Commun. Math. Sci.
Vol. 7, No. 1, pp. 175–187

ON THE DEGREE PROPERTIES OF GENERALIZED RANDOM GRAPHS*

YI Y. SHI† AND HONG QIAN‡

Abstract. A generalization of the classical Erdös and Rényi (ER) random graph is introduced and investigated. A *generalized random graph* (GRG) admits different values of probabilities for its edges rather than a single probability uniformly for all edges as in the ER model. In probabilistic terms, the vertices of a GRG are no longer statistically identical in general, giving rise to the possibility of complex network topology. Depending on their surrounding edge probabilities, vertices of a GRG can be either "homogeneous" or "heterogeneous". We study the statistical properties of the degree of a single vertex, as well as the degree distribution over the whole GRG. We distinguish the degree distribution for the entire random graph ensemble and the degree frequency for a particular graph realization, and study the mathematical relationship between them. Finally, the connectivity of a GRG, a property which is highly related to the degree distribution, is briefly discussed and some useful results are derived.

Key words. Random graph, degree distribution, connectivity, giant component.

AMS subject classifications. 05C80; 05C40.

1. Introduction

The study of graphs has a long history. Graph theory has become one of the important branches of discrete mathematics and provides powerful tools for solving application problems such as computational algorithms and network optimization. Random graphs arise from introducing probabilistic ideas into graph theory. They have provided new perspectives on real world networks and analytical tools for analyzing systems with uncertainties, complementary to the standard graph theory. In recent years, with the dramatically increasing capacity of computing and storage, researchers are beginning to explore the properties and underlying principles of large-scale systems such as the Internet and biological networks [4, 19]. This new research trend has triggered a revitalization of the random graph theory and taken the field to a new era. Still, as has been recognized by some experts, we are far from capturing and explaining many of the universal, fascinating features shared by most of the large networked systems in real world. A rigorous study of large-scale random graphs is still in its infancy.

The best known random graph model is the classical random graph $\mathcal{G}(N,p)$ proposed by Erdös and Rényi (ER) in 1950's [6, 11]. This model considers a graph with N vertices and assumes a uniform probability p for each pair of vertices to form an edge. With varying p, different topological properties arise in the graph. In probabilistic terms, an ER random graph assigns a set of identical independent Bernoulli random variables with parameter p to the edges of a graph. Thus, a random graph from this point of view is an ensemble of matrices consisting of random variables $\{\mathbf{e}_{ij} | 1 \le i,j \le N, i \ne j\}$. The simplicity of the ER model makes it possible, on one hand, for extensive analytical studies, but on the other hand limits its application to many complex networks from the real world. For example, it has been observed

*Received: October 8, 2007; accepted (in revised version): January 1, 2009.

†Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA (shiyiyi@amath.washington.edu).

‡Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA. (qian@amath.washington.edu)

recently that many complex networks, such as the Internet and protein-protein interactions in biological cells, have a "power-law" degree distribution [2, 3, 10, 21]. This differs significantly from the Poisson degree distribution dictated by the classical ER random graph.

In order to approach various situations in real random networks, we suggest a more general form of random graphs. Rather than using a uniform probability p, we assign an arbitrary probability value $p_{ij} \in [0,1]$ on the edge associated with vertices i and j. This generalized random graph (GRG) model is simple in presentation, but significantly complex in analysis, in particular when the p_{ij}'s are connected with a set of Bernoulli random variables that are not independent. We shall classify random graphs within our model into two types, *homogeneous* and *heterogeneous*, and study their degree properties respectively. Many papers in this field give only vague definitions about the degree distribution in random graphs. In fact, a random graph model, depending on its complexity, could generate a hierarchy consisting of several levels in terms of conditional probabilities and conditional expectations. The degree distributions at various levels are conceptually different, thus should be treated separately. As we shall show, under some conditions these degree distributions have a definitive relationship and could be used interchangeably. In general there is an issue of ergodicity in many large random graphs when $N \to \infty$.

The connectivity is another important dimension to explore the topology of random graphs. Many theoretical papers have worked on the connectivity of classical ER random graphs, as well as graphs with given degree sequences, and several beautiful results have been established [9, 11]. The most fascinating phenomenon is the existence of a critical point at which a phase transition occurs, i.e., a giant component appears in the graph instead of many isolated small components. In the later part of this paper, we will investigate the connectivity of the GRG model. We shall concentrate our effort for the case of homogeneous random graphs and derive some convenient criteria. More general conclusions can be obtained for GRGs only if more specific details are given.

2. Model proposal

A standard graph with N vertices can be completely specified by the adjacency matrix E: a symmetric $N \times N$ matrix consisting of 0's and 1's with all diagonal entries being 0. A matrix entry e_{ij} with value 1 represents an edge between vertex pair (i,j). A generalized random graph is defined as follows: a Bernoulli random variable $\mathbf{e}_{ij}(p_{ij})$ is assigned to the vertex pair (i,j). A Bernoulli random variable $\mathbf{e}(p)$ has $\Pr\{\mathbf{e}=1\}=p$ and $\Pr\{\mathbf{e}=0\}=1-p$. The matrix $\mathbf{M}=[\mathbf{e}_{ij}(p_{ij})]_{N\times N}$, therefore, is a random matrix, which corresponds to a random graph ensemble. Note that each realization of $\mathbf{M}$ should be symmetric with all the diagonal entries being 0. Theoretically, this is the universal representation for all random graphs; we shall denote it by $\mathcal{G}(N,\mathbf{M})$. In the present paper, we assume that the entries $\mathbf{e}_{ij}$ in matrix $\mathbf{M}$ are *mutually independent* unless otherwise noted. Random graphs with dependent $\mathbf{e}_{ij}$'s exhibit greater topological complexity which will be the subject of forthcoming publications. From $\mathcal{G}(N,\mathbf{M})$, different graph realizations could be obtained by randomly sampling the Bernoulli random variable $\mathbf{e}_{ij}$ on each vertex pair (i,j). All possible graph realizations can be characterized by their adjacency matrices $E_1,E_2,\dots,E_m$, where $m=2^{N(N-1)/2}$. We use $G(N,E)$ to denote the graph realization G with adjacency matrix $E=[e_{ij}]_{N\times N}$. The following two facts hold:

FACT 2.1. $\Pr\{G(N,E)|\mathbf{M}\}=\prod_{i,j} p_{ij}^{e_{ij}}(1-p_{ij})^{1-e_{ij}}$.

