Probability in the Engineering and Informational Sciences, **23**, 2009, 661–674. Printed in the U.S.A.
doi:10.1017/S0269964809990064

RANDOM INTERSECTION GRAPHS WITH TUNABLE DEGREE DISTRIBUTION AND CLUSTERING

Maria Deijfen
Department of Mathematics
Stockholm University
106 91 Stockholm, Sweden
E-mail: mia@math.su.se

Willemien Kets
Santa Fe Institute
1399 Hyde Park Road
Santa Fe, NM 87501 and
CentER, Tilburg University
P.O. Box 90153
5000 LE Tilburg, Netherlands
E-mail: willemien.kets@santafe.edu

A random intersection graph is constructed by assigning independently to each vertex a subset of a given set and drawing an edge between two vertices if and only if their respective subsets intersect. In this article a model is developed in which each vertex is given a random weight and vertices with larger weights are more likely to be assigned large subsets. The distribution of the degree of a given vertex is characterized and is shown to depend on the weight of the vertex. In particular, if the weight distribution is a power law, the degree distribution will be as well. Furthermore, an asymptotic expression for the clustering in the graph is derived. By tuning the parameters of the model, it is possible to generate a graph with arbitrary clustering, expected degree, and—in the power-law case—tail exponent.

1. INTRODUCTION

During the last decade there has been a large interest in the study of large complex networks; see, for example, Dorogovtsev and Mendes [7] and Newman, Strogatz, and Watts [16] and the references therein. Due to the rapid increase in computer power,

 0269-9648/09 $25.00

it has become possible to investigate various types of real networks such as social contact structures, telephone networks, power grids, the Internet, and the World Wide Web. The empirical observations reveal that many of these networks have similar properties. For instance, they typically have power-law degree sequences, that is, the fraction of vertices with degree k is proportional to $k^{-\tau}$ for some exponent $\tau > 1$. Furthermore, many networks are highly clustered—meaning roughly that there is a large number of triangles and other short cycles. In a social network, this is explained by the fact that two people who have a common friend often meet and become friends, creating a triangle in the network. A related explanation is that human populations are typically divided into various subgroups—working places, schools, associations, and so forth—which gives rise to high clustering in the social network, since members of a given group typically know each other; see Palla, Derényi, Farkas, and Vicsek [17] for some empirical observations.

Real-life networks are generally very large, implying that it is a time-consuming task to collect data to delineate their structure in detail. This makes it desirable to develop models that capture essential features of the real networks. A natural candidate to model a network is a random graph, and, to fit with the empirical observations, such a graph should have a heavy-tailed degree distribution and considerable clustering. We will quantify the clustering in a random graph by the conditional probability that three given vertices constitute a triangle, given that two of the three possible links between them exist. Other (empirical) definitions occur in the literature—see, for example Newman [14]—but they all capture essentially the same thing.

Obviously, the classical Erdős–Rényi graph will not do a good job as a network model, since the degrees are asymptotically Poisson distributed. Moreover, existing models for generating graphs with a given degree distribution—see, for example, Molloy and Reed [12,13]—typically have zero clustering in the limit. In this article, we propose a model, based on the so-called random intersection graph, in which both the degree distribution and the clustering can be controlled. More precisely, the model makes it possible to obtain arbitrary prescribed values for the clustering and to control the mean and the tail behavior of the degree distribution.

1.1. Description of the Model

The random intersection graph was introduced by Singer [18] and Karoński, Scheinerman, and Singer-Cohen [11] and has been further studied and generalized by Fill, Scheinerman, and Singer-Cohen [8], Godehardt and Jaworski [9], Stark [19], and Jaworksi, Karoński, and Stark [10]. Newman [14] and Newman and Park [15] discussed a similar model. In its simplest form the model is defined as follows.

1. Let $\mathcal{V} = \{1, \ldots, n\}$ be a set of n vertices and $\mathcal{A}$ a set of m elements. For $p \in [0, 1]$, construct a bipartite graph $B(n, m, p)$ with vertex sets $\mathcal{V}$ and $\mathcal{A}$ by including each one of the nm possible edges between vertices from $\mathcal{V}$ and elements from $\mathcal{A}$ independently with probability p.

2. The random intersection graph $G(n,m,p)$ with vertex set $\mathcal{V}$ is obtained by connecting two distinct vertices $i,j \in \mathcal{V}$ if and only if there is an element $a \in \mathcal{A}$ such that both i and j are adjacent to a in $B(n,m,p)$.

When the vertices in $\mathcal{V}$ are thought of as individuals and the elements of $\mathcal{A}$ as social groups, this gives rise to a model for a social network in which two individuals are joined by an edge if they share at least one group. In the following, we frequently borrow the terminology from the field of social networks and refer to the vertices as individuals and the elements of $\mathcal{A}$ as groups, with the understanding that the model is, of course, much more general.

To get an interesting structure, the number of groups m is typically set to $m = \lfloor n^\alpha \rfloor$ for some $\alpha > 0$; see Karoński et al. [11]. We will assume this form for m in the following. Let D_i be the degree of vertex $i \in \mathcal{V}$ in $G(n,m,p)$. The probability that two individuals do not share a group in $B(n,m,p)$ is $(1-p^2)^m$. It follows that the edge probability in $G(n,m,p)$ is $1-(1-p^2)^m$ and, hence, the expected degree is

$$\begin{aligned}\mathbf{E}[D_i] &= (n-1)(1-(1-p^2)^m) \\ &= (n-1)\left(mp^2 + O(m^2p^4)\right).\end{aligned}$$

To keep the expected degree bounded as $n \to \infty$, we let $p = \gamma n^{-(1+\alpha)/2}$ for some constant $\gamma > 0$. We then have that $\mathbf{E}[D_i] \to \gamma^2$.

