Tilburg University

Random Intersection Graphs with Tunable Degree Distribution and Clustering

Deijfen, M.; Kets, W.

Publication date:
2007

Citation for published version (APA):
Deijfen, M., & Kets, W. (2007). *Random Intersection Graphs with Tunable Degree Distribution and Clustering*.
(CentER Discussion Paper; Vol. 2007-008). Microeconomics.

RANDOM INTERSECTION GRAPHS WITH TUNABLE DEGREE DISTRIBUTION AND CLUSTERING

By Maria Deijfen, Willemien Kets

Random intersection graphs with tunable degree distribution and clustering

Maria Deijfen[*]   Willemien Kets[†]

January 2007

Abstract

A random intersection graph is constructed by independently assigning 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 paper 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 determined 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 so 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.

Keywords: Random intersection graphs, degree distribution, power law distribution, clustering, social networks.

JEL classification: C65, Z13.

1   Introduction

During the last decade there has been a large interest in the study of large complex networks; see e.g. Dorogovtsev and Mendes (2003) and Newman et al. (2006) and the references therein. Due to the rapid increase in computer power, 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

---

[*]Address: Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: mia@math.su.se. Tel: +46-8-164591.

[†]Address: Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: w.kets@uvt.nl. Tel: +31-13-4662478.

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 etc – which gives rise to high clustering in the social network, since members of a given group typically know each other; see Palla et al. (2005) 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 there is an edge between two vertices given that they have a common adjacent vertex. Other definitions occur in the literature, see e.g. Newman (2003), 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 e.g. Molloy and Reed (1995, 1998) – typically have zero clustering in the limit. In this paper, we propose a model, based on the so-called random intersection graph, where 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 in Singer (1995) and Karoński et al. (1999), and has been further studied and generalized in Fill et al. (2000), Godehardt and Jaworski (2002), Stark (2004) and Jaworksi et al. (2006). Newman (2003) and Newman and Park (2003) discuss 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

2

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. (1999). 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

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

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 (2004; Theorem 2) shows 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 0, a Poisson distribution or a compound 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 current 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 in Godehardt and Jaworski (2002). The degree of a given vertex in such a graph is analyzed in Jaworski et al. (2006), where conditions on the group distribution are specified under which the degree is asymptotically Poisson distributed.

In the current paper, 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. The model is inspired by a generalization of the Erdős-Rényi random graph studied in Britton et al. (2006), in which the vertices are equipped with random weights to obtain more arbitrary degree distributions. See also Yao et al. (2005) for a related model (with deterministic weight) aimed specifically at producing power law degree distributions. 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 before, take $\mathcal{V} = \{1, \ldots, n\}$ to be a set of n vertices and $\mathcal{A}$ a set of m elements. Also, let $\{W_i\}$ be an i.i.d. sequence of positive random variables with distribution F, where F is assumed to have 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 before 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 where the success of a vertex (measured by its degree) depends on some specific feature of the vertex; see e.g. Palla et al. (2005) 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 which, as we will see, makes it possible to exactly characterize the degree distribution and the clustering in the resulting graph.

1.2 Results

Our main results concern the degree distribution and the clustering in the graph $G(n, m, F)$. As for the degree distribution, first note that, conditional on W_i and W_j, the probability that there is an edge between two individuals $i, j \in \mathcal{V}$ in $G(n, m, F)$ is

$$1 - (1 - p_i p_j)^m = \beta \gamma^2 W_i W_j n^{-1} + O(W_i^2 W_j^2 n^{-2}).$$

By summing the expectations of the edge indicators over j, it is easy to see that, at least when the weights have finite second moment, the expected degree of individual i given its weight W_i is $\beta \gamma^2 W_i$ (recall that weight distributions with finite mean are assumed to be scaled so that the mean equals 1). The following theorem, which is a generalization of Theorem 2 in Stark (2004), gives a full characterization of the degree distribution for different values of α.

