1$ most of the time. However, once in a while, the node which is removed happens to be playing a crucial role in the graph structure despite its low population. Then its removal can trigger large changes in the structure and catastrophic drops in s_1 and l. Alternatively it can sometimes happen that the new node added can trigger a catastrophe because of the new graph structure it creates. The catastrophes and the mechanisms which cause them are the subject of the next section.

16.6 Catastrophes and recoveries in the organized phase

Figure 16.12 shows the same run as that of Figure 16.9b for $n = 1$ to n =50,000. In this long run one can see several sudden, large drops in s_1: *catastrophes* in which a large fraction of the s species become extinct. Some of the drops seem to take the system back into the random phase, others are followed by *recoveries* in which s_1 rises back towards its maximum value s. The recoveries are comparatively slower than the catastrophes, which in fact occur in a single time step.

In order to understand what is happening during the catastrophes and subsequent recoveries we begin by examining the possible changes that an addition or a deletion of a node can make to the core of the dominant ACS.

Deletion of a node

We have already seen how the deletion of a node can change the core – recall the discussion of keystone nodes in section 3: the removal of a keystone node results in a zero overlap between the cores of the dominant ACS before and after the removal. A zero core overlap means that a single graph update event (in which one of the least populated species is replaced by a randomly connected one) has caused a major reorganization of the dominant ACS: the cores of the dominant ACS before and after the event (if an ACS still exists) have not even a single link in common. We will call such events *core-shifts*.

In an actual run a keystone node can only be removed if it happens to be one of the nodes with the least X_i. However the core nodes are often 'protected' by having higher X_i. Why is that?

$\mathbf{X}$ is an eigenvector of C with eigenvalue λ_1. Therefore, when $\lambda_1 \neq 0$ it follows that for nodes of the dominant ACS, $X_i = (1/\lambda_1)\sum_j c_{ij}X_j$. If node i of the dominant ACS has only one incoming link (from the node j, say) then $X_i = X_j/\lambda_1$; we can say that X_i is 'attenuated' with respect to X_j by a factor λ_1. The periphery of an ACS is a tree like structure emanating

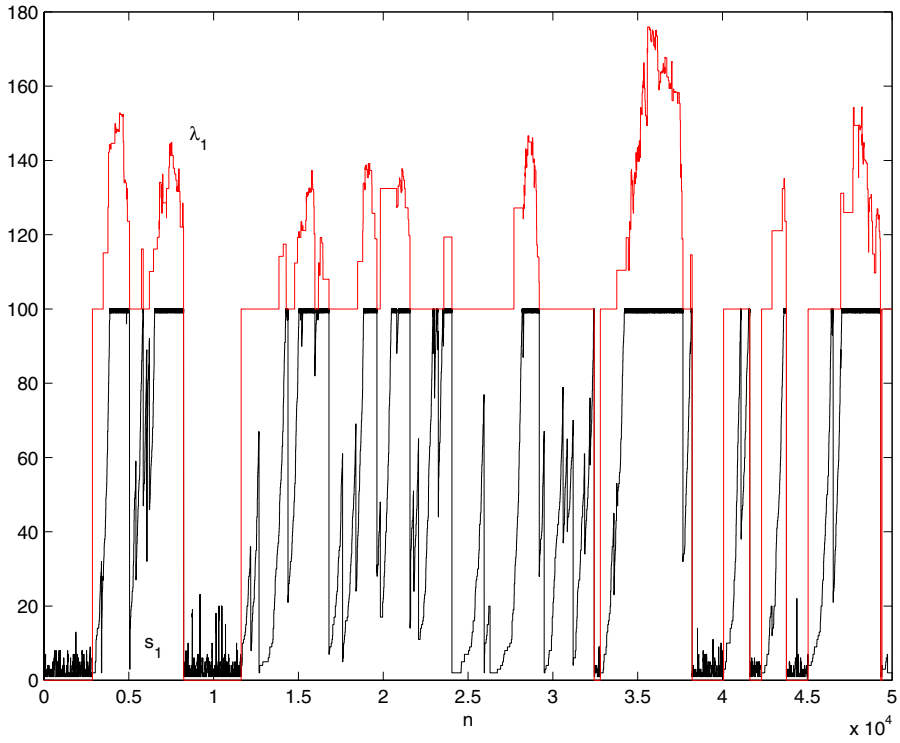


Figure 16.12: The same run displayed in Figure 16.9b over a longer timescale, till $n = 50000$. This displays repeated rounds of crashes and recoveries.

from the core, and for small p most periphery nodes have a single incoming link. For instance the graph in Figure 16.10c, whose $\lambda_1 = 1.31$, has a chain of nodes $44 \to 45 \to 24 \to 29 \to 52 \to 89 \to 86 \to 54 \to 78$. The farther down such a chain a periphery node is, the lower is its X_i because of the cumulative attenuation. For such an ACS with $\lambda_1 > 1$ the 'leaves' of the periphery tree (such as node 78) will typically be the species with least X_i while the core nodes will have larger X_i.