Stark [14, Thm. 2] showed that in a random intersection graph with the above choice of p, the distribution of the degree of a given vertex converges to a point mass at zero, a compound Poisson distribution or a Poisson distribution depending on whether $\alpha < 1$, $\alpha = 1$, or $\alpha > 1$. This means that the current model cannot account for the power-law degree distributions typically observed in real networks.

In the above formulation of the model, the number of groups that a given individual belongs to is binomially distributed with parameters m and p. A generalization of the model, allowing for an arbitrary group distribution, is described by Godehardt and Jaworski [9]. The degree of a given vertex in such a graph is analyzed by Jaworski et al. [10], for which conditions on the group distribution are specified under which the degree is asymptotically Poisson distributed.

In the current article, we are interested in obtaining graphs where non-Poissonian degree distributions can be identified. To this end, we propose a generalization of the original random intersection graph where the edge probability p is random and depends on weights associated with the vertices. Other work in this spirit include, for instance, Chung and Lu [4,5], Yao, Zhang, Chen, and Li [20], Britton, Deijfen, and Martin-Löf [3], Bollobás, Janson, and Riordan [1] and Deijfen, van den Esker, van der Hofstad, and Hooghiemstra [6]. The model is defined as follows:

1. Let n be a positive integer and define $m = \lfloor \beta n^\alpha \rfloor$ with $\alpha, \beta > 0$. As earlier, take $\mathcal{V} = \{1, \ldots, n\}$ to be a set of n vertices and $\mathcal{A}$ a set of m elements. Additionally, let $\{W_i\}$ be an independent and identically distributed (i.i.d.) sequence of positive random variables with distribution F, where F is assumed to have

https://doi.org/10.1017/S0269964809990064 Published online by Cambridge University Press

mean 1 if the mean is finite. Finally, for some constant $\gamma > 0$, set

$$p_i = \gamma W_i n^{-(1+\alpha)/2} \wedge 1. \tag{1}$$

Now, construct a bipartite graph $B(n,m,F)$ with vertex sets $\mathcal{V}$ and $\mathcal{A}$ by adding edges to the elements of $\mathcal{A}$ for each vertex $i \in \mathcal{V}$ independently with probability p_i.

2. The random intersection graph $G(n,m,F)$ is obtained as earlier by drawing an edge between two distinct vertices $i,j \in \mathcal{V}$ if and only if they have a common adjacent vertex $a \in \mathcal{A}$ in $B(n,m,F)$.

In the social network setting, the weights can be interpreted as a measure of the social activity of the individuals. Indeed, vertices with large weights are more likely to join many groups and thereby acquire many social contacts. There are several other examples of real networks for which the success of a vertex (measured by its degree) depends on some specific feature of the vertex; see, for example, Palla et al. [17] for an example in the context of protein interaction networks. Furthermore, an advantage of the model is that it has an explicit and straightforward construction that, as we will see, makes it possible to exactly characterize the degree distribution and the clustering in the resulting graph.

1.2. Results

Our results concern the degree distribution and the clustering in the graph $G(n,m,F)$ as $n \to \infty$. More precisely, we will take the parameters α, β, and γ and the weight distribution F to be fixed (independent of n) and then analyze the degree of a given vertex and the clustering in the graph as $n \to \infty$. It turns out that the behavior of these quantities will be different in the three regimes $\alpha < 1$, $\alpha = 1$, and $\alpha > 1$, respectively. The interesting case is $\alpha = 1$, in the sense that this is when both the degree distribution and the clustering can be controlled. The cases $\alpha < 1$ and $\alpha > 1$ are included for completeness.

As for the degree, we begin by observing that if F has finite mean, then the asymptotic mean degree of vertex i, conditional on W_i, is given by $\beta\gamma^2 W_i$ for all values of α.

PROPOSITION 1.1: *Let D_i be the degree of vertex $i \in \mathcal{V}$ in a random intersection graph $G(n,m,F)$ with $m = \lfloor \beta n^\alpha \rfloor$ and p_i as in (1). If F has finite mean, then, for all values of $\alpha > 0$, we have that $\mathbf{E}[D_i|W_i] \to \beta\gamma^2 W_i$ as $n \to \infty$.*

PROOF: We prove the claim for vertex $i = 1$. Define

$$W_j' = W_j \cdot \mathbf{1}_{\{W_j \le n^{1/4}\}} \quad \text{and} \quad W_j'' = W_j \cdot \mathbf{1}_{\{W_j > n^{1/4}\}}$$

and let D' and D'' denote the degree of vertex 1 when $\{W_j\}_{j\neq 1}$ are replaced by $\{W_j'\}$ and $\{W_j''\}$, respectively; that is, D' is the number of neighbors of 1 with weight smaller than or equal to $n^{1/4}$ and D'' is the number of neighbors with weight larger than $n^{1/4}$. Write p_j' and p_j'' for the analog of (1) based on the truncated weights.