Theorem 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).*

(a) *If $\alpha < 1$ and F has finite mean, then, as $n \to \infty$, the degree D_i converges in distribution to a point mass at 0.*

(b) *If $\alpha = 1$ and F has finite moment of order $1 + \varepsilon$ for some $\varepsilon > 0$, 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 $\alpha > 1$ and F has finite moment of order $1 + \varepsilon$ for some $\varepsilon > 0$, then D_i is asymptotically Poisson($\beta \gamma^2 W_i$) distributed.*

To get some intuition for 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 number converges to 0 in probability, so that the degree distribution converges to a point mass at 0, as stated in (a). 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 (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 (c).

4

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 $G(n, m, F)$ – and let $\bar{\mathbf{P}}_n$ be the probability measure of $B(n, m, F)$ conditional on the weights $\{W_1, \ldots, W_n\}$. Given three distrinct vertices $i, j, k \in \mathcal{V}$, the clustering $c_n(G)$ is defined as

$$c_n(G) = \mathbf{E}\left[\bar{\mathbf{P}}_n\left(E_{ij} | E_{ik}, E_{jk}\right)\right] \tag{2}$$

where the expectation is taken over the weights, and we write $c(G) = \lim_{n \to \infty} c_n(G)$. Clearly the vertices are indistinguishable, so $c_n(G)$ does not depend on the particular choice of the vertices i, j and k. However, the clustering does depend on the parameter α, as demonstrated by the following theorem.

Theorem 1.2 *Consider the 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 that*

(a) $c(G) = 1$ *for* $\alpha < 1$;

(b) $c(G) = \mathbf{E}\left[(1 + \beta\gamma W_k)^{-1}\right]$ *for* $\alpha = 1$;

(c) $c(G) = 0$ *for* $\alpha > 1$.

To get some intuition for Theorem 1.2, consider three given individuals $i, j, k \in \mathcal{V}$ and assume that i and k share a group and that j and k share a group. Then, the probability that i and j also have a common group depends on the number of groups that the common neighbor k belongs to. 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 0 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 $\alpha = 1$. For a given weight distribution F (with finite mean), the value of the clustering can then be varied between 0 and 1 by adjusting the parameters β and γ. Furthermore, when $\alpha = 1$, the degree distribution for a given 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 so as well. Since the mean of F is set to 1, the expected asymptotic degree is $\beta\gamma^2$. 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 paper 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 that individual 1 belongs to. If individual 1 is not a member of any group, then clearly $D = 0$, and hence (a) follows if we show that $\mathbf{P}(N = 0) \to 1$ as $n \to \infty$ for $\alpha < 1$. Conditional on W_1, the variable N is binomially distributed with parameters m and p_1 and thus

$$\bar{\mathbf{P}}_n(N = 0) = (1 - p_1)^m = 1 - O(mp_1).$$

By the choice of m and p_1, we have that $mp_1 \leq \beta\gamma W_1 n^{(\alpha-1)/2}$, and, by Markov's inequality,

$$\mathbf{P}\left(W_1 n^{(\alpha-1)/2} > \delta\right) \leq \frac{\mathbf{E}[W_1]}{\delta n^{(1-\alpha)/2}} \quad \text{for any } \delta > 0.$$

If $\alpha < 1$ and W_1 has finite mean, then the right-hand-side above converges to 0. It follows that $\bar{\mathbf{P}}_n(N = 0) \to 1$ in probability. Bounded convergence then gives that $\mathbf{P}(N = 0) = \mathbf{E}[\bar{\mathbf{P}}_n(N = 0)] \to 1$, as desired.