However, when $\lambda_1 = 1$ there is no attenuation. Recall that Proposition 1(iii) shows that at $\lambda_1 = 1$ the core must be a cycle or a set of disjoint cycles, hence each core node has only one incoming link within the dominant ACS. All core nodes have the same value of X_i. As one moves out towards the periphery $\lambda_1 = 1$ implies there is no attenuation, hence each node in the periphery that receives a single link from one of the core nodes will also have the same X_i. Some periphery nodes may have higher X_i if they have more than one incoming link from the core. Iterating this argument as one moves further outwards from the core, it is clear that at $\lambda_1 = 1$ the core is not protected and in fact will always belong to the set of least fit nodes if the dominant ACS spans the graph. We have already seen in section 3 that when $\lambda_1 = 1$ and the core is a single cycle every core node is a keystone node. Thus when $\lambda_1 = 1$ the organization is fragile and susceptible to core-shifts caused by the removal of a keystone node.

Addition of node

We now turn to the effects of the addition of a node to the dominant ACS. We will use the notation $C'_n \equiv C_{n-1} - k$ for the graph of $s-1$ nodes just before the new node at time step n is brought in (and just after the least populated species k is removed from C_{n-1}). Q'_n will stand for the core of C'_n. In the new attractor the new species k may go extinct, i.e., X_k may be zero, or it may survive, i.e., X_k is non-zero. If the new species goes extinct then it remains in the set of least fit nodes and clearly there is no change to the dominant ACS. So we will focus on events in which the new species survives in the new attractor.

Innovations

We define an *innovation* to be a new node for which X_k in the new attractor is nonzero, i.e. a new node which survives till the next graph update [23]. This may seem to be a very weak requirement, yet we will see that it has nontrivial consequences. A description of various types of innovations and their consequences, with examples, is given in [30]. Here we present a graph-theoretic classification of innovations (in terms of a hierarchy, see Figure 16.13).

Remarks to Fig. 16.13: All classes of events except the leaves of the tree are subdivided into two exhaustive and mutually exclusive subclasses (represented by the two branches emanating downwards from the class). The number of events in each class pertain to the run of Figure 16.9b with a total of 9999 graph updates, between $n = 1$ (the initial graph) and $n = 10000$. In that run, out of 9999 node addition events, most (8929 events) are not innovations. The rest (1070 events), which are innovations, are classified according to their graph theoretic structure. The classification is general; it is valid for all runs. X_k is the relative population of the new node in the attractor configuration of (16.1) that is reached in step 1 of the dynamics (see Section 4) immediately following the addition of that node. N stands for the new irreducible subgraph, if any, created by the new node. If the new node causes a new irreducible subgraph to be created, N is the maximal irreducible subgraph that includes the new node. If not, $N = \phi$ (where ϕ stands for the empty set). Q_{in} is the core of the graph just before the addition of the node (just before step 3 of the dynamics in Section 4) and Q_{fin} the core just after the addition of the node. The six leaves of the innovation subtree are numbered from 1 to 6 and correspond to the classes discussed in Section 6. The impact of each kind of innovation on the system dynamics is discussed in the text and in more detail in [30]. Some classes of events happen rarely (e.g., classes numbered 5 and 6) but have a major impact on the dynamics of the system. The precise impact of all these classes of innovations on the system over a short time scale (before the next graph update) as well as their probable impact over the medium term (upto a few thousand graph updates) can be predicted from the graph theoretic structure of N and the rest of the graph at the moment these innovations appear in a run.

The innovations which have the least impact on the populations of the species and the evolution of the graph on a short time scale (of a few graph updates) are ones which do not affect the core of the dominant ACS, if it exists. Such innovations are of three types (see boxes 1-3 in Figure 16.13):

1. Random phase innovations. These are innovations which occur in the random phase when no ACS exists in the graph, and they do not create any new ACSs. These innovations are typically short lived and have little short term or long term impact on the structure of the graph.

