Connected components in random graphs with given expected degree sequences

Fan Chung *† Linyuan Lu *

Abstract

We consider a family of random graphs with a given expected degree sequence. Each edge is chosen independently with probability proportional to the product of the expected degrees of its two endpoints. We examine the distribution of the sizes/volumes of the connected components which turns out depending primarily on the average degree d and the second-order average degree $\tilde{d}$. Here $\tilde{d}$ denotes the ratio of the sum of squares of the expected degree and the sum of the expected degrees of vertices. For example, we prove that the giant component exists if the expected average degree d is at least 1, and there is no giant component if the expected second-order average degree $\tilde{d}$ is at most 1. Examples are given to illustrate that both bounds are best possible.

1 Introduction

The primary subject in the study of random graph theory is the classical random graph $G(n,p)$, as introduced by Erdős and Rényi in 1959 [19]. In $G(n,p)$, every pair of a set of n vertices is chosen to be an edge with probability p. Such random graphs are fundamental and useful for modeling problems in many applications. However, a random graph in $G(n,p)$ has the same expected degree at every vertex and therefore does not capture some of the main behaviors of numerous graphs arising from the real world. It is imperative to consider a versatile and generalized version of random graphs. In this paper, we consider random graphs with given expected degree sequences which contains the classical random graphs as a special case and also include the so-called "power-law" degree distributions occurred in various realistic graphs.

*University of California, San Diego

†Research supported in part by NSF Grant DMS 0100472

It has been observed that many real graphs occurring in the Internet, social sciences, computational biology and nature have degrees obeying a power law [1, 2, 3, 7, 8, 12, 13, 20, 21, 25, 26, 36]. Namely, the fraction of vertices with degree d is proportional to $1/d^\alpha$ for some constant $\alpha > 0$. Although here we consider random graphs with general expected degree distributions, special emphasis will be given to sparse graphs (with average degree a small constant) and to power law graphs (see Section 9). The methods and results that we derive in dealing with random graphs with given expected degree distribution are useful not only for modeling and analyzing realistic graphs but also leading to improvements for problems on classical random graphs [14, 29].

In this paper, we consider the following class of random graphs. We start with a given degree sequence $\mathbf{w} = (w_1, w_2, \ldots, w_n)$. The vertex v_i is assigned vertex weight w_i. The edges are chosen independently and randomly according to the vertex weights as follows. The probability p_{ij} that there is an edge between v_i and v_j is proportional to the product $w_i w_j$ where i and j are not required to be distinct. There are possible loops at v_i with probability proportional to w_i^2. We have

$$p_{ij} = \frac{w_i w_j}{\sum_k w_k}. \tag{1}$$

Throughout the paper we assume that $\max_i w_i^2 < \sum_k w_k$ so that $p_{ij} \leq 1$ for all i and j. We remark that the assumption $\max_i w_i^2 < \sum_k w_k$ implies that the sequence w_i is graphic (in the sense that it satisfies the necessary and sufficient condition for a sequence to be realized by a graph [18]) except that we do not require the w_i's to be integers.

We denote a random graph with a given expected degree sequence $\mathbf{w}$ by $G(\mathbf{w})$. For example, the typical random graph $G(n, p)$ (see [19]) on n vertices and edge density p is just a random graph with expected degree sequence $(pn, pn, \ldots, pn)$. The random graph $G(\mathbf{w})$ is different from the random graphs with a prescribed degree sequence as considered by Molloy and Reed. In [31, 32], Molloy and Reed obtained results on the sizes of connected components for random graphs with prescribed degree sequences which satisfy certain "smoothing"

conditions. There are also a number of evolution models for generating a power-law degree random graphs as in Bollobás, Spencer et al. [11], Cooper and Freeze [17] and Aiello, Chung and Lu [2]. In Section 8, we will describe and compare these models and related results.

Here we give some definitions. The expected average degree d of a random graph G in $G(\mathbf{w})$ is defined to be

$$d = \frac{1}{n}\sum_{i=1}^{n} w_i.$$

For a subset S of vertices, the volume of S, denoted by $\mathrm{Vol}(S)$, is the sum of weights of vertices in S.

$$\mathrm{Vol}(S) = \sum_{v_i \in S} w_i$$

In particular, the volume $\mathrm{Vol}(G)$ of $G(\mathbf{w})$ is just $\sum_i w_i$. The edge probability p_{ij} in (1) can be written as:

$$p_{ij} = \frac{w_i w_j}{\mathrm{Vol}(G)} = w_i w_j \rho$$

where

$$\rho := \frac{1}{\mathrm{Vol}(G)} = \frac{1}{nd}.$$

A connected component C is said to be *ϵ-small* for an $\epsilon < 1/2$ if the volume of C is at most $\epsilon\mathrm{Vol}(G)$. We say that a component is *c-giant* if its volume is at least $c\mathrm{Vol}(G)$, for some small constant $c > 0$. A giant component, if exists, is almost surely unique (to be proved later in Section 7).

For a subset S of vertices, a typical measure is the number of vertices in S that we call the size of S. In the classical random graph $G(n,p)$, a giant component is a connected component having at least $c_1 n$ vertices and at least $c_2 e(G)$ edges for some constants c_1 and c_2, where $e(G)$ denote the total number of edges in G. Our definition of the giant component involves the volume instead of the size of the connected component. In fact, the definition for the giant component using the size of the component simply does not work for random graphs with general degree distributions, as illustrated by the following example.

Example 1: We consider that the weight sequence $\mathbf{w}$ consisting of n^α vertices with weight 2 and the other vertices with weight 0. Here α is a constant satisfying $\frac{1}{2} < \alpha < 1$. The random graph $G(\mathbf{w})$ is the union of a classical random graph $G(n^\alpha, \frac{2}{n^\alpha})$ and some isolated vertices. Therefore, the largest connected component have $\Theta(n^\alpha)$ vertices and $\Theta(e(G))$ edges.

We remark that for the case of expected degrees within a constant factor of each other, the size and the volume of S are of the same order. Also, an upper bound for the volume of a connected component serves as an upper bound of the size of a connected component.

If the average degree $d \geq 1 + \delta$, where δ is a positive constant, we will show that almost surely any ϵ-small connected component has size at most $O(\log n)$ (detailed in Theorems 1-2) and we will call them small components. A general upper bound for the size of the small components will be derived in terms of the average degree d. We will show that this upper bound is asymptotically best possible for certain ranges of d.

Here we state the main results which will be proved in subsequent sections.

Theorem 1 *For any positive $\epsilon < 1$ and $d > \frac{4}{e(1-\epsilon)^2} \approx (1 + 2\epsilon)1.4715\ldots$, in a random graph with a given expected degree sequence, almost surely every connected component either has volume at least ϵn or has size at most $\frac{\log n}{1+\log d-\log 4+2\log(1-\epsilon)}$. where d is the expected average degree. The upper bound $\frac{\log n}{1+\log d-\log 4}$ for small components is asymptotically best possible for large d.*

Theorem 2 *For any positive $\epsilon < 1$ and d satisfying $\frac{1}{1-\epsilon} < d < \frac{2}{1-\epsilon}$, in a random graph with a given expected degree sequence, every connected component almost surely either has volume at least ϵn or has at most $\frac{\log n}{d-1-\log d-\epsilon d}$ vertices, where d is expected average degree. This upper bound $\frac{\log n}{d-1-\log d}$ is asymptotically best possible.*

We consider the second-order average degree $\tilde{d}$ which is the weighted average of the squares of the vertex weights. Namely,

$$\tilde{d} = \sum w_i^2 \rho.$$

Clearly,

$$\tilde{d} = \frac{\sum w_i^2}{\sum w_i} \geq \frac{\sum w_i}{n} = d.$$

For the classical random graphs $G(n,p)$, we have $\tilde{d} = d = np$. It was shown in the seminal paper of Erdős and Rényi [19] that there is a giant component when $np \geq 1+\epsilon$, while there is no giant component when $np \leq 1-\epsilon$. Furthermore, there is a double jump around $np = 1$, where the largest component have size of $\Theta(n^{2/3})$ if $|np-1| = o(n^{-1/3})$. For random graphs $G(\mathbf{w})$ of general degree distribution, the evolution is more complicated. We will show that all components are small if $\tilde{d} < 1-\epsilon$ and there is a giant component if $d > 1+\epsilon$.

Theorem 3 *For a random graph G with a given expected degree sequence, almost surely G has a unique giant component, if the average degree satisfies $d \geq 1+\delta$, where δ is a positive constant. Moreover, we have*

(i). *If $d \geq e$, the volume of the unique giant component is at least*

$$(1 - \frac{2}{\sqrt{de}} + o(1))\mathrm{Vol}(G).$$

(ii). *If $1+\delta \leq d \leq e$, the volume of the unique giant component is at least*

$$(1 - \frac{1+\log d}{d} + o(1))\mathrm{Vol}(G).$$

If the second-order average degree $\tilde{d} \leq 1-\delta$, then almost surely, there is no giant component.

The proof of Theorem 3 can be found in Section 7. It is natural to question the relationship of the degrees to the emergence of the giant component for the range of $\tilde{d} > 1 > d$. The examples in Section 3 show that either case can occur for a general degree distribution when $\tilde{d} > 1 > d$. Therefore the general problems about phase transitions or double jumps for an arbitrary degree sequence becomes mute. It would be interesting, for example, to identify or characterize degree distributions for which the phase transition occurs.

