A Random Graph Model for Power Law Graphs

William Aiello [*] Fan Chung [†‡] Linyuan Lu [§*]

Abstract

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 log-size and log-log growth rate. These parameters capture some universal characteristics of massive graphs. Furthermore, 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 will illustrate the consistency of our model with the behavior of some massive graphs derived from data in telecommunications. We will 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, then 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 structure of the Web has come to light which may enable us to describe graph properties of the Web qualitatively. Kumar et al. [13, 14] and Kleinberg et al. [12] have measured the degree sequences of the Web and shown that it is well approximated by a power law distribution.

[*]AT&T Labs, Florham Park, New Jersey.

[†]University of California, San Diego.

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

[§]University of Pennsylvania, Philadelphia.

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 in [4, 6, 7]. The power law distribution of the degree sequence appears to be a very robust property of the Web despite its dynamic nature. In 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. [1] have shown that the degree sequence of so called *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. In addition, Faloutsos, et al. [10] 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 which hold with high probability in various regimes of our parameters. And finally, we compare the results from the model with the exact connectivity structure for some call graphs computed by Abello et al. [1].

An extended abstract of this paper has appeared in the Proceedings of the Thirtysecond Annual ACM Symposium on Theory of Computing 2000 (see [2]). In this paper, we have included the complete proofs for the main theorems and several additional theorems focused on the second largest components of power graphs in various ranges. In addition, some recent references are provided (also see [11]).

1.1 Power-Law Random Graphs

The study of random graphs dates back to the work of Erdős and Rényi whose seminal papers [8, 9] laid the foundation for the theory of random graphs. There are three standard models for what we will call in this paper *uniform* random graphs [5]. Each has two parameters. One parameters controls the number of nodes in the graph and one controls 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 also has two parameters. The two parameters only roughly delineate the size and density but they are natural and convenient for describing a power law degree sequence. The power law random graph model $P(\alpha, \beta)$ 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.

We remark that 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 bounding error estimates of the variances (which will be further described in a subsequent paper).

1.2 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, i.e., a component of size $\Theta(n)$. Moreover, all smaller components are of size $O(1)$. For $2 < \beta < \beta_0 = 3.4785$ 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 numbers of components of given sizes can be approximated by a power law as well.

1.3 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 [20] studied the connectivity of graphs whose degrees are in an interval $[r,R]$, where $r \geq 3$. Łuczak [16] 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 [17, 18]. They consider a random graph on n vertices with the following degree distribution. The number of vertices of degree $0, 1, 2, \ldots$ are about $\lambda_0 n, \lambda_1 n, \ldots$ respectively, where the λ's sum to 1. It is shown in [17] 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 $n \to \infty$. They also examined the threshold behavior of such graphs. In this paper, we will apply these techniques to deal with the special case that applies to our model.

Several other papers have taken a different approach to modeling power law graphs than the one taken here [3, 6, 7, 12, 14]. 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 this approach is interesting and important it 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, [6, 7] implicitly assume that the probability that a node has a given degree is a continuous function. The authors of [12, 14] will offer a partial analysis in a recent paper [15]. Second, while the models may generate graphs with power law degree sequences, it remains to be seen if they generate graphs which duplicate other structural properties of the Web, the Internet, and call graphs. For example, the model in [6, 7] 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 [14] 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 visa versa. While this may be appropriate for call graphs (e.g., customer service numbers) it remains to be seen whether it models the Web. Thus, while the random graph generation approach holds the promise of accurately predicting a wide variety a 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 equi-probable. The goal is to derive structural properties which hold with probability asymptotically approaching 1. Such an approach, while potentially less accurate than the detailed modelling 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.

We remark that in a subsequent paper[3] several aspects of power law graphs are further examined, including (1) analyzing the evolution of graphs,

(2) the asymmetry of in-degrees and out-degrees, (3) the "scale invariance" of power law graphs.

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. In section 5, we discuss the sizes of the second largest components. In section 6, we compare the results of our model to exact connectivity data for call graphs.