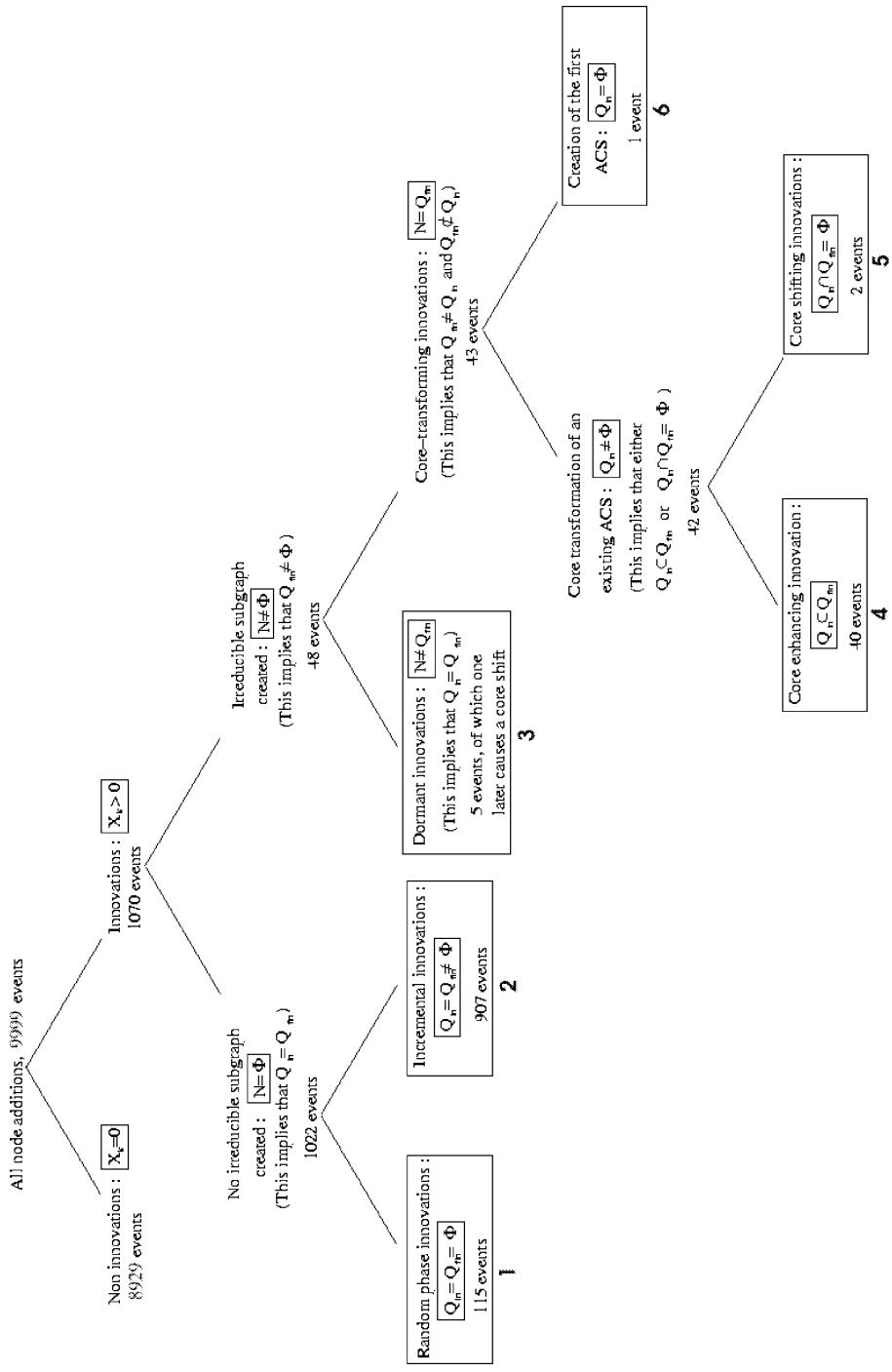


Figure 16.13: A hierarchy of innovations. Each node in this binary tree represents a class of node addition events. Each class has a name; the small box contains the mathematical definition of the class.

2. Incremental innovations. These are innovations which occur in the growth and organized phases, which add new nodes to the periphery of the dominant ACS without creating any new irreducible subgraph. In the short term they only affect the periphery and are responsible for the growth of the dominant ACS. In a longer term they can also affect the core as chains of nodes from the periphery join the core of the dominant ACS.

3. Dormant innovations. These are innovations which occur in the growth and organized phases, which create new irreducible subgraphs in the periphery of the dominant ACS. These innovations also affect only the periphery in the short term. But they have the potential to cause core-shifts later if the right conditions occur (discussed in the next subsection).

Innovations which do immediately affect the core of the existing dominant ACS are always ones which create a new irreducible subgraph. They are also of three types (see boxes 4-6 in Figure 16.13):

4. Core enhancing innovations. These innovations result in the expansion of the existing core by the addition of new links and nodes from the periphery or outside the dominant ACS. They result in an increase of λ_1 of the graph.