2 Basic facts and examples

We will use the following inequality which is a weighted generalization of the Chernoff inequalities for binomial distribution:

Lemma 1 *Let $X_1, \ldots, X_n$ be independent random variables with*

$$Pr(X_i = 1) = p_i, \qquad Pr(X_i = 0) = 1 - p_i$$

For $X = \sum_{i=1}^n a_i X_i$, we have $E(X) = \sum_{i=1}^n a_i p_i$ and we define $\nu = \sum_{i=1}^n a_i^2 p_i$. Then we have

$$Pr(X < E(X) - \lambda) \leq e^{-\lambda^2/2\nu} \quad (2)$$

$$Pr(X > E(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\nu + a\lambda/3)}} \quad (3)$$

where $a = \max\{a_1, a_2, \ldots, a_n\}$.

Inequality (3) is a corollary of a general concentration inequality (see Theorem 2.7 in the survey paper by McDiarmid [30]). Inequality (2) which is a slight improvement of the inequality in [30] can be proved as follows.

Proof: For any $0 \leq p \leq 1$, and $x \geq 0$, we denote $f(x) = px + \ln(1-p+pe^{-x})$ and $g(x) = px^2/2$. Then we have $f(0) = g(0) = 0$, and $f'(0) = g'(0) = 0$. Also,

$$\begin{aligned} f''(x) &= \frac{p(1-p)e^{-x}}{(1-p+pe^{-x})^2} \\ &= \frac{p(1-p)e^{-x}}{(1-p+e^{-x}-(1-p)e^{-x})^2} \\ &\leq \frac{p(1-p)e^{-x}}{(2\sqrt{(1-p)e^{-x}}-(1-p)e^{-x})^2} \\ &\leq \frac{p(1-p)e^{-x}}{(\sqrt{(1-p)e^{-x}})^2} \\ &= p \\ &= g''(x). \end{aligned}$$

Hence $px + \ln(1-p+pe^{-x}) \leq px^2/2$ for any $x \geq 0$.

For any $t > 0$, we have

$$E(e^{-a_i t(X_i - p_i)}) = p_i e^{-ta_i(1-p_i)} + (1-p_i)e^{p_i t a_i} = e^{p_i t a_i + ln(1-p_i+p_i e^{-ta_i})} \leq e^{\frac{p_i (ta_i)^2}{2}}.$$

Hence

$$\begin{aligned} E(e^{-t(X-\sum_{i=1}^{n} a_i p_i)}) &= \prod_{i=1}^{n} e^{-t(X_i - p_i a_i)} \\ &\leq \prod_{i=1}^{n} e^{\frac{p_i (t a_i)^2}{2}} \\ &= e^{\sum_{i=1}^{n} \frac{p_i (t a_i)^2}{2}} \\ &= e^{\frac{t^2 \nu}{2}} \end{aligned}$$

We have

$$\begin{aligned} Pr(X - \sum_{i=1}^{n} a_i p_i < -\lambda) &= Pr(e^{-t(X-\sum_{i=1}^{n} a_i p_i)} > e^{t\lambda}) \\ &\leq E(e^{-t(X-\sum_{i=1}^{n} a_i p_i)}) e^{-t\lambda} \\ &\leq e^{\frac{t^2 \nu}{2} - t\lambda} \\ &= e^{-\frac{\lambda^2}{2\nu}} \end{aligned}$$

by choosing $t = \frac{\lambda}{\nu}$. This completes the proof of Lemma 1.

□

We note that the special case of $a_i = 1$ for all i is the usual inequality that is included in most books in random graph theory and probability (e.g., [24]). As immediate consequences of Lemma 1, the following facts then follow.

Fact 1: For a graph G in $G(\mathbf{w})$, with probability $1 - e^{-c^2/2}$, the number d_i of edges incident to a vertex v_i satisfies

$$d_i > w_i - c\sqrt{w_i}$$

and

$$Prob(d_i < (1+\epsilon)w_i) > 1 - e^{-\epsilon^2 w_i/(2+2\epsilon/3)}.$$

Fact 2: With probability $1 - 2e^{-c^2/2}$, the number $e(G)$ of edges in G, satisfies

$$2e(G) > \mathrm{Vol}(G) - c\sqrt{\mathrm{Vol}(G)}.$$

In the other direction,

$$Prob(2e(G) < (1+\epsilon)\mathrm{Vol}(G)) > 1 - e^{-\epsilon^2 \mathrm{Vol}(G)/(2+2\epsilon/3)}.$$

With probability $1 - \frac{2}{n}$, all vertices v_i satisfy

$$2\sqrt{w_i \log n} \leq d_{v_i} - w_i \leq \frac{2}{3}\log n + \sqrt{(\frac{2}{3}\log n)^2 + 4w_i \log n}.$$

Fact 3: With probability at least $1 - e^{-c}$, the number of edges $e(S)$ between pairs of vertices in S is at least $\frac{1}{2}\mathrm{Vol}(S)^2\rho - \mathrm{Vol}(S)\sqrt{\rho c}$.

Proof: $e(S)$ can be expressed as the sum of independent 0-1 variables $X_{u,v}$, which takes value 1 with probability $w_u w_v \rho$. With probability at least $1 - e^{-c}$, the number of edges between pairs of vertices in S is at least:

$$\begin{aligned} e(S) &= \frac{1}{2}\sum_{u,v\in S} X_{u,v} \\ &\geq E(e(S)) - \sqrt{2E(e(S))c} \\ &= \frac{1}{2}\sum_{u,v\in S} w_u w_v \rho - \sqrt{\sum_{u,v\in S} w_u w_v \rho c} \\ &\approx \frac{1}{2}\mathrm{Vol}(S)^2\rho - \mathrm{Vol}(S)\sqrt{\rho c}. \end{aligned}$$

In the remainder of this section, we will give several examples with proofs which illustrate the sharpness of the main results. These examples are also instrumental for developing methods later on for dealing with random graphs with given expected degree distributions.

Example 2: For the following choices of the weight distribution $\mathbf{w}$ with $d \leq 1$ and $\tilde{d} > 1$, the random graph in $G(\mathbf{w})$ almost surely has no giant component.

Let ϵ be a constant satisfying $1 > \epsilon > 0$. The weight sequence $\mathbf{w}$ of expected degrees will be chosen as follows. For each of the first $n - m$ vertices, the weight is set to $1 - \epsilon$. For each of the other m vertices, the weight is set to x. Here m and x satisfy

$$mx = o(\frac{n}{\log n}) \qquad \text{and } mx^2 > Cn.$$

Here $C > 1$. (For example, we can choose $m = \lceil \log n \rceil$, $x = \sqrt{(1-\epsilon)n/2}$ and $C = 10$.) We have

$$\mathrm{Vol}(G) = (n-m)(1-\epsilon) + mx \approx (1-\epsilon)n.$$

$$d = \frac{\mathrm{Vol}(G)}{n} \approx (1-\epsilon).$$

$$\tilde{d} = \frac{\mathrm{Vol}_2(G)}{\mathrm{Vol}(G)} = \frac{(n-m)(1-\epsilon)^2 + mx^2}{(1-\epsilon)n} > 1 - \epsilon + \frac{C}{1-\epsilon} > 1.$$

Let S_1 be the set of vertices with weight $1-\epsilon$, and S_2 be the set of vertices with weight x. We let G_i denote the induced graph of G on S_i, for $i = 1, 2$.

A classical result in [19] states that almost surely $G(N,p)$ has a giant component if $Np > 1+\epsilon$ and $G(N,p)$ does not have a giant component if $Np < 1-\epsilon$ while all components have sizes of at most $\Theta(\log N)$.

To apply the above results to G_1, we select

$$N = n - m \approx n,$$

$$p = (1-\epsilon)^2\rho \approx (1-\epsilon)\frac{1}{n},$$

Hence, we have $Np \approx (1-\epsilon) < 1$. All components of G_1 have size at most $\Theta(\log N) = \Theta(\log n)$.

We will next show that there is no giant component in G by establishing upper bounds for the sizes and volumes of all components in G. We first construct an auxiliary graph G' from G as follows. A new vertex v is added to G, and is connected to all vertices in S_2 but to no vertex in S_1. The following facts are immediate.

1. Every connected component of G must be contained in some component in G'.

2. G' has a special component C containing v and all vertices in S_2.

3. Components of G' other than C are components of G_1. They can have at most $\Theta(\log n)$ vertices with volume at most $\Theta(\log n)$.

