A Random Graph Model for Power Law Graphs

William Aiello, Fan Chung, and Linyuan Lu

CONTENTS

Research supported in part by NSF Grant No. DMS 98-01446

We propose a random graph model which is a special case of sparse random graphs with given degree sequences which satisfy a power law. This model involves only a small number of parameters, called logsize and log-log growth rate. These parameters capture some universal characteristics of massive graphs. From these parameters, various properties of the graph can be derived. For example, for certain ranges of the parameters, we will compute the expected distribution of the sizes of the connected components which almost surely occur with high probability. We illustrate the consistency of our model with the behavior of some massive graphs derived from data in telecommunications. We also discuss the threshold function, the giant component, and the evolution of random graphs in this model.

1. INTRODUCTION

Is the World Wide Web completely connected? If not, how big is the largest component, the second largest component, etc.? Anyone who has surfed the Web for any length of time will undoubtedly come away feeling that if there are disconnected components at all, they must be small and few in number. Is the Web too large, dynamic and structureless to answer these questions?

Probably yes, if the sizes of the largest components are required to be exact. Recently, however, some of the structure of the Web has come to light which may enable us to describe graph properties of the Web qualitatively. Kumar et al. [1999a; 1999b] and Kleinberg et al. [1999] have measured the degree sequences of the Web and shown that it is well approximated by a power law distribution. That is, the number of nodes, y, of a given degree x is proportional to $x^{-\beta}$ for some constant $\beta > 0$. This was reported independently by Albert, Barabási and Jeong [Albert et al. 1999; Barabási and Albert 1999; Barabási et al. 2000]. The power law distribution of the degree sequence appears to be a very robust property of the Web despite its dynamic nature. In

1058-6458/2001 $0.50 per page

fact, the power law distribution of the degree sequence may be a ubiquitous characteristic, applying to many massive real world graphs. Indeed, Abello et al. [1998] have shown that the degree sequence of *call graphs* is nicely approximated by a power law distribution. Call graphs are graphs of calls handled by some subset of telephony carriers for a specific time period. Faloutsos et al. [1999] have shown that the degree sequence of the Internet router graph also follows a power law.

Just as many other real world processes have been effectively modeled by appropriate random models, in this paper we propose a parsimonious random graph model for graphs with a power law degree sequence. We then derive connectivity results that hold with high probability in various regimes of our parameters. Finally, we compare the results from the model with the exact connectivity structure for some call graphs computed by Abello et al. [1998].

An extended abstract of this paper has appeared in the Proceedings of the Thirtysecond Annual ACM Symposium on Theory of Computing 2000 [Aiello et al. 2000]. In this paper, we have included the complete proofs for the main theorems and additional theorems focused on the second largest components of power graphs in various ranges. In addition, we give some recent references; see also [Hayes 2000].

Power Law Random Graphs

The study of random graphs dates back to the seminal papers of Erdős and Rényi [1960; 1961], which laid the foundation for the theory. There are three standard models for what we will call in this paper *uniform* random graphs [Alon and Spencer 1992]. Each has two parameters, one controlling the number of nodes in the graph and the other the density or number of edges. For example, the random graph model $G(n, e)$ assigns uniform probability to all graphs with n nodes and e edges, while in the random graph model $\mathcal{G}(n, p)$ each edge is chosen with probability p.

Our *power law* random graph model $P(\alpha, \beta)$ also has two parameters. They only roughly delineate the size and density, but they are natural and convenient for describing a power law degree sequence. The model is described as follows. Let y be the number of nodes with degree x. $P(\alpha, \beta)$ assigns uniform probability to all graphs with $y = e^\alpha / x^\beta$ (where self loops are allowed). Note that α is the intercept and β is the (negative) slope when the degree sequence is plotted on a log-log scale.

There is also an alternative power law random graph model analogous to the uniform graph model $\mathcal{G}(n, p)$. Instead of having a fixed degree sequence, the random graph has an expected degree sequence distribution. The two models are basically asymptotically equivalent, subject to a bounding of error estimates for the variances; this will be further discussed in [Aiello et al. $\geq$ 2001].

Our Results

Just as for the uniform random graph model where graph properties are studied for certain regimes of the density parameter and shown to hold with high probability asymptotically in the size parameter, in this paper we study the connectivity properties of $P(\alpha, \beta)$ as a function of the power β which hold almost surely for sufficiently large graphs. Briefly, we show that when $\beta < 1$, the graph is almost surely connected. For $1 < \beta < 2$ there is a giant component, that is, a component of size $\Theta(n)$. Moreover, all smaller components are of size $O(1)$. For $2 < \beta < \beta_0 \approx 3.4785$ there is a giant component and all smaller components are of size $O(\log n)$. For $\beta = 2$ the smaller components are of size $O(\log n / \log\log n)$. For $\beta > \beta_0$ the graph almost surely has no giant component. In addition we derive several results on the sizes of the second largest component. For example, we show that for $\beta > 4$ the number of components of given sizes can be approximated by a power law as well.

Previous Work

Strictly speaking our model is a special case of random graphs with a given degree sequence, for which there is a large literature. For example, Wormald [1981] studied the connectivity of graphs whose degrees are in an interval $[r, R]$, where $r \geq 3$. Łuczak [1992] considered the asymptotic behavior of the largest component of a random graph with given degree sequence as a function of the number of vertices of degree 2. His result was further improved by Molloy and Reed [1995; 1998], who considered a random graph on n vertices with the following degree distribution. The number of vertices of degree $0, 1, 2, \ldots$

are about $\lambda_0 n$, $\lambda_1 n$, ..., respectively, where the λ's sum to 1. In [Molloy and Reed 1995] it is shown that if $Q = \sum_i i(i-2)\lambda_i > 0$ and the maximum degree is not too large, then such random graphs have a giant component with probability tending to 1 as n goes to infinity, while if $Q < 0$ then all components are small with probability tending to 1 as $n \to \infty$. The paper also examines the threshold behavior of such graphs. In this paper, we will apply these techniques to deal with the special case that concerns our model.

Several other papers have taken an approach to modeling power law graphs different from the one taken here [Aiello et al. $\geq$ 2001; Barabási and Albert 1999; Barabási et al. 2000; Kleinberg et al. 1999; Kumar et al. 1999b]. The essential idea of these papers is to define a random process for growing a graph by adding nodes and edges. The intent is to show that the defined processes asymptotically yield graphs with a power law degree sequence with very high probability. While interesting and important, this approach has several difficulties. First, the models are difficult to analyze rigorously, since the transition probabilities are themselves dependent on the the current state. For example, [Barabási and Albert 1999; Barabási et al. 2000] implicitly assume that the probability that a node has a given degree is a continuous function. Kumar et al. [2001] offer a partial analysis of the situation. Second, while the models may generate graphs with power law degree sequences, it remains to be seen if they generate graphs that duplicate other structural properties of the Web, the Internet, and call graphs. For example, the model in [Barabási and Albert 1999; Barabási et al. 2000] cannot generate graphs with a power law other than c/x^3. Moreover, all the graphs can be decomposed into m disjoint trees, where m is a parameter of the model. The (α, β) model in [Kumar et al. 1999b] is able to generate graphs for which the power law for the indegree is different than the power law for the outdegree as is the case for the Web. However, to do so, the model requires that there be nodes that have only indegree and no outdegree and vice versa. While this may be appropriate for call graphs (e.g., customer service numbers) it may not be right for modeling the Web. Thus, while the random graph generation approach holds the promise of accurately predicting a wide variety of structural properties of many real world massive graphs, much work remains to be done.

In this paper we take a different approach. We do not attempt to answer how a graph comes to have a power law degree sequence. Rather, we take that as a given. In our model, all graphs with a given power law degree sequence are equiprobable. The goal is to derive structural properties that hold with probability asymptotically approaching 1. Such an approach, while potentially less accurate than the detailed modeling approach above, has the advantage of being robust: the structural properties derived in this model will be true for the vast majority of graphs with the given degree sequence. Thus, we believe that this model will be an important complement to random graph generation models.