5. Core-shifting innovations. These are innovations which cause an immediate core-shift often accompanied by the extinction of a large number of species.

6. Creation of the first ACS. This is an innovation which creates an ACS for the first time in a graph which till then had no ACSs. The innovation moves the system from the random phase to the growth phase, triggering the self organization of the system around the newly created ACS.

Innovations of types 4, 5 and 6 which affect the core of the dominant ACS will be called *core-transforming innovations*. These innovations cause a substantial change the vector of relative populations in a single graph update. Innovations of type 5 and 6 also make a qualitative change in the structure of the graph and significantly influence subsequent graph evolution. The following theorem makes precise the conditions under which a core transforming innovation can occur.

Core transforming Theorem

Let N (or N_n at time step n) denote the maximal new irreducible subgraph which includes the new species. One can show that N_n will become the new core of the graph, replacing the old core Q_{n-1}, whenever either of the following conditions are true:
(a) $\lambda_1(N_n) > \lambda_1(Q'_n)$ or,
(b) $\lambda_1(N_n) = \lambda_1(Q'_n)$ and N_n is 'downstream' of Q'_n (i.e., there is a path from Q'_n to N_n but not from N_n to Q'_n.)

Such an innovation will fall into category 4 above if $Q_{n-1} \subset N_n$. However, if Q_{n-1} and N_n are disjoint, we get a core-shift and the innovation is of type 5 if Q_{n-1} is non-empty and type 6 otherwise.

16.6.1 Catastrophes, core-shifts and a classification of proximate causes

The large sudden drops visible in Figure 16.12 are now discussed. Our first task is to see if the large drops are correlated to specific changes in the structure of the graph. Let us focus on those events in which more than 50% of the species go extinct. There were 701 such events out of 1.55 million graph updates in a set of runs with $s = 100, p = 0.0025$. Figure 16.14 shows a histogram of core overlaps $Ov(C_{n-1}, C_n)$ for these 701 events. 612 of these have zero core overlap, i.e., they are core-shifts.

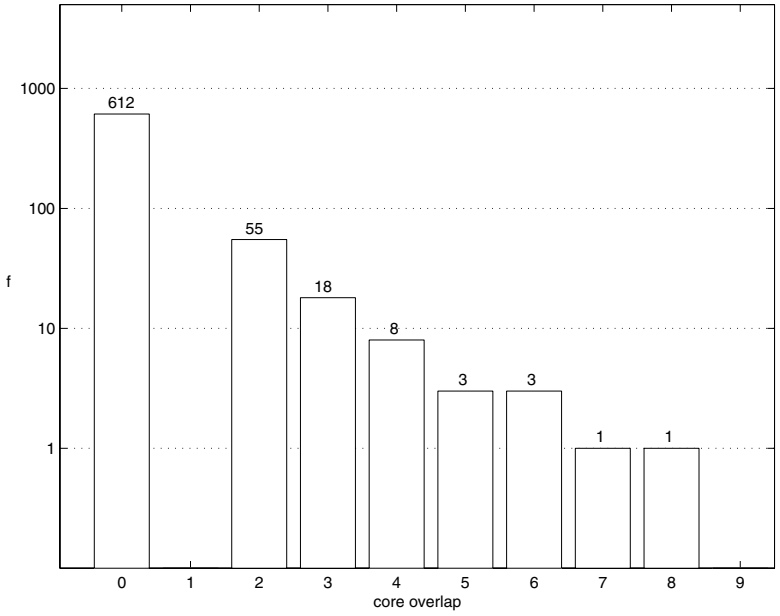


Figure 16.14: Large crashes are predominantly core-shifts. A histogram of core overlaps for the 701 events where s_1 dropped by more than $s/2$ observed in various runs with $s = 100$ and $p = 0.0025$, totalling 1.55 million iterations.

If we now look at only those events in which more than 90% of the species went extinct then we find 235 such events in the same runs, out of which 226 are core-shifts. Clearly most of the large extinction events happen when there is a drastic change in the structure of the dominant ACS – a core-shift.

Classification of core-shifts

Using the insights from the above discussion of the effects of deletion or addition of a node, we can classify the different mechanisms which cause core-shifts. Figure 16.15 differentiates between the 612 core-shifts we observed amongst the 701 crashes.

They fall into three categories [23]: (i) complete crashes (136 events), (ii) takeovers by core-transforming innovations (241 events), and (iii) takeovers by dormant innovations (235 events).