Fact 2.2. *By the law of large numbers, for a sequence of realizations* $E_1, E_2, \ldots, E_n$, *as* $n \to \infty$, $\frac{1}{n}(E_1 + E_2 + \cdots + E_n)$ *converges to* $E[\mathbf{M}] = [p_{ij}]_{N\times N}$ *in probability, where* $E[\mathbf{M}]$ *denotes the regular matrix* $[E[\mathbf{e}_{ij}]]_{N\times N}$.

Among all the properties of random graphs, the most studied one in recent years is the degree distribution. For any single realization $G(N,E)$, we can count the degrees of all the vertices, generate the histogram of degrees, and normalize it to obtain the *degree frequency* of such a particular graph. Unfortunately, this degree frequency is widely called degree distribution. To be more precise, we shall name it the *intra-graph* degree distribution. However, for a single vertex i in $\mathcal{G}(N,\mathbf{M})$, its degree is a random variable, and thus could have many values among different realizations of $\mathcal{G}(N,\mathbf{M})$. We call it the *inter-graph* degree fluctuation. In most empirical studies for complex random networks, e.g., biological networks and computer networks, the data is usually taken only from a single realization of the network. Hence the reported degree distribution is actually an intra-graph degree distribution. Due to the inter-graph fluctuations of degrees, however, the intra-graph degree distribution should not be a fixed function, rather it has certain variance among different graph realizations. The main interest of the present paper is to study such variances in both the degree, as a single random variable, and intra-graph distribution, as a random vector. We shall introduce clear definitions for the "degree distributions" at the different levels of a random graph, as the concept has been vague in the current literature.

3. Inter-graph degree fluctuation

In the random graph model $\mathcal{G}(N,\mathbf{M})$, the degree $\mathbf{d}_i$ of vertex i, which is a discrete random variable, can be expressed as the sum of $N-1$ independent Bernoulli random variables: $\mathbf{d}_i = \sum_{j=1, j\neq i}^{N} \mathbf{e}_{ij}$. If the $\mathbf{e}_{ij}$'s are identical, i.e., have the same success probability p, then the sum $\mathbf{d}_i$ follows a binomial distribution $B(N-1,p)$. By the central limit theorem, $\mathbf{d}_i$ approaches a normal random variable as $N \to \infty$. If the Bernoulli random variables $\mathbf{e}_{ij}$ are not identical, then the distribution of the integer-valued random variable $\mathbf{d}_i$ follows the so-called *Poisson binomial distribution.* This distribution has no explicit expression formula in general. However, when $N \to \infty$ it can be well approximated by either a Poisson or normal distribution under different conditions, as given by the following two lemmas. (We denote $\sum_{j=1,j\neq i}^{N}$ by $\sum_{j=1}^{\prime N}$ for convenience.)

Lemma 3.1. *If* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *are* n *independent Bernoulli random variables with respective success probabilities* $p_1, \ldots, p_n$, *define* $\mathbf{S} = \sum_{i=1}^{n} \mathbf{X}_i$, $\mu = p_1 + p_2 + \cdots + p_n$, *and* $\mathbf{Y}$ *to be a Poisson random variable with mean value* μ. *Then the following inequality holds:*

$$D = \sup_{0\le m\le n} |\Pr(\mathbf{S} \le m) - \Pr(\mathbf{Y} \le m)| \le 2\sum p_i^2. \tag{3.1}$$

Lemma 3.2. *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *be* n *independent Bernoulli random variables with respective success probabilities* $p_1, \ldots, p_n$, $\mathbf{S} = \sum_{i=1}^{n} \mathbf{X}_i$, $\mu = E[\mathbf{S}] = \sum_{j=1}^{n} p_j$, $\sigma^2 = Var[\mathbf{S}] = \sum_{j=1}^{n} p_j(1-p_j)$, *and* $\gamma = E[(\mathbf{S}-\mu)^3] = \sum_{j=1}^{n} p_j(1-p_j)(1-2p_j)$. *Let* $\Phi(x)$ *and* $\phi(x)$ *be the c.d.f. and p.d.f. of the standard normal distribution, respectively, and* $\Gamma(x) = \Phi(x) + \frac{\gamma}{6\sigma^2}(1-x^2)\phi(x)$. *Then the following inequality holds:*

$$\Delta = \sup_{0\le m\le n} \left| \Pr(\mathbf{S} \le m) - \Gamma\left(\frac{m-\mu+\frac{1}{2}}{\sigma}\right) \right| \le \frac{\sigma+3}{4\sigma^3}. \tag{3.2}$$

