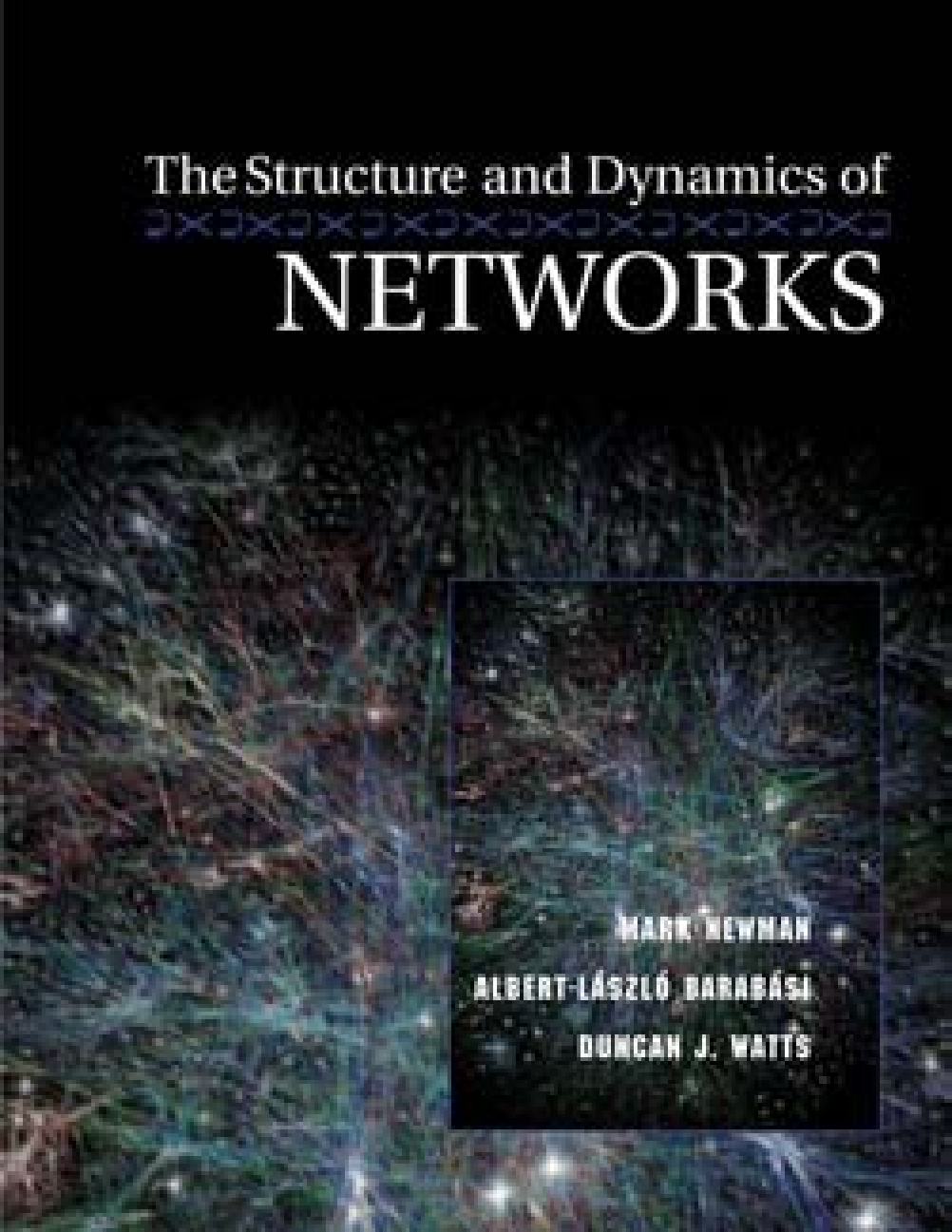
The Structure and Dynamics of NETWORKS



MARK NEWMAN
ALBERT-LÁSZLÓ BARABÁSI
DUNCAN J. WATTS

The Structure and Dynamics of Networks

Princeton Studies in Complexity

Series Editors:

Philip W. Anderson (Princeton University); Joshua M. Epstein (The Brookings Institution); Duncan K. Foley (Barnard College); Simon A. Levin (Princeton University); Martin A. Nowak (Harvard University)

Lars-Erik Cederman, *Emergent Actors in World Politics: How States and Nations Develop and Dissolve*

Robert Axelrod, *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*

Peter S. Albin, *Barriers and Bounds to Rationality: Essays on Economic Complexity and Dynamics in Interactive Systems*. Edited and with an introduction by Duncan K. Foley

Duncan J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*

Scott Camazine, Jean-Louis Deneubourg, Nigel R. Franks, James Sneyd, Guy Theraulaz, Eric Bonabeau, *Self-Organization in Biological Systems*

Peter Turchin, *Historical Dynamics: Why States Rise and Fall*

Mark Newman, Albert-László Barabási, and Duncan J. Watts, eds., *The Structure and Dynamics of Networks*

J. Stephen Lansing, *Perfect Order: Recognizing Complexity in Bali*

The Structure and Dynamics of Networks

Mark Newman
Albert-László Barabási
Duncan J. Watts

Editors

Princeton University Press
Princeton and Oxford

Published by Princeton University Press, 41 William Street,
Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press, 3 Market Place,
Woodstock, Oxfordshire OX20 1SY

Library of Congress Control Number: 2005921569

ISBN-13: 978-0-691-11356-2 (cl. alk. paper)
ISBN-10: 0-691-11356-4 (cl. alk. paper)

ISBN-13: 978-0-691-11357-9 (paper alk. paper)
ISBN-10: 0-691-11357-2 (paper alk. paper)

British Library Cataloging-in-Publication Data is available

The publisher would like to acknowledge the editors of this volume for providing, other than the previously published material, the camera-ready copy from which this book was printed.

Printed on acid-free paper.

pup.princeton.edu

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Networks such as the Internet, the World Wide Web, and social and biological networks of various kinds have been the subject of intense study in recent years. From physics and computer science to biology and the social sciences, researchers have found that a great variety of systems can be represented as networks, and that there is much to be learned by studying those networks. The study of the web, for instance, has led to the creation of new and powerful web search engines that greatly outperform their predecessors. The study of social networks has led to new insights about the spread of diseases and techniques for controlling them. The study of metabolic networks has taught us about the fundamental building blocks of life and provided new tools for the analysis of the huge volumes of biochemical data that are being produced by gene sequencing, microarray experiments, and other techniques.

In this book we have gathered together a selection of research papers covering what we believe are the most important aspects of this new branch of science. The papers are drawn from a variety of fields, from many different journals, and cover both empirical and theoretical aspects of the study of networks. Along with the papers themselves we have included some commentary on their contents, in which we have tried to highlight what we believe to be the most important findings of each of the papers and offer pointers to other related literature. (Note that within the text of our commentary we have for convenience marked in **bold text** citations to papers that themselves are reproduced within this book; we hope this will save the reader some unnecessary trips to the library.)

After a short introduction (Chapter 1), the book opens with a collection of historical papers (Chapter 2) that predate the current burst of interest in networks, but that lay important foundations for the later work. Chapter 3 reproduces a selection of papers on empirical studies of networks in various fields, the raw experimental data on which many theoretical developments build. Then in Chapter 4, which occupies the largest portion of the book, we look at models of networks, focusing particularly on random graph models, small-world models, and models of scale-free networks. Chapter 5 deals with applications of network ideas to particular real-world problems, such as epidemiology, network robustness, and search algorithms. Finally, in Chapter 6 we give a short discussion of the most recent developments and where we see the field going in the next few years. There will of course be many developments that we cannot anticipate at present, and we look forward with excitement to the new ideas researchers come up with as we move into the 21st century.

This field is growing at a tremendous pace, with many new papers appearing every day, so there is no hope of making a compilation such as this exhaustive; inevitably many important and deserving papers have been left out. Nonetheless, we hope that by collecting a representative selection of papers together in one volume, this book will prove useful to students and researchers alike in the field of networks.

A number of people deserve our thanks for their help with the creation of this book. First, our thanks must go to our editor Vickie Kearn and everyone else at Princeton University Press for taking on this project and helping to make it a success. Many thanks also to Ádám Makkai, who translated from the original Hungarian the remarkable short story *Chains* that forms the first article reproduced in the book. And of course we have benefited enormously from conversations with our many erudite colleagues in the field. It is their work that forms the bulk of the material in this volume, and we are delighted to be a part of such an active and inspiring community of scientists.

Mark Newman
Albert-László Barabási
Duncan Watts

Chapter One
Introduction

Networks are everywhere. From the Internet and its close cousin the World Wide Web to networks in economics, networks of disease transmission, and even terrorist networks, the imagery of the network pervades modern culture.

What exactly do we mean by a network? What different kinds of networks are there? And how does their presence affect the way that events play out? In the past few years, a diverse group of scientists, including mathematicians, physicists, computer scientists, sociologists, and biologists, have been actively pursuing these questions and building in the process the new research field of network theory, or the "science of networks" (Barabási 2002; Buchanan 2002; Watts 2003).

Although it is still in a period of rapid development and papers are appearing daily, a significant literature has already accumulated in this new field, and it therefore seems appropriate to summarize it in a way that is accessible to researchers unfamiliar with the topic. That is the purpose of this book. We begin by sketching in this introductory chapter a brief history of the study of networks, whose beginnings lie in mathematics and more recently sociology. We then place the "new" science of networks in context by describing a number of features that distinguish it from what has gone before, and explain why these features are important. At the end of the chapter we give a short outline of the remainder of the book.

1.1 A BRIEF HISTORY OF THE STUDY OF NETWORKS

The study of networks has had a long history in mathematics and the sciences. In 1736, the great mathematician Leonard Euler became interested in a mathematical riddle called the Königsberg Bridge Problem. The city of Königsberg was built on the banks of the Pregel River in what was then Prussia,[1] and on two islands that lie in midstream. Seven bridges connected the land masses, as shown in Figure 1.1. (There are many more than that today.) A popular brain-teaser of the time asked, "Does there exist any single path that crosses all seven bridges exactly once each?" Legend has it that the people of Königsberg spent many fruitless hours trying to find such a path before Euler proved the impossibility of its existence. The proof, which perhaps seems rather trivial to us now, but which apparently wasn't obvious in 1736, makes use of a *graph*—a mathematical object consisting of points, also called *vertices* or *nodes*, and lines, also called *edges* or *links*, which abstracts away

[1] Today Königsberg lies in Russia and is called Kaliningrad.

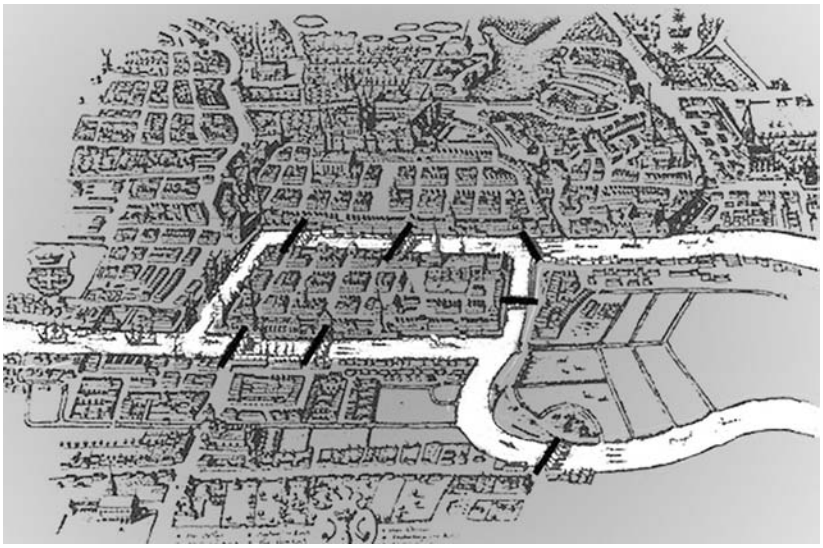


FIGURE 1.1 A map of eighteenth century Königsberg, with its seven bridges highlighted.

all the details of the original problem except for its connectivity. In this graph there are four vertices representing the four land masses and seven edges joining them in the pattern of the Königsberg bridges (Figure 1.2). Then the bridge problem can be rephrased in mathematical language as the question of whether there exists any *Eulerian path* on the network. An Eulerian path is precisely a path that traverses each edge exactly once. Euler proved that there is not, by observing that, since any such path must both enter and leave every vertex it passes through, except the first and last, there can at most be two vertices in the network with an odd number of edges attached. In the language of graph theory, we say that there can at most be two vertices with odd *degree*, the degree of a vertex being the number of edges attached to it.[2] Since all four vertices in the Königsberg graph have odd degree, the bridge problem necessarily has no solution. The problem of the existence of Eulerian paths on networks, as well as the related problem of *Hamiltonian paths* (paths that visit each vertex exactly once), is still of great interest to mathematicians, with new results being discovered all the time.

Many consider Euler's proof to be the first theorem in the now highly developed field of discrete mathematics known as *graph theory*, which in the past three centuries has become the principal mathematical language for describing the properties of networks (Harary 1995; West 1996). In its simplest form, a network is nothing more than a set of discrete elements (the vertices), and a set of connections (the edges) that link the elements, typically in a pairwise fashion. The elements

[2]Within physics some authors have referred to this quantity as the "connectivity" of a vertex, and the reader will see this usage in some of the papers reproduced in this book. As the word connectivity already has another meaning in graph theory, however, this choice of nomenclature has given rise to some confusion. To avoid such confusion, we will stick to standard usage in this book and refer to the degree of a vertex.