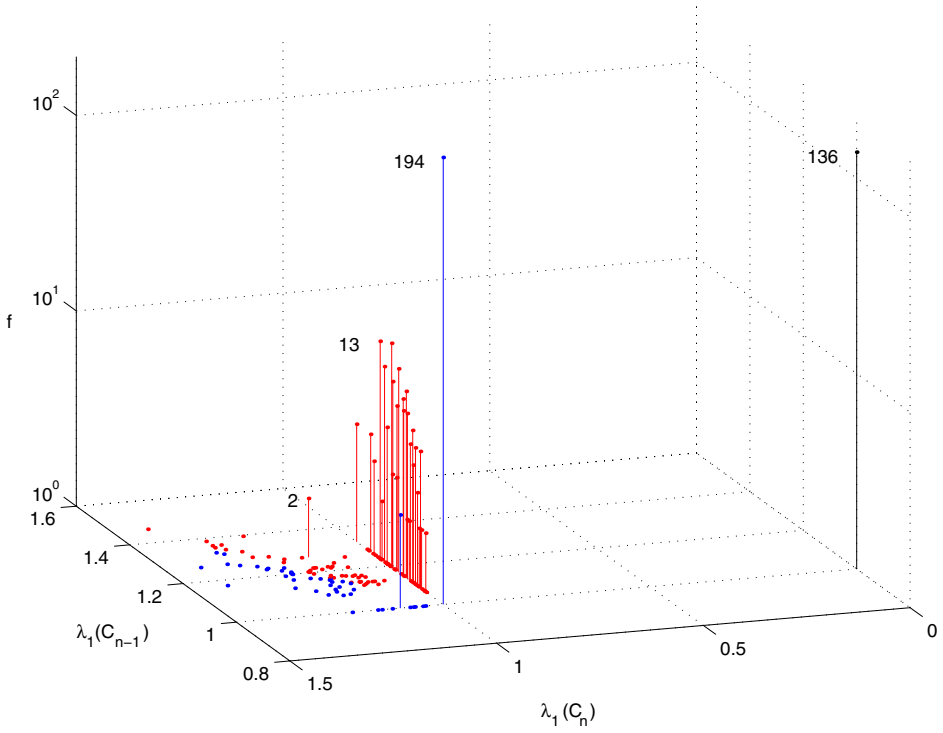


Figure 16.15: Classification of core-shifts into three categories. The graph shows the frequency, f, of the 612 core-shifts observed (see Figure 16.14) in a set of runs with $s = 100$ and $p = 0.0025$ vs. the λ_1 values before, $\lambda_1(C_{n-1})$, and after, $\lambda_1(C_n)$, the core-shift. Complete crashes (black; $\lambda_1(C_{n-1}) = 1$, $\lambda_1(C_n) = 0$), takeovers by core-transforming innovations (blue; $\lambda_1(C_n) \geq \lambda_1(C_{n-1}) \geq 1$) and takeovers by dormant innovations (red; $\lambda_1(C_{n-1}) > \lambda_1(C_n) \geq 1$) are distinguished. Numbers alongside vertical lines represent the corresponding f value.

Complete crashes

A *complete crash* is an event in which an ACS exists before but not after the graph update. Such an event takes the system into the random phase. A complete crash occurs when a keystone node is removed from the graph. For example at $n = 8232$ the graph had $\lambda_1 = 1$ and its core was the simple 3-cycle of nodes 20, 50 and 54. As we have seen above, when the core is a single cycle every core node is a keystone node and is also in the set of least fit nodes. At $n = 8233$ node 54 was removed thus disrupting the 3-cycle. The resulting graph had no ACS and λ_1 dropped to zero. As we have discussed earlier, graphs with $\lambda_1 = 1$ are the ones which are most susceptible to complete crashes. This can be seen in Figure 16.15: every complete crash occurred from a graph with $\lambda_1(C_{n-1}) = 1$.

Takeovers by core-transforming innovations

An example of a takeover by a core-transforming innovation is given in Figures 16.10g,h. At $n = 6061$ the core was a single loop comprising nodes 36 and 74. Node 60 was replaced by a new species at $n = 6062$ creating a cycle comprising nodes $60, 21, 41, 19$ and 73, downstream from the old core. The graph at $n = 6062$ has one cycle feeding into a second cycle that is downstream from it. We have already seen in section 3 (see the discussion of Example 4) that for such a graph only the downstream cycle is populated and the upstream cycle and all nodes dependent on it go extinct. Thus the new cycle becomes the new core and the old core goes extinct resulting in a core-shift. This is an example of condition (b) for a core-transforming innovation. For all such events in Figure 16.15, $\lambda_1(Q'_n) = \lambda_1(C_{n-1})$ since k happened not to be a core node of C_{n-1}. Thus these core-shifts satisfy $\lambda_1(C_n) = \lambda_1(N_n) \geq \lambda_1(Q'_n) = \lambda_1(C_{n-1}) \geq 1$ in Figure 16.15.

