Random Graphs

Béla Bollobás

Random Graphs

BÉLA BOLLOBÁS

Department of Pure Mathematics and Mathematical Statistics
University of Cambridge

and

Department of Mathematics
Louisiana State University

1985

Preface

The theory of random graphs was founded by Erdős and Rényi (1959, 1960, 1961a,b) after Erdős (1947, 1959, 1961) had discovered that probabilistic methods were often useful in tackling extremal problems in graph theory. Erdős proved, amongst other things, that for all natural numbers $g \geqslant 3$ and $k \geqslant 3$ there exist graphs with girth g and chromatic number k. Erdős did not construct such graphs explicitly but showed that most graphs in a certain class could be altered slightly to give the required examples.

This phenomenon was not entirely new in mathematics, although it was certainly surprising that probabilistic ideas proved to be so important in the study of such a simple finite structure as a graph. In analysis, Paley and Zygmund (1930a,b, 1932) had investigated random series of functions. One of their results was that if the real numbers c_n satisfy $\sum_{n=0}^{\infty} c_n^2 = \infty$ then $\sum_{n=0}^{\infty} \pm c_n \cos nx$ fails to be a Fourier-Lebesgue series for almost all choices of the signs. To exhibit a sequence of signs with this property is surprisingly difficult: indeed, no algorithm is known which constructs an appropriate sequence of signs from any sequence c_n with $\sum_{n=0}^{\infty} c_n^2 = \infty$. Following the initial work of Paley and Zygmund, random functions were investigated in great detail by Steinhaus (1930), Paley, Wiener and Zygmund (1932), Kac (1949), Hunt (1951), Ryll-Nardzewski (1953), Salem and Zygmund (1954), Dvoretzky and Erdős (1959) and many others. An excellent account of these investigations can be found in Kahane (1963, 1968). Probabilistic methods were also used by Littlewood and Offord (1938) to study the zeros of random polynomials and analytic functions. Some decades later, a simple but crucial combinatorial lemma from their work greatly influenced the study of random finite sets in vector spaces.

The first combinatorial structures to be studied probabilistically were

vii

tournaments, chiefly because random tournaments are intrinsically related to statistics. The study began with Kendall and Babington-Smith (1940) and a concise account of many of the results is given by Moon (1968b). Szele (1943) was, perhaps, the first to apply probabilistic ideas to extremal problems in combinatorics. He observed that some tournament of order n must have at least $n!/2^{n-1}$ Hamilton paths, because the expected number of Hamilton paths is $n!/2^{n-1}$. Once again, it is not easy to construct a tournament of order n with this many Hamilton paths. A little later, Erdős (1947) used a similar argument, based on the expected number of k-cliques in a graph of order n, to show that the Ramsey number $R(k)$ is greater than $2^{k/2}$.

Existence results based on probabilistic ideas can now be found in many branches of mathematics, especially in analysis, the geometry of Banach spaces, number theory, graph theory, combinatorics and computer science. Probabilistic methods have become an important part of the arsenal of a great many mathematicians. Nevertheless, this is only a beginning: in the next decade or two probabilistic methods are likely to become even more prominent. It is also likely that in the not too distant future it will be possible to carry out statistical analyses of more complicated systems. Mathematicians who are not interested in graphs for their own sake should view the theory of random graphs as a modest beginning from which we can learn a variety of techniques and can find out what kind of results we should try to prove about more complicated random structures.

As often happens in mathematics, the study of statistical aspects of graphs was begun independently and almost simultaneously by several authors, namely Ford and Uhlenbeck (1956), Gilbert (1957), Austin, Fagen, Penney and Riordan (1959) and Erdős and Rényi (1959). Occasionally all these authors are credited with the foundation of the theory of random graphs. However, this is to misconceive the nature of the subject. Only Erdős and Rényi introduced the methods which underlie the probabilistic treatment of random graphs. The other authors were all concerned with enumeration problems and their techniques were essentially deterministic.

There are two natural ways of estimating the proportion of graphs having a certain property. One may obtain exact formulae, using Pólya's enumeration theorem, generating functions, differential operators and the whole theory of combinatorial enumeration, and then either consider the task completed or else proceed to investigate the asymptotic behaviour of the exact but complicated formulae, which is often a daunting task. This approach, whose spirit is entirely deterministic, was used in the first three papers mentioned above and has been carried further by numerous authors. Graphical enumeration is

discussed in detail in the well known monograph of Harary and Palmer (1973) and the recent encyclopaedic treatise by Goulden and Jackson (1983). The connection between graphical enumeration and statistical mechanics is emphasized by Temperley (1981). The theory of enumeration is a beautiful, rich and rapidly developing area of combinatorics, but it has very little to do with the theory of random graphs.