The proofs of these two lemmas can be found in [12] and [15], respectively. The first lemma provides the error bound of the Poisson approximation to the Poisson binomial distribution. Under the conditions of $\max_i\{p_i\}\to 0$ and $\mu=\sum_{i=1}^{n}p_i\to\lambda$ as $n\to\infty$, the inequality (3.1) yields $D\le 2\sum p_i^2\le \max_i\{p_i\}\sum p_i\to\lambda\max_i\{p_i\}\to 0$. This reveals the fact that if the p_i's are sufficiently small (on the order of n^{-1}) as $n\to\infty$, the sum $\mathbf{S}$ can be well approximated by the Poisson random variable $\mathbf{Y}$. If the p_i's have medium or large values, then the error bound $2\sum p_i^2$ in Lemma 3.1 will definitely exceed 1 as $n\to\infty$, hence it contains little information on how accurate the Poisson approximation will be. However, in this case the mean μ and variance σ^2 of $\mathbf{S}$ will become large enough such that $\Gamma(x)=\Phi(x)+\frac{\gamma}{6\sigma^2}(1-x^2)\phi(x)\to\Phi(x)$ and $\Delta\le\frac{\sigma+3}{4\sigma^3}\to 0$ as $n\to\infty$, which implies that the sum $\mathbf{S}$ can be approximated by a normal random variable with very small error. Combining these facts and applying them to our random graph model gives the following theorem:

Theorem 3.3. *In the random graph $\mathcal{G}(N,\mathbf{M})$, the degree $\mathbf{d}_i$ of the vertex i follows a Poisson binomial distribution. If $\max_j\{p_{ij}\}\to 0$ and $\sum_{j=1}^{\prime N}p_{ij}\to\lambda$ as $N\to\infty$, then $\mathbf{d}_i$ can be well approximated by a Poisson random variable Poisson(λ); otherwise it can be well approximated by a normal random variable as $N\to\infty$, with mean and variance*

$$\mu_i=E[\mathbf{d}_i]=E\left[\sum_{j=1}^{\prime N}\mathbf{e}_{ij}\right]=\sum_{j=1}^{\prime N}p_{ij}, \tag{3.3}$$

$$\sigma_i^2=Var[\mathbf{d}_i]=Var\left[\sum_{j=1}^{\prime N}\mathbf{e}_{ij}\right]=\sum_{j=1}^{\prime N}(p_{ij}-p_{ij}^2). \tag{3.4}$$

Theorem 3.3 describes the distribution, as well as the mean and variance, of the inter-graph fluctuation of the degree $\mathbf{d}_i$ in our random graph model. More precisely, we could get the following estimates for the value of $\mathbf{d}_i$ in a probabilistic sense:

Theorem 3.4. *In the random graph $\mathcal{G}(N,\mathbf{M})$, the degree $\mathbf{d}_i$ of vertex i satisfies the following inequalities:*

$$\Pr\{\mathbf{d}_i-\mu_i\ge\Delta\}\le e^{-\frac{2\Delta^2}{N}}, \tag{3.5}$$

$$\Pr\{\mathbf{d}_i-\mu_i\le-\Delta\}\le e^{-\frac{2\Delta^2}{N}}. \tag{3.6}$$

where μ_i is given by equation (3.3).

These two inequalities are actually a corollary of McDiarmid's inequality. The proof is not shown, but can be found in [14].

The two theorems above establish the probability laws for the degree of a vertex in a large random network. When N tends to infinity, if the probabilities on the edges $\sim N^{-1}$, then the degree is well approximated by a Poisson distribution; otherwise, it is well apprximated by a normal distribution. In order to get the intra-graph distribution (frequency), we need to know whether the degrees on the different vertices are independent or correlated. This leads to the following theorem.

Theorem 3.5. *In the random graph $\mathcal{G}(N,\mathbf{M})$, the degrees of any two vertices are almost independent in the large limit of N.*

Proof. Consider the degrees $\mathbf{d}_i$ and $\mathbf{d}_j$ of vertices i and j, respectively. Since $\mathbf{d}_i=\sum_{k\neq i}\mathbf{e}_{ik}$ and $\mathbf{d}_j=\sum_{l\neq j}\mathbf{e}_{jl}$, the two summands only have one common term $\mathbf{e}_{ij}$. Denote $\mathbf{X}=\mathbf{e}_{ij}$, $\mathbf{Y}=\sum_{k\neq i,j}\mathbf{e}_{ik}$, $\mathbf{Z}=\sum_{l\neq i,j}\mathbf{e}_{jl}$, then $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ are independent of each other. Therefore the covariance of $\mathbf{d}_i$ and $\mathbf{d}_j$ should be:

$$Cov(\mathbf{d}_i,\mathbf{d}_j)=E[(\mathbf{X}+\mathbf{Y})(\mathbf{X}+\mathbf{Z})]-E[\mathbf{X}+\mathbf{Y}]E[\mathbf{X}+\mathbf{Z}]$$

$$=E[\mathbf{X}^2]-E[\mathbf{X}]^2=Var[\mathbf{X}]=p_{ij}(1-p_{ij}),$$

and their correlation should be

$$Corr(\mathbf{d}_i,\mathbf{d}_j)=\frac{Cov(\mathbf{d}_i,\mathbf{d}_j)}{\sqrt{Var[\mathbf{d}_i]Var[\mathbf{d}_j]}}=\frac{p_{ij}(1-p_{ij})}{\sqrt{Var[\mathbf{d}_i]Var[\mathbf{d}_j]}}$$

$$=\frac{p_{ij}(1-p_{ij})}{\sqrt{(\sum_{j=1}^{n}p_{ij}(1-p_{ij}))(\sum_{i=1}^{n}p_{ji}(1-p_{ji}))}}\approx 0$$

if

$$\left(\sum_{j=1}^{n}p_{ij}(1-p_{ij})\right)\left(\sum_{i=1}^{n}p_{ji}(1-p_{ji})\right)\gg p_{ij}^2(1-p_{ij})^2,$$

which would almost be satisfied as $N\to\infty$. Since $\mathbf{d}_i$ and $\mathbf{d}_j$ both have nearly normal distributions, a very small correlation implies they are almost independent. □