The power law random graph model will be described in detail in the next section. In Sections 3 and 4, our results on connectivity will be derived. Section 5 discusses the sizes of the second largest components. Section 6 compares the results of our model to exact connectivity data for call graphs. A short list of open questions concludes the article.

A subsequent paper [Aiello et al. $\geq$ 2001] examines further several aspects of power law graphs, including their evolution, their "scale invariance", and the asymmetry of in-degrees and out-degrees.

2. A RANDOM GRAPH MODEL

We consider a random graph with the following degree distribution depending on two given values α and β. There are y vertices of degree x, where x and y satisfy

$$\log y = \alpha - \beta \log x.$$

In other words,

$$\left|\{v : \deg v = x\}\right| = y = \frac{e^\alpha}{x^\beta}.$$

Basically, α is the logarithm of the size of the graph and β is the log-log growth rate of the graph.

The number of edges should be an integer. To be precise, the expression above for y should be rounded down to $\lfloor e^\alpha / x^\beta \rfloor$. If we use real numbers instead of rounding down to integers, this may cause some error terms in further computation, but we will see that the error terms can be easily bounded. For simplicity and convenience, we will use real numbers with the understanding that the actual numbers are their integer parts. Another constraint is that the sum of the degrees should be even. This can be assured by adding a vertex of degree 1 if the sum is odd. Furthermore, for simplicity, we assume here that there are no isolated vertices.

We can deduce the following facts for our graph:

(1) The maximum degree of the graph is $e^{\alpha/\beta}$. Note that $0 \le \log y = \alpha - \beta \log x$.

(2) The number n of vertices can be computed as follows: By summing $y(x)$ for x from 1 to $e^{\alpha/\beta}$, we have

$$n = \sum_{x=1}^{e^{\alpha/\beta}} \frac{e^\alpha}{x^\beta} \approx \begin{cases} \zeta(\beta)\, e^\alpha & \text{if } \beta > 1, \\ \alpha e^\alpha & \text{if } \beta = 1, \\ e^{\alpha/\beta}/(1-\beta) & \text{if } 0 < \beta < 1, \end{cases}$$

where $\zeta(t) = \sum_{n=1}^{\infty} n^{-t}$ is the Riemann zeta function.

(3) The number of edges E is given by

$$E = \frac{1}{2} \sum_{x=1}^{e^{\alpha/\beta}} x \frac{e^\alpha}{x^\beta} \begin{cases} \frac{1}{2}\zeta(\beta-1)\, e^\alpha & \text{if } \beta > 2, \\ \frac{1}{4}\alpha e^\alpha & \text{if } \beta = 2, \\ \frac{1}{2} e^{2\alpha/\beta}/(2-\beta) & \text{if } 0 < \beta < 2. \end{cases}$$

The excess of the real numbers in (1)–(3) over their integer parts can be estimated as follows: For the number n of vertices, the error term is at most $e^{\alpha/\beta}$. For $\beta \ge 1$, it is $o(n)$, which is a lower order term. For $0 < \beta < 1$, the error term for n is relatively large. In this case, we have

$$n \ge \frac{e^{\alpha/\beta}}{1-\beta} - e^{\alpha/\beta} = \frac{\beta e^{\alpha/\beta}}{1-\beta}.$$

Therefore, n has same magnitude as $e^{\alpha/\beta}/(1-\beta)$.

The number E of edges can be treated similarly. For $\beta \ge 2$, the error term of E is $o(E)$, a lower order term. For $0 < \beta < 2$, E has the same magnitude as in the formula of item (3). In this paper we deal mainly with the case $\beta > 2$. The case $0 < \beta < 2$ is considered only in the next section, where we refer to $2-\beta$ as a constant. By using real numbers instead of rounding down to their integer parts, we simplify the arguments without affecting the conclusions.

To study the random graph model, we must consider large n. We say a property holds almost surely (a. s.) if the probability that it holds tends to 1 as the number n of the vertices goes to infinity. Thus we consider α to be large but β is fixed.

We use the following random graph model for a given degree sequence:

The model:

1. Form a set L containing $\deg v$ distinct copies of each vertex v.
2. Choose a random matching of the elements of L.
3. For two vertices u and v, the number of edges joining u and v is equal to the number of edges in the matching of L joining copies of u to copies of v.

We remark that the graphs that we are considering are in fact multi-graphs, possibly with loops. This model is a natural extension of the model for k-regular graphs, formed by combining k random matchings. For references and undefined terminology, see [Alon and Spencer 1992; Wormald 1999].

This random graph model is slightly different from the uniform selection model $P(\alpha, \beta)$ described in Section 1.1. However, by using the techniques of [Molloy and Reed 1998, Lemma 1], it can be shown that if a random graph with a given degree sequence a. s. has property P under one of these two models, then it a. s. has property P under the other model, provided some general conditions are satisfied.

3. THE CONNECTED COMPONENTS

Molloy and Reed [1995] showed that for a random graph with $(\lambda_i + o(1))n$ vertices of degree i, where the λ_i are nonnegative values that sum to 1, the giant component emerges when

$$Q := \sum_{i \ge 1} i(i-2)\lambda_i > 0,$$

so long as the maximum degree is less than $n^{1/4-\varepsilon}$. They also show that almost surely there is no giant component when $Q < 0$ and the maximum degree is less than $n^{1/8-\varepsilon}$.

Here we compute Q for our (α, β)-graphs:

$$Q = \sum_{x=1}^{e^{\alpha/\beta}} x(x-2) \left\lfloor \frac{e^\alpha}{x^\beta} \right\rfloor \approx \sum_{x=1}^{e^{\alpha/\beta}} \frac{e^\alpha}{x^{\beta-2}} - 2 \sum_{x=1}^{e^{\alpha/\beta}} \frac{e^\alpha}{x^{\beta-1}}$$
$$\approx \big(\zeta(\beta-2) - 2\zeta(\beta-1)\big) e^\alpha \text{ if } \beta > 3.$$

We are thus led to consider the value $\beta_0 \approx 3.47875$, which is a solution to

$$\zeta(\beta-2) - 2\zeta(\beta-1) = 0.$$

If $\beta > \beta_0$, we have

$$\sum_{x=1}^{e^{\alpha/\beta}} x(x-2)\left\lfloor \frac{e^\alpha}{x^\beta} \right\rfloor < 0.$$

We summarize our results here:

1. When $\beta > \beta_0 = 3.47875\ldots$, the random graph almost surely has no giant component. When $\beta < \beta_0 = 3.47875\ldots$, there is almost surely a unique giant component.
2. When $2 < \beta < \beta_0 = 3.47875\ldots$, almost surely the second largest components have size $\Theta(\log n)$. For any $2 \le x < \Theta(\log n)$, there is almost surely a component of size x.
3. When $\beta = 2$, almost surely the second largest components are of size $\Theta(\log n/\log\log n)$. For any $2 \le x < \Theta(\log n/\log\log n)$, there is almost surely a component of size x.
4. When $1 < \beta < 2$, the second largest components are almost surely of size $\Theta(1)$. The graph is almost surely not connected.
5. When $0 < \beta < 1$, the graph is almost surely connected.
6. For $\beta = \beta_0 = 3.47875\ldots$, the case is complicated. It corresponds to the double jump of a random graph $\mathcal{G}(n, p)$ with $p = 1/n$.
7. For $\beta = 1$, there is a nontrivial probability for either case: that the graph is connected or disconnected.

We remark that for $\beta > 8$, Molloy and Reed's result immediately implies that almost surely there is no giant component. When $\beta \le 8$, additional analysis is needed to deal with the degree constraints. We will prove in Theorem 4.2 that almost surely there is no giant component when $\beta > \beta_0$. In Section 5, we will deal with the range $\beta < \beta_0$. We will show in Theorem 5.1 that almost surely there is a unique giant component when $\beta < \beta_0$. Furthermore, we will determine the size of the second largest component within a constant factor.