To prove (b) and (c), we condition on the weight W_1, which is thus assumed to be fixed in what follows, and show that the generating function of D converges to the generating function of the claimed limiting distribution. To do this, let X_i ($i = 2, \ldots, n$) denote the number of common groups of individual 1 and individual i. 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 X_i's are independent and binomially distributed with parameters N and p_i. 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

$$
\begin{aligned}
\mathbf{E}[t^D] &= \mathbf{E}\left[\prod_{i=2}^{n} \mathbf{E}\left[t^{\mathbf{1}\{X_i \geq 1\}} \big| \{W_i\}, N\right]\right] \\
&= \mathbf{E}\left[\prod_{i=2}^{n} \left(1 + (t-1)\bar{\bar{\mathbf{P}}}_n(X_i \geq 1)\right)\right]
\end{aligned}
$$

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

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

we get that

$$\prod_{i=2}^{n}\Big(1+(t-1)\bar{\bar{\mathbf{P}}}_n(X_i\geq 1)\Big) \;=\; e^{(t-1)N\sum p_i+O\big(N^2\sum p_i^2\big)} \tag{3}$$

$$=\; e^{(t-1)N\sum p_i}+R_n,$$

where

$$R_n=e^{(t-1)N\sum p_i}\Big(e^{O(N^2\sum p_i^2)}-1\Big).$$

Since the product in (3) is the conditional expectation of t^D with $t\in[0,1]$, it takes values between 0 and 1, and, since $e^{(t-1)N\sum p_i}\in(0,1]$, it follows that $R_n\in[-1,1]$. We will show that

(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 $\alpha>1$;

(iii) $R_n\to 0$ in probability for $\alpha\geq 1$.

The limits in (i) and (ii) are the generating functions for the desired compound Poisson and Poisson distribution in part (b) and (c) of the theorem, respectively. Moreover, by bounded convergence, (iii) implies that $\mathbf{E}[R_n]\to 0$. Hence the theorem is proved once (i)-(iii) are established.

Starting with (i) and (ii), we first 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

$$\mathbf{E}\left[e^{(t-1)N\sum p_i}\right]=\mathbf{E}\left[\Big(1+p_1\Big(e^{(t-1)\sum p_i}-1\Big)\Big)^m\right]. \tag{4}$$

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

$$\Big(1+p_1\Big(e^{(t-1)\sum p_i}-1\Big)\Big)^{\lfloor\beta n\rfloor}\to e^{\beta\gamma W_1(e^{\gamma(t-1)}-1)}\quad\text{as }n\to\infty,$$

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

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

$$\Big(1+p_1\Big(e^{(t-1)\sum p_i}-1\Big)\Big)^m=\left(1+\frac{\gamma W_1(t-1)\sum\tilde{p}_i}{n^\alpha}\cdot\frac{e^{(t-1)n^{(1-\alpha)/2}\sum\tilde{p}_i}-1}{(t-1)n^{(1-\alpha)/2}\sum\tilde{p}_i}\right)^{\lfloor\beta n^\alpha\rfloor}.$$

Clearly $\sum\tilde{p}_i\to\gamma$, and, since $(e^x-1)/x\to 1$ as $x\to 0$, it follows that the right hand side above converges to $e^{\beta\gamma^2 W_1(t-1)}$ as $n\to\infty$. By (4) and bounded convergence, this proves (ii).

It remains to show (iii). First recall the definition of R_n and note that, since $t \in [0,1]$, to establish that $R_n \to 0$ it is sufficient to show that $N^2 \sum p_i^2 \to 0$. To do this, define $\xi = \varepsilon(1+\varepsilon)^{-1}$ and write

$$N^2 \sum p_i^2 = \left(Nn^{-(\alpha-1+\xi)/2}\right)^2 n^{\alpha-1+\xi} \sum p_i^2.$$