4. Intra-graph degree distribution

In a single random graph realization $G(N,E)$, its degree frequency can be written as follows:

$$P_G(k)=\frac{1}{N}\sum_{i=1}^{N}\delta(k,d_i)=\frac{1}{N}\sum_{i=1}^{N}\delta\left(k,\sum_{j=1}^{N}e_{ij}\right), \tag{4.1}$$

where d_i is the degree of vertex i in G, a non-negative integer, and $\delta(i,j)$ is the Kronecker delta symbol. $P_G(k)$ is called the intra-graph degree distribution for a given graph realization G. Actually, $P_G(k)$ is just the histogram which represents the frequency of appearance of each degree value. Due to the inter-graph fluctuation of each vertex degree, the function $P_G(k)$ could have different forms for different graph realizations. Therefore, in the random graph ensemble $\mathcal{G}(N,\mathbf{M})$, the intra-graph degree distribution should be a random N-dimensional vector, defined as $\{\mathbf{P}_G(k),k=0,1,\ldots,N-1\}$. For any given degree k, $\mathbf{P}_G(k)$ is a random variable, which could be expressed as follows:

$$\mathbf{P}_G(k)=\frac{1}{N}\sum_{i=1}^{N}\delta(k,\mathbf{d}_i)=\frac{1}{N}\sum_{i=1}^{N}\delta\left(k,\sum_{j=1}^{N}\mathbf{e}_{ij}\right). \tag{4.2}$$

Like the degree $\mathbf{d}_i$, it would also be informative to investigate the mean and variance of the random variable $\mathbf{P}_G(k)$, and their relationship with the random matrix $\mathbf{M}$.

In the random graph ensemble $\mathcal{G}(N,\mathbf{M})$, the degree sequence is written as $(\mathbf{d}_1,\mathbf{d}_2,\ldots,\mathbf{d}_N)$. This as an N-dimensional random vector where the N elements are almost independent of each other by Theorem 3.5. Denote the probability mass function of $\mathbf{d}_i$ by $f_i(k) \in [0,1]$. By Theorem 3.3, $f_i(k)$ could be approximated by a Poisson distribution when μ_i is small and $\max_j\{p_{ij}\} \to 0$:

$$f_i(k) \approx \frac{\mu_i^k}{k!} e^{-\mu_i}, k = 0,1,2,\ldots,N-1, \tag{4.3}$$

or a normal distribution otherwise:

$$f_i(k) \approx \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(k-\mu_i)^2}{2\sigma_i^2}\right), k = 0,1,2,\ldots,N-1, \tag{4.4}$$

where μ_i and σ_i^2 are given by equations (3.3) and (3.4), respectively.

Now let's consider the degree distribution $\mathbf{P}_G(k)$ given by equation (4.2). The function $\delta(k,\mathbf{d}_i)$ is also a Bernoulli random variable with success probability $f_i(k)$, hence $N\mathbf{P}_G(k)$ is the sum of N almost independent Bernoulli random variables. By Theorem 3.3, the random variable $\mathbf{P}_G(k)$ for a given k should follow a Poisson binomial distribution, which could be approximated by normal (or Poisson) distribution with mean and variance

$$E[\mathbf{P}_G(k)] = \frac{1}{N} E\left[\sum_{i=1}^{N} \delta(k,\mathbf{d}_i)\right] = \frac{1}{N}\sum_{i=1}^{N} f_i(k), \tag{4.5}$$

$$Var[\mathbf{P}_G(k)] = \frac{1}{N^2} Var\left[\sum_{i=1}^{N} \delta(k,\mathbf{d}_i)\right] \approx \frac{1}{N^2}\sum_{i=1}^{N} f_i(k)(1-f_i(k)). \tag{4.6}$$

The first equation shows the expected value of $\mathbf{P}_G(k)$. Since each $f_i(k) \in [0,1]$, the second equation implies that as $N \to \infty$,

$$Var[\mathbf{P}_G(k)] = \frac{1}{N^2}\sum_{i=1}^{N} f_i(k)(1-f_i(k)) \le \frac{1}{N^2}\sum_{i=1}^{N} \frac{1}{4} = \frac{1}{4N} \to 0.$$

Therefore, considering both the inter-graph degree fluctuation and intra-graph degree distribution, we can define a single degree distribution function $P(k)$ for the random graph ensemble $\mathcal{G}(N,\mathbf{M})$. Naturally this degree distribution $P(k)$ should be defined as the expected value of $\mathbf{P}_G(k)$ over all possible graph realizations in the ensemble, which can be obtained from equation (4.5).

With all of the facts given above, we are now in the position to state the following ergodic property of degree distributions:

Theorem 4.1. *In the random graph $\mathcal{G}(N,\mathbf{M})$, the intra-degree distribution $\mathbf{P}_G(k)$ of any graph realization will converge to the degree distribution $P(k)$ of the graph ensemble in the large limit of the graph size N. Moreover, $P(k)$ can be computed using equation (4.5).*

In dealing with large-scale random networks in the real world, one often counts only the degree frequency for a single network realization, and uses it as the degree distribution for the whole network ensemble. This theorem serves as the theoretical basis for the feasibility of this substitution. Note that the key assumption behind this theorem is that all the edge indicators $\mathbf{e}_{ij}$ are independent. The ergodic property no longer holds if this assumption is violated.

5. Random graph classification

In the previous two sections, we established the model $\mathcal{G}(N,\mathbf{M})$ of generalized random graphs and obtained some useful results concerning their degree properties. From these results, we know that the degree fluctuation of a given vertex i is almost determined by the two quantities in equation (3.3) and equation (3.4), which are both related to the probabilities p_{ij} associated with vertex i. This fact suggests a natural classification for the generalized random graphs: if the probability set $\{p_{ij}|j=1,2,\ldots,N,j\neq i\}$ around vertex i follows identical statistical properties over all the vertices in the graph, then in the statistical sense these vertices are homogeneous. Otherwise, they are heterogeneous and should be analyzed separately. In the following we distinguish between these two types of random graphs and make the respective analysis.