4. THE SIZES OF CONNECTED COMPONENTS IN CERTAIN RANGES FOR β

For $\beta > \beta_0 = 3.47875\ldots$, almost surely there is no giant component. This range is of special interest since it is quite useful later for describing the distribution of small components.

Theorem 4.1. *For (α, β)-graphs with $\beta > 4$, the distribution of the number of connected components is as follows:*

1. *For each vertex v of degree $d = \Omega(1)$, let τ be the size of the connected component containing v. Then*
$$\Pr\left(\left|\tau - \frac{d}{c_1}\right| > \frac{2\lambda}{c_1}\sqrt{\frac{dc_2}{c_1}}\right) \le \frac{2}{\lambda^2},$$
where
$$c_1 = 2 - \frac{\zeta(\beta-2)}{\zeta(\beta-1)} \quad and \quad c_2 = \frac{\zeta(\beta-3)}{\zeta(\beta-1)} - \left(\frac{\zeta(\beta-2)}{\zeta(\beta-1)}\right)^2$$
are constants and where $\lambda = d^\varepsilon$ with ε an arbitrary small positive number and d a (slowly) increasing function of n. In other words, the vertex v almost surely belongs to a connected component of size
$$\frac{d}{c_1} + O(d^{1/2+\varepsilon}).$$
2. *The number of connected components of size x is almost surely at least*
$$(1 + o(1))\frac{e^\alpha}{c_1^{\beta-1} x^\beta}.$$
and at most
$$c_3 \frac{e^\alpha \log^{\beta/2-1} n}{x^{\beta/2+1}},$$
where
$$c_3 = \frac{4^{1+\beta} c_2}{(\beta-2) c_1^{1+\beta}}$$
is a constant depending only on β.
3. *A connected component of the (α, β)-graph almost surely has size at most*
$$e^{2\alpha/(\beta+2)}\alpha = \Theta(n^{2/(\beta+2)} \log n).$$

In our proof of this result we use the second moment, whose convergence depends on $\beta > 4$. In fact for $\beta \le 4$ the second moment diverges as the size of the graph goes to infinity, so our method no longer applies.

Theorem 4.1 strengthens the following result — which can be derived from [Molloy and Reed 1995, Lemma 3] — for the range of $\beta > 4$.

Theorem 4.2. *For $\beta > \beta_0 = 3.47875\ldots$, a connected component of the (α, β)-graph almost surely has size at most $Ce^{2\alpha/\beta}\alpha = \Theta(n^{2/\beta} \log n)$, where $C = 16/c_1^2$ is a constant depending only on β.*

The proof of Theorem 4.2, which we briefly describe here because it is needed in proving Theorem 4.1, uses the branching process method. Pick any vertex v in the graph, expose its neighbors, then the neighbors of its neighbors, repeating until the entire component is exposed. We expose only one vertex at each stage. At stage i, let L_i the set of vertices exposed and X_i be the random variable that counts the number of vertices in L_i. We mark all vertices in L_i as either live or dead. A vertex in L_i whose neighbors have not all been exposed yet is marked live. One whose neighbors have all been exposed is marked dead. Let O_i be the set of live vertices and Y_i the random variable that is the number of vertices in O_i. At each step we mark exact one dead vertex, so the total number of dead vertices at the i-th step is i. We have $X_i = Y_i + i$. Initially we assign $L_0 = O_0 = \{v\}$. Then at stage $i \geq 1$, we do the following:

1. If $Y_{i-1} = 0$, stop and output X_{i-1}.
2. Otherwise, randomly choose a live vertex u from O_{i-1} and expose its neighbors in N_u. Then mark u dead and mark each vertex live if it is in N_u but not in L_{i-1}. We have
$$\begin{aligned} L_i &= L_{i-1} \cup N_u, \\ O_i &= (O_{i-1} \setminus \{u\}) \cup (N_u \setminus L_{i-1}). \end{aligned}$$

Suppose that v has degree d. Then $X_1 = d+1$, and $Y_1 = d$. Eventually Y_i will hit 0 if i is large enough. Let τ denote the stopping time of Y, namely, $Y_\tau = 0$. Then $X_\tau = Y_\tau + \tau = \tau$ measures the size of the connected component. We first compute the expected value of Y_i and then use Azuma's Inequality [Molloy and Reed 1995] to prove Theorem 4.2.

Suppose that vertex u is exposed at stage i. Then $N_u \cap L_{i-1}$ contains at least one vertex v, which was exposed to reach u. However, $N_u \cap L_{i-1}$ may contain more than one vertex. We call an edge from u to a vertex in L_{i-1} other than v a *backedge*. Backedges cause the exploration to stop sooner, especially when the component is large. However in our case $\beta > \beta_0 = 3.47875\ldots$, the contribution of backedges is quite small. We set $Z_i = \#\{N_u\}$ and $W_i = \#\{N_u \cap L_{i-1}\} - 1$, so Z_i measures the degree of the vertex exposed at stage i, while W_i measures the number of backedges. By definition,
$$Y_i - Y_{i-1} = Z_i - 2 - W_i.$$

We have
$$\begin{aligned} E(Z_i) &= \sum_{x=1}^{e^{\alpha/\beta}} x \frac{x(e^\alpha/x^\beta)}{E} = \frac{e^\alpha}{E} \sum_{x=1}^{e^{\alpha/\beta}} x^{2-\beta} \\ &= \frac{\zeta(\beta-2) + O(n^{3/\beta-1})}{\zeta(\beta-1) + O(n^{2/\beta-1})} \\ &= \frac{\zeta(\beta-2)}{\zeta(\beta-1)} + O(n^{3/\beta-1}). \end{aligned}$$

Now we bound W_i. Suppose there are m edges exposed at stage $i-1$. Then the probability that a new neighbor is in L_{i-1} is at most m/E. We have
$$\begin{aligned} E(W_i) &\leq \sum_{x=1}^{\infty} x\Big(\frac{m}{E}\Big)^x = \frac{m/E}{(1-m/E)^2} \\ &= \frac{m}{E} + O\Big(\Big(\frac{m}{E}\Big)^2\Big), \end{aligned} \tag{4–1}$$
provided that $m/E = o(1)$.

When $i \leq Ce^{2\alpha/\beta}\alpha$, m is at most $ie^{\alpha/\beta} \leq Ce^{3\alpha/\beta}\alpha$. Hence,
$$\frac{m}{E} = O(n^{3/\beta-1}\log n) = o(1).$$

We have
$$\begin{aligned} E(Y_i) &= Y_1 + \sum_{j=2}^{i} E(Y_j - Y_{j-1}) \\ &= d + \sum_{j=2}^{i} E(Z_j - 2 - W_j) \\ &= d + (i-1)\Big(\frac{\zeta(\beta-2)}{\zeta(\beta-1)} - 2\Big) - i\,O(n^{3/\beta-1}\log n) \\ &= d - c_1(i-1) + i\,o(1). \end{aligned}$$

Proof of Theorem 4.2. Since $|Y_j - Y_{j-1}| \leq e^{\alpha/\beta}$, by Azuma's martingale inequality, we have
$$\Pr\big(|Y_i - E(Y_i)| > t\big) \leq 2e^{-t^2/(2ie^{2\alpha/\beta})},$$
where $i = (16/c_1^2)e^{2\alpha/\beta}\log n$ and $t = \frac{1}{2}c_1 i$. Since
$$\begin{aligned} E(Y_i) + t &= d - c_1(i-1) + io(1) + \tfrac{1}{2}c_1 i \\ &= -\tfrac{1}{2}c_1 i + d + c_1 + io(1) < 0, \end{aligned}$$
we have
$$\begin{aligned} \Pr\big(\tau > (16/c_1^2)e^{\alpha/\beta}\log n\big) &= \Pr\big(\tau > i\big) \leq \Pr(Y_i \geq 0) \\ &\leq \Pr\big(Y_i > E(Y_i) + t\big) \\ &\leq 2\exp -t^2/2ie^{2\alpha/\beta} = \frac{2}{n^2}. \end{aligned}$$

Hence, the probability that there exists a vertex v such that v lies in a component of size greater than

$$\frac{16}{c_1^2} e^{2\alpha/\beta} \log n$$

is at most

$$n\frac{2}{n^2} = \frac{2}{n} = o(1). \qquad \square$$

The proof of Theorem 4.1 uses the methodology above as a starting point while introducing the calculation of the variance of the above random variables.