Since $\mathbf{E}[N] = \mathbf{E}[mp_1] \leq \beta\gamma n^{(\alpha-1)/2}$, by Markov's inequality, we have for any $\delta > 0$ that

$$\mathbf{P}\left(Nn^{-(\alpha-1+\xi)/2} > \delta\right) \leq \frac{\beta\gamma}{\delta n^{\xi/2}}$$

and it follows that $Nn^{-(\alpha-1+\xi)/2} \to 0$ in probability as $n \to \infty$. To see that $n^{\alpha-1+\xi} \sum p_i^2 \to 0$ as well, note that, since $p_i^2 \leq \gamma^2 W_i^2 n^{-(1+\alpha)}$, we have that

$$n^{\alpha-1+\xi} \sum p_i^2 \leq \gamma^2 \frac{\sum W_i^2}{n^{2-\xi}}.$$

Hence it suffices to show that $\sum W_i^2/n^{2-\xi} \to 0$. Obviously, $W_i \leq \max_{k \leq n}\{W_k\}$, so that

$$\frac{\sum W_i^2}{n^{2-\xi}} \leq \left(\frac{\sum W_i}{n}\right)\left(\frac{\max\{W_i\}}{n^{1-\xi}}\right).$$

By the law of large numbers, we have that $\sum W_i/n \to 1$ and, recalling the definition of ξ, for any $\delta > 0$, we have that

$$\begin{aligned}
\mathbf{P}\left(\max\{W_i\} > \delta n^{1-\xi}\right) &\leq n\mathbf{P}(W_i > \delta n^{1-\xi}) \\
&= n\mathbf{P}(W_i^{1+\epsilon} > \delta^{1+\varepsilon}n).
\end{aligned}$$

Here, $n\mathbf{P}(W_i^{1+\epsilon} > \delta^{1+\varepsilon}n) \to 0$, since W_i has finite $1 + \varepsilon$ moment. It follows that $\sum W_i^2/n^{2-\xi} \to 0$ in probability, and the proof of (iii) is complete. $\qquad\square$

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 groups. It will be convenient to generalize 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 that*

(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 (a), the probability that three given individuals i, j and k do not share a 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

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

To prove (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 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 $1 - (1 - p_j p_k)^{m-2} = m p_j p_k + O(m^2 p_j^2 p_k^2)$. 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

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

Part (c) is derived analogously.

As for (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 $r = p_k(p_i + p_j - p_i p_j)$, and, conditional on that there is exactly one group to which all three individuals i, j and k belong (the probability of this is $m p_i p_j p_k (1 - p_i p_j p_k)^{m-1} = O(m p_i p_j p_k)$), 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

$$\begin{aligned}
\bar{\mathbf{P}}_n(E_{ijk} E_{ik,jk}) &= O(m^2 p_i p_j p_k r) \\
&= O\left(\frac{W_i^2 W_j^2 W_k^2}{n^{(5+\alpha)/2}}\right).
\end{aligned}$$

$\square$

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

9

Proof of Theorem 1.2. Recall the definition (2) of the clustering $c_n(G)$ 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 (a), applying the estimates of Lemma 3.1 and merging the error terms yields

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

By Markov's inequality, when $\alpha < 1$, we have for any $\delta > 0$ that

$$\mathbf{P}(W_i W_j W_k n^{-(3-\alpha)/2} > \delta) \leq \frac{\mathbf{E}[W_i W_j W_k]}{\delta n} \to 0,$$

since W_i, W_j and W_k are independent and have finite mean. This means that $W_i W_j W_k n^{-(3-\alpha)/2}$ goes to 0 in probability, and, similarly, $W_i W_j W_k n^{-(3+\alpha)/2} \to 0$ in probability. Furthermore, it is easy to see that $n^{(\alpha-1)/2} \to 0$ for $\alpha < 1$ as $n \to \infty$. Hence the fraction in (5) converges to 1 in probability for $\alpha < 1$, and it follows from bounded convergence that $c_n(G) \to 1$.

To prove part (b), note that for $\alpha = 1$, it follows from (5) and the above reasoning that $\liminf c_n(G) \geq \mathbf{E}\left[(1 + \beta\gamma W_k)^{-1}\right]$, and so it suffices to show that $\limsup c_n(G) \leq \mathbf{E}\left[(1 + \beta\gamma W_k)^{-1}\right]$. Applying Lemma 3.1 with $\alpha = 1$ and simplifying, we get that

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

Since the weights are independent with finite mean, Markov's inequality can be used to conclude that $W_i W_j W_k n^{-1}$ converges to 0 in probability, and the desired conclusion follows from bounded convergence.