Now we will use the branching process starting from v to reveal the component C. Here, for a subset S, we define the i-neighborhood $\Gamma_i(S) = \{u : d(u,S) = i\}$. We have

$$\Gamma_1(v) = S_2.$$

For each $u \in S_1$, the probability that $u \in \Gamma(S_2)$ is

$$1 - (1-(1-\epsilon)x\rho)^m \approx (1-\epsilon)mx\rho \quad \text{since } mx\rho = o(1).$$

The size of $\Gamma(S_2)$ can be upper bounded by a sum of $n-m$ independent 0-1 variables. The probability of each random variable with value one is about $(1-\epsilon)mx\rho$. These random variables are independent to each other. Let $X = \sum_i X_i$ be a sum of independent 0-1 variables. Using Lemma 1, we have

$$Pr(X - E(X) > \lambda) < e^{-\frac{\lambda^2}{2(E(X)+\lambda/3)}}.$$

by choosing

$$\lambda = E(X) \approx (n-m)(1-\epsilon)mx\rho \approx mx.$$

With probability at least $1 - e^{-3mx/8} = 1 - o(1)$, the size of $\Gamma(S_2)$ is at most $2mx$. Note that $\Gamma_2(v) = \Gamma(S_2)$ are completely contained in S_1, and so are the $i-$ neighborhoods $\Gamma_i(v)$ for all $i \geq 2$. Since in G_1, any branching process can expand at most $\Theta(\log n)$ vertices, the total size of C can have at most

$$2mx\Theta(\log n) + m + 1 = \Theta(mx \log n).$$

The volume of $C \setminus \{v\}$ is at most

$$2mx\Theta(\log n)(1-\epsilon) + mx = \Theta(mx \log n).$$

Hence each component in G can have volume at most $\Theta(mx \log n) = o(n)$. Thus, there is no giant component in G.

Example 3: For the following choice of the weight distribution $\mathbf{w}$ with $d < 1$ and $\tilde{d} > 1$, the random graph $G(\mathbf{w})$ almost surely has a giant component.

Let M be a very large but fixed constant. For each of the first $\lceil \frac{(M-1)n}{M} \rceil$ vertices, the weight is set to be $x = o(1)$. For the other $\frac{n}{M}$ vertices, each weight is set to $1+\epsilon$. In this example, we have

$$\begin{aligned}
\mathrm{Vol}(G) &\approx \frac{(M-1)n}{M}x + \frac{1+\epsilon}{M}n = \frac{1+\epsilon+o(1)}{M}n, \\
d &= \frac{\mathrm{Vol}(G)}{n} = \frac{1+\epsilon+o(1)}{M} << 1, \\
\tilde{d} &= \frac{\mathrm{Vol}_2(G)}{\mathrm{Vol}(G)} = 1+\epsilon-o(1) > 1.
\end{aligned}$$

Note that $G(\mathbf{w})$ contains a classical random graph $G(N,p)$, where $N = \frac{n}{M}$, and $p = \frac{M(1+\epsilon+o(1))}{n}$. Since $Np = \frac{n}{M}\frac{M(1+\epsilon+o(1))}{n} = 1+\epsilon+o(1) > 1$, $G(N,p)$ has a giant component of size $\Theta(N) = \Theta(n)$. The component of G containing this connected subset will have at least $\Theta(n)$ vertices and at least $\Theta(\mathrm{Vol}(G))$ volume.

3 The expected number of components of size k

In this section, we consider the probability of the existence of a connected component of size k. This is useful later for proving the uniqueness of the giant component.

Suppose that we have a subset of vertices $S = \{v_{i_1}, v_{i_2} \ldots, v_{i_k}\}$ with weights $w_{i_1}, w_{i_2}, \ldots, w_{i_k}$. The probability that there is no edge leaving S is

$$\begin{aligned} \prod_{v_i \in S, v_j \notin S} & (1 - w_i w_j \rho) \\ \approx & \; e^{-\rho \sum_{v_i \in S, v_j \notin S} w_i w_j} \\ = & \; e^{-\rho \mathrm{Vol}(S)(\mathrm{Vol}(G) - \mathrm{Vol}(S))} \end{aligned} \tag{4}$$

where $\rho = \frac{1}{\sum_{i=1}^{n} w_i} = \frac{1}{nd}$. We next consider the edges inside S. If S is a connected component, the induced subgraph on S contains at least one spanning tree T. The probability of containing a spanning tree T is

$$Pr(T) = \prod_{(v_{i_j} v_{i_l}) \in E(T)} w_{i_j} w_{i_l} \rho.$$

Hence the probability of the existence of a connected spanning graph on S is at most

$$\sum_T Pr(T) = \sum_T \prod_{(v_{i_j} v_{i_l}) \in E(T)} w_{i_j} w_{i_l} \rho,$$

where T ranges over all spanning trees on S.

By a generalized version of celebrated matrix-tree Theorem [34], the above sum equals the determinant of any $k-1$ by $k-1$ principal sub-matrix of the matrix $D - A$, where A is the weight matrix

$$A = \begin{pmatrix} 0 & w_{i_1}w_{i_2}\rho & \cdots & w_{i_1}w_{i_k}\rho \\ w_{i_2}w_{i_1}\rho & 0 & \cdots & w_{i_2}w_{i_k}\rho \\ \vdots & \vdots & \ddots & \vdots \\ w_{i_k}w_{i_1}\rho & w_{i_k}w_{i_2}\rho & \cdots & 0 \end{pmatrix}$$

and D is the diagonal matrix $diag(w_{i_1}(\mathrm{Vol}(S) - w_{i_1})\rho, \ldots, w_{i_k}(\mathrm{Vol}(S)w_{i_k} - w_{i_k})\rho)$. By evaluating the determinant, we conclude that

$$\sum_T P(T) = w_{i_1}w_{i_2}\cdots w_{i_k}\mathrm{Vol}(S)^{k-2}\rho^{k-1}. \tag{5}$$

Let X_k be the random variable of the number of the components with size k. By combining (4) and (5), we have proved the following:

Lemma 2 *The expected value $E(X_k)$ of the number of connected components of size k is at most*

$$E(X_k) \quad \leq \quad \sum_S w_{i_1}w_{i_2}\cdots w_{i_k}\mathrm{Vol}(S)^{k-2}\rho^{k-1}e^{-\mathrm{Vol}(S)(1-\mathrm{Vol}(S)/\mathrm{Vol}(G))} \tag{6}$$

where the sum ranges over all sets S of k vertices.

Lemma 3 *For a positive $\epsilon < 1$, The expected value $E(Y_k)$ of the number of ϵ-small connected components on size k is at most*

$$E(Y_k) \quad \leq \quad \sum_S w_{i_1}w_{i_2}\cdots w_{i_k}\mathrm{Vol}(S)^{k-2}\rho^{k-1}e^{-\mathrm{Vol}(S)(1-\epsilon)} \tag{7}$$

where the sum ranges over all set S of k vertices with $\mathrm{Vol}(S) < \epsilon\mathrm{Vol}(G)$.

4 Proof of Theorem 1

Suppose that G is a random graph with a given expected degree sequence $\mathbf{w}$. In addition, we assume that the expected average degree d satisfies $d > 1$. We want to show that the expected number $E(Y_k)$ of ϵ-small components of size k is small if k is sufficiently large.

We follow the notation in Section 2. From Lemma 3, it suffices to upper bound

$$f(k) = \sum_S w_{i_1}w_{i_2}\cdots w_{i_k}\mathrm{Vol}(S)^{k-2}\rho^{k-1}e^{-\mathrm{Vol}(S)(1-\epsilon)}$$

By using the fact that the function $x^{2k-2}e^{-x(1-\epsilon)}$ achieves its maximum value at $x = (2k-2)/(1-\epsilon)$, we have

$$\begin{aligned} f(k) &= \sum_S w_{i_1} w_{i_2} \cdots w_{i_k} \mathrm{Vol}(S)^{k-2} \rho^{k-1} e^{-\mathrm{Vol}(S)(1-\epsilon)} \\ &\leq \sum_S \frac{\rho^{k-1}}{k^k} \mathrm{Vol}(S)^{2k-2} e^{-\mathrm{Vol}(S)(1-\epsilon)} \\ &\leq \sum_S \frac{\rho^{k-1}}{k^k} (\frac{2k-2}{1-\epsilon})^{2k-2} e^{-(2k-2)} \\ &\leq \frac{n^k}{k!} \frac{\rho^{k-1}}{k^k} (\frac{2k-2}{1-\epsilon})^{2k-2} e^{-(2k-2)} \\ &\leq \frac{1}{4\rho(k-1)^2} (n\rho)^k (\frac{2}{1-\epsilon})^{2k} e^{-k} \\ &\leq \frac{1}{4\rho(k-1)^2} (\frac{4}{de(1-\epsilon)^2})^k \end{aligned}$$

The above inequality is useful when $d > \frac{4}{e(1-\epsilon)^2}$ which is the assumption for Theorem 1. When k satisfies $\frac{\log n}{1+\log d - \log(4) - 2\epsilon} < k < \frac{2\log n}{1+\log d - \log(4) - 2\epsilon}$, we have

$$f(k) \leq \frac{1}{4n\rho(k-1)^2} = O(\frac{1}{\log n}).$$

When k satisfies $\frac{2\log n}{1+\log d - \log(4) - 2\epsilon} \leq k \leq n$, we have

$$f(k) \leq \frac{1}{4n^2\rho(k-1)^2} = O(\frac{1}{n\log n}).$$

We write $k_0 = \frac{\log n}{1+\log d - \log(4) - 2\epsilon}$. The probability that a small component of size $k > k_0$ is at most

$$\sum_{k>k_0} f(k) \leq \frac{\log n}{1+\log d - \log(4) - 2\epsilon} \times o(\frac{1}{\log n}) + n \times o(\frac{1}{n\log n}) = o(1).$$

Therefore, almost surely the size of a ϵ-small component is at most $k_0 = \frac{\log n}{1+\log d - \log 4 - 2\epsilon}$. We have proved the first part of Theorem 1.

To show that the above upper bound for the size of a small component is asymptotically best possible for large d, we consider the following example.

