

# A spatial model for social networks

Ling Heng Wong\*, Phillipa Pattison, and Garry Robins

*Department of Psychology,*

*The University of Melbourne.*

May 19, 2005

## Abstract

We study spatial embeddings of random graphs in which nodes are randomly distributed in geographical space. We let the edge probability between any two nodes to be dependent on the spatial distance between them and demonstrate that this model captures many generic properties of social networks, including the “small-world” properties, skewed degree distribution, and most distinctively the existence of community structures.

**MSC classifications:** 91D30, 90B10, 82B99.

**Keywords:** Social networks, small world, spatial model, community structure, homophily.

## 1 Introduction

Complex social networks arise in a wide range of contexts, for example as corporate partnership networks [22], scientist collaboration networks [30], company director networks [41], film actors networks [3], sexual contact networks [26], etc. Indeed, a lot of attention has been given by both physical and social scientists in recent years to model these networks so as to gain better understandings of their general structures as well as their various functions like information flow [18], locating individuals [1], disease spread [26], etc. For a review of recent efforts, see for example [40], [2] and [31]. While there is an apparent increase in the number of network models in the literature, not all of these models have taken full advantage of the sociological and psychological insights on how social networks may be formed.

### 1.1 Spatial characteristics of social ties

The principle of *homophily*, or in essence “birds of a feather flock together,” has been firmly established by many empirical studies [24]. While we clearly tend to befriend those who are like us, there are many situations where having a lot of friends like us

---

\*Corresponding author. Current address: Department of Psychology, School of Behavioural Science, The University of Melbourne, VIC 3010, Australia. e-mail: lingw@unimelb.edu.au. Tel: +61 3 8344 6362. Fax: +61 3 9347 6618.

is simply because we are *stuck* with people who are like us in the first place. For example if you are a millionaire and all your friends are millionaires, it might simply be because you were born into an elite family and live in an elite area so you only know millionaires in your life, even though you do not actively choose to befriend millionaires over non-millionaires. Therefore, it is useful to divide homophily into two main types: *baseline* homophily and *inbreeding* homophily [24]. Baseline homophily is attributed to the fact that we have a *limited potential tie pool* due to factors like demography and foci of activities [13]. Inbreeding homophily is conceptualised as any other kind of homophily measured over that potential tie pool — this may include homophily regarding gender, religion, social class, education, and other intra-personal or behavioural characteristics. While many network models have taken inbreeding homophily into account [51, 42, 33, 32, 43, 46], they have generally assumed that there are no baseline homophily effects, i.e. the potential tie pool for all actors equals the *entire* population. However, this is obviously not very realistic and baseline homophily effects can potentially have profound consequences on the structure of social networks.

A basic source of baseline homophily is the *geographical space*. As a matter of simple opportunity and/or the need to minimise efforts to form and maintain a social tie [54], we can expect that we tend to form ties with those who are geographically close to us. Thus, intuitively, this creates a very strong constraint on our potential tie pool. In fact, there is ample empirical evidence that demonstrates this claim. The earliest studies of which we are aware of date back to Festinger *et al.* [14] and Caplow and Forman [9] both on student housing communities. The results showed that in these rather homogeneous communities, spatial arrangement of student rooms/units was an important factor in predicting whether two dwellers have at least weak ties. Many other network studies also reached similar results, for example see [4, 5]. More recently, Wellman [48] and Mok *et al.* [27] re-analysed Wellman’s earlier dataset on Torontorian personal communities [49, 50] and noted that most personal friendships were indeed “local,” contrary to the beliefs that recent technological advances have freed us from spatial constraints. For instance, in [49] it was found that on average 42% of “frequent contact” ties live within a mere 1 mile radius of a typical person, while the rest of his/her ties could be directed to anywhere in the rest of the world.