Now, conditional on the weights, the probability that there is an edge between 1 and j is $1-(1-p_1p_j)^m$. To see that $\mathbf{E}[D''] \to 0$ as $n \to \infty$, we observe that

$$1-(1-p_ip_j'')^m \leq mp_1p_j'' = \beta\gamma W_1 n^{(\alpha-1)/2}p_j''.$$

Summing the expectation of the right-hand side over $j \neq 1$, keeping W_1 fixed, gives (recall the truncation at 1 in (1))

$$\mathbf{E}[D''] \leq \beta\gamma n^{(1+\alpha)/2}\mathbf{E}[p_k''] \leq \beta\gamma\big(\gamma\mathbf{E}[W_k''] + n^{(1+\alpha)/2}\mathbf{P}\big(\gamma W_k \geq n^{(1+\alpha)/2}\big)\big),$$

where both terms on the right-hand side converge to 0 as $n \to \infty$ since F has finite mean. As for D', we have

$$1-(1-p_1p_j')^m = \beta\gamma^2 W_1 W_j' n^{-1} + O(W_1^2(W_j')^2n^{-2}).$$

The sum over $j \neq 1$ of the expectation of the first term equals $\beta\gamma^2W_1\mathbf{E}[W_k']$, where $\mathbf{E}[W_k'] \to \mathbf{E}[W_k] = 1$ (since F has finite mean) and the sum of the expectation of the second term converges to zero (since $(W_j')^2 \leq n^{1/2}$). Since $D_0 = D' + D''$, this proves the proposition. ■

The following theorem, which is a generalization of Theorem 2 in Stark [19], gives a full characterization of the degree distribution for different values of α.

THEOREM 1.1: *Consider the degree D_i of vertex $i \in \mathcal{V}$ in a random intersection graph $G(n,m,F)$ with $m = \lfloor \beta n^\alpha \rfloor$ and p_i as in (1) and assume that F has finite mean.*

- *(a) If $\alpha < 1$, then D_i converges in distribution to a point mass at 0 as $n \to \infty$.*
- *(b) If $\alpha = 1$, then D_i converges in distribution to a sum of a Poisson($\beta\gamma W_i$) distributed number of Poisson(γ) variables, where all variables are independent.*
- *(c) If $\alpha > 1$, then D_i is asymptotically Poisson($\beta\gamma^2 W_i$) distributed.*

To understand Theorem 1.1, note that the expected number of groups that individual i belongs to is roughly $\beta\gamma W_i n^{(\alpha-1)/2}$. If $\alpha < 1$ and W_i has finite mean, this converges to zero in probability, so that the degree distribution converges to a point mass at zero, as stated in part (a) (the group size, however, goes to infinity, explaining why the expected degree is still positive in the limit). For $\alpha = 1$, the number of groups that individual i is a member of is Poisson($\beta\gamma W_i$) distributed as $n \to \infty$, and the number of other individuals in each of these groups is approximately Poisson(γ) distributed, which explains part (b). Finally, for $\alpha > 1$, individual i belongs to infinitely many groups as $n \to \infty$. This means that the edges indicators will be asymptotically independent, giving rise to the Poisson distribution specified in part (c).

Moving on to the clustering, write E_{ij} for the event that individuals $i, j \in \mathcal{V}$ have a common group in the bipartite graph $B(n,m,F)$; that is, E_{ij} is equivalent to the event that there is an edge between vertices i and j in $G(n,m,F)$. Let $\bar{\mathbf{P}}_n$ be the probability

https://doi.org/10.1017/S0269964809990064 Published online by Cambridge University Press

measure of $B(n, m, F)$ conditional on the weights $\{W_1, \dots, W_n\}$. For distinct vertices $i, j, k \in \mathcal{V}$, define

$$\bar{c}^{(n)}_{i,j,k} = \bar{\mathbf{P}}_n \left(E_{ij} | E_{ik}, E_{jk}\right); \tag{2}$$

that is, $\bar{c}^{(n)}_{i,j,k}$ is the edge probability between i and j in $G(n, m, F)$, given that they are both connected to k, conditional on the weights. To quantify the asymptotic clustering in the graph, we will use

$$c(G) := \lim_{n\to\infty} \mathbf{E}\big[\bar{c}^{(n)}_{i,j,k}\big],$$

where the expectation is taken over the weights; that is, $c(G)$ is the limiting probability that three given vertices constitute a triangle conditional on that two of the three possible edges between them exist (the vertices are indistinguishable, so indeed $c(G)$ does not depend on the particular choice of i, j, and k). This should be closely related to the limiting quotient of the number of triangles and the number of triples with at least two edges present, which is one of the empirical measures of clustering that occur in the literature; see, for example, Newman [14]. Establishing this connection rigorously however requires additional arguments.

The asymptotic behavior of $\bar{c}^{(n)}_{i,j,k}$ is specified in the following theorem. By bounded convergence, it follows that $c(G)$ is obtained as the mean of the in-probability-limits.

THEOREM 1.2: *Let $\{i, j, k\}$ be three distinct vertices in a random intersection graph $G(n, m, F)$ with $m = \lfloor \beta n^{\alpha} \rfloor$ and p_i as in* (1). *If F has finite mean, then we have the following:*

(a) $\bar{c}^{(n)}_{i,j,k} \to 1$ *in probability for* $\alpha < 1$*;*

(b) $\bar{c}^{(n)}_{i,j,k} \to (1 + \beta\gamma W_k)^{-1}$ *in probability for* $\alpha = 1$*;*

(c) $\bar{c}^{(n)}_{i,j,k} \to 0$ *in probability for* $\alpha > 1$.