Proof of Theorem 4.1. We follow the notation and previous results of Section 4. Under the assumption $\beta > 4$, we consider

$$\begin{aligned}
\mathrm{Var}(Z_i) &= \sum_{x=1}^{e^{\alpha/\beta}} x^2 \frac{x(e^\alpha/x^\beta)}{E} - E(Z_i)^2 \\
&= \frac{e^\alpha}{E} \sum_{x=1}^{e^{\alpha/\beta}} x^{3-\beta} - E(Z_i)^2 \\
&= \frac{\zeta(\beta-3) + O(n^{4/\beta-1})}{\zeta(\beta-1) + O(n^{2/\beta-1})} - \Big(\frac{\zeta(\beta-2)}{\zeta(\beta-1)}\Big)^2 \\
&\qquad + O(n^{3/\beta-1}) \\
&= \frac{\zeta(\beta-3)}{\zeta(\beta-1)} - \Big(\frac{\zeta(\beta-2)}{\zeta(\beta-1)}\Big)^2 + O(n^{4/\beta-1}) \\
&= c_2 + o(1),
\end{aligned}$$

since $\beta > 4$.

We need to compute the covariance. There are models for random graphs in which the edges are in dependently chosen. Then, Z_i and Z_j are independent. However, in the model based on random matchings, there is a small correlation. For example, $Z_i = x$ slightly effects the probability of $Z_j = y$. Namely, $Z_j = x$ has slightly less chance, while $Z_j = y \neq x$ has slightly more chance. Both differences can be bounded by

$$\frac{1}{E-1} - \frac{1}{E} \leq \frac{2}{E^2}.$$

Hence

$$\begin{aligned}
\mathrm{Covar}(Z_i, Z_j) &\leq E(Z_i)E(2/E^2) \\
&= O\Big(\frac{1}{n}\Big) \quad \text{if } i \neq j.
\end{aligned}$$

Now we will bound W_i. Suppose that there are m edges exposed at stage $i-1$. Then the probability that a new neighbor is in L_{i-1} is at most m/E. We have

$$\begin{aligned}
\mathrm{Var}(W_i) &\leq \sum_{x=1}^{\infty} x^3 \left(\frac{m}{E}\right)^x - E(W_i)^2 \\
&= \frac{m/E(m/E+1)}{(1-m/E)^3} - O\Big(\Big(\frac{m}{E}\Big)^2\Big) \\
&= \frac{m}{E} + O\Big(\Big(\frac{m}{E}\Big)^2\Big),
\end{aligned}$$

$$\begin{aligned}
\mathrm{Covar}(W_i, W_j) &\leq \sqrt{\mathrm{Var}(W_i)\,\mathrm{Var}(W_j)} \\
&\leq \frac{m}{E} + O\Big(\Big(\frac{m}{E}\Big)^2\Big),
\end{aligned}$$

$$\mathrm{Covar}(Z_i, W_j) \leq \sqrt{\mathrm{Var}(Z_i)\,\mathrm{Var}(W_j)} = O\left(\sqrt{\frac{m}{E}}\right).$$

When $i = O(e^{\alpha/\beta})$, $m \leq ie^{\alpha/\beta} = O(e^{2\alpha/\beta})$, we have

$$\begin{aligned}
E(Y_i) &= d + (i-1)\Big(\frac{\zeta(\beta-2)}{\zeta(\beta-1)} - 2\Big) + iO(n^{3/\beta-1}) + i\frac{m}{E} \\
&= d - (i-1)c_1 + O(n^{4/\beta-1}) \\
&= d - (i-1)c_1 + o(1)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(Y_i) &= \mathrm{Var}\left(d + \sum_{j=2}^{i}(Y_j - Y_{j-1})\right) \\
&= \mathrm{Var}\left(\sum_{j=2}^{i}(Z_j - W_j)\right) \\
&= \sum_{j=2}^{i}\big(\mathrm{Var}(Z_j) + \mathrm{Var}(W_j)\big) \\
&\quad + \sum_{2\leq j\neq k\leq i}\big(\mathrm{Covar}(Z_j, Z_k) - \mathrm{Covar}(Z_j, W_k) \\
&\qquad + \mathrm{Covar}(W_j, W_k)\big) \\
&= ic_2 + i\,o(1) + i^2\big(O(1/n) + O(\sqrt{e^{(2/\beta-1)\alpha}}) \\
&\qquad + O(e^{(2/\beta-1)\alpha})\big) \\
&= ic_2 + i\,o(1) + i\big(O(e^{(2/\beta-1/2)\alpha}) + O(e^{(3/\beta-1)\alpha})\big) \\
&= ic_2 + i\,o(1).
\end{aligned}$$

Chebyshev's inequality gives

$$\Pr\big(|Y_i - E(Y_i)| > \lambda\sigma\big) < \frac{1}{\lambda^2},$$

where σ is the standard deviation of Y_i, and $\sigma = \sqrt{ic_2} + o(\sqrt{i})$. Set

$$i_1 = \left\lfloor \frac{d}{c_1} - \frac{2\lambda}{c_1}\sqrt{\frac{dc_2}{c_1}} \right\rfloor, \quad i_2 = \left\lceil \frac{d}{c_1} + \frac{2\lambda}{c_1}\sqrt{\frac{dc_2}{c_1}} \right\rceil.$$

Then

$$\begin{aligned} E(Y_{i_1})-\lambda\sigma &= d-(i_1-1)c_1+o(1)-\big(\lambda\sqrt{c_2 i_1}+o(\sqrt{i_1})\big) \\ &\geq 2\lambda\sqrt{\frac{dc_2}{c_1}}-\lambda\sqrt{c_2\frac{d}{c_1}}-o(\sqrt{d}) \\ &= \lambda\sqrt{\frac{dc_2}{c_1}}-o(\sqrt{d}) > 0. \end{aligned}$$

Hence,

$$\begin{aligned} \Pr(\tau < i_1) &\leq \Pr(Y_{i_1} \leq 0) \\ &\leq \Pr\big(Y_{i_1} < E(Y_{i_1}) - \lambda\sigma\big) \leq \frac{1}{\lambda^2}. \end{aligned}$$

Similarly,

$$\begin{aligned} E(Y_{i_2})+\lambda\sigma &= d-(i_2-1)c_1+o(1)+\big(\lambda\sqrt{c_2 i_2}+o(\sqrt{i_2})\big) \\ &\geq -2\lambda\sqrt{\frac{dc_2}{c_1}}+\lambda\sqrt{c_2\frac{d}{c_1}}+o(\sqrt{d}) \\ &= -\lambda\sqrt{\frac{dc_2}{c_1}}+o(\sqrt{d}) < 0. \end{aligned}$$

Hence,

$$\begin{aligned} \Pr(\tau > i_2) &\leq \Pr(Y_{i_2} > 0) \\ &\leq \Pr\big(Y_{i_2} > E(Y_{i_2}) + \lambda\sigma\big) \leq \frac{1}{\lambda^2}. \end{aligned}$$

Therefore

$$\Pr\left(\left|\tau - \frac{d}{c_1}\right| > \frac{2\lambda}{c_1}\sqrt{\frac{dc_2}{c_1}}\right) \leq \frac{2}{\lambda^2}.$$

For a fixed v and λ a function slowly increasing to infinity, the preceding inequality implies that almost surely we have $\tau = d/c_1 + O(\lambda\sqrt{d})$.

