Probab. Theory Relat. Fields 110, 277–285 (1998)

Probability Theory and Related Fields

Some large deviation results for sparse random graphs

Neil O'Connell

BRIMS, Hewlett-Packard Labs, Filton Road, Stoke Gifford, Bristol BS12 6QZ, UK

Received: 14 November 1996 / In revised form: 15 August 1997

Summary. We obtain a large deviation principle (LDP) for the relative size of the largest connected component in a random graph with small edge probability. The rate function, which is not convex in general, is determined explicitly using a new technique. The proof yields an asymptotic formula for the probability that the random graph is connected.

We also present an LDP and related result for the number of isolated vertices. Here we make use of a simple but apparently unknown characterisation, which is obtained by embedding the random graph in a random directed graph. The results demonstrate that, at this scaling, the properties 'connected' and 'contains no isolated vertices' are not asymptotically equivalent. (At the threshold probability they are asymptotically equivalent.)

Mathematics Subject Classification (1991): 60F10, 05C80

1 Introduction

The central object of study in this paper is the random graph $\mathscr{G}(n,p)$, with $p = O(1/n)$. The random graph $\mathscr{G}(n,p)$ is constructed on n vertices by including each of the $n(n-1)/2$ potential edges independently with probability p. Much is known about the first order properties of such graphs. For example, if $p = c/n$ and $c > 1$, the size (in vertices) X_n of the largest connected component is asymptotically an, where $a > 0$ satisfies $a = 1 - \mathrm{e}^{-ac}$. (The precise statement is that the sequence X_n/n converges in probability to a.) The other components are all of order $\log n$ in size. This is the so-called 'giant component', and it only appears when $c > 1$. As with many other first order results for random graphs, little is known about the nature of fluctu-

Keywords: Giant component – Connectivity – Isolated vertices.

ations from the mean, other than some crude probability estimates. The main result of this paper is a large deviation principle for the sequence X_n/n. We also obtain an explicit expression for the rate function using a new technique. Some informal discussion and illustration of this technique is presented in [4].

We also present an LDP for the number of isolated vertices in the random graph $\mathcal{G}(n,c/n)$. Here we use a seemingly unknown characterisation of the distribution of the number of isolated vertices. As a kind of corollary (formally it is, technically it isn't) we demonstrate that the properties 'connected' and 'contains no isolated vertices' are not asymptotically equivalent at this scaling. At the threshold probability – that is, for the random graph $\mathcal{G}(n,p)$ with

$$p = \frac{\log n}{n} + \frac{c}{n}$$

– the two properties *are* asymptotically equivalent, with probability approaching $e^{-e^{-c}}$. This is a famous result due to Erdos and Rényi (see, for example, [5]).

The standard reference on random graphs is the book of Bollobás [1]; the lecture notes of Spencer [5] provide a useful introduction. For an overview of the main results on sparse random graphs, see [2]. The giant component is the subject of recent paper by Janson et al. [3], where some very sharp results are presented.

2 Preliminaries

For completeness we will record here some definitions and basic facts. Let Z_n be a sequence of random variables taking values in $\{0,1,\ldots,n\}$. A *rate function* on $[0,1]$ is a lower semicontinuous mapping I: $[0,1] \to [0,\infty]$ such that for all $y \in [0,\infty)$ the level set $\{x\colon\ I(x) \le y\}$ is closed in $[0,1]$. We say the sequence Z_n/n satisfies the LDP in $[0,1]$ with rate function I if, for all Borel sets B in $[0,1]$,

$$\begin{aligned} -\inf_{x\in B^\circ} I(x) &\le \liminf_n \tfrac{1}{n}\log P(Z_n/n \in B) \\ &\le \limsup_n \tfrac{1}{n}\log P(Z_n/n \in B) \\ &\le -\inf_{x\in \bar{B}} I(x) \ . \end{aligned}$$

An easy fact that we will make use of is the following: if I is continuous, and

$$\lim_n \tfrac{1}{n}\log P(Z_n = [xn]) = -I(x) \ , \tag{1}$$

uniformly for $x \in (0,1]$, then Z_n/n satisfies the LDP in $[0,1]$ with rate function I.