## 1.2 General features of social networks

Before embarking on specifying the model, we shall review some of the general features of social networks. Not many current models simultaneously displays all of these. We suggest that, by including baseline spatial homophily into our network model, one can reproduce all the following features, at least in broad terms:

1. **Low tie density.** The number of possible ties in a network is theoretically quadratic to the number of actors, but most networks realise only a tiny fraction of these ties. The cognitive ability of human places an upper bound on the number of ties one may maintain [11]. On the other hand, other factors corresponding to baseline homophily can also play a role [13];
2. **Short average geodesic distances.** Geodesic distance between two actors is defined to be the length of the shortest connection between them. In large social networks, it is believed that the typical geodesic distance between any two actors remains small. This property was demonstrated empirically by Stanley Milgram

in his classical experiment in the 1960s [25], contributing to the popular saying that no one on this earth is separated from you by more than six “handshakes”;

3. **High level of clustering.** Clustering is defined to be the average probability that two friends of an actor are themselves friends. Equivalently, it is a measure of how having a mutual friend will heighten the conditional probability that the two friends of an actor will be friends themselves. In their well-known article [47], Watts and Strogatz demonstrated the importance of *short-cuts* in social networks that simultaneously display high clustering and short average geodesic distances. Such an idea of short-cuts dates back to Granovetter’s arguments on the strength of weak ties [16];
4. **Positively skewed actor degree distribution.** The degree of an actor is the number of social ties he/she has. In many social networks, a majority of actors have relatively small degrees, while a small number of actors may have very large degrees. This feature is displayed in a wide range of social networks. While it is still debated whether generic social networks have power-law, exponential, or other degree distributions, or indeed whether there is any *generic* distribution at all [17], there is no doubt that degree distributions are in general positively skewed;
5. **Existence of communities.** In many cases, clustering does not occur evenly over the entire network. We can often observe subgroups of actors who are highly connected within themselves but loosely connected to other subgroups which are themselves highly inter-connected. We call these highly-connected subgroups *communities* [28]. A long tradition in social network analysis has developed a range of algorithms to identify these cohesive subsets of nodes [52].

An example of a social network that displays all of the above properties is a well-known alliance network of 16 tribes in the Eastern Central Highlands of New Guinea [39]<sup>1</sup>. The network is depicted in Fig. 1 where nodes correspond to tribes and ties correspond to alliances between the relevant tribes. First of all, the density of the network is fairly low ( $\rho = 0.24$ ) given the small size of the network. The degree distribution is positively skewed (skewness statistic = 0.99). The network is a “small world” in which it has low median geodesic distance (2) and high level of clustering coefficient (clustering coefficient = 0.63). Most importantly, two distinct communities can be easily observed in Fig. 1: one is disjoint from the rest and is fully connected (i.e. Nodes 1, 2, 15, and 16) and the other is highly connected within itself (i.e. Nodes 3, 6, 7, 8, 11, and 12).

### 1.3 Random graph models

We here represent social networks by non-directed graphs. A non-directed graph is defined to be a pair  $(\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$  is the *node set* denoting the individual actors in the network, and  $\mathcal{E} = \{e_1, e_2, \dots, e_M\}$ , with each edge  $e_i$  being an unordered pair of nodes  $e_i = (v_r, v_s)$  ( $r \neq s$  and  $v_{r,s} \in \mathcal{V}$ ), is the *edge set* denoting the social ties among the actors. A compact way to represent a graph is through its

---

<sup>1</sup>The full data set is included as a sample data set in the standard social network analysis program UCINET, which is available at <http://www.analytictech.com/ucinet.htm>.

adjacency matrix  $\mathbf{X} = [x_{ij}]$ ,  $i, j \in \{1, 2, \dots, N\}$  such that  $x_{ij} = 1$  iff  $(v_i, v_j) \in \mathcal{E}$ , otherwise  $x_{ij} = 0$ . In general, the size of the set  $\mathcal{V}$  (or equivalently the dimensions of  $\mathbf{X}$ ) is fixed but whether an edge  $(v_i, v_j)^t \in \mathcal{E}$  (or equivalently  $x_{ij} = 1$  in  $\mathbf{X}$ ) is determined by a random process. Such a random process is defined so as to reflect the underlying social dynamics.

The simplest model for social networks is the Erdős-Rényi or *Bernoulli* random graph model, initiated independently by Paul Erdős and Alfred Rényi [12] and Anatol Rapaport [38], where the random process is a *Bernoulli* trial, i.e.

$$x_{ij} = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1 - p \end{cases}, \quad (1)$$

where the constant  $p \in [0, 1]$  is called the *edge probability*. In other words  $x_{ij}$  are identically and independently distributed (i.i.d.) Bernoulli random variables. Due to its simplicity, it is amenable to rigorous treatments. It is fairly straightforward to show that the average geodesic distance between any two nodes is  $\langle r_{ij} \rangle \sim \ln N$ , a feature that, as discussed above, resembles the property in some real networks. However the level of clustering in this model can be shown to vanish as  $N \rightarrow \infty$  and it is one of the major short-comings of this simple model in modelling social networks. Further, Erdős and Rényi showed that there is a critical edge probability  $p_c \approx N^{-1}$  at which there always exists a connected component containing a significant proportion of nodes in the network (almost surely) [7]. Such component is known as the giant component. For an extensive review of these results refer to Bollobás [7] and Janson *et al.* [20].