Almost all components generated by vertices of degree x have size about d/c_1. One such component can have at most about $1/c_1$ vertices of degree d. Hence, the number of components of size d/c_1 is at least $c_1 e^{\alpha/\beta}/d^\beta$. Let $d = c_1 x$. Then the number of components of size x is at least

$$\frac{e^{\alpha/\beta}}{c_1^{\beta-1}x^\beta}\big(1+o(1)\big).$$

The argument above actually gives the following result. The size of every component whose vertices have degree at most d_0 is almost surely $Cd_0^2\log n$, where $C = 16/c_1^2$ is the same constant as in Theorem 4.2. Set $x = Cd_0^2\log n$ and consider the number of components of size x. A component of size x almost surely contains at least one vertex of degree greater than d_0.

For each vertex v with degree $d \geq d_0$, by part 1 in the statement of Theorem 4.1, we have

$$\Pr\left(\left|\tau - \frac{d}{c_1}\right| > \frac{2\lambda_d}{c_1}\sqrt{\frac{dc_2}{c_1}}\right) \leq \frac{2}{\lambda_d^2}.$$

Letting

$$\lambda_d = \frac{c_1 C d_0^2 \log n}{4}\sqrt{\frac{c_1}{c_2 d}},$$

we have

$$\begin{aligned} \Pr(\tau \geq Cd_0^2\log n) &\leq \Pr\left(\tau > \frac{d}{c_1} + \frac{2\lambda_d}{c_1}\sqrt{\frac{dc_2}{c_1}}\right) \\ &\leq C_3\frac{d}{d_0^4\log^2 n}, \end{aligned}$$

where $C_3 = 32c_2/(c_1^3C^2) = c_1c_2/8$ is a constant depending only on β. Since there are only e^α/d^β vertices of degree d, the number of components of size at least x is at most

$$\begin{aligned} \sum_{d=d_0}^{e^{\alpha/\beta}} \frac{e^\alpha}{d^\beta}C_3\frac{d}{d_0^4\log^2 n} &\leq \frac{C_3e^\alpha}{d_0^4\log^2 n}\sum_{d=d_0}^{\infty}\frac{1}{d^{\beta-1}} \\ &\leq \frac{C_3e^\alpha}{d_0^4\log^2 n}\frac{2}{\beta-2}\frac{1}{d_0^{\beta-2}} \\ &= \frac{2C_3e^\alpha}{(\beta-2)d_0^{\beta+2}\log^2 n} \\ &= c_3\frac{e^\alpha \log^{\beta/2-1} n}{x^{\beta/2+1}}, \end{aligned}$$

where

$$c_3 = \frac{2C_3}{(\beta-2)}C^{1+\beta/2} = \frac{4^{1+\beta}c_2}{(\beta-2)c_1^{1+\beta}}.$$

For $x = e^{2\alpha/(\beta+2)}\alpha$, the preceding inequality implies that the number of components of size at least x is at most $o(1)$. In other words, almost surely no component has size greater than $e^{2\alpha/(\beta+2)}\alpha$. This completes the proof of Theorem 4.1. □

5. ON THE SIZE OF THE SECOND LARGEST COMPONENT

For $\beta < \beta_0 = 3.47875\ldots$, we consider the giant component as well as the size of the second largest component.

Theorem 5.1. *Consider an (α,β)-graph with $\beta < \beta_0 = 3.47875\ldots$.*

1. *There is a unique giant component of size $\Theta(n)$.*

2. *When $2 < \beta < \beta_0$, almost surely the size of the second largest component is $\Theta(\log n)$.*
3. *When $\beta = 2$, almost surely the size of the second largest component is $\Theta(\log n / \log\log n)$.*
4. *When $1 \leq \beta < 2$, almost surely the size of the second largest component is $\Theta(1)$.*
5. *When $0 < \beta < 1$, almost surely the graph is connected.*

Proof. When $\beta < \beta_0$, the branching process method is no longer feasible when vertices of large degrees are involved. Thus, we cannot apply Azuma's martingale inequality for bounding Y_i as we did in earlier proofs. We will modify the branching process method as follows.

(a) Choose a number x_β (to be specified later depending on β).
(b) Start with Y_0^* live vertices and $Y_0^* \geq C \log n$. All other vertices are unmarked.
(c) At the i-th step, choose one live vertex u and exposed its neighbors. If the degree of u is less than or equal to x_β, proceed as in Section 4, by marking u dead and all vertices $v \in N_u$ live (provided v is not marked before). If the degree of u is greater than x_β, mark exactly one vertex $v \in N_u$ live and others dead, provided v is unmarked. In both cases u is marked dead.

The main idea is to show that Y_i^*, a truncated version of Y_i, is well-concentrated around $E(Y_i^*)$. Although it is difficult to directly derive such a result for Y_i because of vertices of large degrees, we will be able to bound the distribution Y_i^*. Indeed, we will show that the set of marked vertices (live or dead) grows to a giant component if Y_0^* exceeds a certain bound. We consider three ranges for β.

Case 1: $2 < \beta < \beta_0$. We consider the positive constant

$$Q = \frac{1}{E} \sum_{x=1}^{e^{\alpha/\beta}} x(x-2) \left\lfloor \frac{e^\alpha}{x^\beta} \right\rfloor.$$

There is a constant integer x_0 satisfying

$$\frac{1}{E} \sum_{x=1}^{x_0} x(x-2) \left\lfloor \frac{e^\alpha}{x^\beta} \right\rfloor > \frac{Q}{2}.$$

We choose δ satisfying

$$\frac{\delta}{(1-\delta)^2} = \frac{Q}{4}.$$

If the component has more than δE edges, it must have $\Theta(n)$ vertices since $\beta > 2$. So it is a giant component and we are done. We may assume that the component has no more than δE edges.

We now choose $x_\beta = x_0$ and apply the modified branching process. Then, Y_i^* satisfies:

- $Y_0^* \geq \lceil C \log n \rceil$, where $C = 130 x_0^2 / Q$ is a constant depending only on β.
- $-1 \leq Y_i^* - Y_{i-1}^* \leq x_0$.
- Let W_i be the number of backedges as defined in Section 4. By inequality (4–1) and the assumption that the number of edges m in the component is at most δn, we have
$$E(W_i) \leq \frac{\delta}{(1-\delta)^2} = \frac{Q}{4}.$$
Hence,
$$\begin{aligned} E(Y_i^* - Y_{i-1}^*) &\approx \frac{1}{E} \sum_{x=1}^{x_0} x(x-2) \left\lfloor \frac{e^\alpha}{x^\beta} \right\rfloor - E(W_i) \\ &\geq \frac{Q}{2} - \frac{Q}{4} = \frac{Q}{4}. \end{aligned}$$

By Azuma's martingale inequality,

$$\begin{aligned} \Pr\Big(Y_i^* \leq \frac{Qi}{8}\Big) &\leq \Pr\Big(Y_i^* - E(Y_i^*) \leq -\frac{Qi}{8}\Big) \\ &< \exp - \frac{(Qi/8)^2}{2ix_0^2} = o(n^{-1}) \end{aligned}$$

provided that $i > C \log n$.

The preceding inequality implies that with probability at least $1 - o(n^{-1})$, we have $Y_i^* > Qi/8 > 0$ when $i > \lceil C \log n \rceil$. Since Y_i^* decreases by at most 1 at each step, Y_i^* cannot be zero if $i \leq \lceil C \log n \rceil$. So $Y_i^* > 0$ for all i. In other words, a. s. the branching process will not stop. However, it is impossible to have $Y_n^* > 0$—a contradiction. Thus we conclude that the component must have at least δn edges. So it is a giant component. We note that if a component has more than $\lceil C \log n \rceil$ edges exposed, then almost surely it is a giant component. In particular, any vertex with degree more than $\lceil C \log n \rceil$ is almost surely in the giant component. Hence, the second components have size of at most $\Theta(\log n)$.