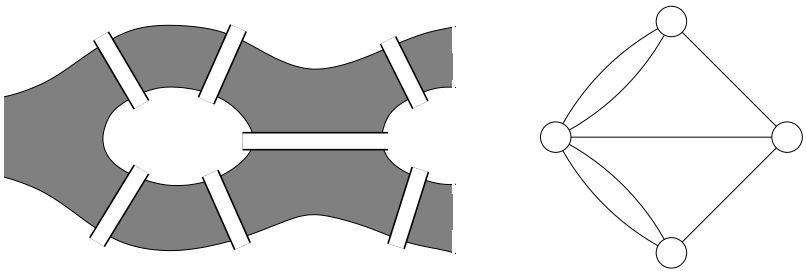


FIGURE 1.2 Left: a simplified depiction of the pattern of the rivers and bridges in the Königsberg bridge problem. Right: the corresponding network of vertices and edges.

and their connections can be almost anything—people and friendships (Rapoport and Horvath 1961), computers and communication lines **(Faloutsos *et al.* 1999)**, chemicals and reactions **(Jeong *et al.* 2000; Wagner and Fell 2001)**, scientific papers and citations (**Price 1965**; Redner 1998)[3]—causing some to wonder how so broad a definition could generate anything of substantive interest. But its breadth is precisely why graph theory is so powerful. By abstracting away the details of a problem, graph theory is capable of describing the important topological features with a clarity that would be impossible were all the details retained. As a consequence, graph theory has spread well beyond its original domain of pure mathematics, especially in the past few decades, to applications in engineering (Ahuja *et al.* 1993), operations research (Nagurney 1993), and computer science (Lynch 1996). Nowhere, however, has graph theory found a more welcome home than in sociology.

Starting in the 1950s, in response to a growing interest in quantitative methods in sociology and anthropology, the mathematical language of graph theory was coopted by social scientists to help understand data from ethnographic studies (Wasserman and Faust 1994; Degenne and Forsé 1999; Scott 2000). Much of the terminology of social network analysis—actor centrality, path lengths, cliques, connected components, and so forth—was either borrowed directly from graph theory or else adapted from it, to address questions of status, influence, cohesiveness, social roles, and identities in social networks. Thus, in addition to its role as a language for describing abstract models, graph theory became a practical tool for the analysis of empirical data. Also starting in the 1950s, mathematicians began to think of graphs as the medium through which various modes of influence—information and disease in particular—could propagate (**Solomonoff and Rapoport 1951**; **Erdős and Rényi 1960**). Thus the structural properties of networks, especially their connectedness, became linked with behavioral characteristics like the expected size of an epidemic or the possibility of global information transmission. Associated

[3]Throughout this book citations highlighted in **bold text** refer to papers that are reproduced within this book.

with this trend was the notion that graphs are properly regarded as stochastic objects (**Erdős and Rényi 1960**; Rapoport 1963), rather than purely deterministic ones, and therefore that graph properties can be thought of in terms of probability distributions—an approach that has been developed a great deal in recent years.

1.2 THE "NEW" SCIENCE OF NETWORKS

So what is there to add? If graph theory is such a powerful and general language and if so much beautiful and elegant work has already been done, what room is there for a new science of networks? We argue that the science of networks that has been taking shape over the last few years is distinguished from preceding work on networks in three important ways: (1) by focusing on the properties of real-world networks, it is concerned with empirical as well as theoretical questions; (2) it frequently takes the view that networks are not static, but evolve in time according to various dynamical rules; and (3) it aims, ultimately at least, to understand networks not just as topological objects, but also as the framework upon which distributed dynamical systems are built. As we will see in Chapter 3, elements of all these themes predate the recent explosion of interest in networks, but their synthesis into a coherent research agenda is new.

Modeling real-world networks

The first difference between the old science of networks and the new is that, social network analysis aside, traditional theories of networks have not been much concerned with the structure of naturally occurring networks. Much of graph theory qualifies as pure mathematics, and as such is concerned principally with the combinatorial properties of artificial constructs. Pure graph theory is elegant and deep, but it is not especially relevant to networks arising in the real world. Applied graph theory, as its name suggests, is more concerned with real-world network problems, but its approach is oriented toward design and engineering. By contrast, the recent work that is the topic of this book is focused on networks as they arise naturally, evolving in a manner that is typically unplanned and decentralized. Social networks and biological networks are naturally occurring networks of this kind, as are networks of information like citation networks and the World Wide Web. But the category is even broader, including networks—like transportation networks, power grids, and the physical Internet—-that are intended to serve a single, coordinated purpose (transportation, power delivery, communications), but which are built over long periods of time by many independent agents and authorities. Social network analysis, for its part, is strongly empirical, but tends to be descriptive rather than constructive in nature. With the possible exception of certain types of random graph models (Holland and Leinhardt 1981; Strauss 1986; Anderson *et al.* 1999), network analysis in the social sciences has largely avoided modeling, preferring simply to describe the properties of networks as observed in collected data.

In contrast to traditional graph theory on the one hand, and social network analysis on the other, the work described in this book takes a view that is both theoretical and empirical. In order to develop new graph-theoretic models that can account for the structural features of real-world networks, we must first be able to say what those features are and hence empirical data are essential. But adequate theoretical models are equally essential if the significance of any particular empirical finding is to be correctly understood. Just as in traditional science, where theory and experiment continually stimulate one another, the science of networks is being built on the twin foundations of empirical observation and modeling.

That such an obvious requirement for scientific validity should have made its first appearance in the field so recently seems surprising at first, but is understandable given the historical difficulty of obtaining high quality, large-scale network data. For most of the past fifty years, the collection of network data has been confined to the field of social network analysis, in which data have to be collected through survey instruments that not only are onerous to administer, but also suffer from the inaccurate or subjective responses of subjects. People, it turns out, are not good at remembering who their friends are, and the definition of a "friend" is often quite ambiguous in the first place.

For example, the General Social Survey[4] requests respondents to name up to six individuals with whom they discuss "important matters." The assumption is that people discuss matters that are important to them with people who are important to them, and hence that questions of this kind—so-called "name generators"—are a reliable means of identifying strong social ties. However, a recent study by Bearman and Parigi (2004) shows that when people are asked about the so-called "important matters" they are discussing, they respond with just about every topic imaginable, including many that most of us wouldn't consider important at all. Even worse, some topics are discussed with family members, some with close friends, some with coworkers, and others with complete strangers. Thus, very little can be inferred about the network ties of respondents simply by looking at the names generated by the questions in the General Social Survey. Bearman and Parigi also find that some 20% of respondents name no one at all. One might assume that these individuals are "social isolates"—people with no one to talk to—yet nearly 40% of these isolates are married! It is possible that these findings reveal significant patterns of behavior in contemporary social life—perhaps many people, even married people, really do not have anyone to talk to, or anything important to talk about. But apparently the respondent data are so contaminated by diverse interpretations of the survey instrument, along with variable recollection and even laziness, that any inferences about the corresponding social network must be regarded with skepticism.

The example of the General Social Survey is instructive because it typifies the uncertainties associated with traditional, survey-based collection of network data. If people have difficulty identifying even their closest confidants, how can one expect to extract reliable information concerning more subtle relations? And if, in response to this obstacle, survey instruments become more elaborate and spe-

[4] See `http://www.norc.uchicago.edu/projects/gensoc.asp`.

cific, then as the size of the surveyed population increases, the work required of the researcher to analyze and understand the resulting volume of raw data becomes prohibitive. A better approach would be to record the activities and interactions of subjects directly, thus avoiding recall problems and allowing us to apply consistent criteria to define relationships. In the absence of accurate recording technologies, however, such direct observation methods are even more onerous than the administration of surveys.

Because of the effort involved in compiling them, social network datasets rarely document populations of more than a hundred people and almost never more than a thousand. And although other kinds of (nonsocial) networks have not suffered from the same difficulties, empirical examples prior to the last decade have been few—probably because other network-oriented disciplines have lacked the empirical focus of sociology. The lack of high quality, large-scale network data has, in turn, delayed the development of the kind of statistical models with which much of the work in this book is concerned. Such models, as we will see, can be very successful and informative when applied to large networks, but tend to break down, or simply don't address the right questions, when applied to small ones. As an example, networks of contacts between terrorists have been studied recently by, for instance, Krebs (2002), but they are poor candidates for statistical modeling because the questions of interest in these networks are not statistical in nature, focusing more on the roles of individuals and small groups within the network as a whole. The traditional tools of social network analysis—centrality indices, structural measures, and measures of social capital—are more useful in such cases.

Recent years, however, have witnessed a dramatic increase in the availability of network datasets that comprise many thousands and sometimes even millions of vertices—a consequence of the widespread availability of electronic databases and, even more important, the Internet. Not only has the Internet focused popular and scientific attention alike on the topic of networks and networked systems, but it has led to data collection methods for social and other networks that avoid many of the difficulties of traditional sociometry. Networks of scientific collaborations, for example, can now be recorded in real time through electronic databases like Medline and the Science Citation Index (**Newman 2001a**; Barabási *et al.* 2002), and even more promising sources of network data, such as email logs (Ebel *et al.* 2002; Guimerà *et al.* 2003; Tyler *et al.* 2003) and instant messaging services (Smith 2002; Holme *et al.* 2004), await further exploration. Being far larger than the datasets of traditional social network analysis, these networks are more amenable to the kinds of statistical techniques with which physicists and mathematicians are familiar. As the papers in Chapter 3 of this volume demonstrate, real networks, from citation networks and the World Wide Web to networks of biochemical reactions, display properties—like local clustering and skewed degree distributions—that were not anticipated by the idealized models of graph theory, and that have forced the development of new modeling approaches, some of which are introduced in Chapter 4.

Networks as evolving structures

A second distinguishing feature of the work described in this book is that, whereas in the past both graph theory and social network analysis have tended to treat networks as static structures, recent work has recognized that networks evolve over time (**Barabási and Albert 1999**; Watts 1999). Many networks are the product of dynamical processes that add or remove vertices or edges. For instance, a social network of friendships changes as individuals make and break ties with others. An individual with many acquaintances might, by virtue of being better connected or better known, be more likely to make new friends than someone else who is less well connected. Or individuals seeking friends might be more likely to meet people with whom they share a common acquaintance. The ties people make affect the form of the network, and the form of the network affects the ties people make. Social network structure therefore evolves in a historically dependent manner, in which the role of the participants and the patterns of behavior they follow cannot be ignored.

Similar statements apply to other kinds of networks as well: processes operating at the local level both constrain and are constrained by the network structure. A principal objective of the new science of networks (as dealt with by a number of papers in Chapter 4), is an understanding of how structure at the global scale (say, the connectivity of the network as a whole) depends on dynamical processes that operate at the local scale (for example, rules governing the appearance and connections of new vertices).

Networks as dynamical systems

The final feature that distinguishes the research described in this book from previous work is that traditional approaches to networks have tended to overlook or oversimplify the relationship between the structural properties of a networked system and its behavior. A lot of the recent work on networks, by contrast, takes a dynamical systems view according to which the vertices of a graph represent discrete dynamical entities, with their own rules of behavior, and the edges represent couplings between the entities. Thus a network of interacting individuals, or a computer network in which a virus is spreading, not only has topological properties, but has dynamical properties as well. Interacting individuals, for instance, might affect one another's opinions in reaching some collective decision (voting in a general election, for example), while an outbreak of a computer virus may or may not become an epidemic depending on the patterns of connections between machines. Which outcomes occur, how frequently they occur, and with what consequences, are all questions that can only be resolved by thinking jointly about structure and dynamics, and the relationship between the two.

Questions of this nature are not easily tackled, however; dynamical problems lie at the forefront of network research, where there are many unanswered questions. One class of problems on which some progress has been made, and

which is addressed in Section 5.1, is that of contagion dynamics. Whether we are interested in the spread of a disease or the diffusion of a technological innovation, it is frequently the case that contagion occurs over a network. Not only physical but also social contacts can significantly influence the probability that a particular disease or piece of information will be transmitted, and also what effect it will have. In traditional mathematical epidemiology, as well as research on the diffusion of information, it is usually assumed that all members of the population have equal likelihood of interaction with all others. Clearly this assumption requires modification once we take network structure into account. As the papers in Section 5.1 demonstrate, the particular structure of the network through which a contagious agent is transmitted can have a dramatic impact on outcomes at the level of entire populations.

1.3 OVERVIEW OF THE VOLUME

Mirroring the themes introduced above, this volume is divided into a number of parts, each of which is preceded by an introduction that outlines the general theme and summarizes the contributions of the papers in that part. Chapter 2 sets the stage by presenting a selection of papers that we feel are important historical antecedents to contemporary research. Although recent work on networks takes a distinctly different approach from traditional network studies, a careful reading of Chapter 2 reveals that many of the basic themes were anticipated by mathematicians and social scientists years or even decades earlier. Given their age, some of these contributions seem remarkably familiar and modern, occasionally to the extent that recent papers almost exactly replicate previous results. Power-law distributions, random networks with local clustering, the notion of long-range shortcuts, and the small-world phenomenon were all explored and analyzed well before the new science of networks reconstituted the same ideas in the language of mathematical physics.

Chapter 3 emphasizes the empirical side of the new science of networks, and Chapter 4 presents some of the foundational modeling ideas that have generated a great deal of subsequent interest and activity. By exploring some tentative applications of the ideas introduced in Chapters 3 and 4, Chapter 5 takes the reader to the cutting edge of network science, the relationship between network structure and system dynamics. From disease spreading and network robustness to search algorithms, Chapter 5 is a potpourri of topics at this poorly understood but rapidly expanding frontier. Finally, Chapter 6 provides a short discussion of what we see as some of the most interesting directions for future research. We hope the reader will be encouraged to strike out from where the papers in this volume leave off, adding his or her own ideas and results to this exciting and fast-developing field.

Chapter Two
Historical developments

The study of networks has had a long history in mathematics and the sciences, stretching back at least as far as Leonhard Euler's 1736 solution of the Königsberg Bridge Problem discussed in Chapter 1. In this chapter we present a selection of historical publications on the subject of networks of various kinds. Of particular interest to us are papers from mathematical graph theory and from the literature on social networks. For example, the classic model of a network that we know of as the *random graph*, and which is discussed in greater detail in Section 4.1, was first described by the Russian mathematician and biologist Anatol Rapoport in the early 1950s, before being rediscovered and analyzed extensively by Paul Erdős and Alfréd Rényi in a series of papers in the late 1950s and early 1960s. Around the same time, a social scientist and a mathematician, Ithiel de Sola Pool and Manfred Kochen, in collaboration gave a beautiful and influential discussion of the "small-world effect" in an early preprint on social networks.