5.1. Homogeneous random graph.

The best known homogeneous random graph is the classical ER random graph. This model assumes all the probabilities take a same value p, thus making all the vertices identical in the statistical sense. Based on this model, a further extension could be made while the homogeneity over vertices is still maintained. Suppose that in the random graph $\mathcal{G}(N,\mathbf{M})$ the success probabilities p_{ij} for the $\mathbf{e}_{ij}$'s are independently sampled from a given distribution, say $F_p(x)$, where $x\in[0,1]$. From the symmetric property of $\mathbf{M}$, only $\frac{N(N-1)}{2}$ independent samples are needed. Under this model, although the probability sets $\{p_{ij}|j=1,2,\ldots,N,j\neq i\}$ are not identical for different vertices, they still have the same statistical properties when N is large since each set represents a large number of independent samples from the same distribution. Actually, by the law of large numbers, we immediately have the following results regarding the degree of vertex i as $N\to\infty$:

$$E[\mathbf{d}_i]=\sum_{j=1}^{'N}p_{ij}\approx N\int_0^1 xF_p(x)dx=NE[p], \tag{5.1}$$

$$\begin{aligned} Var[\mathbf{d}_i]=\sum_{j=1}^{'N}(p_{ij}-p_{ij}^2)&\approx N\int_0^1 (x-x^2)F_p(x)dx \\ &=N\left(E[p]-E[p^2]\right), \end{aligned} \tag{5.2}$$

showing that the degree of each vertex has the same mean and variance, thus approximately follows the same normal or Poisson distribution. This homogeneity among vertices resembles the situation in the classical ER random graph. Consequently, the degree distribution $P(k)$ of this type of homogeneous random graph has the simple form

$$P(k)=\frac{1}{N}\sum_{i=1}^{N}f_i(k)=f_i(k), \tag{5.3}$$

where $f_i(k)$ is given by either equation (4.3) or (4.4), i.e. the degree distribution of the random graph model is the same as the degree fluctuation of any vertex.

This type of random graph shares similar properties with the classical ER random graph, but with more complexity in the matrix $\mathbf{M}$. Actually, the ER random graph can be viewed as a special case of this model if we take $F_p(x)=\delta(x,p)$, where δ is the Kronecker delta symbol. The independent sampling from a given distribution for probabilities p_{ij}'s generates the homogeneity among vertices in the graph, yielding the approximately normal (or Poisson) degree distribution. On the contrary, if we observe that a random graph model has a degree distribution different from the normal (or Poisson) distribution, then the conclusion could be drawn that there must be an inhomogeneity among vertices. We classify such graphs to be heterogeneous random graphs as discussed below.

5.2. Heterogeneous random graph. The random graph $\mathcal{G}(N,\mathbf{M})$ is a heterogeneous random graph if the probability sets $\{p_{ij}|j=1,2,\dots,N,j\neq i\}$ around all vertices are not statistically identical. From the previous subsection, we know that a random graph with degree distribution different from normal (or Poisson), e.g., power-law, exponential, etc., must be a heterogeneous graph. However, the reverse statement is not true. A heterogeneous graph could also have a normal (or Poisson) degree distribution with appropriate choices of p_{ij}. Generally, heterogeneous random graphs are very hard to analyze due to the complexity of the entries in $\mathbf{M}$. The approximation formula for the degree distribution $P(k)$, given by equation (4.5), should be the best result we can extract from the general matrix $\mathbf{M}$.

However, there is a widely-used class of heterogeneous random graphs allowing us to make further analysis. This is the "intrinsic fitness model" introduced by Caldarelli, et al. [5, 7, 9]. In this model, each vertex i in the graph is assigned a "intrinsic fitness" number $x_i\in[0,\infty)$ independently sampled from a certain distribution $g_x(x)$, and the probability p_{ij} is defined to be $p_{ij}=h(x_i,x_j)$ for each pair of (i,j), where $h(a,b)$ is in [0,1], symmetric under exchange of arguments, i.e. $h(a,b)=h(b,a)$, and monotonically increasing with respect to each argument. For example, $h(a,b)$ could be ab, $a+b$, etc. As long as the intrinsic fitness x_i's assigned to the vertices are different, there would be a heterogeneity among all the vertices. With this particular model, the degree d_i of vertex i has mean and variance

$$E[\mathbf{d}_i]=\sum_{j=1}^{'N} p_{ij}\approx N\int_0^\infty h(x_i,y)g_x(y)dy=\mu(x_i), \tag{5.4}$$

$$\begin{aligned} Var[\mathbf{d}_i]=\sum_{j=1}^{'N}(p_{ij}-p_{ij}^2)\approx N\int_0^\infty (h(x_i,y)-h^2(x_i,y))g_x(y)dy \\ =\sigma^2(x_i). \end{aligned} \tag{5.5}$$

Consequently, the degree distribution $P(k)$ for the random graph $\mathcal{G}(N,\mathbf{M})$ can be expressed as:

$$P(k)=\frac{1}{N}\sum_{i=1}^{N} f_i(k)\approx\int_0^\infty \frac{\mu(x)^k}{k!}e^{-\mu(x)}g_x(x)dx, \tag{5.6}$$

or

$$P(k)=\frac{1}{N}\sum_{i=1}^{N} f_i(k)\approx\int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2(x)}}\exp\left(\frac{(k-\mu(x))^2}{2\sigma^2(x)}\right)g_x(x)dx, \tag{5.7}$$

depending on the value of $\mu(x)$, where $\mu(x)$ and $\sigma^2(x)$ are given by equations (5.4) and (5.5), respectively.