3 The largest connected component

Write $X(n,p)$ for the size (in vertices) of the largest connected component in the random graph $\mathcal{G}(n,p)$. It is well-known (see, for example, [5]) that, for $c > 1$, the sequence $X(n,c/n)/n$ converges in probability, as $n \to \infty$, to the unique positive solution to the equation $a = 1 - \mathrm{e}^{-ca}$, which we will denote by a_c. If $c \leq 1$, then $X(n,c/n)/n$ converges in probability to zero. For convenience we will set $a_c = 0$ for $0 < c \leq 1$.

The main result of this section is the following. Set $m(y) = \log(1 - \mathrm{e}^{-y})$.

Theorem 3.1 *For $c > 0$, the sequence $X(n,c/n)/n$ satisfies the* LDP *in* $[0,1]$ *with rate function given by*

$$I_c(x) = -kxm(cx) + kx\log x + (1-kx)\log(1-kx) + cx - k(k+1)cx^2/2$$

for $x_k \leq x \leq x_{k-1}$, where $x_0 = 1$ and

$$x_k = \sup\left\{x : \frac{x}{1-kx} = 1 - \mathrm{e}^{-cx}\right\} .$$

Note that if $c \leq 1$, then $x_1 = 0$ and

$$I_c(x) = -xm(cx) + x\log x + (1-x)\log(1-x) + cx(1-x)$$

on $[0,1]$, and this is a convex rate function. If $c > 1$, the rate function I_c is not convex. A plot of I_3 is shown in Fig. 1.

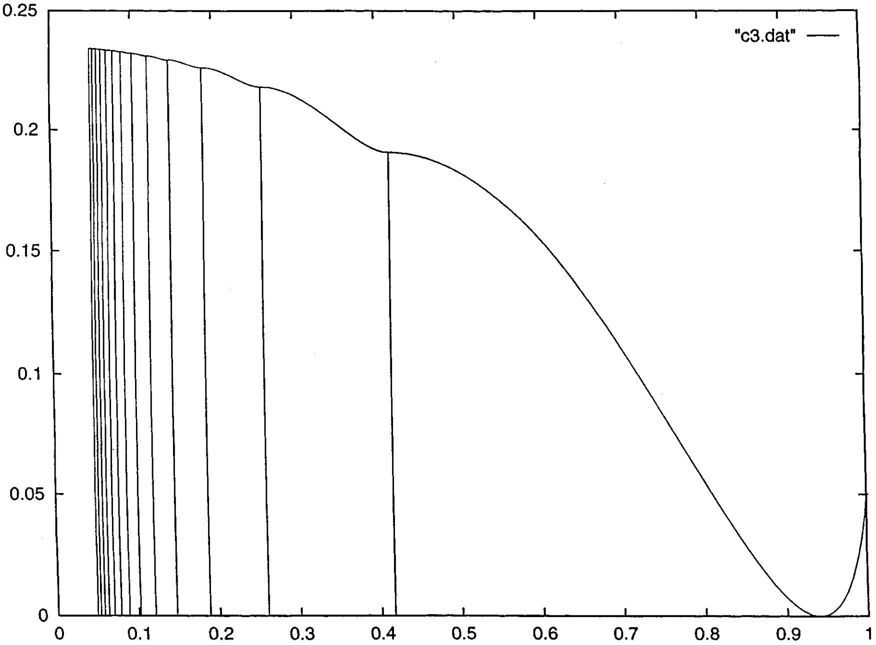

Fig. 1. A plot of the rate function I_3

The interpretation here is that the most likely way for $\mathscr{G}(n, c/n)$ to have a largest connected component of size $\approx xn$, when $x_k \leq x \leq x_{k-1}$, is for it to have exactly k connected components of that size.