Thus, while much of this book is devoted to recent work on networks in the physics and applied mathematics communities, many of the crucial ideas that have motivated that work were well known, at least to some, many decades earlier. The articles reproduced in this chapter provide an overview of some of the original work on these topics and set the scene for the material that appears in the following chapters.

Karinthy (1929)

The first publication reproduced in this chapter is in fact not a scientific paper at all, but a translation of a short story, a work of fiction, originally published in Hungarian in 1929. Certainly this is an unusual way to start a volume of scientific reprints, but, as the reader will see, this brief story, published more than seventy years ago, describes beautifully one of the fundamental truths about network structure that has driven scientific research in the field for the last few decades, the concept known today as the "small-world effect," or "six degrees of separation."

The writer Frigyes Karinthy (1887–1938) became an overnight sensation in Hungary following the publication in 1912 of his first book, a volume of literary caricature, which is required reading in Hungarian schools even today. Karinthy's 1929 volume of short stories, entitled *Everything is Different*, did not receive the same warm welcome from the literary establishment. Friends and critics alike believed the book to be little more than a scheme for making some quick money

by stringing together a set of short pieces with scant respect for coherence or flow. Among the pieces in this collection, however, is one gem of a story, entitled "Chains," in which the writer raised in a fictional context some of the questions that network theory would be struggling with for much of the rest of the century. Without any pretensions to scientific rigor or proof, Karinthy tackled and suggested answers to one of the deep problems in the theory of networks.

In "Chains" Karinthy argues, as Jules Verne did fifty years earlier, that the world is getting smaller. Unlike Verne in "Around the World in Eighty Days," however, Karinthy proposes to demonstrate his thesis not by physical means—circumnavigating the globe—but by a *social* argument. He claims that people are increasingly connected to each other via their acquaintances, and that the dense web of friendship surrounding each person leads to an interconnected world in which everyone on Earth is at most *five acquaintances away from anyone else.*

To back up this remarkable claim, Karinthy demonstrates that it is possible to connect a Nobel prize winner to himself via a chain of just five acquaintances. He also points out, however, that this may not be an entirely fair example, because famous people with many social connections can be more easily connected to others, an insight whose relevance in the study of networks has only been fully appreciated quite recently. To show that his "five degrees of separation" claim also applies to less prominent people than Nobel laureates, he connects a worker in a Ford factory to himself, again via five acquaintances. Finally, he argues that the changing nature of human acquaintance patterns is a consequence of human exploration, of the demolition of geographical boundaries, and of new technologies that allow us to stay in touch even when we are thousands of miles apart.

The idea of chains of acquaintances linking distant individuals has been revisited many times in the decades since Karinthy's story. Jane Jacobs in her influential 1961 book *The Death and Life of Great American Cities* recalls:

> When my sister and I first came to New York from a small city, we used to amuse ourselves with a game we called Messages. The idea was to pick two wildly dissimilar individuals—say a head hunter in the Solomon Islands and a cobbler in Rock Island, Illinois—and assume that one had to get a message to the other by word of mouth; then we would each silently figure out a plausible, or at least possible, chain of persons through which the message could go. The one who could make the shortest plausible chain of messengers won. The head hunter would speak to the head man of his village, who would speak to the trader who came to buy cobra, who would speak to the Australian patrol officer when he came through, who would tell the man who was next slated to go to Melbourne on leave, etc. Down at the other end the cobbler would hear from his priest, who got it from the mayor, who got it from a state senator, who got it from the governor, etc. We soon had these close-to-home messengers down to a routine for almost everybody we could conjure up.
>
> — Jacobs (1961), pp. 134–135

Jacobs settles on an unusually long chain, however; the path in her example is at least nine links long, not counting the links that are presumably missing between the man heading to Melbourne and the governor.

Solomonoff and Rapoport (1951)

Scientific interest in the structure of networks began to develop in earnest in the 1940s and 1950s. Perhaps the most profound thinker in the field during this period was Anatol Rapoport, a Russian immigrant to the United States who worked not in sociology but in mathematical biology. Trained first as a pianist in Vienna, Rapoport turned to mathematics after realizing that a successful career as a concert performer would require the support of a wealthy patron, which he didn't have (Spencer 2002). He was unusual in developing an interest for mathematical biology at a time when mathematicians and biologists hardly spoke to each other, and he developed startling and prescient views about many topics that fall into the area we now call complex systems. In particular, he was decades ahead of his time in his views on the properties and importance of networks, developing methods that concentrated, as we often do today, on general statistical properties of networks, rather than individual properties of network nodes or edges. In a 1961 paper with William H. Horvath, he wrote,

> The theoretician's interest, however, is seldom focused on a particular large sociogram [i.e., network]. Rather, the interesting features of large sociograms are revealed in their gross, typical properties. Thus one seeks to define classes of sociograms, or else describe them by a few well-chosen parameters. It is perhaps natural to consider statistical parameters, since one is interested in trends or averages, or distributions rather than particulars.
>
> — Rapoport and Horvath (1961)

This remarkable statement could easily serve as a manifesto for the revolution in the study of networks that has recently taken place, four decades later, in the physics and mathematics communities.

In this chapter, we reproduce the important 1951 paper by Rapoport and Ray Solomonoff which presents the first systematic study of what we would now call a *random graph*. The paper is important both because it introduces the random graph for the first time and because it demonstrates one of the most crucial properties of the model: as the ratio of the number of edges to vertices in the graph is increased, the network reaches a point at which it undergoes an abrupt change from a collection of disconnected vertices to a connected state in which, in modern parlance, the graph contains a *giant component*.

The paper starts by considering a graph composed of a collection of vertices randomly connected to one another by edges (or axons, to use the paper's neurologically inspired terminology). The authors discuss three natural systems in which such networks might appear: neural networks, the social networks of physical contacts that are responsible for the spread of epidemic disease, and a network problem rooted in genetics.

Solomonoff and Rapoport define a quantity called the *weak connectivity*, which is the expected number of vertices reachable through the network from a randomly chosen vertex. In the modern terminology of networks, the weak connectivity is the average component size in the network. Solomonoff and Rapoport then derive an iteration relation for the weak connectivity by reasoning about the behavior of a simple component-finding algorithm which is equivalent to what we would today call a burning algorithm or breadth-first search. This result leads them to conclude that the average component size depends crucially on the mean degree a, where the degree again is the number of edges connected to a vertex. They show that for $a < 1$ the network is broken into many small isolated islands, but that when the mean degree exceeds $a = 1$ a giant component forms that contains a finite fraction of all the vertices in the network. Thus, although they did not use this language, Solomonoff and Rapoport predicted[1] in 1951 the existence of a phase transition from a fragmented network for $a < 1$ to one dominated by a giant component for $a > 1$.

Erdős and Rényi (1960)

Despite the early contributions of Solomonoff and Rapoport, random graph theory did not really take off until the late 1950s and early 1960s, when several important papers on the subject appeared almost simultaneously (Ford and Uhlenbeck 1957; Erdős and Rényi 1959, 1960; Gilbert 1959). Among these, the most influential, and the most relevant to current work, were the papers by Paul Erdős and Alfréd Rényi, who are considered the fathers of the modern theory of random graphs. Between 1959 and 1968 Erdős and Rényi published eight papers on random graphs that set the tone for network research for many decades to come. The next paper reproduced in this section **(Erdős and Rényi 1960)** is probably the most important of these. It deals with the evolution of the structure of random graphs as the mean degree is increased.

In this paper, the authors showed that many properties of random graphs emerge not gradually but suddenly, when enough edges are added to the graph. They made use of the following definition: if the probability of a graph having property Q approaches 1 as the size of the graph $N \to \infty$, then we say that *almost every* graph of N vertices has the property Q. They studied the behavior of a variety of different properties as a function of the probability p of the existence of an edge between any two vertices, and showed that for many properties there is a critical probability $p_c(N)$ such that if $p(N)$ grows more slowly than $p_c(N)$ as $N \to \infty$ then almost every graph with connection probability $p(N)$ fails to have the property Q. Conversely, if $p(N)$ grows faster than $p_c(N)$ then almost every graph has the property Q. Thus the probability that a graph with N nodes and

[1] This result, and indeed the invention of the random graph itself, is usually attributed to Erdős and Rényi (1959), but it is clear that Solomonoff and Rapoport had many of the crucial results almost a decade earlier. Erdős and Rényi appear not to have been aware of Solomonoff and Rapoport's work, and rediscovered their results independently. Erdős and Rényi's work also went much farther than that of Solomonoff and Rapoport and maintained a substantially higher level of rigor.

exponent	$-\infty$	-2	$-\frac{3}{2}$	$-\frac{4}{3}$	-1	$-\frac{2}{3}$	$-\frac{1}{2}$
subgraph							

TABLE 2.1 The threshold probabilities at which different subgraphs appear in a random graph. For $pN^{3/2} \to 0$ the graph consists only of isolated nodes or pairs connected by edges. When $p \sim N^{-3/2}$ trees with 3 edges appear and at $p \sim N^{-4/3}$ trees with 4. If $p \sim N^{-1}$ trees of all sizes are present, as well as cycles of all lengths. When $p \sim N^{-2/3}$ the graph contains complete subgraphs of 4 vertices and for $p \sim N^{-1/2}$ there are complete subgraphs of 5 vertices. As the exponent approaches 0, the graph contains complete subgraphs of increasing order.

connection probability $p = p(N)$ has property Q satisfies

$$\lim_{N\to\infty} P_{N,p}(Q) = \begin{cases} 0 & \text{if } p(N)/p_c(N) \to 0, \\ 1 & \text{if } p(N)/p_c(N) \to \infty. \end{cases} \tag{2.1}$$

As an example, let us consider one of the first cases discussed by Erdős and Rényi, namely the appearance of a given subgraph within a random graph. For low values of the edge probability p, the graph is very sparse and the likelihood of finding, for example, a single vertex connected to two others is very low. One might imagine that in general the probability of there being such a structure somewhere on the graph would increase slowly with increasing p, but Erdős and Rényi prove that this is not the case. Instead, the probability of finding a connected trio of vertices is negligible if $p < cN^{-1/2}$ for some constant c, but tends to 1 as N becomes large if $p > cN^{-1/2}$. In other words, almost all graphs contain a connected trio of vertices if the number of links is greater than a constant times $N^{1/2}$, but almost none of them do if the number of links is less than this.

Erdős and Rényi generalized this result further to show that the probability of occurrence of a tree of k vertices in the graph (i.e., a connected set of k vertices containing no loops) tends to 1 on large graphs with more than a constant times $N^{(k-2)/(k-1)}$ edges. They also extended their method to cycles, that is, closed loops of vertices in which every two consecutive vertices, and only these, are joined by an edge. Cycles also show a threshold behavior—they are present with probability 1 above some critical value of p as the graph becomes large. In Table 2.1 we summarize some of the thresholds found in the evolution of random graphs.

An area of study closely related to random graphs is *percolation theory* (Stauffer and Aharony 1992; Bunde and Havlin 1994, 1996), which has been the object of attention within the physics community for many years, since the introduction in the 1950s of the original percolation model by Hammersley and others (Broadbent and Hammersley 1957; Hammersley 1957). In bond percolation models, one studies the properties of the system in which the bonds on a lattice or network are either occupied or not with some occupation probability p, asking

questions such as what the mean sizes are of the clusters of lattice sites connected together by occupied bonds, and whether or not there exists a "spanning cluster" in the limit of large system size (i.e., a cluster that connects opposite sides of the lattice via a path of occupied bonds). It is clear that the random graph model is equivalent to bond percolation on a complete graph (i.e., a graph in which every vertex is connected to every other), and hence the methods developed for studying percolation can be applied to random graphs also. In particular, there has been much effort devoted to the study of the behavior of percolation models close to the phase transition at which a spanning cluster forms, which in random graph language is the point at which a *giant component* appears. It is known, for example, that many properties display *universal* behavior close to the phase transition, behavior that is dependent on the system dimension but not on the details of the lattice. And even the dimension dependence vanishes above the so-called *upper critical dimension*, giving way to generic behavior that can be extracted using simple mean-field theories. Since a complete graph is a formally infinite-dimensional object in the limit of large system size, the behavior of the random graph near the phase transition therefore falls into this *mean-field universality class*, and many results for the random graph, such as values for critical exponents can then be extracted from mean-field theory.

Pool and Kochen (1978)

In the late 1950s, around the same time that Erdős and Rényi were beginning their work on random graphs, the sociological community started developing an interest in applications of graph theory. The next paper reproduced in this section is the influential article on patterns of social contacts by the political scientist Ithiel de Sola Pool and the mathematician Manfred Kochen **(Pool and Kochen 1978)**. This paper was actually written in 1958, and circulated for many years in preprint form. In it Pool and Kochen addressed for the first time many of the questions that the field would be struggling with for the next few decades, and yet they felt that they had not dealt satisfactorily with the issues and so didn't submit their work for publication in a journal. It was not until twenty years after its first appearance that the authors consented to the publication of this important work in the new journal *Social Networks*, on page 1 of volume 1. Pool and Kochen's work provided the inspiration for, among other things, the famous "small-world" experiments conducted in the 1960s by Stanley Milgram (Milgram 1967; **Travers and Milgram 1969**), which are the subject of the following paper in this book. Hence it is appropriate that we here reproduce Pool and Kochen's work ahead of Milgram's, even though Milgram's bears the earlier date of publication.

In the introduction to their paper, Pool and Kochen formulate some of the questions that have come to define the field of social networks:

i) How many other people does each individual in a network know? In other words what is the person's degree in the network? (Pool and Kochen refer to this quantity as the "acquaintance volume.")

ii) What is the distribution of the degrees? What are their mean and their largest and smallest values?

iii) What kinds of people have large numbers of contacts? Are these the most influential people in the network?

iv) How exactly are the contacts organized? What is the structure of the network?

In addition to these general questions about individuals and about the network as a whole, Pool and Kochen looked also at questions about paths between pairs of individuals:

i) What is the probability that two people chosen at random from the population will know each other?

ii) What is the chance that they have a friend in common?

iii) What is the chance that the shortest chain between them requires two intermediates? Or more than two?