2 A random graph model

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

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

In other words, we have

$$| \{v \ : \ deg(v) = x\} | = y = \frac{e^\alpha}{x^\beta}$$

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

We note that the number of edges should be an integer. To be precise, the above expression for y should be rounded down to $\lfloor \frac{e^\alpha}{x^\beta} \rfloor$. If we use real numbers instead of rounding down to integers, it may cause some error terms in further computation. However, we will see that the error terms can be easily bounded. For simplicity and convenience, we will use real numbers with the understanding 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 old if needed. Furthermore, for simplicity, we here assume that there is no isolated vertices.

We can deduce the following facts for our graph:
(1) The maximum degree of the graph is $e^{\frac{\alpha}{\beta}}$. Note that $0 \leq \log y = \alpha - \beta \log x$.
(2) The vertices number n can be computed as follows: By summing $y(x)$ for x from 1 to $e^{\frac{\alpha}{\beta}}$, we have

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

where $\zeta(t) = \sum_{n=1}^{\infty} \frac{1}{n^t}$ is the Riemann Zeta function.
(3) The number of edges E can be computed as follows:

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

(4) The differences of the real numbers in (1)-(3) and their integer parts can be estimated as follows: For the number n of vertices, the error term is at most $e^{\frac{\alpha}{\beta}}$. For $\beta \geq 1$, 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

$$n \geq \frac{e^{\frac{\alpha}{\beta}}}{1-\beta} - e^{\frac{\alpha}{\beta}} = \frac{\beta e^{\frac{\alpha}{\beta}}}{1-\beta}$$

Therefore, n has the same magnitude as $\frac{e^{\frac{\alpha}{\beta}}}{1-\beta}$. The number E of edges can be treated in a similarly way. For $\beta \geq 2$, the error term of E is $o(E)$, a lower order term. For $0 < \beta < 2$, E has the same magnitude as in formula of item (3). In this paper, we mainly deal with the case $\beta > 2$. The only place that we deal with the case $0 < \beta < 2$ is 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.

In order to consider the random graph model, we will need to consider large n. We say that some property almost surely (a. s.) happens if the probability that the property holds tends to 1 as the number n of the vertices goes to infinity. Thus we consider α to be large but where 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, which is formed by combining k random matching. For references and undefined terminology, the reader is referred to [5, 21].

We note that this random graph model is slightly different from the uniform selection model $P(\alpha, \beta)$ as described in section 1.1. However, by

using techniques in Lemma 1 of [18], 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 [17] showed that for a random graph with $(\lambda_i + o(1))n$ vertices of degree i, where λ_i are non-negative values which sum to 1, the giant component emerges when $Q = \sum_{i\geq 1} i(i-2)\lambda_i > 0$, provided that the maximum degree is less than $n^{1/4-\epsilon}$. They also show that almost surely there is no giant component when $Q = \sum_{i\geq 1} i(i-2)\lambda_i < 0$ and maximum degree less than $n^{1/8-\epsilon}$.

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

$$\begin{aligned} Q &= \sum_{x=1}^{e^{\frac{\alpha}{\beta}}} x(x-2)\lfloor \frac{e^\alpha}{x^\beta} \rfloor \\ &\approx \sum_{x=1}^{e^{\frac{\alpha}{\beta}}} \frac{e^\alpha}{x^{\beta-2}} - 2\sum_{x=1}^{e^{\frac{\alpha}{\beta}}} \frac{e^\alpha}{x^{\beta-1}} \\ &\approx (\zeta(\beta-2) - 2\zeta(\beta-1))e^\alpha \text{ if } \beta > 3 \end{aligned}$$

Hence, we consider the value $\beta_0 = 3.47875\ldots$, which is a solution to

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

If $\beta > \beta_0$, we have

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

We first summarize the results here:

1. When $\beta > \beta_0 = 3.47875\ldots$, the random graph a. s. has no giant component. When $\beta < \beta_0 = 3.47875\ldots$, there is a. s. a unique giant component.

2. When $2 < \beta < \beta_0 = 3.47875\ldots$, the second largest components are a. s. of size $\Theta(\log n)$. For any $2 \leq x < \Theta(\log n)$, there is almost surely a component of size x.