Next we show that the second largest has size at least $\Theta(\log n)$. We consider the vertices v of degree $x = c\alpha$, where c is some constant. There is a positive probability that all neighboring vertices of v have degree 1. In this case, we get a connected component of size $x+1 = \Theta(\log n)$. The probability of this is about

$$\left(\frac{1}{\zeta(\beta-1)}\right)^{c\alpha}.$$

There are $e^\alpha/(c\alpha)^\beta$ vertices of degree x. Thus the probability that none of them has the preceding property is about

$$\begin{aligned}\left(1-\frac{1}{\zeta(\beta-1)^{c\alpha}}\right)^{\frac{e^\alpha}{(c\alpha)^\beta}} &\approx \exp\left(-\frac{1}{\zeta(\beta-1)^{c\alpha}}\frac{e^\alpha}{(c\alpha)^\beta}\right)\\ &= \exp\left(-\frac{\left(e/\zeta(\beta-1)^c\right)^\alpha}{(c\alpha)^\beta}\right)\\ &= o(1),\end{aligned}$$

where

$$c = \begin{cases} 1 & \text{if } \beta \geq 3,\\ \dfrac{1}{-2\log(\beta-2)} & \text{if } 3 > \beta > 2.\end{cases}$$

In other words, a. s. there is a component of size $c\alpha + 1 = \Theta(\log n)$. Therefore, the second largest component has size $\Theta(\log n)$. Moreover, the argument still holds if we replace $c\alpha$ by any small number. Hence, small components exhibit a continuous behavior.

Case 2: $\beta = 2$. We choose $x_\beta = 10\alpha$. We note that a component with more than $2E/3$ edges must be unique. We will prove that almost surely the unique component contains all vertices with degree greater than $101\alpha^2$. So it contains $(1-o(1))E$ edges and it is the giant component.

We further modify the branching process by starting from $Y_0^* \geq \lceil 101\alpha^2\rceil$ vertices. If the component has more than $\frac{2}{3}E$ edges, we are done. Otherwise, the expected number of backeges is small.

$$E(W_i) \leq \frac{2/3}{(1-2/3)^2} = 6$$

from inequality (4–1). Hence, Y_i^* satisfies

$$\begin{aligned} Y_0^* &\geq \lceil 101\alpha^2\rceil,\\ -1 \leq Y_i^* - &Y_{i-1}^* \leq 10\alpha,\\ E(Y_i^* - Y_{i-1}^*) &\approx \frac{1}{E}\sum_{x=1}^{10\alpha} x(x-2)\left\lfloor\frac{e^\alpha}{x^\beta}\right\rfloor - E(W_i)\\ &> 10-2-6 = 2.\end{aligned}$$

By Azuma's martingale inequality,

$$\begin{aligned}\Pr(Y_i^* \leq i) &\leq \Pr(Y_i^* - E(Y_i^*) \leq -i\\ &< \exp\left(-\frac{i^2}{i(10\alpha)^2}\right) = o(n^{-1})\end{aligned}$$

provided that $i \geq 101\alpha^2$.

This inequality implies that with proability at least $1-o(n^{-1})$, we have $Y_i^* \geq i > 0$ when $i > \lceil 101\alpha^2\rceil$. Since Y_i^* decreases at most by 1 at each step, Y_i^* cannot be zero if $i \leq \lceil 101\alpha^2\rceil$. So $Y_i^* > 0$ for all i. In other words, a. s. the branching process will not stop. However, it is impossible to have $Y_n^* > 0$—a contradiction. Thus we conclude that the component must have at least $\frac{2}{3}E$ edges. We note that a. s. all vertices with degree more than $\lceil 101\alpha^2\rceil$ are in the unique component with at least $\frac{2}{3}E$ edges, hence the giant component.

The probability that a random vertex is in the giant component is at most

$$\frac{1}{E}\sum_{x=1}^{101\alpha^2} x\frac{e^\alpha}{x^2} \approx \frac{2\log\alpha}{\alpha}.$$

The probability that there are $2.1\alpha/\log\alpha$ vertices not in the giant component is at most

$$\left(\frac{2\log\alpha}{\alpha}\right)^{2.1\alpha/\log\alpha} = e^{-(2.1+o(1))\alpha} = o(n^{-2}).$$

Since there is at most n connected components, we conclude that a. s. a connected component of size greater that

$$2.1\frac{\alpha}{\log\alpha} = \Theta\left(\frac{\log n}{\log\log n}\right)$$

must be the giant component.

Now we find a vertex v of degree x with $x \leq 0.9\alpha/\log\alpha$. The probability that all its neighbors are of degree 1 is $(1/\alpha)^x$. The probability that no such vertex exists is at most

$$\begin{aligned}\left(1-\left(\frac{1}{\alpha}\right)^x\right)^{e^\alpha/x^2} &\approx \exp\left(-\left(\frac{1}{\alpha}\right)^x\frac{e^\alpha}{x^2}\right)\\ &= \exp-\frac{e^{0.1\alpha}}{x^2} = o(1).\end{aligned}$$

Hence, almost surely there is a vertex of degree $x \leq 0.9\alpha/\log\alpha$ that, which forms a connected component of size $x+1$. This proves that a. s. the second largest component has size $\Theta(\log n/\log\log n)$.

Case 3: $0 < \beta < 2$. We use the modified branching process by choosing

$$x_\beta = \exp\frac{(5-2\beta)\alpha}{(6-2\beta)\beta}.$$

If a component has more than $2E/3$ edges, it is the unique giant component and we are done. Otherwise,

$$E(W_i) \le \frac{2/3}{(1-2/3)^2} = 6.$$

Hence, Y_i^* satisfies

$$\begin{aligned} Y_0^* &\ge \frac{5}{C^2}\exp\frac{(2-\beta)\alpha}{(3-\beta)\beta},\\ -1 \le Y_i^* - Y_{i-1}^* &\le \exp\frac{(5-2\beta)\alpha}{(6-2\beta)\beta},\\ E(Y_i^* - Y_{i-1}^*) &\approx \frac{1}{E}\sum_{x=1}^{\exp\frac{(5-2\beta)\alpha}{(6-2\beta)\beta}} x(x-2)\left\lfloor\frac{e^\alpha}{x^\beta}\right\rfloor - E(W_i)\\ &\approx Ce^{\alpha/(2\beta)}. \end{aligned}$$

Here C is a constant depending only on β.

By Azuma's martingale inequality,

$$\begin{aligned} \Pr\left(Y_i^* \le \tfrac{1}{2}Ce^{\alpha/(2\beta)}i\right) &< \Pr\left(Y_i^* - E(Y_i^*) \le -\tfrac{1}{2}Ce^{\alpha/(2\beta)}i\right)\\ &< \exp\left(-\frac{(\frac{1}{2}Ce^{\alpha/(2\beta)}i)^2}{i\left(\exp\frac{(5-2\beta)\alpha}{(6-2\beta)\beta}\right)^2}\right)\\ &= o(n^{-1}) \end{aligned}$$

provided that

$$i \ge \frac{5}{C^2}\exp\frac{(2-\beta)\alpha}{(3-\beta)\beta}.$$

This inequality shows that with probability at least $1 - o(n^{-1})$, we have $Y_i^* > \frac{1}{2}Ce^{\alpha/(2\beta)}i > 0$ provided that

$$i > \frac{5}{C^2}\exp\frac{(2-\beta)\alpha}{(3-\beta)\beta}.$$

Since Y_i^* decreases at most by 1 at each step, Y_i^* cannot be zero if

$$i \le \frac{5}{C^2}\exp\frac{(2-\beta)\alpha}{(3-\beta)\beta}.$$

So $Y_i^* > 0$ for all i. In other words, a. s. the branching processing will not stop. However, it is impossible to have $Y_n^* > 0$ — a contradiction. So, a. s. all vertices with degree more than

$$\frac{5}{C^2}\exp\frac{(2-\beta)\alpha}{(3-\beta)\beta}$$

are in the giant component. The probability that a random vertex is in the giant component is at most

$$\frac{1}{E}\sum_{x=1}^{\frac{5}{C^2}\exp\frac{(2-\beta)\alpha}{(3-\beta)\beta}} x\frac{e^\alpha}{x^\beta} = \Theta\Big(\exp -\frac{(2-\beta)\alpha}{(3-\beta)\beta}\Big).$$

The probability that all

$$2\left\lfloor\frac{3-\beta}{2-\beta}\right\rfloor + 1$$

vertices are not in the giant vertex is at most

$$\Theta\left(\exp\Big(-\frac{(2-\beta)\alpha}{(3-\beta)\beta}\Big)\right)^{2\left\lfloor\frac{3-\beta}{2-\beta}\right\rfloor+1} = o(n^{-2}).$$

Since there are at most n connected components, we conclude that a. s. a connected component of size greater that

$$2\left\lfloor\frac{3-\beta}{2-\beta}\right\rfloor = \Theta(1)$$

must be the giant component.