The other approach was introduced by Erdős and Rényi and is expounded in this volume. It has only the slightest connection with enumeration. One is not interested in exact formulae but rather in approximating a variety of exact values by appropriate probability distributions and using probabilistic ideas, whenever possible. As shown by Erdős and Rényi, this probabilistic approach is often more powerful than the deterministic one.

It is often helpful to imagine a random graph as a living organism which evolves with time. It is born as a set of n isolated vertices and develops by successively acquiring edges at random. Our main aim is to determine at what stage of the evolution a particular property of the graph is likely to arise. To make this more precise, we shall consider the properties of a 'typical' graph in a probability space consisting of graphs of a particular type. The simplest such probability space consists of all graphs with a given set of n labelled vertices and M edges, and each such graph is assigned the same probability. Usually we shall write G_M for a random element of this probability space. Then, if H is any graph with the given vertex set and M edges, then

$$P(G_M = H) = 1 \bigg/ \binom{N}{M} \text{ where } N = \binom{n}{2}.$$

In most cases we shall have a sequence of probability spaces. For each natural number n there will be a probability space consisting of graphs with exactly n vertices. We shall be interested in the properties of this space as $n \to \infty$. In this situation we shall say that a *typical element of our space has a property Q* when the probability that a random graph on n vertices has Q tends to 1 as $n \to \infty$. We also say that *almost every (a.e)* graph has property Q. Thus almost every G_M has property Q if the proportion of graphs with this property tends to 1 as $n \to \infty$. Once we are given such probability spaces of graphs, numerous natural questions arise. Is a typical graph connected? Is it k-connected? Is the chromatic number at least k? Does almost every graph contain a triangle? Does it have diameter at most d? Is almost every graph Hamiltonian?

The greatest discovery of Erdős and Rényi was that many important properties of graphs appear quite suddenly. If we pick a function $M = M(n)$ then, in many cases, either almost every graph G_M has property Q or else

almost every graph fails to have property Q. In this vague sense, we have a $0-1$ law. The transition from a property being very unlikely to it being very likely is usually very swift. To make this more precise, consider a *monotone* (increasing) property Q, i.e. one for which a graph has Q whenever one of its subgraphs has Q. For many such properties there is a *threshold function* $M_0(n)$. If $M(n)$ grows somewhat slower than $M_0(n)$, then almost every G_M fails to have Q. If $M(n)$ grows somewhat faster than $M_0(n)$, then almost every G_M has the property Q. For example, $M_0(n) = \frac{1}{2}n \log n$ is a threshold function for connectedness in the following sense: if $\omega(n) \to \infty$, no matter how slowly, then almost every G is disconnected for $M(n) = \frac{1}{2}n (\log n - \omega(n))$ and almost every G is connected for $M(n) = \frac{1}{2}n (\log n + \omega(n))$.

As the proportion M/N of edges increases, where, as always, $N = \binom{n}{2}$ is the total number of possible edges, the shape of a typical graph G_M passes through several clearly identifiable stages, in which many of the important parameters of a typical graph are practically determined. When M is neither too small nor too large, then the most important property is that, for every fixed k, any k vertices in a typical graph have about the same number of neighbours. Thus a typical random graph is rather similar to an ideal regular graph whose automorphism group is transitive on small sets of vertices. Of course, there is no such non-trivial regular graph, and in many applications random graphs are used precisely because they approximate an ideal regular graph.

This book is the first systematic and extensive account of a substantial body of results from the theory of random graphs. Considerably shorter treatments of random graphs can be found in various parts of Erdős and Spencer (1974), in Chapter VIII of Marshall (1971), in Chapter VII of Bollobás (1979a), and in the review papers by Grimmett (1980), Bollobás (1981f) and Karoński (1982). Despite over 750 references, several topics are not covered in any detail and we can make no claim to completeness. Perhaps the greatest omission is the extensive theory of random trees, about which the reader could get some idea by consulting the beautiful book by Moon (1970a) and the papers by Moon and Meir listed in the references. I might justify this particular omission because the tools used in its study are often those of enumeration rather than probability theory. However, here and in general, the choice of topics is mainly a reflection of my own interests.