Takeovers by dormant innovations

We have earlier discussed dormant innovations, which create an irreducible structure in the periphery of the dominant ACS which does not affect its core at that time. For example the 2-cycle comprising nodes 36 and 74 formed at $n = 4696$. At a later time such a dormant innovation can result in a core-shift if the old core gets sufficiently weakened.

In this case the core has become weakened by $n = 5041$, when it has $\lambda_1 = 1.24$. The structure of the graph at this time is very similar to the graph in Figure 16.7a. Just as node 3 in Figure 16.7a was a keystone node, here nodes 44, 85, and 98 are keystone nodes because removing any of them results in a graph like Figure 16.7b, consisting of two 2-cycles, one downstream from the other.

Indeed at $n = 5041$, node 85 is hit and the resulting graph at $n = 5042$ has a cycle (26 and 90) feeding into another cycle (36 and 74). Thus at $n = 5042$ nodes 36 and 74 form the new core with only one other downstream node, 11, being populated. All other nodes become depopulated resulting in a drop in s_1 by 97. A dormant innovation can takeover as the new core only following a keystone extinction which weakens the old core. In such an event the new core necessarily has a lower (but nonzero) λ_1 than the old core, i.e., $\lambda_1(C_{n-1}) > \lambda_1(C_n) \geq 1$ (see Figure 16.15).

Note that 85 is a keystone node, and the graph is susceptible to a core-shift *because* of the innovation which created the cycle 36-74 earlier. If the cycle between 36 and 74 were absent, 85 would *not* be a keystone species by our definition, since its removal would still leave part of the core intact (nodes 26 and 90).

16.6.2 Recoveries

After a complete crash the system is back in the random phase. In $O(s)$ graph updates each node is removed and replaced by a randomly connected node, resulting in a graph as random as the initial graph. Then the process starts again, with a new ACS being formed after an average of $1/p^2 s$ time steps and then growing to span the entire graph after, on average, $(1/p)\ln(s/s_0)$ time steps, where s_0 is the size of the initial ACS that forms in this round (typically $s_0 = 2$).

After other catastrophes, an ACS always survives. In that case the system is in the growth phase and immediately begins to recover, with s_1 growing exponentially on a timescale $1/p$. Note that these recoveries happen because of innovations (mainly of type 2 and 4, and some of type 3).

16.6.3 Correlation between graph theoretic nature of perturbation and its short and long term impact

In previous sections we have analysed several examples of perturbations to the system. These can be broadly placed in two classes based on their effect on s_1:
(i) 'Constructive perturbations': these include the birth of a new organization (an innovation of type 6), the attachment of a new node to the core (an innovation of type 4) and an attachment of a new node to the periphery of the dominant ACS (an innovation of type 2).
(ii) 'Destructive perturbations': these include complete crashes and takeovers by dormant innovations (both caused by the loss of a keystone node), and takeovers by core-transforming innovations (innovations of type 5). Note that the word 'destructive' is used only in the sense that several species go extinct on a short time scale (a single graph update in the present model) after such a perturbation. In fact, over a longer time scale (ranging from a few to several hundred graph updates in the run of Figure 16.9b), the 'destructive' takeovers by innovations generally trigger a new round of 'constructive' events like incremental innovations (type 2) and core enhancing innovations (type 4).

Note that the maximum upheaval is caused by those perturbations that introduce new irreducible structures in the graph (innovations of type 4, 5 and 6) or those that destroy the existing irreducible structure. For example the creation of the first ACS at $n = 2854$ triggered the growth phase, a complete change in the effective dynamics of the system. Other examples of large upheavals are core-shifts caused by a takeover by a core-transforming innovation at $n = 6061$, takeover by a dormant innovation at $n = 5041$, and a complete crash at $n = 8233$. In sections 2 and 3 we have mentioned that irreducibility is related to the existence of positive feedback and cooperation, and the 'magnitude' of the feedback is measured by λ_1. While the present model is a highly simplified model of evolving networks, we expect that this qualitative feature, namely, the correlation between the dynamical impact of a perturbation and its 'structural' character embodied in its effect on the 'level of feedback' in the underlying graph, will hold for several other complex systems.

16.7 Concluding remarks

In this article we have attempted to show that a certain class of dynamical systems, those in which graphs coevolve with other dynamical variables living on them (in our example, living on the nodes of the graph), possess rich dynamical behaviour which is analytically and computationally tractable. Even in the highly idealized model discussed here, this behaviour is reminiscent of what happens in real life — birth of organizational structure characterized by interdependence of components, cooperation of parts of the organization giving way to competition, robust organizations becoming fragile, crashes and recoveries, innovations causing growth as well as collapse, etc.