3. When $\beta = 2$, a. s. the second largest components are of size $\Theta(\frac{\log n}{\log\log n})$. For any $2 \leq x < \Theta(\frac{\log n}{\log\log n})$, there is almost surely a component of size x.

4. When $1 < \beta < 2$, the second largest components are a. s. of size $\Theta(1)$. The graph is a. s. not connected.

5. When $0 < \beta < 1$, the graph is a. s. connected.

6. For $\beta = \beta_0 = 3.47875\ldots$, this is a very complicated case. It corresponds to the double jump of random graph $\mathcal{G}(n,p)$ with $p = \frac{1}{n}$. For $\beta = 1$, there is a nontrivial probability for either cases 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 $\beta \leq 8$, additional analysis is needed to deal with the degree constraints. We will prove in Theorem 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 3 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. We will prove the following:

Theorem 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 connected component containing v. Then*

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

where $\lambda = d^{\epsilon}$. In other words, the vertex v a. s. belongs to a connected component of size $\frac{d}{c_1} + O(d^{\frac{1}{2}+\epsilon})$, where $c_1 = 2 - \frac{\zeta(\beta-2)}{\zeta(\beta-1)}$, $c_2 = \frac{\zeta(\beta-3)}{\zeta(\beta-1)} -$

$\left(\frac{\zeta(\beta-2)}{\zeta(\beta-1)}\right)^2$ *are two constants,* ϵ *is an arbitrary small positive number and* d *is a (slowly) increasing function of* n*.*

2. *The number of connected components of size* x *is a. s. at least*

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

and at most

$$c_3\frac{e^{\alpha}\log^{\frac{\beta}{2}-1}n}{x^{\frac{\beta}{2}+1}}$$

where $c_3=\frac{4^{1+\beta}c_2}{(\beta-2)c_1^{1+\beta}}$ *is a constant only depending on* β*.*

3. *A connected component of the* (α,β)*-graph a. s. has the size at most*

$$e^{\frac{2\alpha}{\beta+2}}\alpha=\Theta(n^{\frac{2}{\beta+2}}\log n)$$

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

Theorem 1 strengthens the following result (which can be derived from Lemma 3 in [17]) for the range of $\beta>4$.

Theorem 2 *For* $\beta>\beta_0=3.47875\ldots$*, a connected component of the* (α,β)*-graph a. s. has the size at most*

$$Ce^{\frac{2\alpha}{\beta}}\alpha=\Theta(n^{\frac{2}{\beta}}\log n)$$

where $C=\frac{16}{c_1^2}$ *is a constant only depending on* β*.*

The proof for Theorem 2 is by using branching process method. We here briefly describe the proof since it is needed for the proof of Theorem 1. Pick any vertex v in our graph, expose its neighbors, and 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 be 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 by either "live" or "dead". A vertex in L_i, whose neighbors have not been all exposed yet, is marked "live". A vertex, whose neighbors have all been exposed, is marked "dead". Let O_i be the set of live vertices and Y_i be the random variable that is the number of vertices

in O_i. Each step we mark exact one dead vertex, so the total number of dead vertices at 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 $Y_{i-1} = 0$, then we 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 $L_i = L_{i-1} \cup N_u$, and $O_i = (O_{i-1} \setminus \{u\}) \cup (N_u \setminus L_{i-1})$.

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 [17] to prove Theorem 2.

Suppose that the 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 L_{i-1} (that is not v) a "backedge". We note that "backedges" causes the exploration to stop more quickly, especially when the component is large. However in our case $\beta > \beta_0 = 3.47875\ldots$, the contribution of "backedges" is quite small. We denote $Z_i = \#\{N_u\}$ and $W_i = \#\{N_u \cap L_{i-1}\} - 1$. Z_i measures the degree of the vertex exposed at stage i, while W_i measures the number of "backedges". By definition, we have

$$Y_i - Y_{i-1} = Z_i - 2 - W_i.$$

We have

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

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 $\frac{m}{E}$. We have

$$E(W_i) \leq \sum_{x=1}^{\infty} x \left(\frac{m}{E}\right)^x$$

$$
\begin{aligned}
&= \frac{\frac{m}{E}}{(1-\frac{m}{E})^2} \qquad (*)\\
&= \frac{m}{E} + O((\frac{m}{E})^2)
\end{aligned}
$$

provided $\frac{m}{E} = o(1)$.