Example 4: We consider a random graph with the following weights as the expected degree sequence. Here we assume that $d > 10$.

There are $n^{2/3}$ vertices with weights $(d-1)n^{1/3} + 1$. The rest of $n - n^{2/3}$ vertices have weights 1. The average weight (degree) is exactly d.

Let S_1 denote the set of vertices with weight 1, and S_2 denote the set of vertices with weight $(d-1)n^{1/3}+1$. Let G_i be the induced graph of G on S_i, for $i = 1, 2$. The graph G_2 is the classical random graph $G(N, p)$ with $N = n^{2/3}$ and $Np = n^{2/3}((d-1)n^{1/3}+1)^2/(nd) = \Theta(\sqrt{N})$. Almost surely G_2 is connected. In fact, G_2 is contained in the giant component of G. Let c denote the fraction of vertices, which is not in the giant component. We claim that c is bounded away from 0.

To prove the claim, we consider a special branching process. We first reveal all edges in G_2. Then we examine the neighborhood of S_2 in S_1, the 2-neighborhood of S_2, and so on, which grows into the giant component of G. For any vertex $u \in S_1$, the probability of u in $\Gamma(S_2)$ is

$$1 - \left(1 - \frac{(d-1)n^{1/3}+1}{nd}\right)^{n^{2/3}} \approx 1 - e^{-1+\frac{1}{d}}.$$

The size of $\Gamma(S_2)$ can be well approximated by the binomial distribution with $N = n - n^{2/3}$ and $p = 1 - e^{-1+\frac{1}{d}}$. Thus with high probability, its size is about $(1 - e^{-1+\frac{1}{d}})n$. We will estimate the size of $\Gamma_i(S_2)$ by induction. Suppose $|\Gamma_i(S_i)|$ is highly concentrated on $a_i n$ for some constant a_i, for $i \geq 2$. Let $c_i = 1 - \sum_{k=1}^{i} a_i$. For any vertex u not in $\cup_{j \leq i} \Gamma_i(S_i)$, the probability of $u \in \Gamma_{i+1}(S_2)$ is

$$1 - \left(1 - \frac{1}{nd}\right)^{a_i n} \approx 1 - e^{-\frac{a_i}{d}}.$$

The size of $\Gamma_{i+1}(S_2)$ can be well approximated by the binomial distribution with $N = c_i n$ and $p = 1 - e^{-\frac{a_i}{d}}$. By the definition of a_i. We have

$$a_{i+1} = c_i(1 - e^{-\frac{a_i}{d}}).$$

$$c_{i+1} = c_i - a_{i+1} = c_i e^{-\frac{a_i}{d}}.$$

Hence

$$c_{i+1} = c_1 \prod_{k=1}^{i} e^{-\frac{a_k}{d}} = (1 - e^{-1+\frac{1}{d}})e^{-\frac{1-c_i}{d}}.$$

By the above recurrence for c_i, we see that $c = \lim_{i->\infty} c_i$ exists and satisfies

$$c = (1 - e^{-1+\frac{1}{d}})e^{-\frac{1-c}{d}}.$$

By some elementary analysis, the above equation has a unique solution of c in $[0,1]$ for $d > 1$ and the solution for c increases as d increases. Since we choose $d > 10$, c is bounded away from zero. The claim is proved.

The size of the second largest component can be estimated as follows. After removing the giant component from G, the remain graph is a classical random graph $G(t,p)$ with $t = cn$ and $p = \frac{1}{nd} = \frac{c}{dt}$. By the classical result of Erdős and Rényi [19], the largest component of $G(t, \frac{c}{dt})$ with $d < 1$ has size about

$$\frac{\log n - 5/2 \log\log n}{\frac{c}{d} - 1 - \log \frac{c}{d}} = \frac{(1+o(1)) \log n}{\log d - \log c - 1 + \frac{c}{d}}$$

The constant $\frac{1}{\log d - \log c - 1 + \frac{c}{d}}$ is asymptotically close to $\frac{1}{1 + \log d - \log 4}$ when d is large and ϵ is arbitrarily small. This completes the proof for Theorem 1.

Remark: When $d > 1$, the classical random graph $G(n, \frac{d}{n})$ has small connected components except for the giant component. In [19], it was shown that the size of the second largest connected components is approximately the same of the size of the largest connected component of $G(m, \frac{c}{m})$. Here c is the unique solution of $ce^{-c} = de^{-d}$ for c in $(0,1)$, and $m = \frac{c}{d}n$. From [19], the largest component of $G(m, \frac{c}{m})$ has size about

$$\frac{\log m - 5/2 \log\log m}{c - 1 - \log c} = \frac{(1+o(1)) \log n}{d - 1 - \log d}$$

which is smaller than the upper bound in the statement of Theorem 1.

5 Proof of Theorem 2

In this section, we consider $\frac{1}{1-\epsilon} < d < \frac{2}{1-\epsilon}$. The methods in the proof for Theorem 1 for establishing the upper bound of $f(k)$ no longer works and a different estimate for $f(k)$ is needed here. We will derive an upper bound for the expected number $E(Y_k)$ of connected components of size k by using inequality (2). First, we split $f(k)$ into two parts as follows:

$$f(k) \quad = \quad f_1(k) + f_2(k)$$

where

$$\begin{aligned} f_1(k) &= \sum_{\mathrm{Vol}(S)<dk} w_{i_1} w_{i_2} \cdots w_{i_k} \mathrm{Vol}(S)^{k-2} \rho^{k-1} e^{-\mathrm{Vol}(S)(1-\epsilon)} \\ f_2(k) &= \sum_{\mathrm{Vol}(S)\geq dk} w_{i_1} w_{i_2} \cdots w_{i_k} \mathrm{Vol}(S)^{k-2} \rho^{k-1} e^{-\mathrm{Vol}(S)(1-\epsilon)} \end{aligned}$$

To bound $f_1(k)$, we note that $x^{2k-2}e^{-x(1-\epsilon)}$ is an increasing function when $x < (2k-2)/(1-\epsilon)$. Thus we have

$$\mathrm{Vol}(S)^{2k-2} e^{-\mathrm{Vol}(S)} \leq (dk)^{2k-2} e^{-dk(1-\epsilon)}$$

since $\mathrm{Vol}(S) < dk < (2k-2)/(1-\epsilon)$. This implies

$$\begin{aligned} f_1(k) &= \sum_{\mathrm{Vol}(S)<dk} w_{i_1} \cdots w_{i_k} \mathrm{Vol}(S)^{k-2} \rho^{k-1} e^{-\mathrm{Vol}(S)(1-\epsilon)} \\ &\leq \sum_{\mathrm{Vol}(S)<dk} \frac{\rho^{k-1}}{k^k} \mathrm{Vol}(S)^{2k-2} e^{-\mathrm{Vol}(S)(1-\epsilon)} \\ &\leq \sum_{\mathrm{Vol}(S)<dk} \frac{\rho^{k-1}}{k^k} (dk)^{2k-2} e^{-dk(1-\epsilon)} \\ &\leq \binom{n}{k} \frac{\rho^{k-1}}{k^k} (dk)^{2k-2} e^{-dk(1-\epsilon)} \\ &\leq \frac{n^k}{k!} \frac{\rho^{k-1}}{k^k} (dk)^{2k-2} e^{-dk(1-\epsilon)} \\ &\leq \frac{1}{d^2 k^2 \rho} (n\rho)^k d^{2k} e^{-(d(1-\epsilon)-1)k} \\ &= \frac{n}{dk^2} \left(\frac{d}{e^{d(1-\epsilon)-1}}\right)^k \end{aligned}$$

Next, we consider bounding $f_2(k)$ from above. Note that $x^{k-2}e^{-x(1-\epsilon)}$ is a decreasing function when $x > (k-2)/(1-\epsilon)$. We have

$$\mathrm{Vol}(S)^{k-2} e^{-\mathrm{Vol}(S)(1-\epsilon)} \leq (dk)^{k-2} e^{-dk(1-\epsilon)}$$

since $\mathrm{Vol}(S) \geq dk > \frac{k-2}{1-\epsilon}$. We have

$$\begin{aligned} f_2(k) &= \sum_{\mathrm{Vol}(S)\geq dk} w_{i_1} w_{i_2} \cdots w_{i_k} \mathrm{Vol}(S)^{k-2} \rho^{k-1} e^{-\mathrm{Vol}(S)(1-\epsilon)} \\ &\leq \sum_{\mathrm{Vol}(S)\geq dk} w_{i_1} \cdots w_{i_k} \rho^{k-1} (dk)^{k-2} e^{-dk(1-\epsilon)} \\ &\leq \sum_{S} w_{i_1} w_{i_2} \cdots w_{i_k} \rho^{k-1} (dk)^{k-2} e^{-dk(1-\epsilon)} \\ &< \frac{\mathrm{Vol}(G)^k}{k!} \rho^{k-1} (dk)^{k-2} e^{-dk(1-\epsilon)} \\ &\leq \frac{1}{d^2k^2\rho} d^k e^{-(d(1-\epsilon)-1)k} \\ &\leq \frac{n}{dk^2} (\frac{d}{e^{(d(1-\epsilon)-1)}})^k \end{aligned}$$

Hence,

$$f(k) = f_1(k) + f_2(k) \leq \frac{2n}{dk^2} (\frac{de}{e^{d(1-\epsilon)-1}})^k.$$

When $\frac{\log n}{d(1-\epsilon)-1-\log d} < k < \frac{2\log n}{d(1-\epsilon)-1-\log d}$, we have

$$f(k) \leq \frac{2}{dk^2} = O(\frac{1}{\log^2 n}).$$

When $\frac{2\log n}{d(1-\epsilon)-1-\log d} \leq k \leq n$, we have

$$f(k) \leq \frac{2}{ndk^2} = O(\frac{1}{n\log^2 n}).$$

By setting $k_1 = \frac{\log n}{d(1-\epsilon)-1-\log d}$, the probability that a small component of size $k > k_1$ is at most

$$\sum_{k>k_1} f(k) \leq \frac{\log n}{d(1-\epsilon)-1-\log d} \times O(\frac{1}{\log^2 n}) + n \times O(\frac{1}{n\log^2 n}) = o(1).$$

Therefore, almost surely the size of a small component is at most $k_1 = \frac{\log n}{d(1-\epsilon)-1-\log d}$.