5.3. Discussion on the correlation effect. One important feature distinguishing the "intrinsic fitness model" from the homogeneous random graph model is that the p_{ij}'s around a given vertex are no longer independent samples from a distribution, but have certain kind of correlation. If we imagine each p_{ij} is one realization of a random variable $\mathbf{p}_{ij}$, then for the homogeneous random graph, all $\mathbf{p}_{ij}$'s are i.i.d. random variables following the same distribution $F_p(x)$. But for the intrinsic fitness model, $\mathbf{p}_{ij} = h(\mathbf{x}_i, \mathbf{x}_j)$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are i.i.d. random variables following the distribution $g_x(x)$. Hence for a given vertex i, the $\mathbf{p}_{ij}$'s associated with it would have correlations depending on the relation function $h(a,b)$. A vertex with a large fitness number tends to make the probabilities around it mostly high, while a vertex with a small fitness number tends to make them mostly low.

Introducing the idea of correlation would enrich our understanding of random graph classification. In all the previous discussion, we have assumed that the Bernoulli random variables $\mathbf{e}_{ij}$ representing edges in random graphs are *mutually independent*. Under this assumption, we classify random graphs into two types, "homogeneous" and "heterogeneous", solely based on the statistical properties of the $\mathbf{p}_{ij}$'s. A homogeneous random graph could have i.i.d. $\mathbf{p}_{ij}$'s, or identical, dependent $\mathbf{p}_{ij}$'s with the same type of correlation for every vertex, which would still keep the homogeneity among vertices. A heterogeneous random graph could have identical, dependent $\mathbf{p}_{ij}$'s with different types of correlation as in the case in the intrinsic fitness model, or independent, non-identical $\mathbf{p}_{ij}$'s, or even $\mathbf{p}_{ij}$'s which are neither identical nor independent. If we further introduce correlations to the Bernoulli random variables $\mathbf{e}_{ij}$, then the situation will become much more complicated. First, the Poisson approximation and the extended version of the central limit theorem dictated by Theorem 3.3 do not hold any more. Hence the degree fluctuation does not necessarily follow a normal (or Poisson) distribution, rather it could have any form depending on the correlations among $\mathbf{e}_{ij}$'s. Secondly, the ergodic property will break down. In the large limit of graph size N, the intra-graph degree distribution will not always converge to the degree distribution of the random graph ensemble. The simplest example to support this argument is that the $\mathbf{e}_{ij}$'s have perfect positive correlations. Then the graph realization only has two possibilities: a completely connected graph, or a graph with no edges. The intra-graph degree distribution will switch between two delta functions, and would never converge no matter how large the graph size is. In the random graph field, there has not been much study so far on the idea of edge or probability correlation, mainly due to the difficulty of defining the correlation quantitatively. However the richness of new problems makes it a promising research topic, and we will carry on a more thorough analysis in future publications.

6. Connectivity

The main focus of this paper is on the degree properties of generalized random graphs. Another interesting but more sophisticated topological property is the connectivity of random graphs. For the classical ER random graph $\mathcal{G}(N,p)$, a thorough analysis of its connectivity has been provided in [6, 11]. The main results for this simplest random graph model are as follows:

(i) if $Np < 1$, then almost surely the graph is disconnected and composed of isolated trees;

(ii) if $Np > 1$, then almost surely the graph has a giant component;

(iii) if $Np > \log N$, then almost surely the graph is totally connected.

For more realistic random graph models where edge probabilities are not identical, the analysis of connectivity could be very difficult since there is no general pattern shared by all the vertices. However, as shown by Molloy and Reed [17, 18], for the class of graphs with a same intra-graph degree distribution $P_G(k)$ as defined in equation (4.1), the connectivity of the random graph class can be determined solely by $P_G(k)$, as stated in the following lemma.

Lemma 6.1. *In the ensemble of graphs with the same degree sequence $P_G(k)$, if $\sum k(k-2)P_G(k)>0$, then such graphs almost surely have a giant component; if $\sum k(k-2)P_G(k)<0$, then such graphs almost surely are disconnected and composed of isolated small components.*

This criterion has also been obtained by Newman using a generating function method [20]. Note that the random graph model studied by these authors is quite different from what we study in this paper. In their model, they classify all the possible graphs with N vertices into different groups with the criterion that graphs in the same group has the same intra-graph degree distribution $P_G(k)$. For a given $P_G(k)$, the graph is chosen uniformly at random from that corresponding group. This is where the randomness comes from in this so-called "configuration" model. In our model $\mathcal{G}(N,\mathbf{M})$, however, the predetermined parameter is the random matrix $\mathbf{M}$. Under this matrix $\mathbf{M}$, different graph realizations may have different intra-graph degree distributions $P_G(k)$, thus would be classified into different groups in the configuration model. Fortunately, by Theorem 4.1 the variance of $P_G(k)$ among different realizations tend to be 0 as the graph size $N\to\infty$. Therefore, in the large limit of N, we could place all the realizations of random graph $\mathcal{G}(N,\mathbf{M})$ into a same group according to the degree distribution $P(k)$ of $\mathcal{G}(N,\mathbf{M})$. It seems that by this means the same criterion in Lemma 6.1 could be applied to infer the connectivity of our generalized random graph model $\mathcal{G}(N,\mathbf{M})$. However, this is not always true. In fact, this criterion is still valid and could even be simplified for the homogeneous random graphs, but is not applicable for the heterogeneous random graphs.

The homogeneous random graph has the nice feature that all the vertices have identical statistical properties. As shown in equation (5.3), all the vertices have the same degree fluctuation $f_i(k)$, which is also equal to the degree distribution $P(k)$ of the random graph model. This property allows us to treat the component size distribution problem as a homogeneous branching process, and apply the same generating function method used by Newman in [20]. The phase transition also occurs at the critical point $\sum k(k-2)P(k)=0$. Since the $P(k)$ in the homogeneous random graph model has a particular form, we can obtain a simpler criterion.