## 1.4 The overview

The main advantage of the Erdős-Rényi model is its simplicity. Although it does not predict some of the generic features outlined above, it serves as a good foundation to build more realistic models. In this paper, we study a generalisation of Erdős-Rényi random graph model for social networks which incorporates a simple *baseline spatial homophily* effect in the formation of individual network ties. We note that this class of models is for a *single snapshot* of a network, thus *temporal network dynamics* are not taken into account. In Section 2, we shall specify the model and examine some of its basic properties. In Section 3, we shall outline our simulation methods and discuss the main results. In Section 4, we shall demonstrate the application of our model in a particular social network. In Section 5, we shall discuss the implications of the results and describe ongoing research on this and more generalised models.

## 2 Spatial random graph model

First of all, we embedded the nodes of graphs in the Euclidean space  $\mathbb{R}^2$  with a *distance* function  $d$  defined to map any unordered pair of nodes to a real number, i.e.  $d : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ . Further,  $d$  satisfies the standard triangle inequalities and the positivity condition. Every node  $v_i$  is assigned a coordinate  $(x_i, y_i)^t$  according to some spatial distribution and we define  $\chi = \{(x_1, y_1)^t, (x_2, y_2)^t, \dots\}$  to be the location vector that specifies the spatial locations of *all* nodes in a network. Further we shall use the shorthand  $d_{ij} = d((x_i, y_i)^t, (x_j, y_j)^t)$  to denote the distance between nodes  $v_i$  and  $v_j$ . Next, we assume that the spatial locations of the nodes are randomly scattered in

space (and therefore all locations are mutually independent as well). In other words, we assumed that the points are distributed in space according to a homogeneous Poisson point process (see for example [10, 21]). Recall that a Poisson point process with rate  $\rho < \infty$  in the  $d$ -dimensional Euclidean space  $\mathbb{R}^d$  is a process such that:

- for all disjoint subsets  $A_1, A_2, \dots, A_k \subset \mathbb{R}^d$ , the random variable denoting the number of point in the each subset,  $N(A_1), N(A_2), \dots, N(A_k)$ , are independently distributed;
- $N(A)$  has Poisson distribution; and
- $E[N(A)] = \rho|A|$  for all  $A \subset \mathbb{R}^d$ .

To model the baseline homophily effect, we let the edge probability between two nodes to be dependent on the spatial distance between them. That is, given a  $\chi$ ,  $P(x_{ij} = 1|\chi) = f(d_{ij})$ , where  $f : \mathbb{R} \rightarrow [0, 1]$ . Motivated by the discussion in the last section, we shall consider the case where  $f$  is a simple step function, i.e.

$$P(x_{ij} = 1|\chi) = \begin{cases} p + p_b, & \text{if } d_{ij} \leq H, \\ p - \Delta, & \text{if } d_{ij} > H, \end{cases} \quad (2)$$

where  $p$  is the average density of the network,  $H$  is the *neighbourhood radius*,  $p_b$  is the *proximity bias* which specifies the sensitivity to geographical space by the actors on establishing social links. Thus  $p_b$  probabilistically controls the locations of potential tie pool. Of course we have assumed that such a spatial sensitivity is the same for all actors in the network. Further,  $\Delta = \Delta(p_b, H|\chi)$  is some correction term. The correction term  $\Delta$  is introduced to maintain the expected average density to be a constant  $p$ , given  $\chi$ , for all feasible values of  $H$  and  $p_b$ , i.e.

$$E \left[ \frac{1}{N-1} \sum_{i < j} x_{ij} \middle| \chi \right] = p, \quad (3)$$

and there is no other substantive purpose. Without maintaining constant density, it will be difficult to isolate the effects of  $p_b$  and  $H$  on the graph's structural properties because of the confounding effects from the varying expected number of edges in the model.