To see that this upper bound is best possible, we consider the random graph $G(n, \frac{d}{n})$ with $d < 1$, which is a random graph with a given expected degree sequence having equal weights d. By the classical result of Erdős and Rényi [19], the largest component of $G(n, \frac{d}{n})$ has size about $\frac{\log n - 5/2 \log\log n}{d-1-\log d}$ as required.

6 Neighborhood expansions

In this section we will prove two key lemmas on neighborhood expansions.

Lemma 4 *Let A, B be two disjoint sets. Suppose that each vertex in B has weight at most c_1. If $\mathrm{Vol}(A) = o(\frac{1}{c_1}\mathrm{Vol}(G))$ and $\mathrm{Vol}(A)\mathrm{Vol}(B) \geq 8\mathrm{Vol}(G)\log n$, then with probability at least $1-\frac{1}{n}$, $\Gamma(A)\cap B$ has at least $(\frac{1}{2}-o(1))\mathrm{Vol}(A)\mathrm{Vol}(B)\rho$ vertices.*

Proof: For any $v_j \in B$, let X_j be the indicator random variable for the event that v_j has exactly one edge joining to A. The probability for $X_j = 1$ is

$$\begin{aligned} Pr(X_j = 1) &= \sum_{v_i \in A} w_i w_j \rho \prod_{v_{i'} \in A, i' \neq i} (1 - w_{i'} w_j \rho) \\ &\geq \sum_{v_i \in A} w_i w_j \rho (1 - \mathrm{Vol}(A) w_j \rho) \\ &\geq \mathrm{Vol}(A) w_j \rho - \mathrm{Vol}(A)^2 w_j^2 \rho^2. \end{aligned}$$

We note that

$$\begin{aligned} Pr(X_j = 1) &= \sum_{v_i \in A} w_i w_j \rho \prod_{v_{i'} \in A, i' \neq i} (1 - w_{i'} w_j \rho) \\ &\leq \sum_{v_i \in A} w_i w_j \rho \\ &= \mathrm{Vol}(A) w_j \rho. \end{aligned}$$

$\Gamma(A) \cap B$ has at least $\sum_{v_j \in B} X_j$ vertices. Now we apply lemma 1 to $X = \sum_{v_j \in B} X_j$ and we have

$$E(\sum_{v_j \in B} X_j) \geq \mathrm{Vol}(A)\mathrm{Vol}(B)\rho - \mathrm{Vol}(A)^2 \mathrm{Vol}_2(B)\rho^2$$

where

$$\nu = \sum_{v_j \in B} Pr(X_j = 1) \leq \mathrm{Vol}(A)\mathrm{Vol}(B)\rho.$$

We choose $\lambda = \frac{1}{2}\mathrm{Vol}(A)\mathrm{Vol}(B)\rho$, then

$$\begin{aligned} E(\sum_{v_j \in B} X_j) - \lambda &\geq \frac{1}{2}\mathrm{Vol}(A)\mathrm{Vol}(B)\rho - \mathrm{Vol}(A)^2\mathrm{Vol}_2(B)\rho^2 \\ &\geq \frac{1}{2}\mathrm{Vol}(A)\mathrm{Vol}(B)\rho(1 - \mathrm{Vol}(A)c_1\rho) \\ &= (1 - o(1))\frac{1}{2}\mathrm{Vol}(A)\mathrm{Vol}(B)\rho \end{aligned}$$

Also,

$$\begin{aligned} e^{-\frac{\lambda^2}{2\nu}} &\leq e^{-\frac{(\mathrm{Vol}(A)\mathrm{Vol}(B)\rho)^2}{8\mathrm{Vol}(A)\mathrm{Vol}(B)\rho}} \\ &\leq e^{-\frac{1}{8}\mathrm{Vol}(A)\mathrm{Vol}(B)\rho} \\ &\leq \frac{1}{n} \end{aligned}$$

By Lemma 1, with probability at least $1 - \frac{1}{n}$, $\Gamma(A) \cap B$ has at least $(1 - o(1))\frac{1}{2}\mathrm{Vol}(A)\mathrm{Vol}(B)\rho$ vertices. □

The next lemma is useful for bounding from below the growth rate of volumes in the branching process.

Lemma 5 *Let A, B be two disjoint sets. Suppose that each vertex in B has weight at most c_1. Let x, y be two positive constants and we assume that*

$$yc_1 \log n \leq \mathrm{Vol}(A) = o(\frac{\mathrm{Vol}(G)}{c_1})$$

and

$$\mathrm{Vol}_2(B)\rho \geq 1 + 2x.$$

Then with probability at least $1 - n^{-\frac{(1+x)^2 y}{2(1+2x)}}$, we have

$$\mathrm{Vol}(\Gamma(A) \cap B) \geq (1 + x - o(1))\mathrm{Vol}(A).$$

Proof: For any $v_j \in B$, let X_j be the indicator random variable for the event that v_j has exactly one edge joining to A as in the proof of Lemma 4. We have

$$\mathrm{Vol}(A)w_j\rho \geq Pr(X_j = 1) \geq \mathrm{Vol}(A)w_j\rho - \mathrm{Vol}(A)^2 w_j^2 \rho^2.$$

$\Gamma(A) \cap B$ has at least $\sum_{v_j \in B} X_j$ vertices. Applying Lemma 1 to $X = \sum_{v_j \in B} w_j X_j$, we have

$$E(\sum_{v_j \in B} w_j X_j) \geq \mathrm{Vol}(A)\mathrm{Vol}_2(B)\rho - \mathrm{Vol}(A)^2\mathrm{Vol}_3(B)\rho^2$$

where

$$\nu = \sum_{v_j \in B} w_j^2 Pr(X_j = 1) \leq \mathrm{Vol}(A)\mathrm{Vol}_3(B)\rho.$$

We choose $\lambda = \frac{x}{1+2x}\mathrm{Vol}(A)\mathrm{Vol}_2(B)\rho$. By using the fact that $\mathrm{Vol}_3(B) \leq c_1\mathrm{Vol}_2(B)$, we have

$$\begin{aligned} E(\sum_{v_j \in B} w_j X_j) - \lambda &\geq \frac{1+x}{1+2x}\mathrm{Vol}(A)\mathrm{Vol}_2(B)\rho - \mathrm{Vol}(A)^2\mathrm{Vol}_3(B)\rho^2 \\ &\geq \frac{1+x}{1+2x}\mathrm{Vol}(A)\mathrm{Vol}_2(B)\rho(1 - \frac{1+2x}{1+x}\mathrm{Vol}(A)c_1\rho) \\ &\geq (1+x-o(1))\mathrm{Vol}(A) \end{aligned}$$

Also,

$$\begin{aligned} e^{-\frac{\lambda^2}{2\nu}} &\leq e^{-\frac{(\frac{1+x}{1+2x}\mathrm{Vol}(A)\mathrm{Vol}_2(B)\rho)^2}{2\mathrm{Vol}(A)\mathrm{Vol}_3(B)\rho}} \\ &\leq e^{-\frac{(1+x)^2}{2(1+2x)^2 c_1}\mathrm{Vol}(A)\mathrm{Vol}_2(B)\rho} \\ &\leq e^{-\frac{(1+x)^2 y}{2(1+2x)}\log n} \\ &= n^{-\frac{(1+x)^2 y}{2(1+2x)}} \end{aligned}$$

We apply lemma 1 and the proof is complete. □

7 Proof of Theorem 3

The most difficult part of this paper is the proof of Theorem 3, involving the existence of a giant component. The straightforward method of branching process works well for almost regular graphs but not as useful here. The probabilistic bounds for Poisson trials are less effective when we have some vertices of large weights while possibly a large number of vertices are possibly of much smaller weights. Alternative methods are needed.

Before proving Theorem 3, we will briefly sketch two main ideas. The first idea is an easy reduction to the existence of a connected subset of size $\Theta(\log n)$. The second idea is to focus on key structures in the graph by identifying two subsets of the vertices. One subset has a large volume but consists of vertices with small weights. The other set consists of vertices with large weights, and

is augmented into a connected subset of weights $\Theta(\log n)$. Then we will apply a refined and truncated version of branching process to show that the second subset will grow into a giant component.

Proof of theorem 3: We will state a series of reductions and useful lemmas.