Theorem 6.2. *For the homogeneous random graph model $\mathcal{G}(N,\mathbf{M})$, if the average degree of any vertex is more than 2, then almost surely the graph has a giant component; if the average degree is less than 1, then almost surely the graph is disconnected and has many small components.*

Proof. If $f_i(k)$ is better approximated by equation (4.4), then

$$\sum_{k=0}^{N}k(k-2)P(k)=\sum_{k=0}^{N}k(k-2)f_i(k)\approx\sum_{k=0}^{N}k(k-2)\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(k-\mu_i)^2}{2\sigma_i^2}\right)$$

$$\approx\int_0^\infty x(x-2)\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)dx\approx\sigma_i^2+\mu_i^2-2\mu_i$$

$$=N(E[p]-E[p^2])+N^2E[p]^2-2NE[p]=N(NE[p]^2-E[p]-E[p^2]).$$

Therefore, the condition $\sum k(k-2)P(k)>0$ corresponds to $NE[p]^2>E[p]+E[p^2]$, i.e., $NE[p]>1+\frac{E[p^2]}{E[p]}$, and the condition $\sum k(k-2)P(k)<0$ corresponds to $NE[p]^2<E[p]+E[p^2]$, i.e., $NE[p]<1+\frac{E[p^2]}{E[p]}$. Since $0\le\frac{E[p^2]}{E[p]}\le 1$, we can make the conclusion that if $NE[p]=E[\mathbf{d}_i]>2$, then $\sum k(k-2)P(k)>0$; if $NE[p]=E[\mathbf{d}_i]<1$, then $\sum k(k-2)P(k)<0$.

If $f_i(k)$ is better approximated by equation (4.3), then

$$\begin{aligned}\sum_{k=0}^{N}k(k-2)P(k)&=\sum_{k=0}^{N}k(k-2)f_i(k)\approx\sum_{k=0}^{N}k(k-2)\frac{\mu_i^k}{k!}e^{-\mu_i}\\&=\sum_{k=0}^{N}[k(k-1)-k]\frac{\mu_i^k}{k!}e^{-\mu_i}=\sum_{k=2}^{N}\frac{\mu_i^k}{(k-2)!}e^{-\mu_i}-\sum_{k=1}^{N}\frac{\mu_i^k}{(k-1)!}e^{-\mu_i}\\&=\mu_i^2\sum_{k=2}^{N}\frac{\mu_i^{k-2}}{(k-2)!}e^{-\mu_i}-\mu_i\sum_{k=1}^{N}\frac{\mu_i^{k-1}}{(k-1)!}e^{-\mu_i}=\mu_i^2-\mu_i=\mu_i(\mu_i-1).\end{aligned}$$

Therefore, if $\mu_i=NE[p]>1$, then $\sum k(k-2)P(k)>0$; if $\mu_i=E[p]<1$, then $\sum k(k-2)P(k)<0$.

Combining the results in the two cases above, we can draw a common conclusion that if the average degree is more than 2, then almost surely the graph has a giant component; if the average degree is less than 1, then almost surely the graph is disconnected and has many small components. □

Unlike the classical ER random graph, there exits a gap $E[\mathbf{d}_i]=NE[p]\in(1,2)$ where we cannot make an obvious conclusion and should defer to the original criterion $\sum k(k-2)P(k)>0$. This is due to the intricacy introduced in our model, where the probability can take many values rather than a unique one. To explain the situation in this gap, we consider an example where two graphs with the same average degree have different connectivity situations. Let $E[\mathbf{d}_i]=NE[p]=1.5$. The first graph is constructed as follows: in the matrix $\mathbf{M}$ of that graph, each row has only two non-zero entries, one is a Bernoulli random variable with success probability 0.5, the other is a fixed number 1. The positions of these two entries are chosen uniformly and independently at random in each row. For this graph, the maximum degree is 2 and some of the vertices have degree 1. Hence $\sum k(k-2)P(k)<0$ and almost surely this graph is disconnected and has many small components. The second graph is the classical ER random graph with $Np\to1.5$ as $N\to\infty$. By the result (ii) in the beginning of this section, this graph almost surely has a giant component.

The heterogeneous random graph does not have any pattern to follow, and thus may generate a lot of complexity. For a given degree distribution $P(k)$, there could be many random graphs $\mathcal{G}(N,\mathbf{M})$ with different matrices $\mathbf{M}$ sharing this same degree distribution. The failure of building a 1-to-1 correspondence between $P(k)$ and $\mathbf{M}$ is the major reason why we can not replicate the criterion used in the configuration model. To see why this is true, we use a simple example as an illustration. Suppose the following degree distribution is given: $P(k)=1$ for $k=3$ and $P(k)=0$ for all other values of k, i.e. all the vertices in the graph have degree 3. We will construct two graphs, one is connected, the other is not, while both having the given degree distribution $P(k)$. The first graph has $N=4m$ vertices and m isolated components. Each component is a complete graph K_4 with 4 vertices, thus the degree of each vertex is 3. The second graph is constructed based on the first one with certain modification.

We firstly erase one edge in each component K_4, then connect all the vertices in the graph with degree 2, but not in the same component, to form a ring. This graph also has the same degree distribution, but is connected. If we choose the appropriate matrix $\mathbf{M}$ to represent the first graph and apply the criterion $\sum k(k-2)P(k)>0$, then the conclusion would be the graph almost surely has a giant component, which is obviously incorrect. Note that this example does not contradict with the original criterion for the configuration model, since the first graph is actually a rare event in the group of all graphs with degree distribution $P(k)$ and the probability of the appearance of that graph tends to 0 as $N\to\infty$. Almost all the other graphs with such degree distribution $P(k)$ would have a giant component, and the second graph is one of them.