To calculate  $\Delta$ , we first determine the number of all possible edges shorter than the neighbourhood radius  $H$  in the network embedded in  $\chi$ , call this number  $S_{\leq H}(\chi)$ ; when  $N$  is sufficiently large (ignoring boundary effects),  $S_{\leq H}(\chi) \approx N\pi\rho^2/2$ . The number of possible edges longer than the neighbourhood radius is therefore  $S_{> H}(\chi) = \binom{N}{2} - S_{\leq H}(\chi)$ . By definition the expected density within neighbourhoods is  $p + p_b$ , so the expected number of all realised edges within all neighbourhoods is  $(p + p_b)S_{\leq H}$ . It follows that, in order to maintain density to equal  $p$ , the expected number of realised edges outside the neighbourhood ought to be  $\binom{N}{2}p - (p + p_b)S_{\leq H}$ . As a result,

$$p - \Delta = \frac{1}{S_{> H}} \left[ \binom{N}{2}p - (p + p_b)S_{\leq H} \right] = p - \frac{S_{\leq H}}{\binom{N}{2} - S_{\leq H}} p_b. \quad (4)$$

<sup>2</sup> Since the probability  $P(x_{ij} = 1|\chi)$  is bounded by 0 and 1,  $0 \leq p + p_b \leq 1$  and  $0 \leq p - \Delta \leq 1$ . These conditions defines upper and lower bounds for  $p_b$  given a  $p$  and

<sup>2</sup>If the expected number of nodes within the neighbourhood is the same across all nodes (e.g. when we have a homogeneous point process with periodic boundary condition), then let the expected

$H$ . We note that at the a particular set of values where  $p = \left[ \binom{N}{2} - S_{\leq}(\chi) \right] / \binom{N}{2}$  and  $p_b = 1 - p$ , i.e. when there is probability 1 that all pair of nodes less than distance  $H$  apart will be joint and probability 0 otherwise, this model becomes the so-called *random geometric graph model* [37]<sup>3</sup>. A typical instance of spatial random graph can be found in Fig. 2.

It is convenient to express our model in the exponential form which fits into the exponential random graph network modelling framework. More details of the framework can be found in Wasserman and Pattison [15, 53]. The modelling framework has recently received renewed attention in the physics community, see [8] and [34]. Let  $x$  be an instance of random graph with spatial locations of nodes specified by  $\chi$ . Then the general form of the probability function for obtaining  $x$  is given by

$$P(X = x|\chi) = \frac{1}{Z(\chi)} \exp[-\mathcal{H}(x|\chi)], \quad (6)$$

where  $\mathcal{H}(x|\chi)$  is the Hamiltonian of graph  $x$  given the locations of nodes specified by  $\chi$  and

$$Z(\chi) = \sum_x \exp[-\mathcal{H}(x|\chi)] \quad (7)$$

is the partition function of the model. The Hamiltonian can take any (sensible) form to reflect the dependency of edges. The simplest choice is  $-\mathcal{H}(x) = \theta_0 L(x|\chi)$ , where  $L(x|\chi) = L(x)$  is the number of edges in the graph  $x$  independent of  $\chi$  and  $\theta_0$  is its associated parameter. By tuning  $\theta_0$ , we can change the expected density of a typical graph in the model. This gives us the classical Erdős-Rényi random graph model [12] as introduced in Section 1. To define  $\mathcal{H}(x|\chi)$  for our simple spatial model, we first define  $L_{\leq}(x|\chi) = \sum_{i < j, d_{ij} \leq H} x_{ij}$  and  $L_{>}(x|\chi) = \sum_{i < j, d_{ij} > H} x_{ij}$  be the number of edges shorter and longer than  $H$  respectively, then the Hamiltonian for our model can be written as

$$-\mathcal{H}(x|\chi) = \theta_{\leq} L_{\leq}(x|\chi) + \theta_{>} L_{>}(x|\chi), \quad (8)$$

where  $\theta_{\leq}$  and  $\theta_{>}$  are parameters whose values can be calculated directly from Eq. 2, i.e.