Pool and Kochen start by discussing the difficulty of determining the number of social contacts people have. There are two primary problems: ambiguity about what exactly constitutes a social contact, and the fact that people are not very good at estimating the number of their acquaintances even if the definition of an acquaintance is clear. Typically most people underestimate their acquaintance volume.

Given the limited and unreliable nature of network data, Pool and Kochen resort to mathematical models. Inspired by Rapoport's work **(Solomonoff and Rapoport 1951)**, they base their work on the random graph, using this simple model to make conjectures about the characteristics of social networks.

This paper discusses for the first time in scientific terms the phenomenon we now call the small-world effect. Starting with the assumption that each person has about 1,000 acquaintances, they predict that most pairs of people on Earth can be connected via a path that goes through just two intermediate acquaintances. They give arguments reminiscent of those found in Karinthy's 1929 short story, reproduced in this chapter, to make this counterintuitive claim more plausible. They also consider the possibility that community groupings and social stratification within the network would affect their conclusions. But, after some laborious calculation, they conclude, apparently somewhat to their own surprise, that social strata have only a small effect on the average distance between individuals.

Travers and Milgram (1969)

Although network ideas were already becoming popular among sociologists in the 1950s and 1960s, it was an experimentalist, Stanley Milgram, who propelled the field into the public consciousness in the late 1960s with his famous small-world experiment. Milgram, at that time working at Harvard and influenced by the thinking of Harrison White and Ithiel Pool, both also in the Boston area, was inspired

to devise an experiment that could test Pool and Kochen's surprising conjectures about path lengths between individuals in social networks.

Milgram published several papers about his small-world experiments. The earliest and best known is a 1967 piece that he wrote for the popular newsstand magazine *Psychology Today* (Milgram 1967). Although entertaining and thought provoking, this is not a rigorous piece of scientific writing, and many of the details of his work are left out of the discussion. After the first set of experiments, Milgram started collaborating with Jeffrey Travers, and repeated the experiments with new subjects and more detailed quantitative analyses. The 1969 article Milgram coauthored with Travers contains a clear and thorough explanation of these new experiments, and it is this second paper that we reproduce here.

Milgram's experiments started by selecting a target individual and a group of starting individuals. A package was mailed to each of the starters containing a small booklet or "passport" in which participants were asked to record some information about themselves. Then the participants were to try and get their passport to the specified target person by passing it on to someone they knew on a first-name basis who they believed either would know the target, or might know somebody who did. These acquaintances were then asked to do the same, repeating the process until, with luck, the passport reached the designated target. At each step participants were also asked to send a postcard to Travers and Milgram, allowing the researchers to reconstruct the path taken by the passport, should it get lost before it reached the target. Travers and Milgram recruited 296 starting individuals, 196 from Omaha, Nebraska and the other 100 from Boston. The target was a stockbroker who lived in Sharon, Massachusetts, a small town outside Boston.

In the end, 64 of the 296 chains reached the target, 29% of those that started out. The number of intermediate acquaintances between source and target varied from 1 to 11, the median being 5.2. Five intermediate acquaintances means that there were six steps along the chain, a result that has passed into popular myth in the phrase "six degrees of separation," which was the title of a 1990 Broadway play by John Guare in which one of the characters discusses the small-world effect.

To what degree can we trust the results of Milgram's experiments? Are we indeed just six steps from anyone else on average, or could the real result be closer to three as predicted by Pool and Kochen? Or perhaps the average separation is larger than six? This question is discussed in some detail by Travers and Milgram. The letters were more likely to get lost if they took a longer path from source to target, and hence the completed chains that Travers and Milgram used to estimate the average chain length are probably biased toward the shorter lengths. As Travers and Milgram describe in a footnote, however, White (1970) calculated a correction to the raw results to allow for this effect and found that the change in the figures was not large: the correction increases the average separation from 6 to 8. But there are other effects acting in the opposite direction also, potentially making the mean separation shorter than six. In particular, there is no guarantee that Travers and Milgram's subjects would have found the shortest path through the network to the target person. They forwarded the letter to the person of their acquaintance

who they *thought* was closest to the target person, but they could easily have had another acquaintance who—unknown to them—was acquainted directly with the target. Thus the real separation between participants could be much shorter than that recorded by the experiment.

Price (1965)

At about the same time that Milgram was developing his first small-world experiment, the empirical study of networks was being taken up in another very different branch of the scientific community, information science. Derek de Solla Price's 1965 article "Networks of Scientific Papers," which appeared in the journal *Science*, is a hidden treasure, largely unknown within the mathematics and physics communities. In this paper, Price studies one of the oldest of information networks, the network of citations between scientific papers, in which each vertex represents a paper and a directed edge from one paper to another indicates that the first paper cites the second in its bibliography. Price indeed appears to have been one of the first to suggest that we view the pattern of citations as a network at all, and to present detailed statistical analyses of this network, for which he made use of the databases of citations that were just starting to become available, thanks to the work of Eugene Garfield and others.

Since citation networks are directed, each paper in such a network has both an out-degree (the numbers of papers that it cites) and an in-degree (the number of papers in which it is cited). Price studied the distributions of both in- and out-degrees and found that both have power-law tails, with exponents of about -2 and -3, respectively. Networks with power-law degree distributions are now known to occur in a number of different settings and are often called "scale-free networks" (see Chapter 3 and Section 4.3).

The quality of citation data has improved markedly in the years since Price's pioneering work, and particularly since the advent of computer tabulation of data, and a number of more recent studies have improved upon Price's results. Of particular interest is the paper by Redner (1998), in which the author independently discovered Price's power law using two large databases of citations of physics papers. Redner investigated the citation frequency of 783 339 papers published in 1981 and cited over 6 million times between 1981 and 1997, using data collected by the Institute for Scientific Information, the commercial enterprise that grew out of Garfield's early work on citation. The careful analysis presented in the paper shows that the in-degree of the citation network does indeed have a power-law tail, with an exponent roughly equal to -3. A second data set compiled from the bibliographies of 24 296 papers published in the journal *Physical Review D* between 1975 and 1994 shows similar results.

The paper reproduced here is not Derek Price's only contribution to the study of citation networks. A decade later he published a second remarkable paper in which he proposed a possible mechanism for the generation of the power

laws seen in the citation distribution (Price 1976). Building on previous work by Simon (1955), he proposed that papers that have many citations receive further citations in proportion to the number they already have. He called this process "cumulative advantage," and gave a mathematical model of it, which he solved to demonstrate that it does indeed give rise to power-law distributions as observed in the data. The cumulative advantage process is more commonly known today under the name "preferential attachment," and is widely accepted as the explanation for the occurrence of power-law degree distributions in networks as diverse as the World Wide Web, social networks, and biological networks.

De Castro and Grossman (1999)

Our final paper in this chapter deals with another network formed by the patterns of scientific publication, and while it is a relatively recent work, having been published in 1999, we feel it belongs here in this historical section, as it summarizes an idea that has been current in the mathematics community for some decades but has rarely been studied formally.

Paul Erdős was a stunningly prolific mathematician who lived from 1913 to 1996. During his long life, he authored over 1500 papers with more than 500 coauthors, including his papers on random graphs with fellow Hungarian Alfréd Rényi, which are discussed earlier in this section. His staggering output, together with his pivotal role in the development of the theory of networks, prompted some of his colleagues to see him as a central node of the worldwide collaboration network of mathematicians and other scientific researchers.

Consider the network whose vertices are mathematicians and scientists, with an edge between any two vertices if the researchers they represent have coauthored one or more papers together. For each vertex we define the *Erdős number* to be the length of the shortest path from that vertex to Paul Erdős along the edges of the network. As de Castro and Grossman describe it,

> Paul Erdős himself has Erdős number 0, and his co-authors have Erdős number 1. People not having Erdős number 0 or 1 who have published with someone with Erdős number 1 have Erdős number 2, and so on. Those who are not linked in this way to Paul Erdős have Erdős number ∞.

For many years now, it has been a popular cocktail-party pursuit among mathematicians to calculate their Erdős number, or more strictly an upper bound on their Erdős number, since it is rarely possible to be certain one has considered all possible paths through the network. Most mathematicians, and many in other subjects as well, have no difficulty establishing a fairly low upper bound on their Erdős number. As de Castro and Grossman conjecture, "Most mathematical researchers of the twentieth century have a finite (and rather small) Erdős number."

In the paper, de Castro and Grossman argue in favor of this conjecture by charting paths through the collaboration network to Erdős from a wide variety of starting individuals. As well as mathematicians, they derive upper bounds on

the Erdős numbers of Nobel Prize winners in physics, economics, biology, and chemistry. And since it seems likely that most scientists in those fields could be connected to the corresponding Nobel laureates in a small number of steps, it is reasonable to suppose that most scientists have small Erdős numbers.

This exercise of course constitutes another demonstration of the small-world effect, this time in the context of the scientific community. With the recent increase in interest in networks, the Erdős number has been elevated from mathematical anecdote to the subject of serious (if playful) scientific inquiry. **Newman** (**2001a**, 2001b, 2001c), for instance, has studied in detail collaboration networks from a variety of subjects, including networks of biologists, physicists, and computer scientists, while Barabási and coworkers (2002) have focused on understanding the time evolution of collaboration networks, using data for publications in mathematics and neuroscience.

CHAIN-LINKS

by
Frigyes Karinthy

We were arguing energetically about whether the world is actually evolving, headed in a particular direction, or whether the entire universe is just a returning rhythm's game, a renewal of eternity. "There has to be something of crucial importance," I said in the middle of debate. "I just don't quite know how to express it in a new way; I hate repeating myself."

Let me put it this way: Planet Earth has never been as *tiny* as it is now. It shrunk – relatively speaking of course – due to the quickening pulse of both physical and verbal communication. This topic has come up before, but we had never framed it quite this way. We never talked about the fact that anyone on Earth, at my or anyone's will, can now learn in just a few minutes what I think or do, and what I want or what I would like to do. If I wanted to convince myself of the above fact: in couple of days I could be — *Hocus pocus*! — where I want to be.

Now we live in fairyland. The only slightly disappointing thing about this land is that it is smaller than the real world has ever been.

Chesterton praised a tiny and intimate, small universe and found it obtuse to portray the Cosmos as something *very big*. I think this idea is peculiar to our age of transportation. While Chesterton rejected technology and evolution, he was finally forced to admit that the fairyland he dreamed of could only come about through the scientific revolution he so vehemently opposed.

Everything returns and renews itself. The difference now is that the *rate* of these returns has increased, in both space and time, in an unheard-of fashion. Now my thoughts can circle the globe in minutes. Entire passages of world history are played out in a couple of years.

Something must result from this chain of thoughts. If only I knew what! (I feel as if I knew the answer to all this, but I've forgotten what it was or was overcome with doubt. Maybe I was *too close* to the truth. Near the North Pole, they say, the needle of a compass goes haywire, turning around in circles. It seems as if the same thing happens to our beliefs when we get too close to God.)

A fascinating game grew out of this discussion. One of us suggested performing the following experiment to prove that the population of the Earth is closer together now than they have ever been before. We should select any person from the 1.5 billion inhabitants of the Earth – anyone, anywhere at all. He bet us that, using no more than *five* individuals, one of whom is a personal acquaintance, he could contact the selected individual using nothing except the network of personal acquaintances. For example, "Look, you know Mr. X.Y., please ask him to contact his friend Mr. Q.Z., whom he knows, and so forth."

"An interesting idea!" — someone said — "Let's give it a try. How would you contact Selma Lagerlöf?"[1]

---

[1] Swedish novelist Selma Lagerlöf (1858–1940), who received the Nobel Prize for literature in 1909, was a champion of the return of Swedish romanticism with a mystical overtone. She also wrote novels for children.

"Well now, Selma Lagerlöf," the proponent of the game replied, "Nothing could be easier." And he reeled off a solution in two seconds: "Selma Lagerlöf just won the Nobel Prize for Literature, so she's bound to know King Gustav of Sweden, since, by rule, he's the one who would have handed her the Prize. And it's well known that King Gustav loves to play tennis and participates in international tennis tournaments. He has played Mr. Kehrling,[2] so they must be acquainted. And as it happens I myself also know Mr. Kehrling quite well." (The proponent was himself a good tennis player.) "All we needed this time was two out of five links. That's not surprising since it's always easier to find someone who knows a famous or popular figure than some run-of-the-mill, insignificant person. Come on, give me a harder one to solve!"

I proposed a more difficult problem: to find a chain of contacts linking myself with an anonymous riveter at the Ford Motor Company — and I accomplished it in four steps. The worker knows his foreman, who knows Mr. Ford himself, who, in turn, is on good terms with the director general of the Hearst publishing empire. I had a close friend, Mr. Árpád Pásztor, who had recently struck up an acquaintance with the director of Hearst publishing. It would take but one word to my friend to send a cable to the general director of Hearst asking him to contact Ford who could in turn contact the foreman, who could then contact the riveter, who could then assemble a new automobile for me, should I need one.

And so the game went on. Our friend was absolutely correct: nobody from the group needed more than five links in the chain to reach, just by using the method of acquaintance, any inhabitant of our Planet.

---

[2] Béla Kehrling, (1891–1937) was a noted Hungarian sportsman, soccer, ping-pong and tennis player. In tennis, he emerged victorious in 1923 in Gothenberg, Sweden, both indoors and in the open; he placed third in the Wimbledon doubles. He also played soccer and ice hockey.

And this leads us to another question: Was there ever a time in human history when this would have been impossible? Julius Caesar, for instance, was a popular man, but if he had got it into his head to try and contact a priest from one of the Mayan or Aztec tribes that lived in the Americas at that time, he could not have succeeded — not in five steps, not even in three hundred. Europeans in those days knew less about America and its inhabitants than we now know about Mars and its inhabitants.

So something is going on here, a process of contraction and expansion which is beyond rhythms and waves. Something coalesces, shrinks in size, while something else flows outward and grows. How is it possible that all this expansion and material growth can have started with a tiny, glittering speck that flared up millions of years ago in the mass of nerves in a primitive human's head? And how is it possible that by now, this continuous growth has the inundating ability to reduce the entire physical world to ashes? Is it possible that power can conquer matter, that the soul makes a mightier truth than the body, that life has a meaning that survives life itself, that good survives evil as life survives death, that God, after all, is more powerful than the Devil?