Fact 4: Suppose $d > 1+\delta$, and there is a connected subset containing more than $C \log n$ vertices, where $C = max\{\frac{2}{\delta - \log \delta}, 10\}$. Then there is a giant component.

Proof: If a connected component has at least $C \log n$ vertices, it can not be a small component by the following observation. We have

$$C \geq \frac{2}{\delta - \log \delta} > \frac{1}{d - 1 - \log d - \epsilon_1 d}$$

for some $\epsilon_1 > 0$ when $1 + \delta < d \leq 2$. We also have

$$C \geq 10 > \frac{1}{1 + \log d - \log 4 + 2\log(1 - \epsilon_2)}$$

for some $\epsilon_2 > 0$ when $2 < d$. By theorem 1 and 2, this component is a giant component. □

From Fact 4, it is enough to find a connected subset of size $C \log n$ in order to show the existence of a giant component.

We first consider the range $d > 4 + \delta$.

The weights $\mathbf{w}$ are ordered non-increasingly $w_1 \geq w_2 \geq \cdots \geq w_n$. (Ties are being broken arbitrarily.) We claim the following:

Claim A: There exists an i_0 satisfying

1. $n^{1/3} \leq i_0 \leq n$.
2. $w_{i_0} \geq \sqrt{\frac{(1+\frac{\delta}{8})\mathrm{Vol}(G)}{i_0}}$.

Proof of Claim A:

Suppose the contrary. We have

$$w_i \leq \sqrt{\frac{(1 + \frac{\delta}{8})\mathrm{Vol}(G)}{i}} \quad \text{for all } n^{1/3} \leq i \leq n.$$

We consider

$$\begin{aligned}\mathrm{Vol}(G) &= \sum_{i=1}^{n^{1/3}} w_i + \sum_{i=n^{1/3}}^{n} w_i \\ &\leq n^{1/3}n^{1/2} + \sum_{i=n^{1/3}}^{n} \sqrt{\frac{(1+\frac{\delta}{8})\mathrm{Vol}(G)}{i}} \\ &\leq o(n) + 2\sqrt{(1+\frac{\delta}{8})\mathrm{Vol}(G)n}\end{aligned}$$

Hence we have

$$\mathrm{Vol}(G) \leq o(n) + 4(1+\frac{\delta}{8})n,$$

which contradicts $\mathrm{Vol}(G) = nd \geq (4+\delta)n$. Claim A is proved.

Now we consider the subgraph G_1 on the first i_0 vertices. For any pairs of vertices (v_i, v_j), $i, j \leq i_0$, the probability that it is an edge of G_1 is at least

$$w_i w_j \rho \geq w_{i_0}^2 \rho \geq \frac{(1+\frac{\delta}{8})}{i_0}.$$

From [19], G_1 has a giant component of size $\Theta(i_0) = \Omega(n^{1/3})$. Theorem 1-2 show that almost surely any connected component with size of $\Omega(\log n)$ is a giant component. Hence G has a giant component in this case.

We now consider the range $1+\delta \leq d \leq 4+\delta$. We claim the following.

Claim B: The vertex set can be partitioned into two sets S and T, satisfying

(a) There is a positive number c_1, so that the vertices in S has weight at most c_1 and the vertices in T has weight at least c_1.

(b) $\mathrm{Vol}(S) \geq \frac{1+d}{2}n$.

(c) T has at least $n^{1/3}$ vertices.

(d) If T has t vertices, we have

$$tc_1^2\rho \geq \frac{\delta^2}{32(1+\delta)}.$$

Proof of Claim B.

We denote $c_2 = \frac{\delta^2}{32(1+\delta)}$. Recall that the weights $\mathbf{w}$ are in a non-increasing order. Let n_0 be a fixed index satisfying $\sum_{i=n_0}^{n} w_i = (1+o(1))\frac{1+d}{2}n$. We can find an i_0 satisfying

(i) $n^{1/3} \leq i_0 \leq n_0$, and,

(ii) $w_{i_0} \geq \sqrt{\frac{c_2 \mathrm{Vol}(G)}{i_0}}$,

since, we have otherwise

$$\begin{aligned}
\mathrm{Vol}(G) &= \sum_{i=1}^{n^{1/3}} w_i + \sum_{i=n^{1/3}}^{n_0} w_i + \sum_{i=n_0}^{n} w_i \\
&\leq n^{1/3} n^{1/2} + \sum_{i=n^{1/3}}^{n_0} \sqrt{\frac{c_2 \mathrm{Vol}(G)}{i}} + \frac{1+d}{2}n \\
&\leq o(n) + 2\sqrt{c_2 \mathrm{Vol}(G) n_0} + \frac{1+d}{2}n.
\end{aligned}$$

Hence, we have $c_2 \geq \frac{(d-1)^2}{16d}$, which is a contradiction to the definition of c_2.

We choose $c_1 = w_{i_0}$ and let T denote the set of the first i_0 vertices while S be the complement of T. It is straightforward to verify all conditions. The proof of claim B is complete.

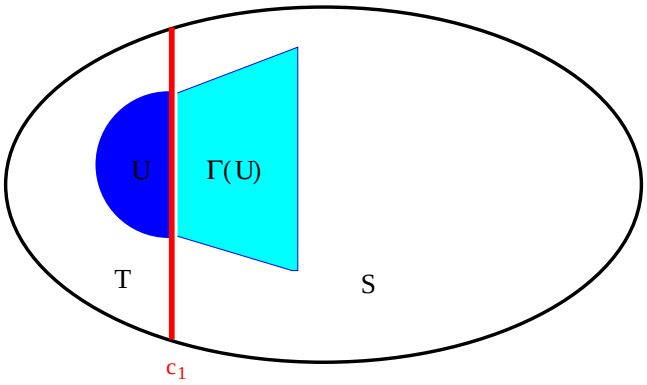

Figure 1: The method — G is partitioned into two parts S and T. T has vertices with weights at least c_1 and contains a connected subset U of size $\Theta(\log n)$. Each vertex of S has weight at most c_1.

Now we consider a special branching process. We consider all pairs with both ends in T. Since $tc_1^2\rho \geq \frac{\delta^2}{32(1+\delta)}$, the probability of each pair (v_i, v_j) being an edge is $w_i w_j \rho$ which is greater than or equal to $\frac{\delta^2}{32(1+\delta)}\frac{1}{t}$. For a constant $c = \frac{\delta^2}{32(1+\delta)}$, the induced subgraph of G on T contains a random graph $G(t, \frac{c}{t})$ as a subgraph. By the classical result of Erdős and Rényi [19], the largest component of $G(t, \frac{c}{t})$ have size about $g(c)(\log t - \frac{5}{2}\operatorname{loglog} n)$. where

$$g(c) = \frac{1}{c - 1 - \log c}.$$

Hence, almost surely there is a connected subset (denoted by U) in T with size at least $f(\delta)\log t$, where $f(\delta) = \frac{1}{2}g(\frac{\delta^2}{32(1+\delta)})$ is a positive constant only depending on δ. We have

$$\mathrm{Vol}(U) \geq c_1 f(\delta)\log t \geq \frac{c_1 f(\delta)}{3}\log n.$$

Let U' denote the connected component containing U. If $\mathrm{Vol}(U') \geq \Theta(\frac{\mathrm{Vol}(G)}{c_1})$, then U' is a giant component and we are done. We may assume $\mathrm{Vol}(U') = o(\frac{\mathrm{Vol}(G)}{c_1})$ which implies that

$$\mathrm{Vol}(U) = o(\frac{\mathrm{Vol}(G)}{c_1})$$

and for all i

$$\mathrm{Vol}(\Gamma_i(U)) = o(\frac{\mathrm{Vol}(G)}{c_1}).$$

If $c_1 \geq \frac{24C}{f(\delta)}$, then we have $\mathrm{Vol}(U) \geq 8C\log n$. We can apply lemma 4 by choosing $A = U$ and $B = S$. We have

$$\mathrm{Vol}(A)\mathrm{Vol}(B) \geq 8C\log n\frac{d+1}{2}n > 8\mathrm{Vol}(G)\log n.$$

By lemma 4, $\Gamma(A)\cap S$ has more than $(4C - o(1))\log n$ vertices. By Fact 4, the giant component exists.

If $c_1 \leq \frac{24C}{f(\delta)}$, we will use lemma 5 repeatedly. We will inductively prove that

$$\mathrm{Vol}(\Gamma_i(U)\cap S) \geq (1 + x - o(1))^i \mathrm{Vol}(U), \tag{8}$$

for all i from 1 to $i' = \log_x \frac{Cf(\delta)}{3}$ satisfying $\mathrm{Vol}(\Gamma_{i'}(U)\cap S) > 8C\log n$. Here $x = \frac{(d-1)^2}{16d} \geq \frac{\delta^2}{16(1+\delta)}$. By applying the same argument to $\mathrm{Vol}(\Gamma_{i'}(U)\cap S)$, we conclude that there is always a giant component.

Here is the proof for (8).