Proof of Theorem 3.1. Let $q(n,p)$ denote the probability that $\mathscr{G}(n,p)$ is connected. For $1 \leq k \leq n$ we have

$$\begin{aligned}&\binom{n}{k}(1-c/n)^{k(n-k)}q(k,c/n)P(X(n-k,c/n)<k)\\&\quad\leq P(X(n,c/n)=k)\\&\quad\leq \binom{n}{k}(1-c/n)^{k(n-k)}q(k,c/n)P(X(n-k,c/n)\leq k)\ .\end{aligned}\tag{2}$$

The upper bound is just Boole's inequality; the lower bound is the probability of having *exactly one* component of size k, and none exceeding that size.

We will assume $c > 1$ until it is stated otherwise. Set

$$l_n(x,c) = \binom{n}{[xn]}(1-c/n)^{[xn](n-[xn])}q([xn],c/n)\ .$$

Since $X(n, c/n)/n \to a_c$ in probability we have that for any neighbourhood A of a_c,

$$\lim_n \max_{x\in A} \tfrac{1}{n}\log P(X(n,c/n)=[xn]) = 0\ . \tag{3}$$

Combining this with the inequalities (2) we see that

$$\limsup_n \tfrac{1}{n}\log l_n(a_c,c) \leq 0\ ,$$

and

$$\liminf_n \max_{x\in A} \tfrac{1}{n}\log l_n(x,c) \geq 0\ .$$

To proceed we need a technical lemma.

Lemma 3.1 *For $x \in (0,1]$ we have*

$$\lim_{\epsilon\searrow 0}\limsup_n \sup_{y:\ |x-y|<\epsilon}\left|\frac{1}{n}\log\frac{q([xn],c/n)}{q([yn],c/n)}\right| = 0\ .$$

Proof. If a graph on l vertices is connected then, for each $k < l$, there exists a subset of k vertices such that the restriction of the graph to those vertices is connected. (For example, one can choose the k vertices by performing a random walk on the graph until its range is k.) It follows that

$$q(l,p) \leq \binom{l}{k}q(k,p)\ .$$

Now let $l = [xn]$, $k = [yn]$ and $p = c/n$, and consider the normalised logarithmic limit to get that (for ϵ sufficiently small)

$$\limsup_n \sup_{x-\epsilon<y<x}\frac{1}{n}\log\frac{q([xn],c/n)}{q([yn],c/n)} \leq xh(1-\epsilon/x)$$

for $0 < y < x < 1$. For a lower bound we observe that, again for $l > k$,

$$\begin{aligned} q(l,p) &\geq \left[1-(1-p)^{l}\right]\left[1-(1-p)^{l-1}\right]\cdots\left[1-(1-p)^{k}\right]q(k,p) \\ &\geq \left[1-(1-p)^{k}\right]^{l-k}q(k,p) \end{aligned}$$

Here we are using, for the first inequality, the natural embedding of $\mathscr{G}(j-1,p)$ in $\mathscr{G}(j,p)$: $\mathscr{G}(j,p)$ is obtained from $\mathscr{G}(j-1,p)$ by adding a vertex and attaching it to each existing vertex independently with probability p; If $\mathscr{G}(j-1,p)$ is connected and the new vertex is not isolated, then $\mathscr{G}(j,p)$ is connected. Now let $l=[xn]$, $k=[yn]$ and $p=c/n$ as before, and consider the normalised logarithmic limit to get that

$$\liminf_{n} \inf_{x-\epsilon<y<x} \frac{1}{n}\log\frac{q([xn],c/n)}{q([yn],c/n)} \geq \epsilon\log\left(1-\mathrm{e}^{-c(x-\epsilon)}\right) \ .$$

The result follows. □

The above lemma allows us to conclude that

$$\lim_{n}\tfrac{1}{n}\log l_n(a_c,c)=0 \ ,$$

and hence that

$$\lim_{n}\tfrac{1}{n}\log q([a_c n],c/n)=a_c m(ca_c) \ ,$$

for some function m on $[0,\infty)$ satisfying

$$h(a_c)+a_c m(ca_c)-ca_c(1-a_c)=0 \ , \tag{4}$$

where

$$h(x)=-x\log x-(1-x)\log(1-x) \ .$$

Since $c>1$ is arbitrary, and the range of $c\mapsto ca_c$ is $(0,\infty)$, we can solve (4) to deduce that $m(y)=\log(1-\mathrm{e}^{-y})$. We have thus shown that, for $0<x<1$,