I am embarrassed to admit — since it would look foolish — that I often catch myself playing our well-connected game not only with human beings, but with objects as well. I have become very good at it. It's a useless game, of course, but I think I'm addicted to it, like a gambler who, having lost all of his money, plays for dried beans without any hope of real gain — just to see the four colors of the cards. The strange mind-game that clatters in me all the time goes like this: how can I link, with three, four, or at most five links of the chain, trivial, everyday things of life. How can I link one phenomenon to another? How can I join the relative and the ephemeral with steady, permanent things — how can I tie up the part with the whole?

It would be nice to just live, have fun, and take notice *only* of the utility of things: how much pleasure or pain they cause me. Alas, it's not possible. I hope that this game will help me find

something else in the eyes that smile at me or the fist that strikes me, something beyond the urge to draw near to the former and shy away from the latter. One person loves me, another hates me. Why? Why the love and the hatred?

There are two people who do not understand one another, but I'm supposed to understand both. How? Someone is selling grapes in the street while my young son is crying in the other room. An acquaintance's wife has cheated on him while a crowd of hundred and fifty thousand watches the Dempsey match, Romain Roland's[3] last novel bombed while my friend Q changes his mind about Mr. Y. Ring-a-ring o' roses, a pocketful of posies. How can one possibly construct any chain of connections between these random things, without filling thirty volumes of philosophy, making only reasonable suppositions. The chain starts with the matter, *and its last link leads to me,* as the source of everything.

Well, just like this gentleman, who stepped up to my table in the café where I am now writing. He walked up to me and interrupted my thoughts with some trifling, insignificant problem and made me forget what I was going to say. Why did he come here and disturb me? The first link: he doesn't think much of people he finds scribbling. The second link: this world doesn't value scribbling nearly as much as it used to just a quarter of a century ago. The famous worldviews and thoughts that marked the end of the 19th century are to no avail today. Now we disdain the intellect. The third link: this disdain is the source of the hysteria of fear and terror that grips Europe today. And so to the fourth link: the order of the world has been destroyed.

Well, then, let a New World Order appear! Let the new Messiah of the world come! Let the God of the universe show himself once more through the burning rosehip-bush! Let there be peace, let there be war, let there be revolutions, so that — and here is

---

[3] Romain Roland, the noted French novelist, lived from 1866 until 1944. He was awarded the Nobel Prize for literature in 1915. Nearly all of his works were translated into Hungarian, just as in the case of Selma Lagerlöf.

the fifth link — it cannot happen again that someone should dare disturb me when I am at play, when I set free the phantoms of my imagination, when I think!

Translated from Hungarian and annotated by
Adam Makkai
Edited by ***Enikö Jankó***

BULLETIN OF
MATHEMATICAL BIOPHYSICS
VOLUME 13, 1951

CONNECTIVITY OF RANDOM NETS

RAY SOLOMONOFF AND ANATOL RAPOPORT
DEPARTMENT OF PHYSICS AND COMMITTEE ON MATHEMATICAL BIOLOGY
THE UNIVERSITY OF CHICAGO

The weak connectivity γ of a random net is defined and computed by an approximation method as a function of a, the axone density. It is shown that γ rises rapidly with a, attaining 0.8 of its asymptotic value (unity) for $a = 2$, where the number of neurons in the net is arbitrarily large. The significance of this parameter is interpreted also in terms of the maximum expected spread of an epidemic under certain conditions.

Numerous problems in various branches of mathematical biology lead to the consideration of certain structures which we shall call "random nets." Consider an aggregate of points, from each of which issues some number of outwardly directed lines (axones). Each axone terminates upon some point of the aggregate, and the probability that an axone from one point terminates on another point is the same for every pair of points in the aggregate. The resulting configuration constitutes a *random net.*

The existence of a *path* in a random net from a point A to a point B implies the possibility of tracing directed lines from A through any number of intermediate points, on which these lines terminate, to B.

We shall say that B is t axones removed from A, if t is the smallest number of axones contained in any of the paths from A to B. Point A itself is zero axones removed from A. All the other points upon which the axones of A terminate are one axone removed. The points upon which the axones from these latter points terminate, and which are not one or zero axones removed, are two axones removed, etc.

The notion of a random net may be generalized, if it is not assumed that the probability of direct connection between every pair of points in the net is the same. In that case it is necessary to define this probability for every pair of points. This can be done, for example, in terms of the distance between them or in some other way.

If the connections are not equiprobable, we shall speak of a net with a bias.

The following examples illustrate problems in which the concept of a net, defined by the probability of the connections among its points, seems useful.

1. *A problem in the theory of neural nets.* Suppose the points of a net are neurons. What is the probability that there exists a path between an arbitrary pair of neurons in the net? If the net has bias, what is the probability that there exists a path between a specified pair? In particular, what is the probability that a neuron is a member of a cycle (i.e., there exists a path from the neuron to itself through any positive number of internuncials)? Or, one may ask, what is the probability that there exists a path from a given neuron to every other neuron in the net?

2. *A problem in the theory of epidemics.* Suppose a number of individuals in a closed population contract a contagious disease, which lasts a finite time and then either kills them or makes them immune. If the probability of transmission is defined for each pair of individuals, what is the expected number of individuals which will contract the disease at a specified time? In particular, what is the

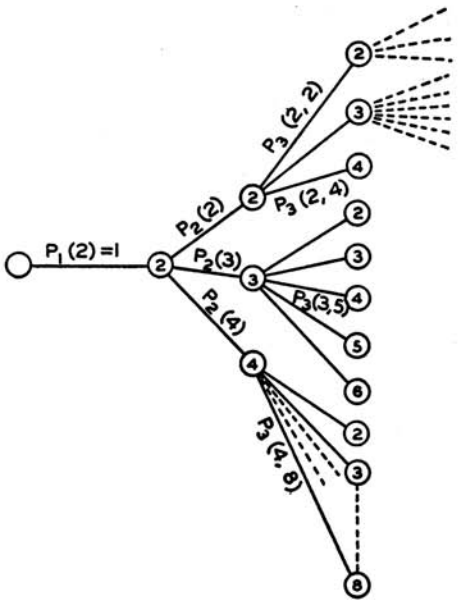


FIGURE 1. The probability tree for the number of ancestors of a single individual.

expected number of individuals which will eventually (after an infinite time) contract the disease? Or else, what is the probability that the entire population will succumb? Note that if the probability of transmission is the same for each pair of individuals, we are dealing with a random net.

3. *A problem in mathematical genetics.* Given the probability of mating between each pair of individuals in a population (as a function of their distance, or kinship, or the like), what is the expected number of ancestors of a given order for each individual? Clearly, the less the expected number of ancestors, the greater the genetic homogeneity of the population.

Each of these problems can be formalized by constructing a "probability tree." As an example, a tree for the genetic problem is illustrated in **Figure 1.**

We note that the tree consists of "nodes" connected by lines. The nodes can be designated by "first order," "second order," etc., depending on their distance from the "root." The number at the node indicates a possible number of ancestors of a given order. The lines connecting the nodes are labeled with the corresponding probabilities. Thus $p_1(2) = 1$, since it is certain that an individual has exactly two ancestors of the first order (parents). However, the parents may have been siblings or half-siblings. Therefore it is possible that the number of grandparents is 2, 3, or 4. The corresponding probabilities are $p_2(2)$, $p_2(3)$, and $p_2(4)$. The probability of having a certain number of great-grandparents depends on how many grandparents one has had. Consequently, those probabilities must be designated by $p_3(i,j)$ where $i = 2, \cdots 4$ and $j = 2, \cdots 8$. In general, the probability of having a certain number of ancestors of order k will depend on how many ancestors of each of the smaller orders one has had. If, however, we simplify the problem by supposing that the probability of having a certain number of ancestors of the kth order depends only on how many ancestors of the $(k-1)$th order one has, then the probability that an individual has exactly n ancestors of the mth order will be given by

$$P_m(n) = \sum_{r=2}^{2^m} \cdots \sum_{j=2}^{8} \sum_{i=2}^{4} p_2(2,i)\, p_3(i,j)\, p_4(j,k) \cdots p_m(r,n). \qquad (1)$$

The expected number of ancestors of the mth order will then be

$$E(m) = \sum_{n=2}^{2^m} nP(n). \qquad (2)$$

Clearly, a similar tree can be constructed for the neural net problem. Here the numbers at the nodes of the kth order would designate the possible number of neurons k axones removed from a given neuron. The p's would designate the corresponding transition probabilities from a certain number of neurons $(k-1)$ axones removed to a certain number k axones removed, etc. If N is the number of neurons in the aggregate, clearly, a neuron B is at most N axones removed from a neuron A, or else there exists no path from A to B. Hence $E(N)$ represents the expected number of neurons in the aggregate to which there exist paths from an arbitrary neuron, if the neurons are not in any way distinguished from each other. This expected number we shall call the *weak connectivity* of a random net and will designate it by γ.

The contagion problem could be formulated in similar terms. Here weak connectivity would represent the expected number of individuals which will contract the disease eventually. If we define Γ, the *strong connectivity* as the probability that from an arbitrary point in a random net there exist paths to every other point, then Γ will represent the probability that the entire population will succumb in the epidemic described above. In this case, the number of "axones" represents the number of individuals infected by a carrier before he recovers or dies.

The weak connectivity of a random net. We shall compute the weak connectivity of a neural net in terms of certain approximations whose justification will be given in subsequent papers. It will be assumed that:

1. The number of axones per neuron a is constant throughout the net. This constant (the axone density) need not be an integer, since it may equally well be taken as the average number of axones per neuron.

2. Connections are equiprobable, i.e., an axone synapses upon one or another neuron in the aggregate with equal probability.

A. Shimbel (1950) has formulated the problem in terms of the following differential-difference equation

$$dx/dt = [N - x(t)][x(t) - x(t-\tau)]. \tag{3}$$

Here $x(t)$ is a function related to the expected number of neurons t axones removed from an arbitrary neuron, and τ is related to the axone density. Then the problem of finding γ is equivalent to the

problem of finding $x(\infty)$. A somewhat generalized form of equation (3) is given also by M. Puma (1939). The solution of the equation is, however, not given.

An approximate expression for γ where N is large was derived by one of the authors (Rapoport, 1948) where the number of axones per neuron is exactly one. This case will be generalized here to a axones per neuron, which are supposed constant through out the aggregate.

The axone-tracing procedure. Let us start with an arbitrarily selected neuron A and consider the set of all neurons removed by not more than t axones from A. Let x be the expected number of these neurons. Then evidently $x = x(N, a, t)$ depends on the total number of neurons in the net, on the axone density, and on t. Moreover, the weak connectivity of the net can be expressed as

$$\gamma(N, a) = x(N, a, N)/N. \tag{4}$$

Since N and a are fixed, we shall refer to the expected number of points removed from A by not more than t axones by $x(t)$. Note that t is a positive integer.

We seek a recursion formula for $x(t)$ which will give us an approximate determination of that function. To give a rigorous treatment of the problem, one would need to deal with distribution functions instead of expected values. For example, $p(i,t)$, denoting the probability that there are *exactly* i neurons not more than t axones removed from A, would determine the distribution for t. Successive distributions (for $t + 1$, etc.) would then depend on previous *distributions*, instead of merely upon the first moments of these distributions (expected values). The "probability tree" method does take these relations into account. An "exact" approach to the problem will be given in a subsequent paper. Meanwhile, however, we shall develop an approximation method in which it will be assumed that the expected value $x(t)$ depends only upon previous expected values, and, of course, upon the parameters of the net.

The recursion formula. We now seek an expression for $x(t + 1) - x(t)$. This is evidently the expected number of neurons *exactly* $(t + 1)$ axones removed from A. We shall make use of the following formula, which may be readily verified. Let s marbles be placed independently and at random into N boxes. Then the expected number of boxes occupied by one or more marbles will be given by

$$N[1 - (1 - 1/N)^s]. \tag{5}$$

In our axone-tracing procedure there are $a[x(t) - x(t-1)]$ axones of the *newly* contacted neurons to be traced on each step. Then the total number of neurons contacted on the $(t+1)$th tracing will be, according to formula (5),

$$N[1-(1-1/N)^{a[x(t)-x(t-1)]}]. \tag{6}$$

But of these neurons the fraction $x(t)/N$ has already been contacted. Hence the expected number of newly contacted neurons will be given by

$$x(t+1)-x(t)=[N-x(t)][1-(1-1/N)^{a[x(t)-x(t-1)]}], \tag{7}$$

which is our desired recursion formula.

Determination of γ. Let us set

$$y(t)=N-x(t). \tag{8}$$

Then equation (7) may be written as

$$y(t+1)=y(t)(1-1/N)^{a[y(t-1)-y(t)]}, \tag{9}$$

or

$$y(t+1)(1-1/N)^{ay(t)}=y(t)(1-1/N)^{ay(t-1)}. \tag{10}$$

Hence

$$y(t+1)(1-1/N)^{ay(t)}=\text{constant}=K. \tag{11}$$

We proceed to evaluate K. We have

$$y(t+1)=K(1-1/N)^{-ay(t)}. \tag{12}$$

But $y(t)$ represents the expected number of uncontacted points in the tth step. Since before the tracing began one point constituted the set of contacted points, therefore we have

$$y(0)=N-1, \tag{13}$$

and using formula (5),

$$y(1)=(N-1)^{a+1}N^{-a}. \tag{14}$$

Letting $t=0$ in (12), we obtain

$$K=N^{-aN}(N-1)^{aN+1}. \tag{15}$$

Furthermore, since $y(1) \leqq y(0)$ and $(1-1/N)^{-a} > 1$, we have $y(2) \leqq y(1)$, etc., so that $y(t)$ is a non-increasing function of t (this is also intuitively evident from the definition of y). Since $y \geq 0$ for all t, $y(t)$ must approach a limit as t grows without bound. Hence

$$\lim_{t\to\infty} y(t+1) = \lim_{t\to\infty} y(t) = Y. \tag{16}$$

Note that $\gamma = x(N)$ may also be considered as $\lim_{t\to\infty} x(t)/N$. This is so since contacting no new neurons on any tracing implies that no new neurons will be contacted on any subsequent tracings. If we continue to carry out tracings "symbolically," it is evident that at some tracing not greater than the Nth no new neurons will be contacted, and all subsequent tracings will be "dummy" tracings.