Initially, let $A = U$, $B = S$, and $y = \frac{f(\delta)}{3}$. we have

$$\mathrm{Vol}(A) \geq \frac{f(\delta)}{3} c_1 \log n = c_1 y \log n.$$

By using Claim B, we have

$$\begin{aligned}
\mathrm{Vol}_2(B)\rho &\geq \frac{\mathrm{Vol}(B)^2}{n-t}\rho \\
&\geq \frac{(d+1)^2 n^2}{4n^2 d} \\
&= \frac{(d+1)^2}{4d} \\
&= 1 + 4x \\
&> 1 + 2x
\end{aligned}$$

With failure probability at most $n^{-\frac{(1+x)^2 y}{2(1+2x)}}$, we have $\mathrm{Vol}(\Gamma(U) \cap S) \geq (1 + x - o(1))\mathrm{Vol}(\Gamma(U))$.

At the inductive step i, let $A = \Gamma_i(U) \cap S$ and $B = S \setminus (\cup_{j \leq i}(\Gamma_i(U) \cap S))$ and we have

$$\mathrm{Vol}(A) \geq \mathrm{Vol}(U) \geq c_1 y \log n.$$

$$\begin{aligned}
\mathrm{Vol}_2(B)\rho &\geq \mathrm{Vol}_2(S)\rho - \rho \sum_{i' \leq i} \mathrm{Vol}_2(\Gamma_i(U) \cap S) \\
\rho &\geq \mathrm{Vol}_2(S)\rho - \rho i' c_1^2 8C \log n \\
&= 1 + 4x - o(1) \\
&> 1 + 2x
\end{aligned}$$

By lemma 5, with failure probability at most $n^{-\frac{(1+x)^2 y}{2(1+2x)}}$, we have $\mathrm{Vol}(\Gamma_{i+1}(U) \cap S) \geq (1 + x - o(1))^{i+1}\mathrm{Vol}(\Gamma(U))$. The total failure probability is bounded above by

$$\log_x \frac{Cf(\delta)}{3} n^{-\frac{(1+x)^2 y}{2(1+2x)}} = o(1).$$

Thus we finish the inductive proof. Hence, the volume of $\Gamma_i(U)$ will grow in S by a factor of at least $1 + x$ for each i. The process can only stop when the volume is no longer $o(\mathrm{Vol}(G)/c_1)$. Therefore, a giant component will eventually exist.

The above argument using (8) can be used to show the uniqueness of the giant component as well. For any two vertices u and v in two giant connected component, we begin a branching process starting at u but stop at the moment when the volume of its neighbors S_1 reaches $\sqrt{(2+\epsilon)\mathrm{Vol}(G)\log n}$. Then we begin a new branching process starting at v and stop at the moment when the volume of its neighbors S_1 reaches $\sqrt{(2+\epsilon)\mathrm{Vol}(G)\log n}$. Then we see that the probability of two neighbors sets are not connected by any edge is at most

$$\begin{aligned}\prod_{u\in S_1, v\in S_2}(1-w_u w_v\rho) &\leq \prod_{u\in S_1, v\in S_2} e^{-w_u w_v\rho}\\ &= e^{-\sum_{u\in S_1, v\in S_2} w_u w_v\rho}\\ &= e^{-\mathrm{Vol}(S_1)\mathrm{Vol}(S_2)\rho}\\ &\leq e^{-(2+\epsilon)\log n}\\ &= n^{-2-\epsilon}.\end{aligned}$$

The probability that any two vertices belong to the same connected component with probability at least $1-n^{-\epsilon}$. Thus, the giant component is almost surely unique.

Now we consider the volume of the giant component. We want to show the following:

(i) If $d \geq e$, the volume of the giant component is at least $(1-\frac{2}{\sqrt{de}}+o(1))\mathrm{Vol}(G)$.

(ii) If $1+\delta \leq d \leq e$, the volume of the giant component is at least $(1-\frac{1+\log d}{d}+o(1))\mathrm{Vol}(G)$.

Let us consider the case of $d \geq e$. If (i) does not hold, then the giant component is ϵ-small for some ϵ satisfying $\epsilon < 1-\frac{2}{\sqrt{de}}$. By Theorem 1, the size of the giant component is at most $\frac{\log n}{1+\log d-\log 4+2\log(1-\epsilon)}$. Hence there is one vertex with weight w great than or equal to the average:

$$w \geq \frac{\epsilon\mathrm{Vol}(G)}{\frac{\log n}{1+\log d-\log 4+2\log(1-\epsilon)}} \geq c_\epsilon\frac{\mathrm{Vol}(G)}{\log n}.$$

It is easy to check that

$$w^2\rho \gg 1$$

which contradicts our assumption (1). Hence the volume of the giant component is at least $(1 - \frac{2}{\sqrt{de}} + o(1))vol(G)$ if $d \geq e$.

For the case of $d_0 \leq d \leq e$, we again prove by first assuming the contrary that the giant component is ϵ-small for some ϵ satisfying $\epsilon < 1 - \frac{1+\log n}{d}$. By Theorem 2, the size of the giant component is at most $\frac{\log n}{d-1-\log d-\epsilon d}$. Hence there is one vertex with weight w great than or equal to the average:

$$w \geq \frac{\epsilon vol(G)}{\frac{\log n}{d-1-\log d-\epsilon d}} \geq c'_\epsilon \frac{vol(G)}{\log n}.$$

Since $w^2\rho \gg 1$, we again have a contradiction and (ii) is proved.

Now we consider the case of $\tilde{d}$ is smaller than 1. The following claim shows that in this range almost surely all components have volumes of at most $\sqrt{n}\log n$. Therefore there is no giant component in this case.

Claim C: If $\tilde{d} < 1 - \delta$, with probability at least $1 - \frac{d\tilde{d}^2}{C^2(1-\tilde{d})}$, all components have volume at most $C\sqrt{n}$.

Proof of Claim C: Let x be the probability that there is a component having volume greater than $C\sqrt{n}$. Now we choose two random vertices with the probability of being chosen proportional to their weights. Under the condition that there is a component with volume greater than $C\sqrt{n}$, the probability of each vertex in this component is at least $C\sqrt{n}\rho$. Therefore, the probability that the random pair of vertices are in the same component is at least

$$x(C\sqrt{n}\rho)^2 = C^2 xn\rho^2. \tag{9}$$

On the other hand, for any fixed pair of vertices u and v, the probability $P_k(u, v)$ of u and v is connected by path of length $k + 1$ is at most

$$\begin{aligned} P_k(u,v) &\leq \sum_{i_1 i_2 \ldots i_k} (w_u w_{i_1}\rho)\ (w_{i_1} w_{i_2}\rho)\cdots(w_{i_k} w_v \rho) \\ &\leq w_u w_v \rho \tilde{d}^k \end{aligned}$$

The probability that u and v belong to the same component is at most

$$\sum_{k=0}^{n} P_k(u,v) \leq \sum_{k\geq 0} w_u w_v \rho \tilde{d}^k = \frac{1}{1-\tilde{d}} w_u w_v \rho.$$

Since the probabilities of u and v being selected are $w_u\rho$ and $w_v\rho$ respectively, the probability that the random pair of vertices are in the same connected component is at most

$$\sum_{u,v} w_u\rho \; w_v\rho \; \frac{1}{1-\tilde{d}} w_u w_v \rho = \frac{\tilde{d}^2}{1-\tilde{d}}\rho.$$

Combining with (9), we have

$$C^2 x n \rho^2 \leq \frac{\tilde{d}^2}{1-\tilde{d}}\rho$$

which implies

$$x \leq \frac{d\tilde{d}^2}{C^2(1-\tilde{d})}.$$

Claim C is proved.

We have completed the proof for Theorem 4. □

8 Several other models

In the literature, the following model, so called *the configuration model*, is often used to construct a random graph with a prescribed degree sequence. It was first introduced by Bender and Canfield [9], refined by Bollobás [10] and also Wormald [35]. A random graph G with given degrees d_v is formed by first associating to each vertex v a set S_v of d_v nodes, then considering the disjoint union N of S_v and taking a random matching M on N. The number of edges between two vertices u and v is the number of edges in M with one node in S_u and one node in S_v. It is easy to see that the resulting graph (as a multi-graph) has degrees exactly as required.

Molloy and Reed [31, 32] used the configuration model to show that if there are $d_i(n) \approx \lambda_i n$ vertices of degree i, where $\sum_i \lambda_i = 1$ and $\sum i(i-2)\lambda_i > 0$, then the graph almost surely has a giant component if the following conditions are satisfied.

1. The maximum degree is at most $n^{1/4-\epsilon}$.
2. $i(i-2)d_i(n)/n$ tends uniformly to $i(i-2)\lambda_i$.

3. The limit

$$L(\mathcal{D}) = \lim_{n\to\infty} \sum_{i\geq 1} i(i-2)d_i(n)/n$$

exists, and the sum approaches the limit uniformly.

4. The degree sequence is graphic.

The advantage of the configuration model is to generate graphs exactly with the prescribed degrees and it is the primary model for examining regular graphs with constant degrees. There are several disadvantages of the configuration model. The analysis of the configuration model is much more complicated due to the dependency of the edges. A random graph from the configuration model is in fact a multigraph instead of a simple graph. The probability of having multiple edges increases rapidly when the degrees increase. In the papers of Molloy and Reed, the condition on maximum degree with an upper bound of $n^{1/4-\epsilon}$ is required because of occurance of multiple edges in the configuration model. Consequently, this model is restrictive for power-law graphs, where the largest degree can be quite large. Furthermore, additional conditions (e.g., Condition 2 and 3 as in [31, 32]) are often required for the configuration models which are hard to deal with for realistic graphs. In the same way, the classical random graph model $G(n,p)$ is often preferred to the configuration models of random graphs with $p\binom{n}{2}$ edges.