$$\lim_{n}\tfrac{1}{n}\log q([xn],c/n)=xm(cx) \ . \tag{5}$$

Now for any $d>0$, there exists $0<x<1$ and $c>1$ for which $d=cx$, so it follows from (5) that

$$\lim_{n}\tfrac{1}{n}\log q(n,d/n)=m(d) \ . \tag{6}$$

In particular, the convergence (5) can be extended to the case $x=1$ and it follows from Lemma 3.1 (the sequence has a kind of 'approximate equicontinuity') that the limit holds uniformly for $0<x\leq 1$.

The crux of the above argument, and hence the entire proof, is that it reveals an explicit scaling property of the function q, namely that

$$q(xn,c/n)\approx q(a_y k,y/k) \ ,$$

for appropriately chosen y and k.

It follows from (3) that for $x>(1-x)a_{c(1-x)}$ (that is, for $x>x_1$),

$$\lim_{n}\tfrac{1}{n}\log P(X([(1-x)n],c/n)\leq xn)=0 \ ,$$

and so we have

$$\lim_n \tfrac{1}{n} \log P(X(n,c/n) = [xn]) = h(x) + xm(cx) - cx(1-x) =: A(x,c) \tag{7}$$

uniformly on the interval $x_1 < x \le 1$. (Here we are using the easy facts that the limits

$$h(x) = \lim_n \frac{1}{n} \log \binom{n}{[xn]}$$

and

$$-cx(1-x) = \lim_n \tfrac{1}{n} \log \left[(1-c/n)^{[xn](n-[xn])} \right]$$

are uniform.) To determine the rate function on the entire interval we first need another lemma.

Lemma 3.2

$$\lim_n \tfrac{1}{n} \log P(X(n,c/n) < xn) = A(x,c)$$

uniformly on the interval $x_1 < x \le a_c$.

Proof. By considering the respective component sizes we enumerate the possibilities and apply the principle of the largest term to get that

$$\begin{aligned}
&\limsup_n \tfrac{1}{n} \log P(X(n,c/n) < x_1 n) \\
&\le \sup\{-y_1 \log y_1 - y_2 \log y_2 \cdots - y_{k+1} \log y_{k+1} \\
&\qquad + y_1 m(cy_1) + \cdots + y_k m(cy_k) \\
&\qquad - cy_1(1-y_1) - cy_2(1-y_1-y_2) - \cdots - cy_k y_{k+1} : \\
&\qquad k \in \mathbb{Z}_+; 0 < y_i < x_1, \forall i; x_1 < y_1 + \cdots + y_k = 1 - y_{k+1}\}
\end{aligned}$$

It is easy to check, using convexity arguments, that this supremum is acheived on the set $y_1 = y_2 = \cdots = y_k = x_1/k$. (The function $y \mapsto ym(cy) - y \log y$ is concave and the third line is amenable to an elementary inductive argument.) Hence,

$$\begin{aligned}
&\limsup_n \frac{1}{n} \log P(X(n,c/n) < x_1 n) \\
&\le \sup_k \left\{ -x_1 \log(x_1/k) - (1-x_1)\log(1-x_1) + x_1 m(cx_1/k) \right. \\
&\qquad \left. -cx_1 + \frac{k+1}{2k} cx_1^2 \right\} .
\end{aligned}$$

It is now tedious but straightforward to check that this supremum is attained at $k=1$, where it takes the value $A(x_1,c)$. Applying the principle of the largest term once again (using the uniform convergence in (7)), we have

$$\lim_n \tfrac{1}{n} \log P(X(n,c/n) < xn) = A(x,c)$$

for $x_1 < x \le a_c$; the uniformity of this convergence follows from the fact that the argument is monotone (in x) and A is continuous. □

Using this lemma, we can now recursively apply (7) and (2) on successive intervals $(x_k, x_{k-1}]$, $k = 2, 3, \ldots$ to get that

$$\lim_n \frac{1}{n} \log P(X(n, c/n) = [xn]) = \sum_{j=0}^{k-1} (1 - jx) A\left(\frac{x}{1 - jx}, c(1 - jx)\right)$$

uniformly on $(x_k, x_{k-1}]$. It is easily verified that this agrees with the formula for I_c in the statement of the theorem.