From the above arguments, we acquire a brief idea of how connected a generalized random graph would be according to its representation matrix $\mathbf{M}$. To sum up, for the homogeneous random graph, if the average degree is more than 2, then the graph almost surely has a giant component; if the average degree is less than 1, then the graph is almost surely disconnected and has many small components. For the homogeneous random graph with average degree between 1 and 2, as well as the heterogeneous random graph, no general conclusions could be drawn and the connectivity should be studied case by case.

7. Conclusions

This paper introduces a generalized random graph (GRG) model which would fit to most real-world random networks. Based on this model, we have studied the degree properties, namely the inter-graph degree fluctuation and the intra-graph degree distribution. It is pointed out that the degree distribution of the whole graph model and the degree sequence of a single graph realization have different concepts, and due to the inter-graph degree fluctuation on each vertex, they also have different analytical expressions. However, if all the edge indicators are mutually independent, then in a graph with very large size the ergodic property ensures that the intra-graph degree distribution in a single graph realization converges to the degree distribution of the whole graph model. Moreover, we classify the GRGs into two types: "homogeneous" and "heterogeneous", based on the extent of similarity among vertices in a statistical sense. It is found that the homogeneous random graphs have many nice features which resemble the classical ER random graph. Finally, we explore the issue of connectivity in the GRG model. Discussion is mainly focused on the condition of the emergence of a giant component in different types of random graphs, and some simple criteria for the case of homogeneous random graphs are derived .

The structure and dynamics of random graphs has been a rapidly developing research field during the last few decades [8, 13, 23]. While most studies focused on particular models and their corresponding behavior [1, 5, 7], this paper attempts to understand a generalized random graph model. Although only the subject of degree properties and connectivity have been studied, which might be the simplest task among all the features of complex graphs, it is our hope that it could serve as a starting point toward more comprehensive research of this intriguing field, such as network growth and dynamics, etc. We believe that such investigations will eventually help us gain deeper understanding to the complex systems in the real world.

Acknowledgement. We thank Woo H. Kim and Melissa Vellela for many helpful discussions, Professors R.M. O'Malley and K.K. Tung for constant encouragements, and Professors K. Bomsztyk and G.A. Miller for collaborations on biological applications. This work is in part supported by NIH grant GM04134/G232JA.

REFERENCES

[1] W. Aiello, F.R.K. Chung and L. Lu, *A random graph model for power law graphs*, Exper. Math., 10, 53–66, 2001.

[2] A.L. Barabási and R. Albert, *Emergence of scaling in random networks*, Science, 286, 509–512, 1999.

[3] A.L. Barabási, R. Albert and H. Jeong, *Power-law distribution of the World Wide Web*, Science, 287, 2115a, 2000.

[4] R. Albert and A.L. Barabási, *Statistical mechanics of complex networks*, Rev. of Modern Phys., 74, 47–94, 2002.

[5] M. Bogũná and R. Pastor-Satorras, *Class of correlated random networks with hidden variables*, Phys. Rev. E, 68, 036112, 2003.

[6] B. Bollobás, *Random Graphs*, 2nd ed., Cambridge University Press, 2001.

[7] G. Caldarelli, A. Capocci, P. De Los Rios and M.A. Muñoz, *Scale-free networks from varying vertex intrinsic fitness*, Phys. Rev. Lett., 89, 258702, 2002.

[8] F.R.K. Chung and L. Lu, *Complex Graphs and Networks*, AMS Press, 2006.

[9] F.R.K. Chung and L. Lu, *Connected components in random graphs with given expected degree sequence*, Annals of Combin., 6, 125–145, 2002.

[10] E.J. Deeds, O. Ashenberg and E.I. Shakhnovich, *A simple physical model for scaling in protein-protein interaction networks*, Proc. Natl. Acad. Sci. USA, 103, 311–316, 2006.

[11] P. Erdös and A. Rényi, *On random graphs*, Publ. Math. Debrecen, 6, 290–297, 1959.

[12] J.L. Hodges and L.L. Cam, *The Poisson approximation to the Poisson binomial distribution*, Annl. of Math. Stat., 31(3), 737–740, 1960.

[13] P.L. Krapivsky and S. Redner, *Organization of growing random networks*, Phys. Rev. E, 63, 066123, 2001.

[14] C. McDiarmid, *Concentration*, Algorithms Combin., Springer, Berlin, 16, 195–248, 1998.

[15] V.G. Mikhailov, *On a refinement of the central limit theorem for sums of independent random indicators*, Theory Prob. Appl., 38(3), 479–489, 1993.

[16] G.A. Miller, Y.Y. Shi, H. Qian and K. Bomsztyk, *Clustering coefficient of protein-protein interactions*, Phys. Rev. E, 75, 051910, 2007.

[17] M. Molloy and B. Reed, *A critical point for random graphs with a given degree sequence*, Random Structures and Algorithms, 6, 161–179, 1995.

[18] M. Molloy and B. Reed, *The size of the giant component of a random graph with a given degree sequence*, Combin. Prob. and Comput., 7, 295–305, 1998.

[19] M.E.J. Newman, *The structure and function of complex networks*, SIAM Rev., 45, 167–256, 2003.

[20] M.E.J. Newman, *Component sizes in networks with arbitrary degree distributions*, Phys. Rev. E, 76, 045101, 2007.

[21] Y.Y. Shi, G.A. Miller, H. Qian and K. Bomsztyk, *Free-energy distribution of binary protein-protein binding suggests cross-species interactome differences*, Proc. Natl. Acad. Sci. USA, 103, 11527–11532, 2006.

[22] Y.Y. Shi, G.A. Miller, O. Denisenko, H. Qian and K. Bomsztyk, *Quantitative model for binary measurements of protein-protien interactions*, Journ. of Comp. Bio., September 1, 14(7), 1011–1023, 2007.

[23] D.J. Watts, *Small worlds: The Dynamics of Networks Between Order and Randomness*, Princeton University Press, 1999.