Using equations (12) and (15), we see that Y satisfies the transcendental equation

$$Y = (N-1)(1-1/N)^{a(N-Y)}. \tag{17}$$

For large N, this can be approximated by

$$Y \sim N \operatorname{Exp}\{a(Y/N-1)\}. \tag{18}$$

Hence, for large N,

$$Y/N \sim \operatorname{Exp}\{a(Y/N-1)\}. \tag{19}$$

But $\gamma = x(\infty)/N = 1 - Y/N$. Substituting this value into (19), we obtain the transcendental equation which defines γ implicitly as a function of a, namely,

$$\gamma = 1 - e^{-a\gamma}. \tag{20}$$

We note that for $\gamma = 0$, every a is a solution of (20). If $\gamma \neq 0$, then equation (20) can be solved explicitly for a giving

$$a = \frac{-\log(1-\gamma)}{\gamma}. \tag{21}$$

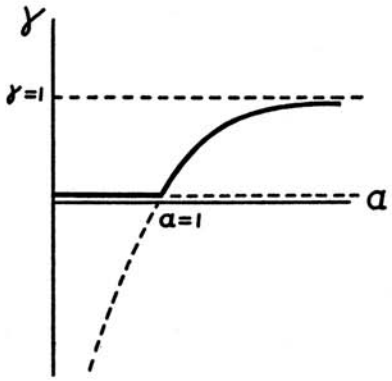


FIGURE 2. Weak connectivity as a function of axone density.

The right side of (21) is analytic in every neighborhood of the origin and tends to unity as γ approaches 0. Expanding that function in powers of γ, we have

$$a = 1 + \gamma/2 + \gamma^2/3 \cdots, \tag{22}$$

which allows us to plot a against γ (cf. Fig. 2). This graph consists of two branches, namely, the entire a-axis and the function (21). Negative values of γ, being physically meaningless, must be discarded. Thus in the region $0 \leqq a \leqq 1$, we have $\gamma \equiv 0$, as is intuitively evident. We must show, however, that for $a > 1$, γ follows the nonzero branch of the graph, otherwise we get the unlikely result that for sufficiently large N the fraction of individuals eventually infected in an epidemic will be negligible, regardless of the number of individuals infected by each carrier of the disease. Actually, the solution $\gamma \equiv 0$ is extraneous for $a > 1$ and appears in our equation because we have let N increase without bound *before* determining the relation between a and γ. In any physical situation N is finite. Hence a physically meaningful procedure is to determine γ as a function of a and N and *then* allow N to increase without bound. Such a function is given by equation (17). Proceeding from that equation we obtain

$$Y/(N-1) = (1-1/N)^{a(N-Y)}, \tag{23}$$

$$\log Y - \log(N-1) = a(N-Y)\log(1-1/N), \tag{24}$$

$$a = \frac{\log Y - \log(N-1)}{(N-Y)\,[\log(N-1) - \log(N)]}. \tag{25}$$

Let us write $Y = N - \phi(N) = N[1-\phi(N)/N]$. Then equation (25) may be written as

$$\begin{aligned} &\log N - \log(N-1) + \log[1-\phi(N)/N] \\ &\quad = a\,\phi(N)\,[\log(N-1) - \log N]. \end{aligned} \tag{26}$$

Since $\phi(N) < N$ for all N, we may expand the last term of the left side of (26) and obtain

$$\begin{aligned} &\log N - \log(N-1) - \phi(N)/N - \tfrac{1}{2}[\phi(N)/N]^2 \\ &\quad - 1/3[\phi(N)/N]^3 \cdots = a\,\phi(N)\,[\log(N-1) - \log N]. \end{aligned} \tag{27}$$

We now expand $\log(N-1) - \log N$ which appears in the right side of (27) and after rearrangements obtain

$$\log N - \log (N-1)$$
$$= \frac{\phi(N)}{N}\left[1 - a + \frac{\phi(N) - a}{2N} + \frac{[\phi(N)]^2 - a}{3N^2} + \cdots\right] \quad (28)$$
$$< \frac{\phi(N)}{N}\left[1 - a + (1 - \phi(N)/N)^{-1}\right].$$

Now if a is fixed and greater than unity, the limit of $\phi(N)/N$ cannot be zero as N increases without bound, because otherwise for N sufficiently large the right side of (28) becomes negative, while the left side is always positive, a contradiction of inequality (28). Therefore, the limit of Y/N, as N increases without bound, cannot be unity for $a > 1$. But this means that $\gamma \neq 0$ if $a > 1$. Hence, for $a > 1$, the non-zero branch of our curve is the only meaningful one.

An examination of the meaningful part of the graph of equation (20) shows that as long as the axone density does not exceed one axone per neuron, $\gamma = 0$, i.e., for very large N, the number of neurons to which there exist paths from an arbitrary neuron is negligible compared with the total number of neurons in the net. On the other hand, as the axone density increases from unity, γ increases rather rapidly, starting with slope 2. Already for $a = 2$, γ reaches about 0.8 of its asymptotic value (unity) and is within a fraction of one per cent of unity for quite moderate a (say > 6). This means that no matter how large the net is, it is practically certain that there will exist a path between two neurons picked at random, provided only the axone density is a few times greater than unity. The interpretation in terms of an epidemic with equiprobable contacts is entirely analogous.

The case a $= 1$. This case was treated by one of the authors (Rapoport, 1948) by a different method. It was shown that for large N, the probability that a neuron was member of a cycle was given by $\sqrt{\pi/2N}$. This gives the probability of the existence of a path from a neuron over any number of internuncials greater than one to itself. But under the assumption of equiprobable connections, this may well represent the probability of the existence of a path from the given neuron to any *other* neuron in the net. Therefore we should have for large N, in the case $a = 1$,

$$\gamma \sim \sqrt{\pi/2N}. \quad (29)$$

For $N = \infty$, γ reduces to zero, as it should according to equation (20). We shall, however, examine the asymptotic behavior of γ for

large N deduced from our approximate method, in order to compare it with the asymptotic behavior (29) deduced from an exact treatment of the special case. Dividing both sides of (17) by N, we may write for $a = 1$

$$Y/N = [(N-1)/N]^{N-Y+1}, \tag{30}$$

whence, since $Y/N = 1 - \gamma$,

$$\begin{aligned} 1-\gamma &= [(N-1)/N]^{N\gamma+1} \\ &= \mathrm{Exp}\{\ln(1-1/N) + N\gamma\ln(1-1/N)\}. \end{aligned} \tag{31}$$

We let $z = N^{-1}$ and examine the behavior of γ for small values of z. Expanding the right side of (31) by power series and retaining only terms of the second order (note that z and γ vanish together), we obtain

$$\begin{aligned} 1-\gamma = 1 &+ [-z - z^2/2 \cdots] + [-\gamma - \gamma z/2 - \cdots] \\ &+ z^2/2 + \gamma^2/2 + \gamma z + \cdots. \end{aligned} \tag{32}$$

Hence,

$$0 = -z + \gamma^2/2 + \gamma z/2 + \cdots. \tag{33}$$

Differentiating with respect to γ, we get

$$dz/d\gamma = \gamma + \gamma/2 \cdot dz/d\gamma + z/2 + \cdots, \tag{34}$$

$$dz/d\gamma \sim (\gamma + z/2)/(1 - \gamma/2). \tag{35}$$

Therefore $dz/d\gamma$ vanishes at $z = 0$, $\gamma = 0$. Differentiating once again with respect to γ, we obtain

$$\left.\frac{d^2z}{d\gamma^2}\right|_{\substack{z=0\\ \gamma=0}} = 1. \tag{36}$$

Hence the power series representing z as a function of γ begins as follows:

$$z = \gamma^2/2 + \cdots. \tag{37}$$

Thus

$$\gamma^2 \sim 2z = 2/N, \tag{38}$$

$$\gamma \sim \sqrt{2/N} \cong 1.41\sqrt{N}. \tag{39}$$

The "exact" result as expressed by (22) gives

$$\gamma \sim 1.2/\sqrt{N}.$$

Thus the approximate method applied to the case $a = 1$ implies an asymptotic behavior of γ for large N which does not depart too sharply from that deduced by the exact method. The limiting value for γ is zero in both cases. The question of how well the limiting values of γ are approached by the approximate method for $a > 1$ remains open.

This investigation is part of the work done under Contract No. AF 19(122)-161 between the U. S. Air Force Cambridge Research Laboratories and The University of Chicago.

LITERATURE

Puma, Marcello. 1939. *Elementi per una teoria matematica del contagio.* Rome: Editoriale Aeronautica.

Rapoport, Anatol. 1948. "Cycle Distributions in Random Nets." *Bull. Math. Biophysics,* **10,** 145–57.

Shimbel, Alfonso. 1950. "Contributions to the Mathematical Biophysics of the Central Nervous System with Special Reference to Learning." *Bull. Math. Biophysics,* **12,** 241–75.

ON THE EVOLUTION OF RANDOM GRAPHS

by

P. ERDŐS and A. RÉNYI

Dedicated to Professor P. Turán at his 50th birthday.

Introduction

Our aim is to study the probable structure of a random graph $\Gamma_{n,N}$ which has n given labelled vertices $P_1, P_2, \ldots, P_n$ and N edges; we suppose that these N edges are chosen at random among the $\binom{n}{2}$ possible edges, so that all $\binom{\binom{n}{2}}{N} = C_{n,N}$ possible choices are supposed to be equiprobable. Thus if $G_{n,N}$ denotes any one of the $C_{n,N}$ graphs formed from n given labelled points and having N edges, the probability that the random graph $\Gamma_{n,N}$ is identical with $G_{n,N}$ is $\frac{1}{C_{n,N}}$. If A is a property which a graph may or may not possess, we denote by $\mathbf{P}_{n,N}(A)$ the probability that the random graph $\Gamma_{n,N}$ possesses the property A, i. e. we put $\mathbf{P}_{n,N}(A) = \frac{A_{n,N}}{C_{n,N}}$ where $A_{n,N}$ denotes the number of those $G_{n,N}$ which have the property A.

An other equivalent formulation is the following: Let us suppose that n labelled vertices $P_1, P_2, \ldots, P_n$ are given. Let us choose at random an edge among the $\binom{n}{2}$ possible edges, so that all these edges are equiprobable. After this let us choose an other edge among the remaining $\binom{n}{2} - 1$ edges, and continue this process so that if already k edges are fixed, any of the remaining $\binom{n}{2} - k$ edges have equal probabilities to be chosen as the next one. We shall study the "evolution" of such a random graph if N is increased. In this investigation we endeavour to find what is the "typical" structure at a given stage of evolution (i. e. if N is equal, or asymptotically equal, to a given function $N(n)$ of n). By a "typical" structure we mean such a structure the probability of which tends to 1 if $n \to +\infty$ when $N = N(n)$. If A is such a property that $\lim_{n\to+\infty} \mathbf{P}_{n,N(n)}(A) = 1$, we shall say that „almost all" graphs $G_{n,N(n)}$ possess this property.

The study of the evolution of graphs leads to rather surprising results. For a number of fundamental structural properties A there exists a function $A(n)$ tending monotonically to $+\infty$ for $n \to +\infty$ such that

$$(1) \qquad \lim_{n\to+\infty} \mathbf{P}_{n,N(n)}(A) = \begin{cases} 0 & \text{if} \quad \lim\limits_{n\to+\infty} \dfrac{N(n)}{A(n)} = 0 \\ 1 & \text{if} \quad \lim\limits_{n\to+\infty} \dfrac{N(n)}{A(n)} = +\infty . \end{cases}$$

If such a function $A(n)$ exists we shall call it a "*threshold function*" of the property A.

In many cases besides (1) it is also true that there exists a probability distribution function $F(x)$ so that if $0 < x < +\infty$ and x is a point of continuity of $F(x)$ then

$$(2) \qquad \lim_{n\to+\infty} \mathbf{P}_{n,N(n)}(A) = F(x) \quad \text{if} \quad \lim_{n\to+\infty} \frac{N(n)}{A(n)} = x .$$

If (2) holds we shall say that $A(n)$ is a „*regular threshold function*" for the property A and call the function $F(x)$ the *threshold distribution function* of the property A.

For certain properties A there exist two functions $A_1(n)$ and $A_2(n)$ both tending monotonically to $+\infty$ for $n\to+\infty$, and satisfying $\lim\limits_{n\to+\infty} \dfrac{A_2(n)}{A_1(n)} = 0$, such that

$$(3) \qquad \lim_{n\to+\infty} \mathbf{P}_{n,N(n)}(A) = \begin{cases} 0 & \text{if} \quad \lim\limits_{n\to+\infty} \dfrac{N(n) - A_1(n)}{A_2(n)} = -\infty \\ 1 & \text{if} \quad \lim\limits_{n\to+\infty} \dfrac{N(n) - A_1(n)}{A_2(n)} = +\infty . \end{cases}$$

Clearly (3) implies that

$$(4) \qquad \lim_{n\to+\infty} \mathbf{P}_{n,N(n)}(A) = \begin{cases} 0 & \text{if} \quad \limsup\limits_{n\to+\infty} \dfrac{N(n)}{A_1(n)} < 1 \\ 1 & \text{if} \quad \liminf\limits_{n\to+\infty} \dfrac{N(n)}{A_1(n)} > 1 . \end{cases}$$

If (3) holds we call the pair $(A_1(n), A_2(n))$ a pair of "*sharp threshold*"-functions of the property A. It follows from (4) that if $(A_1(n), A_2(n))$ is a pair of sharp threshold functions for the property A then $A_1(n)$ is an (ordinary) threshold function for the property A and the threshold distribution function figuring in (2) is the degenerated distribution function

$$F_1(x) = \begin{cases} 0 & \text{for} \quad x \leqq 1 \\ 1 & \text{for} \quad x > 1 \end{cases}$$

and convergence in (2) takes place for every $x \neq 1$. In some cases besides (3) it is also true that there exists a probability distribution function $G(y)$ defined for $-\infty < y < +\infty$ such that if y is a point of continuity of $G(y)$ then

$$\lim_{n\to+\infty} \mathbf{P}_{n,N(n)}(A) = G(y) \quad \text{if} \quad \lim_{n\to+\infty} \frac{N(n) - A_1(n)}{A_2(n)} = y. \tag{5}$$

If (5) holds we shall say that we have a *regular sharp threshold* and shall call $G(y)$ the *sharp-threshold distribution function of the property A*.