When $i \leq Ce^{\frac{2\alpha}{\beta}}\alpha$, m is at most $ie^{\frac{\alpha}{\beta}} \leq Ce^{\frac{3\alpha}{\beta}}\alpha$. Hence,

$$\frac{m}{E} = O(n^{\frac{3}{\beta}-1}\log n) = o(1)$$

We have

$$
\begin{aligned}
E(Y_i) &= Y_1 + \sum_{j=2}^{i} E(Y_j - Y_{j-1})\\
&= d + \sum_{j=2}^{i} E(Z_j - 2 - W_j)\\
&= d + (i-1)\left(\frac{\zeta(\beta-2)}{\zeta(\beta-1)} - 2\right) - iO(n^{\frac{3}{\beta}-1}\log n)\\
&= d - c_1(i-1) + io(1)
\end{aligned}
$$

Proof of Theorem 2: Since $|Y_j - Y_{j-1}| \leq e^{\frac{\alpha}{\beta}}$, by Azuma's martingale inequality, we have

$$Pr(|Y_i - E(Y_i)| > t) \leq 2e^{\frac{-t^2}{2ie^{\frac{2\alpha}{\beta}}}}$$

where $i = \frac{16}{c_1^2}e^{\frac{2\alpha}{\beta}}\log n$, and $t = \frac{c_1}{2}i$. Since

$$E(Y_i) + t = d - c_1(i-1) + io(1) + \frac{c_1}{2}i = -\frac{c_1}{2}i + d + c_1 + io(1) < 0$$

We have

$$
\begin{aligned}
Pr(\tau > \tfrac{16}{c_1^2}e^{\frac{\alpha}{\beta}}\log n) &= Pr(\tau > i) \;\leq\; Pr(Y_i \geq 0)\\
&\leq Pr(Y_i > E(Y_i) + t)\\
&\leq 2e^{\frac{-t^2}{2ie^{\frac{2\alpha}{\beta}}}} \;=\; \tfrac{2}{n^2}
\end{aligned}
$$

Hence, the probability that there exists a vertex v such that v lies in a component of size greater than $\frac{16}{c_1^2}e^{\frac{2\alpha}{\beta}}\log n$ is at most

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

The proof of Theorem 1 uses the methodology above as a starting point while introducing the calculation of the variance of the above random variables.

Proof of Theorem 1

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

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

since $\beta > 4$.

We need to compute the covariants. 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 $CoVar(Z_i, Z_j) \leq E(Z_i)E\frac{2}{E^2} = O(\frac{1}{n})$ if $i \neq j$.

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 $\frac{m}{E}$. We have

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

$$= \frac{m}{E} + O((\frac{m}{E})^2)$$

$$CoVar(W_i, W_j) \leq \sqrt{Var(W_i)Var(W_j)} \leq \frac{m}{E} + O((\frac{m}{E})^2)$$

$$CoVar(Z_i, W_j) \leq \sqrt{Var(Z_i)Var(W_j)} = O(\sqrt{\frac{m}{E}})$$

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

$$\begin{aligned}
E(Y_i) &= d + (i-1)\left(\frac{\zeta(\beta-2)}{\zeta(\beta-1)} - 2\right) + iO(n^{\frac{3}{\beta}-1}) + i\frac{m}{E} \\
&= d - (i-1)c_1 + O(n^{\frac{4}{\beta}-1}) \\
&= d - (i-1)c_1 + o(1) \\
Var(Y_i) &= Var(d + \sum_{j=2}^{i}(Y_j - Y_{j-1})) \\
&= Var(\sum_{j=2}^{i}(Z_j - W_j)) \\
&= \sum_{j=2}^{i}(Var(Z_j) + Var(W_j)) \\
&\quad + \sum_{2\leq j\neq k\leq i}(CoVar(Z_j, Z_k) \\
&\quad -CoVar(Z_j, W_k) + CoVar(W_j, W_k)) \\
&= ic_2 + io(1) + i^2(O(\frac{1}{n}) + O(\sqrt{e^{(\frac{2}{\beta}-1)\alpha}}) \\
&\quad +O(e^{(\frac{2}{\beta}-1)\alpha})) \\
&= ic_2 + io(1) + i(O(e^{(\frac{2}{\beta}-\frac{1}{2})\alpha}) + O(e^{(\frac{3}{\beta}-1)\alpha})) \\
&= ic_2 + io(1)
\end{aligned}$$