From the point of view of the origin of life problem the main conclusions are:

(i) The model shows the emergence of an organization where none exists: a small ACS emerges spontaneously by random processes and then triggers the self-organization of the system.

(ii) A highly structured organization, whose timescale of forming by pure chance is exponentially large (as a function of the size of the system), forms in this model in a very short timescale that grows only logarithmically with the size of the system. In [21] we have speculated that this timescale may be ~ 100 million years for peptide based ACSs, which is in the same ballpark as the timescale on which life is believed to have originated on the prebiotic earth.

We remark that this speculation is not necessarily in conflict with, and is possibly complementary to, some other approaches to the origin of life:

(i) Complex autocatalytic organizations of polypeptides could enter into symbiosis with the autocatalytic citric acid cycle proposed in [31]. The latter would help produce, among other things, amino acid monomers needed by the former; the former would provide catalysts for the latter.

(ii) It is conceivable that membranes (possibly lipid membranes, which have been argued to have their own catalytic dynamics [32]) could form in regions where autocatalytic sets of the kind discussed here existed, thereby surrounding complex molecular organization in an enclosure. These 'cells' may have contained different parts of the ACS, thereby endowing them with different fitnesses. Such an assembly could evolve.

(iii) It is also conceivable that such molecular organizations formed an enabling environment for self replicating molecules such as those needed for an RNA world.

Testing some of these possibilities is a task for future models and experiments. Furthermore, the mathematical ideas and mechanisms discussed here might be relevant for these other approaches also.

The present model has a number of simplifying features which depart from realism but enhance analytical tractability. One is the linearity of the populations dynamics on a fixed graph. Equation (16.1) is nonlinear, but since it originates via a nonlinear change of variables from a linear equation, equation (16.2), its attractors can be easily analysed in terms of the underlying linear system. The attractors are always fixed points, and are just the Perron Frobenius eigenvectors of the adjacency matrix of the graph. This allows us to use (static) graph theoretic results for analysis of the dynamics.

In this context it is helpful to note that while the population dynamics in the present model is essentially linear as long as the graph is fixed, the model feeds the result of the population dynamics into the subsequent graph update (the least populated node is removed). Thus over long time scales over which the graph changes, the 'coupling constants' c_{ij} in equation (16.1) are not constant but implicitly depend upon the x_i, thus making the evolution highly nonlinear. By virtue of the simplifying device of widely separated time scales for the graph dynamics and the population dynamics (the population variables reach their attractor before the graph is modified), what we have is piecewise linear population dynamics. It is essentially linear

between two graph updates, and nonlinear over longer time scales because of the intertwining of population dynamics and graph dynamics. This nonlinearity is essential for all the complex phenomena described above, while the short time scale linearity is an aid in analysis. It would be interesting to explore complex phenomena in models in which the short term population dynamics is also inherently nonlinear. This naturally arises in prebiotic chemistry when the concentration of the reactants (which are assumed buffered here) are dynamical variables in addition to the catalysts and products, as well as in several other fields.

The present model describes a well-stirred reactor; there are no spatial degrees of freedom. This precludes a discussion of the origin of spatial structure and its consequences alluded to in section 1. It is worthwhile to extend the model in that direction. Another issue is the generation of novelty. Here the links of the new node are drawn from a fixed probability distribution. In real systems this distribution depends upon the (history of) states of the system. A further direction for generalization consists in letting the two time scales of the population and graph dynamics, separated by hand in the present model, be endogenous.

Acknowledgements

S. J. acknowledges the Associateship of the Abdus Salam International Centre for Theoretical Physics, Trieste. S. K. acknowledges a Junior Research Fellowship from the Council of Scientific and Industrial Research, India. This work is supported in part by a grant from the Department of Science and Technology, Govt. of India.

References

[1] Albert, R. and Barabasi, A.-L. (2002) Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74**, 47 *(www.arXiv.org/abs/cond-mat/0106096)*.

[2] Dorogovtsev, S. N. and Mendes, J. F. F. (2002) Evolution of networks, *Adv. Phys.* **51** 1079 *(www.arXiv.org/abs/cond-mat/0106144)*.

[3] Strogatz, S. H. (2001) Exploring complex networks, *Nature* **410**, 268-276.

[4] Watts, D. J. (1999) *Small Worlds: The dynamics of Networks between Order and Randomness* (Princeton Univ. Press, Princeton).