As for (c), we apply Lemma 3.1 again to get the bound

$$\begin{aligned}
\bar{\mathbf{P}}_n(E_{ij}|E_{ik}E_{jk}) &\leq \frac{\bar{\mathbf{P}}_n(E_{ijk}) + \bar{\mathbf{P}}_n(E_{ij,ik,jk})}{\bar{\mathbf{P}}_n(E_{ik,jk})} \\
&= \frac{n^{(1-\alpha)/2} + O(W_i W_j W_k n^{-1})}{W_k[\beta\gamma + O(W_i W_j W_k n^{-1})]}.
\end{aligned} \tag{6}$$

Obviously $n^{(1-\alpha)/2} \to 0$ if $\alpha > 1$, and, by Markov's inequality, we have that $W_i W_j W_k n^{-1} \to 0$ in probability. Hence the fraction in (5) converges to 0 in probability and bounded convergence gives that $c_n(G) \to 0$. $\square$
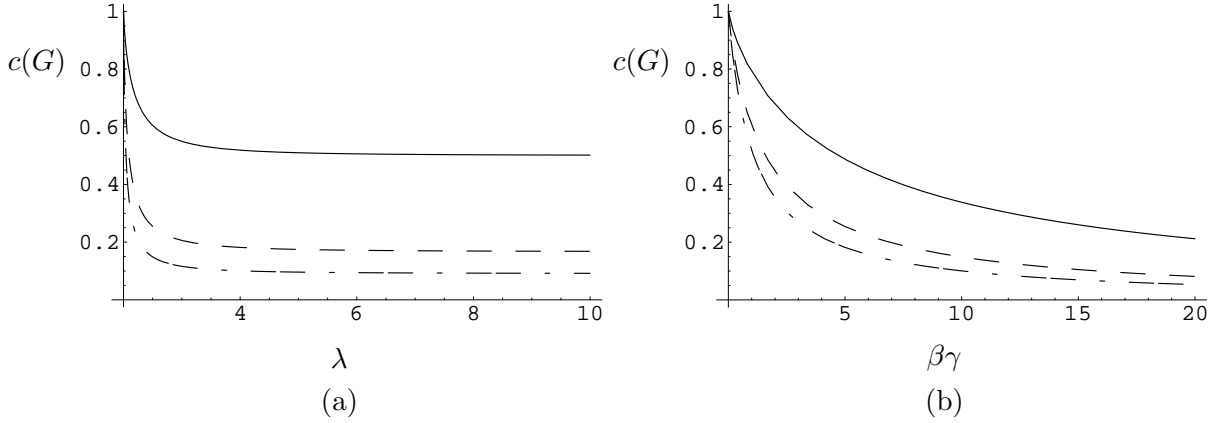
Figure 1: (a) The clustering as a function of λ for different values of $\beta\gamma$: $\beta\gamma = 1$ (—), $\beta\gamma = 5$ ($---$), $\beta\gamma = 10$ ($-\cdot-$). (b) The clustering as a function of $\beta\gamma$ for different values of λ: $\lambda = 2.1$ (—), $\lambda = 2.5$ ($---$), $\lambda = 4$ ($-\cdot-$).