$$\theta_{\leq} = \text{logit}(p + p_b), \quad \text{and} \quad \theta_{>} = \text{logit}(p - \Delta), \quad (9)$$

where  $\text{logit } q = \log[q/(1 - q)]$ . Equivalently, we can also write the Hamiltonian in another form:  $-\mathcal{H}(x|\chi) = \theta L(x|\chi) + \theta' L_{\leq}(x|\chi)$ , where  $L(x|\chi) = \sum_{i < j} x_{ij}$  ( $= L_{\leq}(x|\chi) +$  number of nodes within a node's neighbourhood (not including itself) be  $E[s_{\leq H}]$ , we can then repeat the above and arrive at a simpler form for the correction term,

$$\Delta = \frac{E[s_{\leq H}]}{(N - 1) - E[s_{\leq H}]} p_b. \quad (5)$$

<sup>3</sup>We also later become aware of a generalisation of the random geometric graph model called the *random connection model* where the edge probability is taken to be a general decreasing function of spatial distance,  $g$  [36]. In [36], some rigorous results regarding the giant cluster on the model has been obtained, however, the main variable of their model is the Poisson rate  $\rho$  in space while keeping  $g$  general. However here our focus on the function  $g$ .

$L_>(x|\chi)$ ),  $\theta = \theta_>$  and  $\theta' = \theta_< - \theta_>$ . In the following however, we shall only use the former form of the Hamiltonian.

The simplicity of the Hamiltonian allows us to write down the partition function in closed form,

$$\begin{aligned} Z_N(\chi) = \sum_x \exp[-\mathcal{H}(x)] &= \sum_x \exp(\theta_< L_<(x) + \theta_> L_>(x)) \\ &= \sum_x \exp \left( \theta_< \sum_{i < j, d_{ij} \leq H} x_{ij} + \theta_> \sum_{i < j, d_{ij} > H} x_{ij} \right) \\ &= (1 + e^{\theta_<})^{S_<(x)} (1 + e^{\theta_>})^{\binom{n}{2} - S_<(x)}. \end{aligned}$$

Using this explicit form, one can double-check the constant density in our model for all  $p_b$  and  $H$  given  $\chi$ :

$$\begin{aligned} \binom{N}{2}^{-1} \langle L(x|\chi) \rangle &= \binom{N}{2}^{-1} \frac{1}{Z(\chi)} \sum_x [L_<(x|\chi) + L_>(x|\chi)] \exp[-\mathcal{H}(x|\chi)] \\ &= \binom{N}{2}^{-1} \frac{1}{Z(\chi)} \left( \frac{\partial Z(\chi)}{\partial \theta_<} + \frac{\partial Z(\chi)}{\partial \theta_>} \right) \\ &= \binom{N}{2}^{-1} [S_<(\chi)(p + p_b) + S_>(\chi)(p - \Delta)] = p, \end{aligned}$$

which is what we expected. In theory, we can calculate any statistical average statistical quantities  $\langle Q(x|\chi) \rangle$ , for example the average clustering coefficient etc, from  $Z(\chi)$  by adding an auxiliary term in the Hamiltonian  $\Delta\mathcal{H}(\chi) = yQ(x|\chi)$ . Then,

$$\langle Q(x|\chi) \rangle = \frac{1}{Z(\chi)} \sum_x Q \exp[-\mathcal{H}(x|\chi) - yQ(x|\chi)] \quad (10)$$

$$= \frac{1}{Z(\chi)} \frac{\partial Z(\chi)}{\partial y} \Big|_{y=0} \quad (11)$$

However, as is the case in many other statistical mechanics models, equations like Eq. 11 are very difficult to evaluate exactly as a general approach is not yet available. In this study, we resort to using numerical simulations to explore the properties our model.

### 3 Simulation results and discussions

The Poisson rate  $\rho$  and the neighbourhood radius are relative to each other, so we shall always fix  $\rho = 1$  and vary  $H$  only. On the other hand, the choice of the value of  $H$  is not crucial as long as  $H$  is sufficiently large. When  $H$  is too small, the vast majority of ties connected to a node are inevitably from the outside of the neighbourhood, therefore, the model behaves as the simple Erdős-Rényi random graph model. Here we fix  $H = 3/2$ . The main program of simulations below is to vary the proximity bias  $p_b$  and investigate the effects on the overall structures of the graphs. We used a Markov Chain Monte Carlo method outlined in Snijders [44] for all our simulations. Each individual data point below is a result of a simulation run (250,000 Markov iterations)

of random graphs with 100 nodes on a fixed  $\chi$ . The burn-in phase for each run is about 30,000 iterations and statistics from this phase were removed before further analysis. The estimates of the statistics are then calculated as the simple averages in the post-burn-in phase. And we collect the estimates for six realisations of  $\chi$ .