The audience I had in mind when writing this book consisted mainly of research students and professional mathematicians. The volume should also be of interest to computer scientists. Random graphs are of ever increasing

importance in this field and several of the sections have been written expressly for computer scientists.

The monograph was planned to be considerably shorter and was to be completed by early 1981. However, I soon realized that I had little control over the length which was dictated by the subject matter I intended to explore. Temptations to ease my load by adopting short-cuts have been rife but, for the reader's sake, I have tried to resist them. During its preparation I have given several courses on the material, notably for Part III for the University of Cambridge in 1980/81 and 1983/84. The first seven chapters were complete by the summer of 1982 and were circulated quite widely. I gave a series of lectures on these at the Waterloo Silver Jubilee Conference. These lectures were published only recently (1984d). I have written the book wherever I happened to be: the greatest part at Louisiana State University, Baton Rouge; a few chapters in Cambridge, Waterloo and São Paulo; and several sections in Tokyo and Budapest. I hope never again to travel with 200 lb of paper!

My main consideration in selecting and presenting the material was to write a book which I would like to read myself. In spite of this, the book is essentially self-contained, although familiarity with the basic concepts of graph theory and probability theory would be helpful. There is little doubt that many readers will use this monograph as a compendium of results. This is a pity, partly because a book suitable for that purpose could have been produced with much less effort than has gone into this volume, and also because proofs are often more important than results: not infrequently the reader will derive more benefit from knowing the methods used than from familiarity with the theorems. The list of contents describes fairly the material presented in the book.

The graph theoretic notation and terminology used in the book are standard. For undefined concepts and symbols the reader may consult Bollobás (1978a, 1979a).

The exercises at the end of each chapter vary greatly in importance and difficulty. Most are designed to clarify and complete certain points but a few are important results. The end of a proof, or its absence, is indicated by the symbol $\square$; the *floor* of x (i.e. the greatest integer less than, or equal to, x) is denoted by $\lfloor x \rfloor$; and the *ceiling* of x by $\lceil x \rceil$. With very few exceptions, the various parameters and random variables depend on the number n of vertices of graphs under consideration and the inequalities are only claimed to hold when n is sufficiently large. This is often stated but it is also implicit on many other occasions. The symbols $c_1, c_2, \ldots$ which appear without any explana-

tion, are always independent of n. They may be absolute constants or may depend on other quantities which are independent of n. To assist the reader, it will often be stated which of these is the case.

It is a great pleasure to acknowledge the debt I owe to many people for their help. Several of my friends kindly read large sections of the manuscript. Andrew Thomason from Exeter, Istvan Simon from São Paulo, Masao Furuyama from Kyoto and Boris Pittel from Columbus were especially generous with their help. They corrected many errors and frequently improved the presentation. In addition, I benefited from the assistance of Andrew Barbour, Keith Carne, Geoff Eagleson, Alan Frieze and Jonathan Partington. Many research students, including Keith Ball, Graham Brightwell and Colin Wright, helped me find some of the mistakes; for the many which undoubtedly remain, I apologize. I am convinced that without the very generous help of Andrew Harris in using the computer the book would still be far from finished.

It was Paul Erdős and Alfréd Rényi who, just over 20 years ago, introduced me to the subject of random graphs. My active interest in the field was aroused by Paul Erdős some years later, whose love for it I found infectious; I am most grateful to him for firing my enthusiasm for the subject whose beauty has given me so much pleasure ever since.

Finally, I am especially grateful to Gabriella, my wife, for her constant support and encouragement. Her enthusiasm and patience survived even when mine failed.

Baton Rouge   B.B.
December, 1984

List of Contents

Chapter I

Probability Theoretic Preliminaries

The aim of this chapter is to present the definitions, formulae and results of probability theory we shall need in the main body of the book. Although we assume that the reader has had only a rather limited experience with probability theory and, if somewhat vaguely, we do define almost everything, this chapter is not intended to be a systematic introduction to probability theory. The main purpose is to identify the facts we shall rely on, so only the most important—and perhaps not too easily accessible—results will be proved. Since the book is primarily for mathematicians interested in graph theory, combinatorics and computing, some of the results will not be presented in full generality. It is inevitable that for the reader who is familiar with probability theory this introduction contains too many basic definitions and familiar facts, while the reader who has not studied probability before will find the chapter rather difficult.

There are many excellent introductions to probability theory: Feller (1966), Breiman (1968), K. L. Chung (1974) and H. Bauer (1981), to name only four. The interested reader is urged to consult one of these texts for a thorough introduction to the subject.

