Stefan Bornholdt, Heinz Georg Schuster (Eds.)

Handbook of Graphs and Networks

From the Genome to the Internet

WILEY-VCH

Stefan Bornholdt, Heinz Georg Schuster (Eds.)

Handbook of Graphs and Networks

From the Genome to the Internet

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.

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

- *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

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 $[n] = \{1, 2, \dots, n\}$.

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 $\mathcal{G}_{n,.}$ 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*

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

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.