One of our chief aims will be to determine the threshold respectively sharp threshold functions, and the corresponding distribution functions for the most obvious structural-properties, e. g. the presence in $\Gamma_{n,N}$ of subgraphs of a given type (trees, cycles of given order, complete subgraphs etc.) further for certain global properties of the graph (connectedness, total number of connected components, etc.).

In a previous paper [7] we have considered a special problem of this type; we have shown that denoting by C the property that the graph is connected, the pair $C_1(n) = \frac{1}{2} n \log n$, $C_2(n) = n$ is a pair of strong threshold functions for the property C, and the corresponding sharp-threshold distribution function is $e^{-e^{-2y}}$; thus we have proved[1] that putting $N(n) = \frac{1}{2} n \log n + y n + o(n)$ we have

$$\lim_{n\to+\infty} \mathbf{P}_{n,N(n)}(C) = e^{-e^{-2y}} \qquad (-\infty < y < +\infty). \tag{6}$$

In the present paper we consider the evolution of a random graph in a more systematic manner and try to describe the gradual development and step-by-step unravelling of the complex structure of the graph $\Gamma_{n,N}$ when N increases while n is a given large number.

We succeeded in revealing the emergence of certain structural properties of $\Gamma_{n,N}$. However a great deal remains to be done in this field. We shall call in § 10. the attention of the reader to certain unsolved problems. It seems to us further that it would be worth while to consider besides graphs also more complex structures from the same point of view, i. e. to investigate the laws governing their evolution in a similar spirit. This may be interesting not only from a purely mathematical point of view. In fact, the evolution of graphs may be considered as a rather simplified model of the evolution of certain communication nets (railway, road or electric network systems, etc.) of a country or some other unit. (Of course, if one aims at describing such a real situation, one should replace the hypothesis of equiprobability of all connections by some more realistic hypothesis.) It seems plausible that by considering the random growth of more complicated structures (e. g. structures consisting of different sorts of "points" and connections of different types) one could obtain fairly reasonable models of more complex real growth processes (e. g.

[1] Partial result on this problem has been obtained already in 1939 by P. ERDŐS and H. WHITNEY but their results have not been published.

the growth of a complex communication net consisting of different types of connections, and even of organic structures of living matter, etc.).

§§ 1—3. contain the discussion of the presence of certain components in a random graph, while §§ 4—9. investigate certain global properties of a random graph. Most of our investigations deal with the case when $N(n) \sim cn$ with $c > 0$. In fact our results give a clear picture of the evolution of $\Gamma_{n,N(n)}$ when $c = \frac{N(n)}{n}$ (which plays in a certain sense the role of time) increases. In § 10. we make some further remarks and mention some unsolved problems.

Our investigation belongs to the combinatorical theory of graphs, which has a fairly large literature. The first who enumerated the number of possible graphs with a given structure was A. CAYLEY [1]. Next the important paper [2] of G. PÓLYA has to be mentioned, the starting point of which were some chemical problems. Among more recent results we mention the papers of G. E. UHLENBECK and G. W. FORD [5] and E. N. GILBERT [6]. A fairly complete bibliography will be given in a paper of F. HARARY [8]. In these papers the probabilistic point of view was not explicitly emphasized. This has been done in the paper [9] of one of the authors, but the aim of the probabilistic treatment was there different: the existence of certain types of graphs has been shown by proving that their probability is positive. Random trees have been considered in [14].

In a recent paper [10] T. L. AUSTIN, R. E. FAGEN, W. F. PENNEY and J. RIORDAN deal with random graphs from a point of view similar to ours. The difference between the definition of a random graph in [10] and in the present paper consists in that in [10] it is admitted that two points should be connected by more than one edge ("parallel" edges). Thus in [10] it is supposed that after a certain number of edges have already been selected, the next edge to be selected may be any of the possible $\binom{n}{2}$ edges between the n given points (including the edges already selected). Let us denote such a random graph by $\Gamma^*_{n,N}$. The difference between the probable properties of $\Gamma_{n,N}$ resp. $\Gamma^*_{n,N}$ are in most (but not in all) cases negligible. The corresponding probabilities are in general (if the number N of edges is not too large) asymptotically equal. There is a third possible point of view which is in most cases almost equivalent with these two; we may suppose that for each pair of n given points it is determined by a chance process whether the edge connecting the two points should be selected or not, the probability for selecting any given edge being equal to the same number $p > 0$, and the decisions concerning the different edges being completely independent. In this case of course the number of edges is a random variable, having the expectation $\binom{n}{2} p$; thus if we want to obtain by this method a random graph having in the mean N edges we have to choose the value of p equal to $\frac{N}{\binom{n}{2}}$. We shall denote such a random graph by $\Gamma^{**}_{n,N}$. In many (though not all) of the problems treated in the present paper it does not cause any essential difference if we consider instead of $\Gamma_{n,N}$ the random graph $\Gamma^{**}_{n,N}$.

Comparing the method of the present paper with that of [10] it should be pointed out that our aim is to obtain threshold functions resp. distributions, and thus we are interested in asymptotic formulae for the probabilities considered. Exact formulae are of interest to us only so far as they help in determining the asymptotic behaviour of the probabilities considered (which is rarely the case in this field, as the exact formulae are in most cases too complicated). On the other hand in [10] the emphasis is on exact formulae resp. on generating functions. The only exception is the average number of connected components, for the asymptotic evaluation of which a way is indicated in § 5. of [10]; this question is however more fully discussed in the present paper and our results go beyond that of [10]. Moreover, we consider not only the number but also the character of the components. Thus for instance we point out the remarkable change occuring at $N \sim \frac{n}{2}$. If $N \sim nc$ with $c < 1/2$ then with probability tending to 1 for $n \to +\infty$ all points except a bounded number of points of $\Gamma_{n,N}$ belong to components which are trees, while for $N \sim nc$ with $c > \frac{1}{2}$ this is no longer the case. Further for a fixed value of n the average number of components of $\Gamma_{n,N}$ decreases asymptotically in a linear manner with N, when $N \leqq \frac{n}{2}$, while for $N > \frac{n}{2}$ the formula giving the average number of components is not linear in N.

In what follows we shall make use of the sysmbols O and o. As usually $a(n) = o(b(n))$ (where $b(n) > 0$ for $n = 1, 2, \ldots$) means that $\lim_{n \to +\infty} \frac{|a(n)|}{b(n)} = 0$, while $a(n) = O(b(n))$ means that $\frac{|a(n)|}{b(n)}$ is bounded. The parameters on which the bound of $\frac{|a(n)|}{b(n)}$ may depend will be indicated if it is necessary; sometimes we will indicate it by an index. Thus $a(n) = O_\varepsilon(b(n))$ means that $\frac{|a(n)|}{b(n)} \leqq K(\varepsilon)$ where $K(\varepsilon)$ is a positive constant depending on ε. We write $a(n) \sim b(n)$ to denote that $\lim_{n \to +\infty} \frac{a(n)}{b(n)} = 1$.

We shall use the following definitions from the theory of graphs. (For the general theory see [3] and [4].)

A finite non-empty set V of labelled points $P_1, P_2, \ldots, P_n$ and a set E of different unordered pairs (P_i, P_j) with $P_i \in V$, $P_j \in V$, $i \neq j$ is called a *graph*; we denote it sometimes by $G = \{V, E\}$; the number n is called the *order* (or *size*) of the graph; the points $P_1, P_2, \ldots, P_n$ are called the *vertices* and the pairs (P_i, P_j) the *edges* of the graph. Thus we consider *non-oriented finite graphs without parallel edges and without slings*. The set E may be empty, thus a collection of points (especially a single point) is also a graph.

A graph $G_2 = \{V_2, E_2\}$ is called a *subgraph* of a graph $G_1 = \{V_1, E_1\}$ if the set of vertices V_2 of G_2 is a subset of the set of vertices V_1 of G_1 and the set E_2 of edges of G_2 is a subset of the set E_1 of edges of G_1.

A sequence of k edges of a graph such that every two consecutive edges and only these have a vertex in common is called a *path* of order k.

A cyclic sequence of k edges of a graph such that every two consecutive edges and only these have a common vertex is called a *cycle* of order k.

A graph G is called *connected* if any two of its points belong to a path which is a subgraph of G.

A graph is called a *tree* of order (or size) k if it has k vertices, is connected and if none of its subgraphs is a cycle. A tree of order k has evidently $k - 1$ edges.

A graph is called a *complete graph* of order $\binom{k}{2}$ if it has k vertices and $\binom{k}{2}$ edges. Thus in a complete graph of order k any two points are connected by an edge.

A subgraph G' of a graph G will be called an *isolated subgraph* if all edges of G one or both endpoints of which belong to G', belong to G'. A connected isolated subgraph G' of a graph G is called a *component* of G. The number of points belonging to a component G' of a graph G will be called the *size* of G'.

Two graphs shall be called *isomorphic*, if there exists a one-to-one mapping of the vertices carrying over these graphs into another.

The graph $\overline{G}$ shall be called *complementary graph* of G if $\overline{G}$ consists of the same vertices $P_1, P_2, \ldots, P_n$ as G and of those and only those edges (P_i, P_j) which do not occur in G.

The number of edges starting from the point P of a graph G will be called the *degree* of P in G.

A graph G is called a *saturated even graph of type* (a, b) if it consists of $a + b$ points and its points can be split in two subsets V_1 and V_2 consisting of a resp. b points, such that G contains any edge (P, Q) with $P \in V_1$ and $Q \in V_2$ and no other edge.

A graph is called *planar*, if it can be drawn on the plane so that no two of its edges intersect.

We introduce further the following definitions: If a graph G has n vertices and N edges, we call the number $\frac{2N}{n}$ the "*degree*" of the graph. (As a matter of fact $\frac{2N}{n}$ is the average degree of the vertices of G.) If a graph G has the property that G has no subgraph having a larger degree than G itself, we call G a *balanced* graph.

We denote by $\mathbf{P}(\ldots)$ the probability of the event in the brackets, by $\mathbf{M}(\xi)$ resp. $\mathbf{D}^2(\xi)$ the mean value resp. variance of the random variable ξ. In cases when it is not clear from the context in which probability space the probabilities or respectively the mean values and variances are to be understood, this will be explicitly indicated. Especially $\mathbf{M}_{n,N}$ resp. $\mathbf{D}^2_{n,N}$ will denote the mean value resp. variance calculated with respect to the probabilities $\mathbf{P}_{n,N}$.

We shall often use the following elementary asymptotic formula:

$$\binom{n}{k} \sim \frac{n^k e^{-\frac{k^2}{2n} - \frac{k^3}{6n^2}}}{k!} \text{ valid for } k = o(n^{3/4}) . \tag{7}$$

Our thanks are due to T. GALLAI for his valuable remarks.

§ 1. Thresholds for subgraphs of given type

If N is very small compared with n, namely if $N = o(\sqrt{n})$ then it is very probable that $\Gamma_{n,N}$ is a collection of isolated points and isolated edges, i. e. that no two edges of $\Gamma_{n,N}$ have a point in common. As a matter of fact the probability that at least two edges of $\Gamma_{n,N}$ shall have a point in common is by (7) clearly

$$1 - \frac{\binom{n}{2N}(2N)!}{2^N N! \binom{\binom{n}{2}}{N}} = O\left(\frac{N^2}{n}\right) .$$

If however $N \sim c\sqrt{n}$ where $c > 0$ is a constant not depending on n, then the appearance of trees of order 3 will have a probability which tends to a positive limit for $n \to +\infty$, but the appearance of a connected component consisting of more than 3 points will be still very improbable. If N is increased while n is fixed, the situation will change only if N reaches the order of magnitude of $n^{2/3}$. Then trees of order 4 (but not of higher order) will appear with a probability not tending to 0. In general, the threshold function for the presence of trees of order k is $n^{\frac{k-2}{k-1}}$ $(k = 3, 4, \ldots)$. This result is contained in the following

Theorem 1. *Let $k \geqq 2$ and $l \left(k - 1 \leqq l \leqq \binom{k}{2}\right)$ be positive integers. Let $\mathscr{B}_{k,l}$ denote an arbitrary not empty class of connected balanced graphs consisting of k points and l edges. The threshold function for the property that the random graph considered should contain at least one subgraph isomorphic with some element of $\mathscr{B}_{k,l}$ is $n^{2-\frac{k}{l}}$.*

The following special cases are worth mentioning

Corollary 1. *The threshold function for the property that the random graph contains a subgraph which is a tree of order k is $n^{\frac{k-2}{k-1}}$ $(k = 3, 4, \ldots)$.*

Corollary 2. *The threshold function for the property that a graph contains a connected subgraph consisting of $k \geqq 3$ points and k edges (i. e. containing exactly one cycle) is n, for each value of k.*

Corollary 3. *The threshold function for the property that a graph contains a cycle of order k is n, for each value of $k \geqq 3$.*

Corollary 4. *The threshold function for the property that a graph contains a complete subgraph of order* $k \geqq 3$ *is* $n^{2\left(1-\frac{1}{k-1}\right)}$.

Corollary 5. *The threshold function for the property that a graph contains a saturated even subgraph of type* (a, b) *(i. e. a subgraph* consisting of $a+b$ points $P_1, \ldots, P_a, Q_1, \ldots Q_b$ and of the ab edges (P_i, Q_j) *is* $n^{2-\frac{a+b}{ab}}$.

To deduce these Corollaries one has only to verify that all 5 types of graphs figuring in Corollaries 1—5. are balanced, which is easily seen.