To understand Theorem 1.2, assume that i and k share a group and that j and k share a group. The probability that i and j also have a common group then depends on the number of groups to which the common neighbor k belongs. Indeed, the fewer groups k belongs to, the more likely it is that i and j in fact share the same group with k. Recall that the expected number of groups that k belongs to is roughly $\beta\gamma W_k n^{(\alpha-1)/2}$. If $\alpha > 1$, this goes to zero as $n \to \infty$. Since it is then very unlikely that k belongs to more than one group when n is large, two given edges $\{i, k\}$ and $\{j, k\}$ are most likely generated by the same group, meaning that i and j are connected as well. On the other hand, if $\alpha > 1$, the number of groups that k belongs to is asymptotically infinite. Hence, that i and j each belong to one of these groups does not automatically make it likely that they actually belong to the same group. If $\alpha = 1$, individual k belongs to $\beta\gamma W_k$ groups on average, explaining the expression in part (b) of the theorem.

From Theorem 1.2 it follows that to get a nontrivial tunable clustering, we should choose $\alpha = 1$. Indeed, then we have $c(G) = \mathbf{E}[(1 + \beta\gamma W_k)^{-1}]$, and for a given weight distribution F (with finite mean), $c(G)$ can be varied between zero and 1 by adjusting the parameters β and γ. Furthermore, when $\alpha = 1$, the degree distribution for a given

https://doi.org/10.1017/S0269964809990064 Published online by Cambridge University Press

vertex is asymptotically compound Poisson with the weight of the vertex as one of the parameters—see Theorem 1.1(b)—and it is not hard to see that if F is a power law with exponent τ, then the degree distribution will be as well. Since the mean of F is set to 1, the expected asymptotic degree is $\beta\gamma^2$ by Proposition 1.1. Taken together, this means that when $\alpha = 1$, we can obtain a graph with a given value of the clustering and a power-law degree distribution with prescribed exponent and prescribed mean by first choosing F to be a power law with the desired exponent and then tuning the parameters β and γ to get the correct values of the clustering and the expected degree.

The rest of the article is organized as follows. In Sections 2 and 3, Theorem 1.1 and Theorem 1.2 are proved, respectively. The clustering is analyzed for the important example of a power-law weight distribution in Section 4. Finally, Section 5 provides an outline of possible future work.

2. THE DEGREE DISTRIBUTION

We begin by proving Theorem 1.1.

PROOF OF THEOREM 1.1: We prove the theorem for vertex $i = 1$. Write $D_1 = D$ and denote by N the number of groups to which individual 1 belongs. Conditional on W_1, the variable N is binomially distributed with parameters m and p_1 and, thus,

$$\bar{\mathbf{P}}_n(N = 0) = (1 - p_1)^m \geq 1 - mp_1 \geq 1 - \beta\gamma^2 W_1^{(\alpha-1)/2}.$$

For $\alpha < 1$, the expectation of the last term converges to zero as $n \to \infty$, and it follows from bounded convergence that $\mathbf{P}(N = 0) = \mathbf{E}[\bar{\mathbf{P}}_n(N = 0)] \to 1$. This proves part (a), since clearly $D = 0$ if individual 1 is not a member of any group.

To prove parts (b) and (c), first recall the definition of the weights $\{W_i'\}$ and $\{W_i''\}$—truncated from above and below, respectively, at $n^{1/4}$—and the corresponding degree variables D' and D'' from the proof of Proposition 1.1. We have already showed (in proving Proposition 1.1) that $\mathbf{E}[D''] \to 0$, which implies that D'' converges to zero in probability (indeed, $\mathbf{P}(D'' > 0) \leq \mathbf{E}[D'']$). Hence, it suffices to show that the generating function of D' converges to the generating function of the claimed limiting distribution. To this end, we condition on the weight W_1, which is thus assumed to be fixed in what follows, and let X_i' $(i = 2, \ldots, n)$ denote the number of common groups of individual 1 and individual i when the truncated weights W_i' are used for $i \neq 1$. Since two individuals are connected if and only if they have at least one group in common, we can write $D' = \sum_{i=2}^n \mathbf{1}_{\{X_i' \geq 1\}}$. Furthermore, conditional on N and $\{W_i'\}_{i\geq 2}$, the random variables X_i', $i = 2, \ldots, n$, are independent and binomially distributed with parameters N and $p_i' = \gamma W_i' n^{-(1+\alpha)/2}$. Hence, with $\bar{\bar{\mathbf{P}}}_n$ denoting the probability measure of the bipartite graph $B(n, m, F)$ conditional on both $\{W_i'\}_{i\geq 2}$ and N, the generating function of D' can be written as

$$\mathbf{E}[t^{D'}] = \mathbf{E}\left[\prod_{i=2}^n \mathbf{E}\left[t^{\mathbf{1}\{X_i' \geq 1\}} \middle| \{W_i'\}, N\right]\right] = \mathbf{E}\left[\prod_{i=2}^n \left(1 + (t-1)\bar{\bar{\mathbf{P}}}_n(X_i' \geq 1)\right)\right],$$

https://doi.org/10.1017/S0269964809990064 Published online by Cambridge University Press

where $t \in [0, 1]$. Using the Taylor expansion $\log(1 + x) = x + O(x^2)$ and the fact that

$$\bar{\bar{\mathbf{P}}}_n(X_i' \geq 1) = 1 - (1 - p_i')^N = Np_i' + O(N^2(p_i')^2),$$

we get that

$$\prod_{i=2}^{n} \Big(1 + (t-1)\bar{\bar{\mathbf{P}}}_n(X_i' \geq 1)\Big) \quad = \quad e^{(t-1)N\sum p_i' + O\left(N^2 \sum (p_i')^2\right)}. \tag{3}$$