[5] Dyson, F. (1985) *Origins of Life* (Cambridge Univ. Press Cambridge, UK).

[6] Farmer, J. D., Kauffman, S. and Packard, N. H. (1986) Autocatalytic replication of polymers, *Physica* **D22** 50-67.

[7] Bagley, R. J., Farmer, J. D. and Fontana, W. (1991) Evolution of a metabolism, in *Artificial Life II*, eds. Langton, C. G., Taylor, C., Farmer, J. D. and Rasmussen, S. (Addison Wesley, Redwood City), pp. 141-158.

[8] Kauffman, S. A. (1993) *The Origins of Order* (Oxford Univ. Press).

[9] Bak, P. and Sneppen, K. (1993) Punctuated equilibrium and criticality in a simple model of evolution, *Phys. Rev. Lett.* **71**, 4083-4086.

[10] Fontana, W. and Buss, L. (1994) The arrival of the fittest: Toward a theory of biological organization, *Bull. Math. Biol.* **56**, 1-64.

[11] Harary, F. (1969) *Graph Theory* (Addison Wesley, Reading, MA, USA).

[12] Bang-Jensen, J. and Gutin, G. (2001) *Digraphs: Theory, Algorithms and Applications* (Springer-Verlag, London).

[13] Seneta, E. (1973) *Non-Negative Matrices* (George Allen and Unwin, London).

[14] Berman, A. and Plemmons, R. J. (1994) *Non-negative matrices in the mathematical sciences* (SIAM, Philadelphia).

[15] Rothblum, U. G. (1975) Algebraic eigenspaces of nonnegative matrices, *Linear Algebra and Appl* **12**, 281-292.

[16] Jain, S. and Krishna, S. (1999) Emergence and growth of complex networks in adaptive systems, *Computer Physics Comm.* **121-122**, 116-121.

[17] Eigen, M. (1971) Self-organization of matter and the evolution of biological macromolecules, *Naturwissenschaften* **58**, 465-523.

[18] Kauffman, S.A. (1971) Cellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems, *J. Cybernetics* **1**, 71-96.

[19] Rossler, O. E. (1971) A system theoretic model of biogenesis, *Z. Naturforschung* **26b**, 741-746.

[20] Jain, S. and Krishna, S. (1998) Autocatalytic sets and the growth of complexity in an evolutionary model, *Phys. Rev. Lett.* **81**, 5684-5687.

[21] Jain, S. and Krishna, S. (2001) A model for the emergence of cooperation, interdependence and structure in evolving networks, *Proc. Natl. Acad. Sci. (USA)* **98**, 543-547.

[22] Jain, S. and Krishna, S. (2002) Crashes, recoveries and 'core-shifts' in a model of evolving networks, *Phys. Rev. E* **65**, 026103, *www.arXiv.org/abs/nlin.AO/0107037*.

[23] Jain, S. and Krishna, S. (2001) Large extinctions in an evolutionary model: the role of innovation and keystone species, *Proc. Natl. Acad. Sci. (USA)* **99**, 2055-2060, *www.arXiv.org/abs/nlin.AO/0107038*.

[24] Paine, R. T. (1969) A note on trophic complexity and community stability, *Am. Nat.* **103**, 91-93.

[25] Pimm, S. L. (1991) *The Balance of Nature? Ecological Issues in the Conservation of Species and Communities* (Univ. of Chicago Press, Chicago).

[26] Jordán, F., Takács-Sánta, A. and Molnár, I. (1999) A reliability theoretical quest for keystones, *OIKOS* **86**, 453-462.

[27] Solé, R. V. and Montoya, J. M. (2000) Complexity and fragility in ecological networks, *www.arXiv.org/abs/cond-mat/0011196*.

[28] Joyce, G. F., Schwartz, A. W., Miller, S. L. and Orgel, L. E. (1987) The case for an ancestral genetic system involving simple analogues of the nucleotides, *Proc. Natl. Acad. Sci. (USA)* **84**, 4398-4402.

[29] Joyce, G. F. (1989) RNA evolution and the origins of life, *Nature* **338**, 217-223.

[30] Jain, S. and Krishna, S. (2002) Constructive and destructive effects of 'innovation' in evolving networks, Preprint 2002.

[31] Morowitz, H. J., Kostelnik, J. D., Yang, J. and Cody, G. D. (2000) The origin of intermediary metabolism, *Proc. Natl. Acad. Sci. (USA)* **97**, 7704-7708.

[32] Segré, D., Ben-Eli, D., Deamer, W. D. and Lancet, D. (2001) The lipid world, *Origins of Life and Evol. of the Biosphere* **31**, 119-145.

Index