4 Clustering for a power law weight distribution

When $\alpha = 1$, the clustering is given by $\mathbf{E}[(1+\beta\gamma W_k)^{-1}]$. In general, it is not possible to give an explicit expression for this expectation, but numerical solutions are easily obtained. We now investigate the clustering in more detail in the important case when 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 clustering $c(G)$ is given by the integral

$$\frac{(\lambda - 2)^{\lambda-1}}{(\lambda - 1)^{\lambda-2}} \int_{\frac{\lambda-2}{\lambda-1}}^{\infty} (1 + \beta\gamma x)^{-1} x^{-\lambda} dx. \tag{7}$$

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

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

where ${}_2F_1$ is the hypergeometric function (Abramomowitz and Stegun, 1964). For $\beta\gamma \geq (\lambda - 1)/(\lambda - 2)$, a series expansion of the integrand yields that

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

11

where Φ is the Lerch transcedent. Furthermore, when λ is an integer, the expression for the clustering becomes

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

Since $_2F_1(a,b;c;z)$ is increasing in z, the clustering falls monotonically in $\beta\gamma$. Also, the clustering decreases when λ increases, since more mass is then put on large values of x where the function $(1+\beta\gamma x)^{-1}$ is small. Figure 1 (a) and (b) 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.

5   Future work

There are a number of possible directions for future research. 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 low (high) degree vertices (negative correlation). A next step is thus to quantify the degree correlations in the current model. The fact that individuals share groups most likely induces positive degree correlation, which agrees with empirical observations from social networks; see Newman (2003) and Newman and Park (2003).

Also 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. On the one hand, when the clustering is large, there are many "redundant" edges, which indicates that, for a given edge density, one would expect the average distance between individuals to be larger if the network is highly clustered. On the other hand, when the clustering is large, individuals tend to be organized in groups, and once a path reaches a group, all members of the group are only one step away. This acts to reduce the distances in clustered networks.

Finally, dynamic processes can be expected to behave differently on clustered networks as compared to more tree-like networks. Most work to date, however, has focused on the latter class. The current model makes it possible to vary the clustering and to choose the degree distribution, and it would be interesting to study the behavior of dynamic processes as a function of these properties.

References

Britton, T., Deijfen, M. and Martin-Löf, A. (2006): Generating simple random graphs with prescribed degree distribution, *J. Stat. Phys.* **124**, 1377-1397.

Dorogovtsev, S. and Mendes, J. (2003): *Evolution of Networks, from Biological Nets to the Internet and WWW*, Oxford University Press.

Fill, J., Scheinerman, E. and 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.

Godehardt, E. and Jaworski, J. (2002): Two models of random intersection graphs for classification, in *Exploratory data analysis in empirical research*, eds Schwaiger M. and Opitz, O., Springer, 67-81.

Jaworski, J., Karoński, M. and Stark, D. (2006): The degree of a typical vertex in generalized random intersection graph models, *Discrete Mathematics* **306**, 2152-2165.

Karoński, M., Scheinerman, E. and Singer-Cohen, K. (1999): On random intersection graphs: the subgraphs problem, *Combinatorics Probability & Computing* **8**, 131-159.

Molloy, M. and Reed, B. (1995): A critical point for random graphs with a given degree sequence, *Random Structures and Algorithms* **6**, 161179.

Molloy, M. and Reed, B. (1998): The size of the giant component of a random graph with a given degree sequence, *Combinatorics, Probability and Computing* **7**, 295-305.

Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2002): Random graphs with arbitrary degree distributions and their applications, *Physical Review E* **64**, 026118.

Newman, M. E. J. (2003): Properties of highly clustered networks, *Physical Review E* **68**, 026121.

Newman, M. E. J. and Park J. (2003): Why social networks are different from other types of networks, *Physical Review E* **68**, 036122.

Palla, G., and Derényi, I., Farkas, I. and Vicsek, T. (2005): Uncovering the overlapping community structure of complex networks in nature and society, *Nature* **435**, 814 - 818.

Singer, K. (1995): *Random intersection graphs*, PhD thesis, Johns Hopkins University.

Stark, D. (2004): The vertex degree distribution of random intersection graphs, *Random Structures & Algorithms* **24**, 249-258.

Yao, X., Zhang, C., Chen, J. and Li, Y. (2005): On the scale-free intersection graphs, *Lectures note in computer science* **3481**, 1217-1224.