Defining

$$R_n := \prod_{i=2}^{n} \Big(1 + (t-1)\bar{\bar{\mathbf{P}}}_n(X_i' \geq 1)\Big) - e^{(t-1)N\sum p_i'},$$

we, therefore, have that

$$R_n = e^{(t-1)N\sum p_i}\Big(e^{O(N^2 \sum (p_i')^2)} - 1\Big).$$

Since the product in (3) is the conditional expectation of $t^{D'}$ with $t \in [0, 1]$, it takes values between zero and 1, and since $e^{(t-1)N\sum p_i} \in (0, 1]$, it follows that $R_n \in [-1, 1]$. Furthermore, recalling that $W_i' \leq n^{1/4}$, we have for $\alpha \geq 1$ that

$$N^2 \sum_{i=2}^{n} (p_i')^2 = N^2 \gamma^2 n^{-(1+\alpha)} \sum_{i=2}^{n} (W_i')^2 \leq N^2 \gamma^2 n^{-1/2},$$

implying that $R_n \to 0$ in probability and thus, by bounded convergence, $\mathbf{E}[R_n] \to 0$. Hence, we are done if we show the following:

(i) $\mathbf{E}\left[e^{(t-1)N\sum p_i}\right] \to e^{\beta\gamma W_1 (e^{\gamma(t-1)} - 1)}$ if $\alpha = 1$,

(ii) $\mathbf{E}\left[e^{(t-1)N\sum p_i}\right] \to e^{\beta\gamma^2 W_1 (t-1)}$ if $\alpha > 1$,

where the limits are recognized as the generating functions for the desired compound Poisson and Poisson distribution in parts (b) and (c) of the theorem, respectively. To this end, note that the expectation with respect to N of $e^{(t-1)N\sum p_i'}$ is given by the generating function for N evaluated at the point $e^{(t-1)\sum p_i'}$. Since N is binomially distributed with parameters m and p_1, we have that

$$\mathbf{E}\left[e^{(t-1)N\sum p_i'}\right] = \mathbf{E}\left[\left(1 + p_1\left(e^{(t-1)\sum p_i'} - 1\right)\right)^m\right]. \tag{4}$$

For $\alpha = 1$, we have $m = \lfloor \beta n \rfloor$ and $p_i' = \gamma W_i' n^{-1}$. Recalling that $\mathbf{E}[W_i'] \to \mathbf{E}[W_i] = 1$, it follows that $\sum p_i' \to \gamma$ almost surely. Hence,

$$\left(1 + p_1\left(e^{(t-1)\sum p_i'} - 1\right)\right)^{\lfloor \beta n \rfloor} \to e^{\beta\gamma W_1 (e^{\gamma(t-1)} - 1)} \qquad \text{a.s. as } n \to \infty,$$

and it follows from bounded convergence that the expectation converges to the same limit, proving part (i).

https://doi.org/10.1017/S0269964809990064 Published online by Cambridge University Press

For $\alpha > 1$, define $\tilde{p}'_i = n^{(\alpha-1)/2} p'_i$. With $m = \lfloor \beta n^\alpha \rfloor$ and $p_1 = \gamma W_1 n^{-(1+\alpha)/2} \wedge 1$, we get, after some rewriting, that

$$\left(1 + p_1 \left(e^{(t-1)\sum p'_i} - 1\right)\right)^m = \left(1 + \frac{\gamma W_1 (t-1) \sum \tilde{p}'_i}{n^\alpha} \frac{e^{(t-1)n^{(1-\alpha)/2} \sum \tilde{p}'_i} - 1}{(t-1) n^{(1-\alpha)/2} \sum \tilde{p}'_i}\right)^{\lfloor \beta n^\alpha \rfloor}.$$

By the law of large numbers, $\sum \tilde{p}'_i \to \gamma$ almost surely, and since $(e^x - 1)/x \to 1$ as $x \to 0$, it follows that the right-hand side above converges to $e^{\beta \gamma^2 W_1 (t-1)}$ almost surely as $n \to \infty$. By (4) and bounded convergence, this proves part (ii). ■

3. CLUSTERING

In this section, we prove Theorem 1.2. First, recall that E_{ij} denotes the event that the individuals $i, j \in \mathcal{V}$ share at least one group. It will be convenient to extend this notation. To this end, for $i, j, k \in \mathcal{V}$, denote by E_{ijk} the event that there is at least one group to which all three individuals i, j, and k belong, and write $E_{ij,ik,jk}$ for the event that there are at least three *distinct* groups to which i and j, i and k, and j and k, respectively, belong. Similarly, the event that there are two distinct groups to which individuals i and k, and j and k, respectively belong is denoted by $E_{ik,jk}$. The proof of Theorem 1.2 relies on the following lemma.