Chebyshev's inequality gives

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

where σ is the standard deviation of Y_i, $\sigma = \sqrt{ic_2} + o(\sqrt{i})$ Let $i_1 = \lfloor \frac{d}{c_1} - \frac{2\lambda}{c_1}\sqrt{\frac{dc_2}{c_1}} \rfloor$ and $i_2 = \lceil \frac{d}{c_1} + \frac{2\lambda}{c_1}\sqrt{\frac{dc_2}{c_1}} \rceil$. We have

$$E(Y_{i_1}) - \lambda\sigma = d - (i_1 - 1)c_1 + o(1) - \left(\lambda\sqrt{c_2 i_1} + o(\sqrt{i_1})\right)$$

$$\begin{aligned} &\geq 2\lambda\sqrt{\frac{dc_2}{c_1}} - \lambda\sqrt{c_2\frac{d}{c_1}} - o(\sqrt{d}) \\ &= \lambda\sqrt{\frac{dc_2}{c_1}} - o(\sqrt{d}) \\ &> 0 \end{aligned}$$

Hence,

$$Pr(\tau < i_1) \leq Pr(Y_{i_1} \leq 0) \leq Pr(Y_{i_1} < E(Y_{i_1}) - \lambda\sigma) \leq \frac{1}{\lambda^2}$$

Similarly,

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

Hence,

$$Pr(\tau > i_2) \leq Pr(Y_{i_2} > 0) \leq Pr(Y_{i_2} > E(Y_{i_2}) + \lambda\sigma) \leq \frac{1}{\lambda^2}$$

Therefore

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

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

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

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

The proof 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 $C = \frac{16}{c_1^2}$ is the same constant as in Theorem 2. Let $x = Cd_0^2 \log n$ 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 1, we have

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

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

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

where $C_3 = \frac{32c_2}{c_1^3 C^2} = \frac{c_1 c_2}{8}$ is constant depending only on β. Since there are only $\frac{e^\alpha}{d^\beta}$ vertices of degree d, the number of components of size at least x is at most

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

where $c_3 = \frac{2C_3}{(\beta-2)} C^{1+\frac{\beta}{2}} = \frac{4^{1+\beta} c_2}{(\beta-2) c_1^{1+\beta}}$. For $x = e^{\frac{2\alpha}{\beta+2}}\alpha$, the above 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 $e^{\frac{2\alpha}{\beta+2}}\alpha$. This completes the proof of Theorem 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 3 *For (α, β)-graphs with $\beta < \beta_0 = 3.47875\ldots$, the following holds:*

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 $\Theta(\frac{\log n}{\log\log n})$.*
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 (α, β)-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 can not apply Azuma's martingale inequality for bounding Y_i as in the proofs of the previous sections. We will modify the branching process method as follows.
1. Choose a number x_β (to be specified later depending on β).
2. Start with Y_0^* live vertices and $Y_0^* \geq C \log n$. All other vertices are unmarked.
3. At the i-th step, we choose one live vertex u and exposed its neighbors. If the degree of u is less than or equal to x_β, we 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, we will mark exactly one vertex $v \in N(u)$ live and others dead, provided v is unmarked. In both case 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 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 the following three ranges of β.
Case 1: $2 < \beta < \beta_0$.
We consider $Q = \frac{1}{E}\sum_{x=1}^{e^{\frac{\alpha}{\beta}}} x(x-2)\lfloor\frac{e^\alpha}{x^\beta}\rfloor$. (Note that Q is a positive constant.) There is a constant integer x_0 satisfying $\frac{1}{E}\sum_{x=1}^{x_0} x(x-2)\lfloor\frac{e^\alpha}{x^\beta}\rfloor > \frac{Q}{2}$. 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 the following:

- $Y_0^* \geq \lceil C \log n \rceil$, where $C = \frac{130 x_0^2}{Q}$ is a constant only depending 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 (*) 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, we have

$$\begin{aligned} E(Y_i^* - Y_{i-1}^*) &\approx \frac{1}{E} \sum_{x=1}^{x_0} x(x-2) \lfloor \frac{e^\alpha}{x^\beta} \rfloor - E(W_i) \\ &\geq \frac{Q}{2} - \frac{Q}{4} = \frac{Q}{4}. \end{aligned}$$

By Azuma's martingale inequality, we have

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

provided $i > C \log n$.

The above inequality implies that with probability at least $1 - o(n^{-1})$, $Y_i^* > \frac{Qi}{8} > 0$ when $i > \lceil C \log n \rceil$. Since Y_i^* decreases at most by 1 at each step, Y_i^* can not 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 $Y_n^* > 0$, that is 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 component have size of at most $\Theta(\log n)$.

Next, we will 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

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

Since there are $\frac{e^\alpha}{(c\alpha)^\beta}$ vertices of degree x, the probability that none of them has the above property is about

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

where we have

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

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 above argument 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 $2E/3$ 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 (*). Hence, Y_i^* satisfies:

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

By Azuma's martingale inequality, we have

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

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

The above inequality implies that with proability at least $1 - o(n^{-1})$, $Y_i^* \geq i > 0$ when $i > \lceil 101\alpha^2 \rceil$. Since Y_i^* decreases at most by 1 at each step, Y_i^* can not 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 $Y_n^* > 0$, that is 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 $2.1\frac{\alpha}{\log \alpha}$ vertices not in the giant component is at most

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

Since there is at most n connected components, we conclude that a. s. a connected component of size greater that $2.1\frac{\alpha}{\log \alpha} = \Theta(\frac{\log n}{\log\log n})$ must be the giant component.

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

$$(1 - (\frac{1}{\alpha})^x)^{\frac{e^\alpha}{x^2}} \approx e^{-(\frac{1}{\alpha})^x \frac{e^\alpha}{x^2}} = e^{-\frac{e^{0.1\alpha}}{x^2}} = o(1)$$

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

Case 3: $0 < \beta < 2$.

We use the modified branching process by choosing $x_\beta = e^{\frac{(5-2\beta)\alpha}{(6-2\beta)\beta}}$. If a component has more than $2E/3$ edges, it is the unqiue giant component and we are done. Otherwise, we have

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

Hence, Y_i^* satisfies:

- $Y_0^* \geq \frac{5}{C^2} e^{\frac{(2-\beta)\alpha}{(3-\beta)\beta}}$.
- $-1 \leq Y_i^* - Y_{i-1}^* \leq e^{\frac{(5-2\beta)\alpha}{(6-2\beta)\beta}}$.
- $E(Y_i^* - Y_{i-1}^*) \approx \frac{1}{E} \sum_{x=1}^{e^{\frac{(5-2\beta)\alpha}{(6-2\beta)\beta}}} x(x-2)\lfloor \frac{e^\alpha}{x^\beta} \rfloor - E(W_i)$
  $\approx Ce^{\frac{\alpha}{2\beta}}$

Here C is a constant depending only on β.

By Azuma's martingale inequality, we have

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

provided $i \geq \frac{5}{C^2} e^{\frac{(2-\beta)\alpha}{(3-\beta)\beta}}$.

The above inequality shows that with probability at least $1 - o(n^{-1})$, $Y_i^* > \frac{1}{2}Ce^{\frac{\alpha}{2\beta}}i > 0$ provided $i > \frac{5}{C^2} e^{\frac{(2-\beta)\alpha}{(3-\beta)\beta}}$. Since Y_i^* decreases at most by 1 at each step, Y_i^* can not be zero if $i \leq \frac{5}{C^2} e^{\frac{(2-\beta)\alpha}{(3-\beta)\beta}}$. So $Y_i^* > 0$ for all i. In othe words, a. s. the branching processing will not stop. However, it is impossible to have $Y_n^* > 0$, that is a contradiction. So, a. s. all vertices with degree more than $\frac{5}{C^2} e^{\frac{(2-\beta)\alpha}{(3-\beta)\beta}}$ are in the giant component. The probability that a random vertex is in the giant component is at most