If $c \leq 1$ we can use (2), (3) and (6) to conclude that

$$\lim_n \tfrac{1}{n} \log P(X(n, c/n) = [xn]) = A(x, c)$$

uniformly on $(0, 1]$. This completes the proof of the theorem. □

We have also proved (6):

Theorem 3.2 *For any* $c > 0$,

$$\lim_n \tfrac{1}{n} \log P(\mathcal{G}(n, c/n) \text{ is connected}) = \log(1 - \mathrm{e}^{-c}) \ .$$

Bollobás [2] discusses some related results on connectedness. The closest in spirit to this result is the work of Wright [6] on the enumeration of connected graphs. For example, it is shown that $C(n, n+k)$, the number of connected graphs on n vertices with $n + k$ edges, is asymptotically

$$f_k n^{n+(3k-1)/2} \left\{1 + O(k^{3/2}/n)\right\}$$

for $k = o(n^{1/3})$. (A recursion is given for the sequence f_k.) This scaling has been useful for studying the largest connected component for the critical graph ($c = 1$) – see, for example, [2] – but there is insufficient information here about higher values of k to determine the asymptotics in Theorem 3.2. Bollobás [2, Corollary 5] obtains the universal upper bound

$$C(n, n+k) \leq C_0 2^{-k} n^{n+(3k-1)/2} \ ,$$

for some constant C_0, but for $\mathcal{G}(n, c/n)$ this leads to a trivial upper bound on the probability of connectedness.

We can also deduce the following result on the law of the number of edges in the random graph, given that it is connected.

Corollary 3.3 *Let* E_n^d *denote the number of edges present in the random graph* $\mathcal{G}(n, d/n)$. *For* $0 < c < d$,

$$\limsup_n \frac{1}{n} \log E\left[(c/d)^{E_n^d} \mid \mathcal{G}(n, d/n) \ \text{ is connected}\right] \leq \log\left(\frac{1 - \mathrm{e}^{-c}}{1 - \mathrm{e}^{-d}}\right) \ .$$

Proof. Using the fact that $\mathcal{G}(n, c/n)$ can be represented as the intersection of independent realisations of $\mathcal{G}(n, d/n)$ and $\mathcal{G}(n, c/d)$ on the same set of n vertices, we see that

$$q(n,c/n) \geq q(n,d/n)P\{\mathscr{G}(n,d/n) < \mathscr{G}(n,c/d) | \ \mathscr{G}(n,d/n) \text{ is connected}\}$$
$$= q(n,d/n)E\left[(c/d)^{E_n^d} | \ \mathscr{G}(n,d/n) \text{ is connected}\right] .$$

Here, $<$ denotes 'is a subgraph of'. The statement now follows from Theorem 3.2. □

4 The number of isolated vertices

Denote by $\mathscr{D}(n,q)$ the random directed graph constructed on n vertices, with each of the $n(n-1)$ potential directed edges included independently with probability q. It is clear that the number of isolated vertices in $\mathscr{G}(n,p)$, which we will denote by $V(n,p)$, has the same law as the number of isolated vertices in $\mathscr{D}(n, 1-\sqrt{1-p})$. Now the number of vertices Y in $\mathscr{D}(n,q)$ with no 'incoming' edges has a binomial distribution with parameters n and $(1-q)^{n-1}$; conditional on Y, the number of isolated vertices in $\mathscr{D}(n,q)$ has a binomial distribution with parameters Y and $(1-q)^{n-Y}$.