LEMMA 3.1: *Consider a random intersection graph $G(n, m, F)$ with $m = \lfloor \beta n^\alpha \rfloor$ and p_i defined as in (1). For any three distinct vertices $i, j, k \in \mathcal{V}$, we have the following:*

(a) $\bar{\mathbf{P}}_n(E_{ijk}) = \frac{\beta \gamma^3 W_i W_j W_k}{n^{(3+\alpha)/2}} + O\left(\frac{W_i^2 W_j^2 W_k^2}{n^{3+\alpha}}\right)$;

(b) $\bar{\mathbf{P}}_n(E_{ij,ik,jk}) = \frac{\beta^3 \gamma^6 W_i^2 W_j^2 W_k^2}{n^3} + O\left(\frac{W_i^3 W_j^3 W_k^3}{n^4}\right)$;

(c) $\bar{\mathbf{P}}_n(E_{ik,jk}) = \frac{\beta^2 \gamma^4 W_i W_j W_k^2}{n^2} + O\left(\frac{W_i^2 W_j^2 W_k^3}{n^3}\right)$;

(d) $\bar{\mathbf{P}}_n(E_{ijk} E_{ik,jk}) = O\left(\frac{W_i^2 W_j^2 W_k^2}{n^{(5+\alpha)/2}}\right)$.

PROOF: As for part (a), the probability that three given individuals i, j, and k do not share any group at all is $(1 - p_i p_j p_k)^m$. Using the definitions of m and the edge probabilities $\{p_i\}$, it follows that

$$\bar{\mathbf{P}}_n(E_{ijk}) = 1 - (1 - p_i p_j p_k)^m = \frac{\beta \gamma^3 W_i W_j W_k}{n^{(3+\alpha)/2}} + O\left(\frac{W_i^2 W_j^2 W_k^2}{n^{3+\alpha}}\right).$$

To prove part (b), note that the probability that there is exactly one group to which both i and j belong is $m p_i p_j (1 - p_i p_j)^{m-1} = m p_i p_j + O(m^2 p_i^2 p_j^2)$. Given that i and j share one group, the probability that i and k share exactly one of the *other* $m - 1$ groups is $(m-1) p_i p_k (1 - p_i p_k)^{m-2} = m p_i p_k + O(m^2 p_i^2 p_k^2)$. Finally, the conditional

https://doi.org/10.1017/S0269964809990064 Published online by Cambridge University Press

probability that there is a third group to which both j and k belong given that the pairs i,j and i,k share one group each is $1-(1-p_jp_k)^{m-2} = mp_jp_k + O(m^2p_j^2p_k^2)$. Combining these estimates and noting that scenarios in which i and j or i and k share more than one group have negligible probability in comparison, we get that

$$\begin{aligned}\bar{\mathbf{P}}_n(E_{ij,ik,jk}) &= m^3p_i^2p_j^2p_k^2 + O\big(m^4p_i^2p_j^2p_k^2(p_ip_j + p_ip_k + p_jp_k)\big)\\ &= \frac{\beta^3\gamma^6W_i^2W_j^2W_k^2}{n^3} + O\left(\frac{W_i^3W_j^3W_k^3}{n^4}\right).\end{aligned}$$

Part (c) is derived analogously.

As for part (d), note that the event $E_{ijk}E_{ik,jk}$ occurs when there is at least one group that is shared by all three vertices i, j, and k and a second group shared by either i and k or j and k. Denote by r the probability that individual k and at least one of the individuals i and j belong to a fixed group. Then $r = p_k(p_i + p_j - p_ip_j)$, and conditional on that there is exactly one group to which all three individuals i, j, and k belong (the probability of this is $mp_ip_jp_k(1-p_ip_jp_k)^{m-1} = O(mp_ip_jp_k)$), the probability that there is at least one other group that is shared either by i and k or by j and k is $1-(1-r)^{m-1} = O(mr)$. It follows that

$$\bar{\mathbf{P}}_n(E_{ijk}E_{ik,jk}) = O(m^2p_ip_jp_kr) = O\left(\frac{W_i^2W_j^2W_k^2}{n^{(5+\alpha)/2}}\right).$$

■

Using Lemma 3.1, it is not hard to prove Theorem 1.2.

PROOF OF THEOREM 1.2: Recall definition (2) of $\bar{c}_{i,j,k}^{(n)}$ and note that

$$\bar{\mathbf{P}}_n(E_{ij}|E_{ik}E_{jk}) = \frac{\bar{\mathbf{P}}_n(E_{ijk}\cup E_{ij,ik,jk})}{\bar{\mathbf{P}}_n(E_{ijk}\cup E_{ik,jk})}.$$

As for part (a), applying the estimates of Lemma 3.1 and merging the error terms yields

$$\begin{aligned}\bar{\mathbf{P}}_n(E_{ij}|E_{ik}E_{jk}) &\geq \frac{\bar{\mathbf{P}}_n(E_{ijk})}{\bar{\mathbf{P}}_n(E_{ijk}) + \bar{\mathbf{P}}_n(E_{ik,jk})}\\ &= \frac{1+O(W_iW_jW_kn^{-(3+\alpha)/2})}{1+W_k[\beta\gamma n^{(\alpha-1)/2} + O(W_iW_jW_kn^{-(3-\alpha)/2})]}.\end{aligned} \tag{5}$$

By Markov's inequality and the fact that W_i, W_j, and W_k are independent and have finite mean, it follows that $W_iW_jW_kn^{-(3-\alpha)/2}$ goes to zero in probability when $\alpha < 1$. Similarly, $W_iW_jW_kn^{-(3+\alpha)/2} \to 0$ in probability. Hence, the quotient in (5) converges to 1 in probability for $\alpha < 1$, as claimed.