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

The probability that all $2\lfloor \frac{3-\beta}{2-\beta} \rfloor + 1$ vertices are not in the giant vertex is at most

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

Since there is at most n connected component, we conclude that a. s. a connected component of size greater that $2\lfloor \frac{3-\beta}{2-\beta} \rfloor = \Theta(1)$ 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

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

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

$$(1-\Theta(\frac{e^{\alpha}}{e^{\frac{2\alpha}{\beta}}}))^{e^{\alpha}} \approx e^{-\Theta(e^{2\alpha-\frac{2\alpha}{\beta}})} \approx o(1)$$

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 $\geq M$ 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

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

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

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

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

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

Hence, 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 which 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.

In Figure 2, we plot the number of vertices versus the indegree for the call graph of a particular day. Let $y(i)$ be the number of vertices with indegree i. For each i such that $y(i) > 0$, a $\times$ is marked at the coordinate $(i, y(i))$. As similar plot is shown in Figure 1 for the outdegree. Plots of the number of vertices versus the indegree or outdegree for the call graphs of other days are very similar. For the same call graph in Figure 3 we plot 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 monotonically decrease with increasing degree. Moreover, the call graph is directed, i.e., 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 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 of Figure 2. 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 2 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 2 is approximately 30×10^6. The total number of nodes in the call graph can be estimated by $\zeta(2.1) * e^\alpha = 1.56 * e^\alpha \approx 47 \times 10^6$

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 3[1]. 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 with an (α, β)-graph with $\beta > 4$ (Theorem 1 yields a log-log linear relation between number of components and component size for $\beta > 4$.) This intuitively seems true 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.

[1]This data was 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, Buchsbaum, and Westbrook [1] for computing connected components of massive graphs.

There are numerous questions that 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 referees for their invaluable comments.

References

[1] J. Abello, A. Buchsbaum, and J. Westbrook, A functional approach to external graph algorithms, *Proc. 6th European Symposium on Algorithms,* pp. 332–343, 1998.

[2] W. Aiello, F. Chung, L. Lu, A random graph model for massive graphs, *Proceedings of the Thirtysecond Annual ACM Symposium on Theory of Computing*, (2000), 171-180.

[3] W. Aiello, F. Chung, L. Lu, Random evolution of power law graphs, *Handbook of Massive Data Sets*, Vol. 2, (eds. J. Abello et al.), to appear.

[4] R. Albert, H. Jeong and A. Barabási, Diameter of the World Wide Web, *Nature*, **401**, September 9, 1999.

[5] N. Alon and J. H. Spencer, *The Probabilistic Method*, Wiley and Sons, New York, 1992.

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

[7] A. Barabási, R. Albert, and H. Jeong Scale-free characteristics of random networks: the topology of the world wide web, *Elsevier Preprint* August 6, 1999.

[8] P. Erdős and A. Rényi, On the evolution of random graphs, *Publ. Math. Inst. Hung. Acad. Sci.* **5** (1960), 17–61.

[9] P. Erdős and A. Rényi, On the strength of connectedness of random graphs, *Acta Math. Acad. Sci. Hungar.* **12** (1961), 261-267.

[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationships of the internet topology, *Proceedings of the ACM SIGCOM Conference*, Cambridge, MA, 1999.

[11] Brian Hayes, Graph theory in practice: Part II, *American Scientists*, **88**, (March-April, 2000), 104-109.

[12] J. Kleinberg, S. R. Kumar, P. Raphavan, S. Rajagopalan and A. Tomkins, The web as a graph: Measurements, models and methods, *Proceedings of the International Conference on Combinatorics and Computing*, July 26–28, 1999.

[13] S. R. Kumar, P. Raphavan, S. Rajagopalan and A. Tomkins, Trawling the web for emerging cyber communities, *Proceedings of the 8th World Wide Web Conference*, Toronto, 1999.