Thus, for $s \geq 0$,

$$Es^{V(n,p)} = \sum_{k=0}^{n} \binom{n}{k} r^k (1-r)^{n-k} \left[1-(1-s)(1-p)^{(n-k)/2}\right]^k ,$$

where $r = (1-p)^{(n-1)/2}$. Setting $p = c/n$ and applying the principle of the largest term we get,

$$\begin{aligned} f(s) &:= \lim_n \tfrac{1}{n} \log Es^{V(n,c/n)} \\ &= \sup_{0 \leq x \leq 1} \Big\{ x \log\left[1-(1-s)\mathrm{e}^{-c(1-x)/2}\right] + h(x) - cx/2 \\ &\qquad + (1-x)\log\left[1-\mathrm{e}^{-c/2}\right]\Big\} . \end{aligned}$$

The Gärtner-Ellis theorem does not apply here, because the scaled cumulant generating function $\Lambda(\theta) := f(\mathrm{e}^\theta)$ is not steep.

We can, however, deduce (setting $s = 0$) that

$$\begin{aligned} g(c) &:= \lim_n \tfrac{1}{n} \log P(V(n,c/n) = 0) \\ &= \sup_{0 \leq x \leq 1} \left\{ x \log\left[1-\mathrm{e}^{-c(1-x)/2}\right] + h(x) - cx/2 + (1-x)\log\left[1-\mathrm{e}^{-c/2}\right]\right\} . \end{aligned} \tag{8}$$

Observe that

$$P(V(n,p) = k) = \binom{n}{k}(1-p)^{n-1} \ldots (1-p)^{n-k} P(V(n-k,p) = 0) , \tag{9}$$

and so

$$\lim_n \tfrac{1}{n} \log P(V(n,c/n) = [xn]) = h(x) - cx(1-x/2) + (1-x)g(c(1-x)) .$$

Uniform convergence follows from the next lemma, the proof of which is identical to that of Lemma 3.1 (*exactly* the same bounds are used) and the fact that the mapping $x \mapsto (1-x)g((1-x)c)$ is continuous on $[0,1]$. Set $r(n,p) = P(V(n,p) = 0)$.

Lemma 4.1 *For* $x \in (0,1]$ *we have*

$$\lim_{\epsilon \searrow 0} \limsup_{n} \sup_{y:\,|x-y|<\epsilon} \left| \frac{1}{n} \log \frac{r([xn], c/n)}{r([yn], c/n)} \right| = 0 \ .$$

We have proved:

Theorem 4.1 *For* $c > 0$, *the sequence* $V(n,c/n)/n$ *satisfies the* LDP *in* $[0,1]$ *with rate function given by*

$$J_c(x) = -h(x) + cx(1-x/2) - (1-x)g(c(1-x)) \ .$$

Note that we could have used (9) and the easy fact that $V(n,c/n)/n$ converges in probability to e^{-c} to determine the limit function g, using the technique described in the introduction: $J_c(\mathrm{e}^{-c}) = 0$ implies that

$$g(d) = \log(d/a) - (a-d)^2/(2d) \ ,$$

where $a > 0$ satisfies $1 - \mathrm{e}^{-a} = d/a$. (Note that this also provides an easy alternative to solving the optimisation in (8)!)

Finally, one can verify that $g(d)$ is strictly bigger than $m(d)$, for each $d > 0$. As we remarked in the introduction, the properties ‘connected’ and ‘contains no isolated vertices’ are not asymptotically equivalent at this scaling.

5 References

[1] Bollobás, B.: Random Graphs. Academic Press 1985

[2] Bollobás, B.: The evolution of sparse random graphs. In: Bollobás, B. (ed.) Graph Theory and Combinatorics, Academic Press 1984

[3] Janson, S., Knuth, D., Luczak, T., Pittel, B.: The birth of the giant component. Rand. Struct. Alg. **4**, 233–358 (1993)

[4] O’Connell, N.: From laws of large numbers to large deviation principles. Markov Proc. Rel. Fields (to appear)

[5] Spencer, J.: Nine Lectures on Random Graphs. Ecole d’ Eté de Probabilite’s de Saint-Flour XXI-1991 (P.L. Hennequin, ed.) Lecture Notes in Mathematics 1541, Springer-Verlag. pp 293–347.

[6] Wright, E.M.: The number of connected sparsely edged graphs. J. Graph Theor. **1**, 317–330 (1977)