To prove part (b), note that for $\alpha = 1$, the lower bound (5) for $\bar{c}_{i,j,k}^{(n)}$ converges in probability to $(1 + \beta\gamma W_k)^{-1}$. To obtain an upper bound, we apply Lemma 3.1 with $\alpha = 1$ to get

$$\bar{\mathbf{P}}_n(E_{ij}|E_{ik}E_{jk}) \leq \frac{\bar{\mathbf{P}}_n(E_{ijk}) + \bar{\mathbf{P}}_n(E_{ij,ik,jk})}{\bar{\mathbf{P}}_n(E_{ijk}) + \bar{\mathbf{P}}_n(E_{ik,jk}) - \bar{\mathbf{P}}_n(E_{ijk}E_{ik,jk})} \tag{6}$$
$$= \frac{1 + O(W_iW_jW_kn^{-1})}{1 + W_k[\beta\gamma + O(W_iW_jW_kn^{-1})]}.$$

Here, $W_iW_jW_kn^{-1}$ converges to zero in probability by Markov's inequality, and part (b) follows.

As for part (c), combining the bound in (6) with the estimates in Lemma 3.1 yields

$$\bar{\mathbf{P}}_n(E_{ij}|E_{ik}E_{jk}) \leq \frac{n^{(1-\alpha)/2} + O(W_iW_jW_kn^{-1})}{n^{(1-\alpha)/2} + W_k[\beta\gamma + O(W_iW_jW_kn^{-1})]}.$$

Since $W_iW_jW_kn^{-1} \to 0$ in probability, this bound converges to zero in probability for $\alpha > 1$, as desired. ■

4. CLUSTERING FOR A POWER-LAW WEIGHT DISTRIBUTION

When $\alpha = 1$, the clustering is given by $c(G) = \mathbf{E}[(1 + \beta\gamma W_k)^{-1}]$. Here, we investigate this expression in more detail for the important case that F is a power law. More precisely, we take F to be a Pareto distribution with density

$$f(x) = \frac{(\lambda - 2)^{\lambda-1}}{(\lambda - 1)^{\lambda-2}} x^{-\lambda} \quad \text{for } x \geq \frac{\lambda - 2}{\lambda - 1}.$$

When $\lambda > 2$, this distribution has mean 1, as desired. The asymptotic clustering $c(G)$ is given by the integral

$$\frac{(\lambda - 2)^{\lambda-1}}{(\lambda - 1)^{\lambda-2}} \int_{(\lambda-2)/(\lambda-1)}^{\infty} (1 + \beta\gamma x)^{-1} x^{-\lambda}\, dx.$$

Defining $u := (\lambda - 2)/(x \cdot (\lambda - 1))$, we obtain

$$c(G) = \frac{1}{\beta\gamma} \frac{(\lambda - 1)^2}{(\lambda - 2)} \int_0^1 u^{\lambda-1} \left(1 + \frac{u}{\beta\gamma}\left(\frac{\lambda - 1}{\lambda - 2}\right)\right)^{-1} du$$
$$=: \frac{1}{\beta\gamma\lambda} \frac{(\lambda - 1)^2}{(\lambda - 2)}\, {}_2F_1\left(1, \lambda; 1 + \lambda; -\frac{1}{\beta\gamma}\left(\frac{\lambda - 1}{\lambda - 2}\right)\right),$$

https://doi.org/10.1017/S0269964809990064 Published online by Cambridge University Press

where ${}_2F_1$ is the hypergeometric function. For $\beta\gamma \geq (\lambda-1)/(\lambda-2)$, a series expansion of the integrand yields

$$c(G) = \frac{1}{\beta\gamma}\frac{(\lambda-1)^2}{(\lambda-2)}\sum_{k=0}^{\infty}\left(-\frac{1}{\beta\gamma}\left(\frac{\lambda-1}{\lambda-2}\right)\right)^k\frac{1}{k+\lambda}$$
$$=: \frac{1}{\beta\gamma}\frac{(\lambda-1)^2}{(\lambda-2)}\Phi\left(-\frac{1}{\beta\gamma}\left(\frac{\lambda-1}{\lambda-2}\right), 1, \lambda\right),$$

where Φ is the Lerch transcedent. Furthermore, when λ is an integer, we get

$$c(G) = \frac{(\lambda-2)^{\lambda-1}}{(\lambda-1)^{\lambda-2}}\left[(-\beta\gamma)^{\lambda-1}\ln\left(1+\frac{\lambda-1}{\beta\gamma(\lambda-2)}\right)+\sum_{\ell=1}^{\lambda-1}\frac{(-\beta\gamma)^{\lambda-1-\ell}}{\ell}\left(\frac{\lambda-1}{\lambda-2}\right)^{\ell}\right].$$

Figures 1a and 1b show how the clustering depends on λ and $\beta\gamma$, respectively. For any $c \in (0,1)$ and a given tail exponent λ, we can find a value of $\beta\gamma$ such that the clustering is equal to c. Combining this with a condition on $\beta\gamma^2$, induced by fixing the mean degree in the graph, the parameters β and γ can be specified.