1. NOTATION AND BASIC FACTS

A *probability space* is a triple (Ω, Σ, P), where Ω is a set, Σ is a σ-field of subsets of Ω, P is a non-negative measure on Σ and $P(\Omega) = 1$. In the simplest case Ω is a finite set and Σ is $\mathscr{P}(\Omega)$, the set of all subsets of Ω. Then P is determined by the function $\Omega \to [0, 1]$, $w \to P(\{w\})$, namely

$$P(A) = \sum_{w \in A} P(\{w\}), \qquad A \subset \Omega.$$

1

A real valued *random variable* (r.v.) X is a measurable real-valued function on a probability space, $X : \Omega \to \mathbb{R}$.

Given a real valued r.v. X, its *distribution function* is $F(x) = P(X < x)$, $-\infty < x < \infty$. Thus $F(x)$ is monotone increasing, continuous from the left, $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$. If there is a function $f(t)$ such that $F(x) = \int_{-\infty}^{x} f(t) \, dt$, then $f(t)$ is the *density function* of X. We say that a sequence of r.vs (Y_n) *tends to* X in *distribution* if $\lim_{n \to \infty} P(Y_n < x) = P(X < x) = F(x)$, whenever x is a point of continuity of $F(x)$. The notation for convergence in distribution is $Y_n \overset{d}{\to} X$. Of course, convergence in distribution depends only on the distributions of the r.vs in question.

If h is any real-valued function on $\mathbb{R}$, the *expectation* of $h(x)$ is

$$E(h(X)) = \int_{\Omega} h(X) \, dp = \int_{-\infty}^{\infty} h(x) \, dF(x).$$

In particular, the *mean* of X, usually denoted by μ, is $E(X)$ and the *nth moment* of X is $E(X^n)$. Of course, these need not exist but, as they do exist for the r.vs we are going to consider, we shall assume that they exist. The *variance* of X is $\sigma^2(X) = E\{(X - \mu)^2\} = E(X^2) - \mu^2$ and the standard deviation is the non-negative square root of this.

If X is a non-negative r.v. with mean μ and $t > 0$, then

$$\mu \geqslant P(X \geqslant t\mu) t \mu.$$

Rewriting this slightly we get *Markov's inequality*:

$$P(X \geqslant t\mu) \leqslant 1/t. \tag{1}$$

Now let X be a real-valued r.v. with mean μ and variance σ^2. If $d > 0$, then clearly

$$E\{(X - \mu)^2\} \geqslant P(|X - \mu| \geqslant d) d^2$$

so we have *Chebyshev's inequality*:

$$P(|X - \mu| \geqslant d) \leqslant \sigma^2/d^2. \tag{2}$$

As a special case of this inequality we see that if $\mu \neq 0$, then

$$P(X = 0) \leqslant P(|X - \mu| \geqslant \mu) \leqslant \sigma^2/\mu^2. \tag{2'}$$

In fact, one can do a little better, for by the Cauchy inequality with $\Omega_0 = \{w : X(w) \neq 0\}$ we have

$$E(X)^2 = \left(\int_{\Omega_0} X \, dP \right)^2 \leqslant \left(\int_{\Omega_0} X^2 \, dP \right) \left(\int_{\Omega_0} 1 \, dP \right) = E(X^2) \{1 - P(X = 0)\}.$$

Hence

$$P(X = 0) \leqslant 1 - E(X)^2/E(X^2) = \sigma^2/(\mu^2 + \sigma^2).\qquad(3)$$

Most of the r.vs we encounter are non-negative integer valued, so unless it is otherwise indicated (for example, by the existence of the density function), we assume that the r.v. takes only non-negative integer values. The distribution of such a r.v., X, is given by the sequence

$$p_k = P(X = k), \qquad k = 0, 1, \ldots.$$

Clearly $p_k \geqslant 0$ and $\sum_{k=0}^{\infty} p_k = 1$. Then the mean of X is $\sum_{k=1}^{\infty} kp_k$ and the nth moment is $E(X^n) = \sum_{k=1}^{\infty} k^n p_k$. If $X, X_1, X_2, \ldots$ are non-negative integer valued r.vs then $X_n \overset{d}{\to} X$ if

$$\lim_{n \to \infty} P(X_n = k) = P(X = k) = p_k$$

for every k.