The advantage of the generalized model that we use here is the simplicity without any condition on the degree sequence except for the only assumption (1). Our model does not produce the graph with exact given degree sequence. Instead, it yields a random graph with given expected degree sequence.

Another line of approach which simulates realistic graphs is to generate a vertex/edge at a time, starting from one node or a small graph. Although we will not deal with such models in this paper, we will briefly mention several evolution models. Barabasi and Albert [7] describe the following graph evolution process. Starting with a small initial graph, at each time step they add a new node and an edge between the new node and each of m random nodes in the existing graph, where m is a parameter of the model. The random nodes are

not chosen uniformly. Instead, the probability of picking a node is weighted according to its existing degree (the edges are assumed to be undirected). Using heuristic analysis with the assumption that the discrete degree distribution is differentiable, they derive a power law for the degree distribution with a power of 3, regardless of m. A power law with power 3 for the degree distribution of this model was independently derived and proved by Ballobás et al. [11].

Kumar at el. [28] proposed three evolution models — "linear growth copying", "exponential growth copying", and "linear growth variants". The *Linear growth coping* model adds one new vertex with d out-links at a time. The destination of i-th out-link of the new vertex is either copied from the corresponding out-link of a "prototype" vertex (chosen randomly) or a random vertex. They showed that the in-degree sequence follows the power law. These models were designed explicitly to model the World Wide Web. Indeed, they show that their model has a large number of complete bipartite subgraphs, as has been observed in the WWW graph, whereas several other models do not. This (and the linear growth variants model) has the similar drawback as the first model in [27]. The out-degree of every vertex is always a constant. Edges and vertices in the *exponential growth copying* model increase exponentially.

Aiello et al. described a general random graph evolution process in [3] for generating directed power law graphs with given expected in-degrees and out-degrees. At each time t, a new node is generated and certain edges are added as follows. The end points of new edges can be either the new node or one of the existing nodes. An existing node is selected as the destination (or the origin) with probability proportional to its in-degree (or out-degree). There are four types of edges according to their destinations and origins. A probability space P_t controls the number and the type of edges to be added at time t. Under the assumption that the number of edges added at each time is bounded and P_t has a limiting distribution, Aiello et al. [3] proved this general process generates power law graphs. The power of the power law of out-degree (or in-degree) equals to $2 + \frac{A}{B}$, where A is the expected number of edges per step with the new node as the origin (or the destination) and B is the expected number of

edges per step with an existing node as the origin (or the destination). Recently, Cooper and Frieze [17] independently analyzed the above evolution of adding either new vertices or new edges and derived power law degree distribution for vertices of small degrees.

9 Remarks on power law graphs

In this paper, we examine the sizes of connected components of a random graph with given degree sequences. The results and methods here can be useful to examine power law graphs that arise in various context. A power law graph with power α has the number of vertices of degree k proportional to k^{α}. For example, the collaboration graph consists of 337,000 authors in *Mathematics Review* as vertices and collaborations as edges, as described in the webpage of Jerry Grossman [22] at http://www.oakland.edu/~grossman/trivia.html. From Figure 2, we can see that the degree sequence of the collaboration graph can be approximated by a power law with power 2.2.

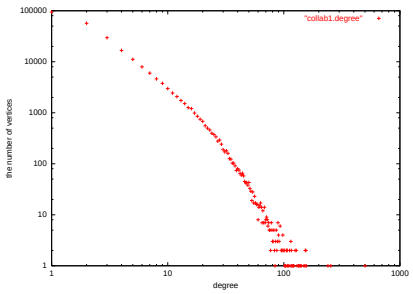

Figure 2: *Degree distribution of the collaboration graph*

If we model realistic graphs by random graphs with power law degree sequences, the results here on the volume of the connected components can be utilized. For example, a rough calculation shows that the results in Theorem 1 is consistent with the actual data on connected components of the collaboration graph (see Figure 1).

To actually model a realistic graph, there are a number of additional factors to be taken into consideration. For example, the so-called "small world"

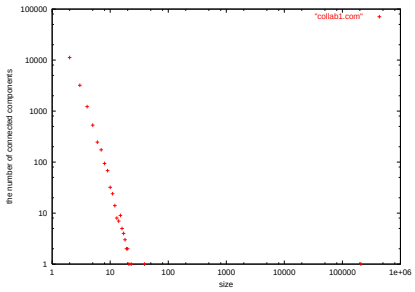

Figure 3: *Connected components distribution*

phenomenon asserts that the distance between two random vertices is small. In a forthcoming paper [15], the authors examine the average distances in a random graph with given expected degree sequences. It is shown that the average distance of a random graph with expected degree sequences almost surely has average distance $(1 + o(1)) \log n / \log \tilde{d}$, provided certain mild conditions are satisfied. (Power law random graphs satisfies such conditions.) There is also a clustering effect that is often found in realistic graphs. A more elaborated model in combining the local structures and global (random) properties will be considered in a subsequent paper [16].

References

[1] L. A. Adamic and B. A. Huberman, Growth dynamics of the World Wide Web, *Nature* , **401**, September 9, 1999, pp. 131.

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

[3] W. Aiello, F. Chung and L. Lu, Random evolution in massive graphs, Extended abstract appeared in *The 42th Annual Symposium on Foundation of Computer Sciences*, October, 2001. Paper version will appear in *Handbook on Massive Data Sets*, Volume 2, (Eds. J. Abello, et. al.).

[4] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, Classes of small-world networks, *Proc. Natl. Acad. Sci. USA*, vol. **97**, no. 21, (2000), 11149-11152.

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

[6] R. B. R. Azevedo and A. M. Leroi, A power law for cells, *Proc. Natl. Acad. Sci. USA*, vol. **98**, no. 10, (2001), 5699-5704.

[7] Albert-László Barabási and Réka Albert, Emergence of scaling in random networks, Science **286** (1999) 509-512.

[8] A. Barabási, R. Albert, and H. Jeong, Scale-free characteristics of random networks: the topology of the world wide web, *Physica* A 272 (1999), 173-187.

[9] E. A. Bender and E. R. Canfield, The asymptotic number of labelled graphs with given degree sequences, *J. Combinat. Theory (A)*, **24**, (1978), 296-307.

[10] B. Bollobás, *Random Graphs*, Academic, New York, 1985.

[11] B. Bollobás, O. Riordan, J. Spencer and G. Tusnády, The Degree Sequence of a Scale-Free Random Graph Process, Random Structures and Algorithms, Vol. **18**, no. 3 (2001), 279-290.

[12] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tompkins, and J. Wiener, "Graph Structure in the Web," *proceedings of the WWW9 Conference*, May, 2000, Amsterdam. Paper version appeared in *Computer Networks* **33**, (1-6), (2000), 309-321.

[13] K. Calvert, M. Doar, and E. Zegura, Modeling Internet topology. *IEEE Communications Magazine*, **35(6)** (1997) 160-163.

[14] Fan Chung and Linyuan Lu, The diameter of random sparse graphs, *Advances in Applied Math.*, **26** (2001), 257-279.

[15] F. Chung and L. Lu, Average distances in random graphs with given expected degree sequences, preprint

[16] Fan Chung and Linyuan Lu, Small world graphs and random graphs, preprint.

[17] C. Cooper and A. Frieze, A general model of web graphs, *http://www.math.cmu.edu/∼af1p/papers.html*.

[18] P. Erdős and T. Gallai, Gráfok előírt fokú pontokkal (Graphs with points of prescribed degrees, in Hungarian), *Mat. Lapok* **11** (1961), 264-274.

[19] P. Erdős and A. Rényi, On random graphs. I, *Publ. Math. Debrecen* **6** (1959), 290-291.

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

[21] N. Gilbert, A simulation of the structure of academic science, *Socialogical Research Online*, **2** (2), 1997.

[22] Grossman, Ion, and De castro, The Erdős Number Project, http://www.oakland.edu/∼grossman/erdoshp.html.

[23] S. Jain and S. Krishna, A model for the emergence of cooperation, interdependence, and structure in evolving networks, *Proc. Natl. Acad. Sci. USA*, vol. **98**, no. 2, (2001), 543-547.

[24] Janson, Łuczak, and Rucinński, *Random Graphs*, John Wiley & Sons, Inc, 2000.

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

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

[27] 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, 1999.

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

[29] Linyuan Lu, The diameter of random massive graphs, *Proceedings of the Twelfth ACM-SIAM Symposium on Discrete Algorithms*, (2001) 912-921.

[30] McDiarmid colin Concentration. *Probabilistic methods for algorithmic discrete mathematics*, 195–248, *Algorithms Combin.*, 16, Springer, Berlin, 1998.

[31] 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.

[32] 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.

[33] M. E. J., Newman, The structure of scientific collaboration networks, *Proc. Natl. Acad. Sci. USA*, vol. **98**, no. 2, (2001), 404-409.

[34] D. West, *Introduction to Graph Theory*, Prentice Hall, 1996.

[35] Some problems in the enumberation of labelled graphs, Doctoral thesis, Newcastle University, 1978.

[36] E. Zegura, K. Calvert, and M. Donahoo, A quantitative comparison of graph-based models for Internet topology. *IEEE/ACM Transactions on Networking*, **5** (6), (1997), 770-783.