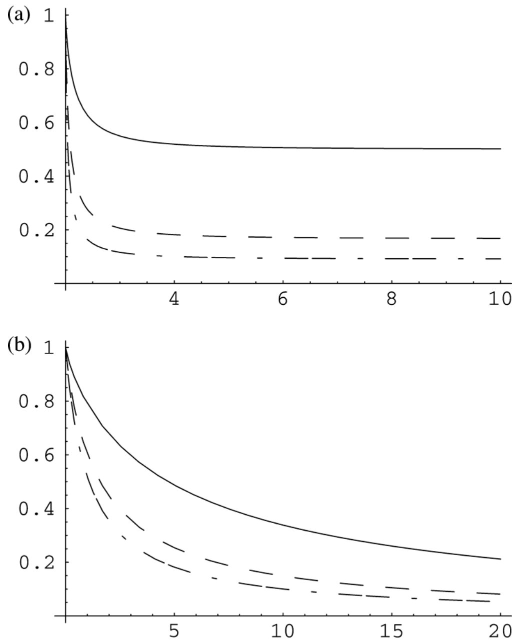

FIGURE 1. Clustering for a power-law distribution. (a) The clustering as a function of λ for different values of $\beta\gamma$: $\beta\gamma = 1$ (—), $\beta\gamma = 5$ (– – –), $\beta\gamma = 10$ (– · –); (b) the clustering as a function of $\beta\gamma$ for different values of λ : $\lambda = 2.1$ (—), $\lambda = 2.5$ (– – –), $\lambda = 4$ (– · –).

https://doi.org/10.1017/S0269964809990064 Published online by Cambridge University Press

5. FUTURE WORK

Apart from the degree distribution and the clustering, an important feature of real networks is that there is typically significant correlation for the degrees of neighboring nodes; that is, either high- (low-) degree vertices tend to be connected to other vertices with high- (low-) degree (positive correlation), or high- (low-) degree vertices tend to be connected to high- (low-) degree vertices (negative correlation). A next step is thus to quantify the degree correlations in the current model. The fact that individuals share groups should indeed induce positive degree correlation, which agrees with empirical observations from social networks; see Newman [14] and Newman and Park [15].

Additionally, other features of the model are worth investigating. For instance, many real networks are "small worlds," meaning roughly that the distances between vertices remain small also in very large networks. It would be interesting to study the relation between the distances between vertices, the degree distribution, and the clustering in the current model.

Finally, dynamic processes behave differently on clustered networks as compared to more treelike networks. Most work to date has focused on the latter class. In Britton, Deijfen, Lagerås, and Lindholm [2] however, epidemics on random intersection graphs without random weights are studied and it is investigated how the epidemic spread is affected by the clustering in the graph. It would be interesting to extend this work to incorporate weights on the vertices, allowing one to tune also the (tail of the) degree distribution and study its impact on the epidemic process.

Acknowledgments

We thank Remco van der Hofstad and Wouter Kager for valuable suggestions that have improved the manuscript.

References

1. Bollobás, B., Janson, S., & Riordan, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms* 31: 3–122.
2. Britton, T., Deijfen, M., Lagerås, A., & Lindholm, M. (2008). Epidemics on random graphs with tunable clustering. *Journal of Applied Probability* 45: 743–756.
3. Britton, T., Deijfen, M., & Martin-Löf, A. (2006). Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics* 124: 1377–1397.
4. Chung, F. & Lu. L (2002). Connected components in random graphs with given degree sequences. *Annals of Combinatorics* 6: 125–145.
5. Chung, F. & Lu. L (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences, USA* 99: 15,879–15,882.
6. Deijfen, M., van den Esker, H., van den Hofstad, R., & Hooghiemstra, G. (2009). A preferential attachment model with random initial degrees. *Arkiv för Matematik* 47: 41–72.
7. Dorogovtsev, S. & Mendes, J. (2003). *Evolution of networks, from biological nets to the Internet and WWW*. Oxford: Oxford University Press.
8. Fill, J., Scheinerman, E., & Singer-Cohen, K. (2000). Random intersection graphs when $m = \omega(n)$: An equivalence theorem relating the evolution of the $G(n, m, p)$ and $G(n, p)$ models. *Random Structures & Algorithms* 16: 156–176.

https://doi.org/10.1017/S0269964809990064 Published online by Cambridge University Press

9. Godehardt, E. & Jaworski, J. (2002). Two models of random intersection graphs for classification. In M. Schwaiger & O. Opitz (eds.), *Exploratory data analysis in empirical research*. New York: Springer, pp. 67–81.
10. Jaworski, J., Karoński, M., & Stark, D. (2006). The degree of a typical vertex in generalized random intersection graph models. *Discrete Mathematics* 306: 2152–2165.
11. Karoński, M., Scheinerman, E., & Singer-Cohen, K. (1999). On random intersection graphs: The subgraphs problem. *Combinatorics, Probability* & *Computing* 8: 131–159.
12. Molloy, M., & Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures* & *Algorithms* 6: 161–179.
13. Molloy, M., & Reed, B. (1998). The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability* & *Computing* 7: 295–305.
14. Newman, M. E. J. (2003). Properties of highly clustered networks. *Physical Review E* 68: 026121.
15. Newman, M. E. J. & Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E* 68: 036122.
16. Newman, M. E. J., Strogatz, S. H., & Watts, D. J. (2002). Random graphs with arbitrary degree distributions and their applications. *Physical Review E* 64: 026118.
17. Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435: 814–818.
18. Singer, K. (1995). Random intersection graphs, Ph.D. thesis, Johns Hopkins University. Baltimore, MD.
19. Stark, D. (2004). The vertex degree distribution of random intersection graphs. *Random Structures & Algorithms* 24: 249–258.
20. Yao, X., Zhang, C., Chen, J., & Li, Y. (2005). *On the scale-free intersection graphs*. Lecture Notes in Computer Science No. 3481. Berlin: Springer, pp. 1217–1224.