Write $\mathscr{L}(X)$ for the distribution (*law*) of a r.v. X. Given integer-valued r.vs X and Y, the *total variation distance* of $\mathscr{L}(X)$ and $\mathscr{L}(Y)$ is

$$d(\mathscr{L}(X), \mathscr{L}(Y)) = \sup\{|P(X \in A) - P(Y \in A)|: A \subset \mathbb{Z}\}.$$

With a slight abuse of notation occasionally we write $d(X, Y)$ or $d(X, \mathscr{L}(Y))$ instead of $d(\mathscr{L}(X), \mathscr{L}(Y))$.

Clearly $X_n \overset{d}{\to} X$ iff $d(X_n, X) \to 0$. Of course, any information about the speed of convergence of $d(X_n, X)$ to 0 is more valuable than the fact that X_n tends to X in distribution.

Given a probability space (Ω, Σ, P) and a set $C \in \Sigma$, $P(C) > 0$, the *probability of a set $A \in \Sigma$ conditional on C* is defined as

$$P(A|C) = P(A \cap C)/P(C).$$

Then $P_C = P(\cdot|C)$ is a probability measure on (Ω, Σ). A r.v. X is said to be taken *conditional on C* if it is considered as a function on (Ω, Σ, P_C); the expectation of this new r.v., denoted by $E(X|C)$, is said to be the expectation of X conditional on C.

Following Feller (1966) we use the notation $(x)_r = x(x-1)\ldots(x-r+1)$. Thus $(n)_n = (n)_{n-1} = n!$ and $\binom{n}{k} = (n)_k/(k)_k$. We define the rth *factorial moment* of a r.v. X as $E_r(X) = E\{(X)_r\}$. Thus if

$$P(X = k) = p_k,$$

then

$$E_r(X) = \sum_{k=r}^{\infty} p_k(k)_r.$$

Note that if X denotes the number of objects in a certain class then $E_r(X)$ is the expected number of *ordered r-tuples* of elements of that class.

The r.vs $X_1, X_2, \ldots$ are said to be *independent* if for each n

$$P(X_i = k_i, i = 1, \ldots, n) = \prod_{i=1}^{n} P(X_i = k_i)$$

for every choice of $k_1, k_2, \ldots, k_n$.

Note that $E(X + Y) = E(X) + E(Y)$ always holds and if $X_1, X_2, \ldots, X_n$ are independent, $E(X_i) = \mu_i$ and $E(X_i - \mu_i)^2 = \sigma_i^2$ then $E(\sum_i X_i) = \sum_i \mu_i$ and

$$\sigma^2\left(\sum_i X_i\right) = E\left[\left\{\sum_i (X_i - \mu_i)\right\}^2\right] = \sum_i \sigma_i^2.$$

In our calculations we shall often need the following rather sharp form of *Stirling's formula* proved by Robbins (1955):

$$n! = \left(\frac{n}{e}\right)^n \sqrt{2\pi n}\, e^{\alpha_n}, \tag{4}$$

where $1/(12n+1) < \alpha_n < 1/12n$.

Throughout the book we use Landau's notation $O\{f(n)\}$ for a term which, when divided by $f(n)$, remains bounded as $n \to \infty$. Similarly $h(n) = o\{g(n)\}$ means that $h(n)/g(n) \to 0$ as $n \to \infty$. Also, $h(n) \sim g(n)$ expresses the fact that $h(n)/g(n) \to 1$ as $n \to \infty$. Thus $h(n) \sim g(n)$ is equivalent to $h(n) - g(n) = o\{g(n)\}$. Note that a weak form of Stirling's formula (4) is $n! \sim (n/e)^n \sqrt{2\pi n}$. If the symbols o, O or $\sim$ are used without a variable, then we mean that the relation holds as $n \to \infty$.

An immediate consequence of (4) is that if $1 \leqslant m \leqslant n/2$, then

$$e^{-1/(6m)} \frac{1}{\sqrt{2\pi}} \left(\frac{n}{m}\right)^m \left(\frac{n}{n-m}\right)^{n-m} \left(\frac{n}{m(n-m)}\right)^{1/2} \leqslant \binom{n}{m}$$

$$\leqslant \frac{1}{\sqrt{2\pi}} \left(\frac{n}{m}\right)^m \left(\frac{n}{n-m}\right)^{n-m} \left(\frac{n}{m(n-m)}\right)^{1/2}. \tag{5}$$

On putting $p = m/n$ and $q = 1 - p$ we find that if $m \to \infty$ and $n - m \to \infty$, then

$$\binom{n}{m} \sim (2\pi)^{-1/2} (p^p q^q)^{-n} (pqn)^{-1/2}.$$