For $1 < \beta < 2$, we fix a vertex v of degree 1. The probability that the other vertex that connects to v is also of degree 1 is

$$\Theta\Big(\frac{e^\alpha}{e^{2\alpha/\beta}}\Big).$$

Therefore the probability that no component has size of 2 is at most

$$\left(1 - \Theta\Big(\frac{e^\alpha}{e^{2\alpha/\beta}}\Big)\right)^{e^\alpha} \approx e^{-\Theta(e^{2\alpha-2\alpha/\beta})} \approx o(1).$$

In other words, the graph a. s. has at least one component of size 2.

For $0 < \beta < 1$, we want to show that the random graph is a. s. connected. Since the size of the possible second largest component is bounded by a constant M, all vertices of degree $\ge M$ are almost surely in the giant component. We only need to show the probability that there is an edge connecting two small degree vertices is small. There are only

$$\sum_{x=1}^{M} x\left\lfloor\frac{e^\alpha}{x^\beta}\right\rfloor \approx Ce^\alpha$$

vertices with degree less than M. For any random pair of vertices (u, v), the probability that there is an edges connecting them is about

$$\frac{1}{E} = \Theta(e^{-2\alpha/\beta}).$$

Hence the probability that there is edge connecting two small degree vertices is at most

$$\sum_{u,v} \frac{1}{E} = (Ce^{\alpha})^2 \Theta(e^{2\alpha/\beta}) = o(1).$$

Thus every vertex is a. s. connected to a vertex with degree $\geq M$, which a. s. belongs to the giant exponent. Hence, the random graph is a. s. connected.

□

6. COMPARISONS WITH REALISTIC MASSIVE GRAPHS

Our (α, β)-random graph model was originally derived from massive graphs generated by long distance telephone calls. These so-called *call graphs* are taken over different time intervals. For the sake of simplicity, we consider all the calls made in one day. Every completed phone call is an edge in the graph. Every phone number that either originates or receives a call is a node in the graph. When a node originates a call, the edge is directed out of the node and contributes to that node's outdegree. Likewise, when a node receives a call, the edge is directed into the node and contributes to that node's indegree.

The particular call graph we used for the statistics in this section correspond to the date August 10, 1998, a typical day. The data were compiled by J. Abello and A. Buchsbaum of AT&T Labs from raw phone call records using, in part, the external memory algorithm of [Abello et al. 1998] for computing connected components of massive graphs.

In Figure 1, we plot the number of vertices versus the indegree and the outdegree for the call graph. Let $y(i)$ be the number of vertices with indegree i. For each i such that $y(i) > 0$, a dot on the left plot is placed at $\big(i, y(i)\big)$. The plot on the right is built in the same way. Plots of the number of vertices versus the indegree or outdegree for the call graphs of other days are very similar.

Figure 2 plots for the same call graph the number of connected components for each possible size.

The degree sequence of the call graph does not obey perfectly the (α, β)-graph model. The number of vertices of a given degree does not even decrease monotonically with increasing degree. Moreover, the call graph is directed: for each edge there is a node that originates the call and a node that receives the call. The indegree and outdegree of a node need not be the same. Clearly the (α, β)-random

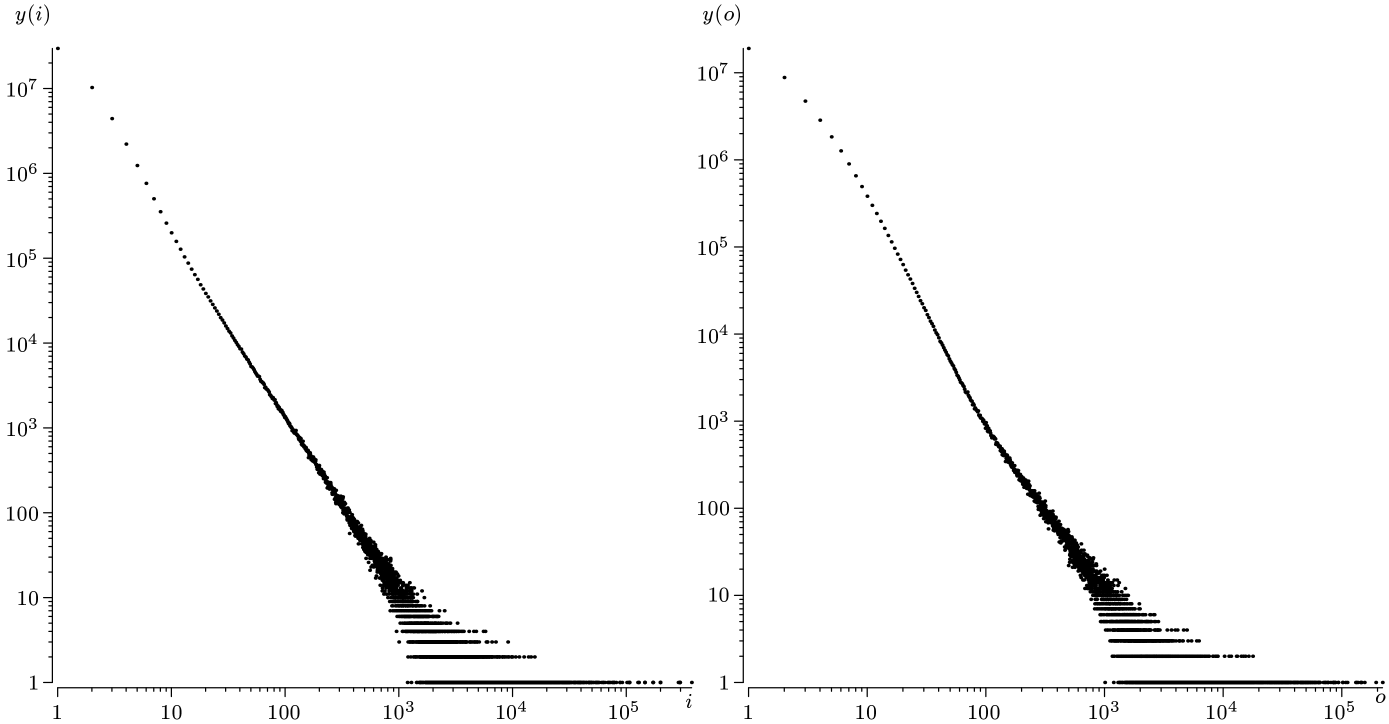

FIGURE 1. Left: number of vertices $y(i)$ versus indegree i, plotted on a log-log scale, for a representative real-life graph. Right: number of vertices versus outdegree o for the same graph.