Proof of Theorem 1. Let $B_{k,l} \geqq 1$ denote the number of graphs belonging to the class $\mathscr{B}_{k,l}$ which can be formed from k given labelled points. Clearly if $P_{n,N}(\mathscr{B}_{k,l})$ denotes the probability that the random graph $\Gamma_{n,N}$ contains at least one subgraph isomorphic with some element of the class $\mathscr{B}_{k,l}$, then

$$\mathbf{P}_{n,N}(\mathscr{B}_{k,l}) \leqq \binom{n}{k} B_{k,l} \frac{\binom{\binom{n}{2}-l}{N-l}}{\binom{\binom{n}{2}}{N}} = O\left(\frac{N^l}{n^{2l-k}}\right). \tag{1.1}$$

As a matter of fact if we select k points (which can be done in $\binom{n}{2}$ different ways) and form from them a graph isomorphic with some element of the class $\mathscr{B}_{k,l}$ (which can be done in $B_{k,l}$ different ways) then the number of graphs $G_{n,N}$ which contain the selected graph as a subgraph is equal to the number of ways the remaining $N-l$ edges can be selected from the $\binom{n}{2}-l$ other possible edges. (Of course those graphs, which contain more subgraphs isomorphic with some element of $\mathscr{B}_{k,l}$ are counted more than once.)

Now clearly if $N = o\left(n^{2-\frac{k}{l}}\right)$ then by

$$\mathbf{P}_{n,N}(\mathscr{B}_{k,l}) = o(1)$$

which proves the first part of the assertion of Theorem 1. To prove the second part of the theorem let $\mathscr{B}_{k,l}^{(n)}$ denote the set of all subgraphs of the complete graph consisting of n points, isomorphic with some element of $\mathscr{B}_{k,l}$. To any $S \in \mathscr{B}_{k,l}^{(n)}$ let us associate a random variable $\varepsilon(S)$ such that $\varepsilon(S)=1$ or $\varepsilon(S)=0$ according to whether S is a subgraph of $\Gamma_{n,N}$ or not. Then clearly (we write in what follows for the sake of brevity $\mathbf{M}$ instead of $\mathbf{M}_{n,N}$)

$$\mathbf{M}\left(\sum_{S \in \mathscr{B}_{k,l}^{(n)}} \varepsilon(S)\right) = \sum_{S \in \mathscr{B}_{k,l}^{(n)}} \mathbf{M}(\varepsilon(S)) = \binom{n}{k} B_{k,l} \frac{\binom{\binom{n}{2}-l}{N-l}}{\binom{\binom{n}{2}}{N}} \sim \frac{B_{k,l}}{k!} \frac{(2N)^l}{n^{2l-k}}. \tag{1.2}$$

On the other hand if S_1 and S_2 are two elements of $\mathscr{G}_{k,l}^{(n)}$ and if S_1 and S_2 do not contain a common edge then

$$\mathbf{M}(\varepsilon(S_1)\,\varepsilon(S_2)) = \frac{\binom{\binom{n}{2}-2l}{N-2l}}{\binom{\binom{n}{2}}{N}}.$$

If S_1 and S_2 contain exactly s common points and r common edges $(1 \leqq r \leqq l-1)$ we have

$$\mathbf{M}(\varepsilon(S_1)\,\varepsilon(S_2)) = \frac{\binom{\binom{n}{2}-2l+r}{N-2l+r}}{\binom{\binom{n}{2}}{N}} = O\left(\frac{N^{2l-r}}{n^{4l-2r}}\right).$$

On the other hand the intersection of S_1 and S_2 being a subgraph of S_1 (and S_2) by our supposition that each S is balanced, we obtain $\frac{r}{s} \leqq \frac{l}{k}$ i.e. $s \geqq \frac{rk}{l}$ and thus the number of such pairs of subgraphs S_1 and S_2 does not exceed

$$B_{k,l}^2 \sum_{j \geq \frac{rk}{l}}{}' \binom{n}{k}\binom{k}{j}\binom{n-k}{k-j} = O\left(n^{2k-\frac{rk}{l}}\right).$$

Thus we obtain

$$\mathbf{M}\left(\Big(\sum_{S \in \mathscr{B}_{k,l}^{(n)}} \varepsilon(S)\Big)^2\right) =$$

(1.3)

$$= \sum_{S \in \mathscr{B}_{k,l}^{(n)}} \mathbf{M}(\varepsilon(S)) + \frac{n!\,B_{k,l}^2}{k!^2(n-2k)!} \frac{\binom{\binom{n}{2}-2l}{N-2l}}{\binom{\binom{n}{2}}{N}} + O\left(\left(\frac{N^l}{n^{2l-k}}\right)^2 \sum_{r=1}^{l}{}' \left(\frac{n^{2-\frac{k}{l}}}{N}\right)^r\right).$$

Now clearly

$$\frac{n!}{k!^2(n-2k)!} \frac{\binom{\binom{n}{2}-2l}{N-2l}}{\binom{\binom{n}{2}}{N}} \leqq \binom{n}{k}^2 \frac{\binom{\binom{n}{2}-l}{N-l}^2}{\binom{\binom{n}{2}}{N}^2}.$$

If we suppose that

$$\frac{N}{n^{2-\frac{k}{l}}} = \omega \to +\infty ,$$

it follows that we have

$$\mathbf{D}^2\Big(\sum_{S \in \mathfrak{B}_{k,l}^{(n)}} \varepsilon(S)\Big) = O\left(\frac{\left(\sum_{S \in \mathfrak{B}_{k,l}^{(n)}} \mathbf{M}(\varepsilon(S))^2\right)}{\omega}\right). \tag{1.4}$$

It follows by the inequality of *Chebysheff* that

$$\mathbf{P}_{n,N}\left(\left|\sum_{S \in \mathfrak{B}_{k,l}^{(n)}} \varepsilon(S) - \sum_{S \in \mathfrak{B}_{k,l}^{(n)}} \mathbf{M}(\varepsilon(S))\right| > \frac{1}{2} \sum_{S \in \mathfrak{B}_{k,l}^{(n)}} \mathbf{M}(\varepsilon(S))\right) = O\left(\frac{1}{\omega}\right)$$

and thus

$$\mathbf{P}_{n,N}\left(\sum_{S \in \mathfrak{B}_{k,l}^{(n)}} \varepsilon(S) \leqq \frac{1}{2} \sum_{S \in \mathfrak{B}_{k,l}^{(n)}} \mathbf{M}(\varepsilon(S))\right) = O\left(\frac{1}{\omega}\right). \tag{1.5}$$

As clearly by (1.2) if $\omega \to +\infty$ then $\sum_{S \in \mathfrak{B}_{k,l}^{(n)}} \mathbf{M}(\varepsilon(S)) \to +\infty$ it follows not only that the probability that $\Gamma_{n,N}$ contains at least one subgraph isomorphic with an element of $\mathfrak{B}_{k,l}$ tends to 1, but also that with probability tending to 1 the number of subgraphs of $\Gamma_{n,N}$ isomorphic to some element of $\mathfrak{B}_{k,l}$ will tend to $+\infty$ with the same order of magnitude as ω^l.

Thus Theorem 1 is proved.

It is interesting to compare the thresholds for the appearance of a subgraph of a certain type in the above sense with probability near to 1, with the number of edges which is needed in order that the graph should have *necessarily* a subgraph of the given type. Such "compulsory" thresholds have been considered by P. TURÁN [11] (see also [12]) and later by P. ERDŐS and A. H. STONE [17]). For instance for a tree of order k clearly the compulsory threshold is $\left[\frac{n(k-2)}{2}\right] + 1$; for the presence of at least one cycle the compulsory threshold is n while according to a theorem of P. TURÁN [11] for complete subgraphs of order k the compulsory threshold is $\frac{(k-2)}{2(k-1)}(n^2 - r^2) + \binom{r}{2}$ where $r = n - (k-1)\left[\frac{n}{k-1}\right]$. In the paper [13] of T. KŐVÁRI, V. T. Sós and P. TURÁN it has been shown that the compulsory threshold for the presence of a saturated even subgraph of type (a, a) is of order of magnitude not greater than $n^{2-\frac{1}{a}}$. In all cases the "compulsory" thresholds in TURÁN's sense are of greater order of magnitude as our "probable" thresholds.

§ 2. Trees

Now let us turn to the determination of threshold distribution functions for trees of a given order. We shall prove somewhat more, namely that if $N \sim \varrho\, n^{\frac{k-2}{k-1}}$ where $\varrho > 0$, then the number of trees of order k contained in $\Gamma_{n,N}$ has in the limit for $n \to +\infty$ a Poisson distribution with mean value $\lambda = \frac{(2\varrho)^{k-1} k^{k-2}}{k!}$. This implies that the threshold distribution function for trees of order k is $1 - e^{-\lambda}$.

In proving this we shall count only *isolated trees* of order k in $\Gamma_{n,N}$, i. e. trees of order k which are isolated subgraphs of $\Gamma_{n,N}$. According to Theorem 1. this makes no essential difference, because if there would be a tree of order k which is a subgraph but not an isolated subgraph of $\Gamma_{n,N}$, then $\Gamma_{n,N}$ would have a connected subgraph consisting of $k+1$ points and the probability of this is tending to 0 if $N = o\left(n^{\frac{k-1}{k}}\right)$ which condition is fulfilled in our case as we suppose $N \sim \varrho\, n^{\frac{k-2}{k-1}}$.

Thus we prove

Theorem 2a. *If* $\lim_{n\to+\infty} \frac{N(n)}{n^{\frac{k-2}{k-1}}} = \varrho > 0$ *and* τ_k *denotes the number of isolated trees of order* k *in* $\Gamma_{n,N(n)}$ *then*

$$\lim_{n\to+\infty} \mathbf{P}_{n,N(n)}(\tau_k = j) = \frac{\lambda^j e^{-\lambda}}{j!} \tag{2.1}$$

or $j = 0, 1, \ldots,$ *where*

$$\lambda = \frac{(2\varrho)^{k-1} k^{k-2}}{k!}. \tag{2.2}$$

For the proof we need the following

Lemma 1. *Let* $\varepsilon_{n1}, \varepsilon_{n2}, \ldots, \varepsilon_{nl_n}$ *be sets of random variables on some probability space; suppose that* $\varepsilon_{ni} (1 \leqq i \leqq l_n)$ *takes on only the values* 1 *and* 0. *If*

$$\lim_{n\to+\infty} \sum_{1\leq i_1<i_2<\ldots<i_r\leq l_n} \mathbf{M}(\varepsilon_{ni_1} \varepsilon_{ni_2} \ldots \varepsilon_{ni_r}) = \frac{\lambda^r}{r!} \tag{2.3}$$

uniformly in r *for* $r = 1, 2, \ldots,$ *where* $\lambda > 0$ *and the summation is extended over all combinations* $(i_1, i_2, \ldots, i_r)$ *of order* r *of the integers* $1, 2, \ldots, l_n$, *then*

$$\lim_{n\to+\infty} \mathbf{P}\left(\sum_{i=1}^{l_n} \varepsilon_{ni} = j\right) = \frac{\lambda^j e^{-\lambda}}{j!} \qquad (j = 0, 1, \ldots) \tag{2.4}$$

i. e. the distribution of the sum $\sum_{i=1}^{l_n} \varepsilon_{ni}$ *tends for* $n \to +\infty$ *to the Poisson-distribution with mean value* λ.

Proof of Lemma 1. Let us put

$$P_n(j) = \mathbf{P}\left(\sum_{i=1}^{l_n} \varepsilon_{n_i} = j\right). \tag{2.5}$$

Clearly

$$\sum_{1 \le i_1 < i_2 < \ldots < i_r \le l_n} \mathbf{M}(\varepsilon_{ni_1} \varepsilon_{ni_2} \ldots \varepsilon_{ni_r}) = \sum_{j=r}^{+\infty} \binom{j}{r} P_n(j) \tag{2.6}$$

thus it follows from (2.3) that

$$\lim_{n \to +\infty} \sum_{j=r}^{+\infty} P_n(j) \binom{j}{r} = \frac{\lambda^r}{r!} \qquad (r = 1, 2, \ldots) \tag{2.7}$$

uniformly in r.

It follows that for any z with $|z| < 1$

$$\lim_{n \to +\infty} \sum_{r=1}^{\infty} \left(\sum_{j=r}^{+\infty} P_n(j) \binom{j}{r} \right) z^r = \sum_{r=1}^{\infty} \frac{(\lambda z)^r}{r!} = e^{\lambda z} - 1. \tag{2.8}$$

But

$$\sum_{r=1}^{\infty} \left(\sum_{j=r}^{+\infty} P_n(j) \binom{j}{r} \right) z^r = \sum_{j=0}^{+\infty} P_n(j) (1+z)^j - 1. \tag{2.9}$$

Thus choosing $z = x - 1$ with $0 < x \leqq 1$ it follows that

$$\lim_{n \to +\infty} \sum_{j=0}^{+\infty} P_n(j) x^j = e^{\lambda(x-1)} \qquad \text{for } 0 < x \leqq 1. \tag{2.10}$$

It follows easily that (2.10) holds for $x = 0$ too. As a matter of fact putting $G_n(x) = \sum_{j=0}^{+\infty} P_n(j) x^j$, we have for $0 < x \leqq 1$

$$|P_n(0) - e^{-\lambda}| \leqq |G_n(x) - e^{\lambda(x-1)}| + |G_n(x) - P_n(0)| + |e^{\lambda(x-1)} - e^{-\lambda}|.$$

As however

$$|G_n(x) - P_n(0)| \leqq x \sum_{j=1}^{+\infty} P_n(j) \leqq x$$

and similarly

$$|e^{\lambda(x-1)} - e^{-\lambda}| \leqq x$$

it follows that

$$|P_n(0) - e^{-\lambda}| \leqq |G_n(x) - e^{\lambda(x-1)}| + 2x.$$

Thus we have

$$\limsup_{n \to +\infty} |P_n(0) - e^{-\lambda}| \leqq 2x;$$

as however $x > 0$ may be chosen arbitrarily small it follows that

$$\lim_{n \to +\infty} P_n(0) = e^{-\lambda}.$$