[14] S. R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, Extracting large-scale knowledge bases from the web, *Proceedings of the 25th VLDB Conference*, Edinburgh, Scotland, September 7–10, 1999.

[15] S. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal, Stochastic models for the Web graph, *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, (2000).

[16] Tomasz Łuczak, Sparse random graphs with a given degree sequence, *Random Graphs*, vol 2 (Poznań, 1989), 165-182, Wiley, New York, 1992.

[17] Michael Molloy and Bruce Reed, A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, Vol. **6**, no. 2 and 3 (1995). 161-179.

[18] Michael Molloy and Bruce Reed, The size of the giant component of a random graph with a given degree sequence, *Combin. Probab. Comput.* **7**, no. (1998), 295-305.

[19] P. Raghavan, personal communication.

[20] N. C. Wormald, The asymptotic connectivity of labelled regular graphs, *J. Comb. Theory* (B) **31** (1981), 156-167.

[21] N. C. Wormald, Models of random regular graphs, *Surveys in Combinatorics*, 1999 (LMS Lecture Note Series 267, Eds J.D.Lamb and D.A.Preece), 239–298.

[22]

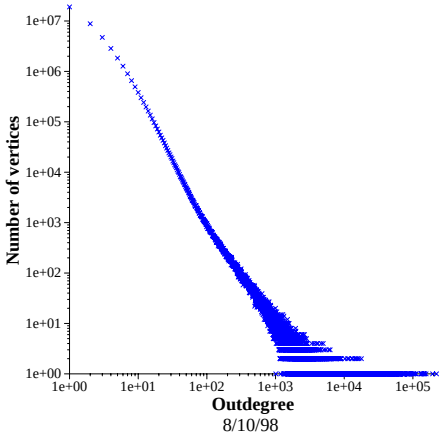

Figure 1: *The number of vertices for each possible outdegree for the call graph of a typical day.*

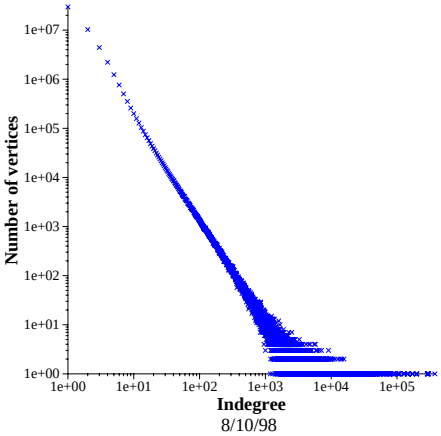

Figure 2: *The number of vertices for each possible indegree for the call graph of a typical day.*

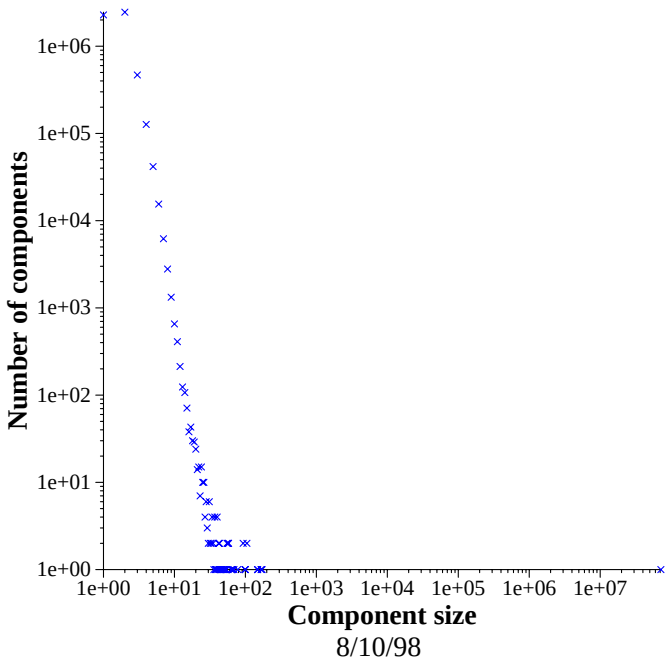

Figure 3: *The number of connected components for each possible component size for the call graph of a typical day.*