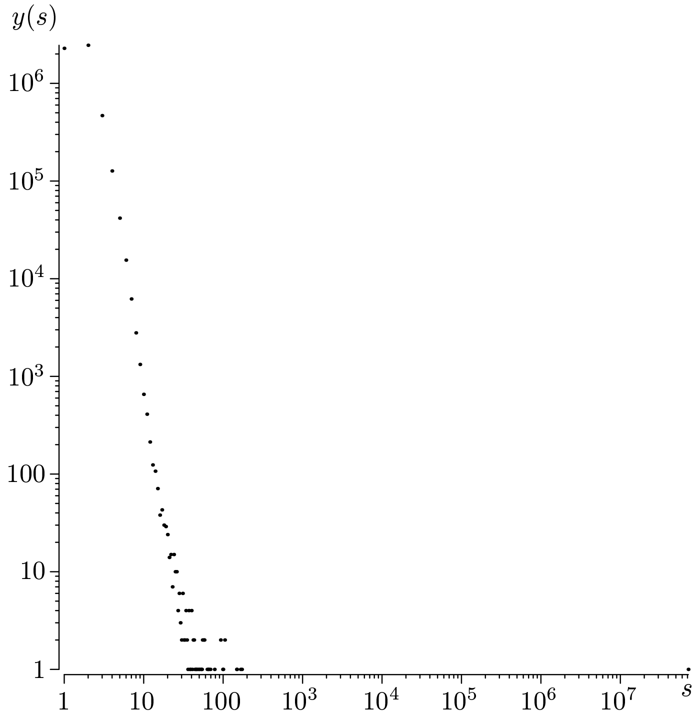

FIGURE 2. Left: number of connected components for each possible component size s for our example graph. Note the giant component on the lower right.

graph model does not capture all of the random behavior of the real world call graph.

Nonetheless, our model does capture some of the behavior of the call graph. To see this we first estimate α and β in Figure 1. Recall that for an (α, β)-graph, the number of vertices as a function of degree is given by $\log y = \alpha - \beta \log x$. By approximating Figure 1 by a straight line, β can be estimated using the slope of the line to be approximately 2.1. The value of e^α for Figure 1 is approximately 30×10^6. The total number of nodes in the call graph can be estimated by $\zeta(2.1)e^\alpha = 1.56e^\alpha \approx 47 \times 10^6$.

For β between 2 and β_0, the (α, β)-graph will have a giant component of size $\Theta(n)$. In addition, a. s. all other components are of size $O(\log n)$. Moreover, for any $2 \geq x \geq O(\log n)$, a component of size x exists. This is qualitatively true of the distribution of component sizes of the call graph in Figure 2. The one giant component contains nearly all of the nodes. The maximum size of the next largest component is indeed exponentially smaller than the size of the giant component. Also, a component of nearly every size below this maximum exists. Interestingly, the distribution of the number of components of size smaller than the giant component is nearly log-log linear. This suggests that after removing the giant component, one is left approximately with an (α, β)-graph with $\beta > 4$. (Theorem 4.1 yields a log-log linear relation between number of components and component size for $\beta > 4$.) This seems intuitively reasonable, since the greater the degree, the fewer nodes of that degree we expect to remain after deleting the giant component. This will increase the value of β for the resulting graph.

7. OPEN QUESTIONS

Numerous questions remain to be studied. For example, what is the effect of time scaling? How does it correspond with the evolution of β? What are the structural behaviors of the call graphs? What are the correlations between the directed and undirected graphs? It is of interest to understand the phase transition of the giant component in the realistic graph. In the other direction, the number of tiny components of size 1 is leading to many interesting questions as well. Clearly, there is much work to be done in our understanding of massive graphs.

ACKNOWLEDGMENTS

We are grateful to J. Feigenbaum, J. Abello, A. Buchsbaum, J. Reeds, and J. Westbrook for their assistance in preparing the figures and for many interesting discussions on call graphs. We are very thankful to the anonymous referee for invaluable comments.

REFERENCES

[Abello et al. 1998] J. Abello, A. L. Buchsbaum, and J. R. Westbrook, "A functional approach to external graph algorithms", pp. 332–343 in *Algorithms—ESA '98* (Venice, 1998), edited by G. Bilardi et al., Lecture Notes in Comp. Sci. **1461**, Springer, Berlin, 1998.

[Aiello et al. 2000] W. Aiello, F. Chung, and L. Lu, "A random graph model for massive graphs", pp. 171–180 in *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing* (Portland, OR, 2000), ACM Press, New York, 2000.

[Aiello et al. ≥ 2001] W. Aiello, F. Chung, and L. Lu, "Random evolution of power law graphs", in *Handbook of massive data sets*, vol. 2, edited by J. Abello et al. To appear.

[Albert et al. 1999] R. Albert, H. Jeong, and A. Barabási, "Diameter of the World Wide Web", *Nature* **401** (September 9, 1999).

[Alon and Spencer 1992] N. Alon and J. H. Spencer, *The probabilistic method*, Wiley, New York, 1992.

[Barabási and Albert 1999] A. Barabási and R. Albert, "Emergence of scaling in random networks", *Science* **286** (October 15, 1999).

[Barabási et al. 2000] A.-L. Barabási, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: The topology of the World Wide Web", *Physica A* **281** (2000). See http://www.nd.edu/~hjeong/./paper.html.

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

[Erdős and Rényi 1961] P. Erdős and A. Rényi, "On the strength of connectedness of a random graph", *Acta Math. Acad. Sci. Hungar.* **12** (1961), 261–267.

[Faloutsos et al. 1999] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology", pp. 251–262 in *ACM SIGCOMM '99 Conference: applications, technologies, architectures, and protocols for computer communications* (Cambridge, MA, 1999), ACM Press, New York, 1999.

[Hayes 2000] B. Hayes, "Graph theory in practice, II", *American Scientist* **88** (March–April 2000), 104–109.

[Kleinberg et al. 1999] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, "The web as a graph: measurements, models, and methods", pp. 1–17 in *Computing and combinatorics* (Tokyo, 1999), edited by T. Asano et al., Lecture Notes in Comp. Sci. **1627**, Springer, Berlin, 1999.

[Kumar et al. 1999a] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the web for emerging cyber communities", in *Proceedings of the 8th World Wide Web Conference* (Toronto, 1999), Elsevier, Amsterdam, 1999.

[Kumar et al. 1999b] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Extracting large-scale knowledge bases from the web", pp. 639–650 in *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases* (Edinburgh, 1999), edited by M. P. Atkinson et al., Morgan Kaufmann, San Francisco, 1999. See http://www.informatik.uni-trier.de/~ley/db/conf/vldb/vldb99.html.

[Kumar et al. 2001] S. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal, "Stochastic models for the Web graph", in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science* (Redondo Beach, CA, 2000), IEEE, Los Alamitos, CA, 2001.

[Łuczak 1992] T. Łuczak, "Sparse random graphs with a given degree sequence", pp. 165–182 in *Random graphs* (Poznań, 1989), vol. 2, edited by A. Frieze and T. Łuczak, Wiley, New York, 1992.

[Molloy and Reed 1995] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence", *Random Structures and Algorithms* **6**:2-3 (1995), 161–179.

[Molloy and Reed 1998] M. Molloy and B. Reed, "The size of the giant component of a random graph with a given degree sequence", *Combin. Probab. Comput.* **7**:3 (1998), 295–305.

[Wormald 1981] N. C. Wormald, "The asymptotic connectivity of labelled regular graphs", *J. Combin. Theory Ser. B* **31**:2 (1981), 156–167.

[Wormald 1999] N. C. Wormald, "Models of random regular graphs", pp. 239–298 in *Surveys in combinatorics* (Canterbury, 1999), edited by J. D. Lamb and D. A. Preece, London Math. Soc. Lecture Note Series **267**, Cambridge Univ. Press, Cambridge, 1999.

William Aiello, AT&T Labs, 180 Park Avenue, Florham Park, NJ 07932, United States (aiello@research.att.com)

Fan Chung, Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0112, United States (fan@ucsd.edu)

Linyuan Lu, Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0112, United States (llu@euclid.ucsd.edu)

Received November 8, 1999; accepted in revised form May 1, 2000