### 3.1 Number of short and long edges

In Fig. 3, we plot the average number of edges shorter than ( $\langle L_{\leq}(x) \rangle$ ) and longer than ( $\langle L_{>}(x) \rangle$ ) the neighbourhood radius; we called them *short* and *long* edges respectively. When  $p_b = 0$ , the opportunity tie pool of each node equals the entire population (except itself). In this case, since there are many more potential long edges than short edges in the graph, the graphs are on average dominated by long edges. As  $p_b$  increase, the potential tie pool of each node concentrate more and more on the population geographically close, so short edges dominate. The two scatter plots in Fig. 3 clearly display linear opposing trends. The linearity is simply a result of our model definition, refer to Eq. 2. Also as  $p_b$  increases, there is growing variance of statistics for fixed  $p_b$ . It is because as the difference between short edge probability and long edge probability grows, the graphs are more and more dependent on the configuration of the *specific instance* of the point process. However, as  $p_b$  approaches  $1 - p$ , the variability decreases again. It is because the large  $p_b$  values place significant constraints on the feasible instances of the point process and thus only a small number of instances  $\chi$  are feasible for large  $p_b$  (see Eq. 4).

### 3.2 Small-world properties

Let us now investigate the effect of  $p_b$  on the global structure of the graphs. First let us define the *geodesic distance*, or graph distance, between two nodes. Note that this distance is independent of the spatial distance between the nodes involved. Given a graph  $X = \{x_{ij}\}$ , let the geodesic distance between  $v_i$  and  $v_j$ ,  $l_{ij}(X)$ , be length of the shortest *self-avoiding path* connecting  $v_i$  and  $v_j$ . If  $v_i$  and  $v_j$  are disconnected, we assign  $l_{ij} = \infty$ . Now, given a  $\chi$ , define

$$w(X) = \left[ \binom{N}{2}^{-1} \sum_{i < j} l_{ij}^{-1} \right]^{-1}, \quad (12)$$

to be the (harmonic) mean of  $l_{ij}(X)$  over all possible pairs on nodes in  $X$ .  $w(X)$  is a simple measure of *connectivity* of  $G$ . Refer to the upper part of Fig. 4 for a plot of  $\langle w(X) \rangle$  against  $p_b$ . From the plot,  $\langle w(X) \rangle$  remains small for a large range of  $p_b$ . Then there appears to be a critical  $p_b$ , as in the Watts-Strogatz model [47], that  $w$  dramatically increases. This indicates a switch from the *simple random graph* regime, where “short-cuts” between neighbourhood are abundant, to the *random geometric graph* regime, where most edges are within each actor’s neighbourhood.

Further, we study the level of clustering in the model. Let  $t_1(X)$  be the number of independent 3-cycles, or triangles, in graph  $X$ , i.e.  $t_1(X) = \sum_{i < j < k} x_{ij}x_{jk}x_{ik}$ . Also let  $s_2(X)$  be the number of 2-stars in graph  $X$ , i.e.  $s_2(X) = \sum_{i < j < k} x_{ij}x_{ik}$ . Then we define the global clustering coefficient to be

$$C(X) = \frac{3t_1(X)}{s_2(X)}. \quad (13)$$

$C(X)$  measures the overall level of clustering in a network or in other words how much on average an actor's friend's friends are also the actor's friends. By definition,  $0 \leq C(X) \leq 1$  for all  $X$ . Refer to Fig. 4 for a plot of the average  $\langle C \rangle$  over all  $X$  in the ensemble as we changes  $p_b$ . This model also display similar behaviours as in the Watts-Strogatz model, and there is a steady increase in its value as  $p_b$  increases. The source of this clustering is entirely *spatial*, i.e. triangles are likely to be formed simply because of the fact that the involved nodes are closed to each other spatially.

Overall, there is a range of  $p_b$  where the model display *simultaneously* relatively low average path length *and* significant level of clustering — a signature of a “small-world” model.

### 3.3 Community structures

Although the clustering coefficient defined above measures the *overall* level of clustering in a graph, there is the question of whether clustering is distributed evenly over the entire graph on average, in other words, do the triangles in the graph tend to clump together or not? The “clumpiness” of triangles can be measured by the number of *higher-order triangles*. A 2-triangle is defined to be the combination of two triangles sharing a common edge (which is called the base edge). In general, a  $k$ -triangle is defined to be the combination of  $k$  triangles all sharing a common base edge and let  $t_k(X)$  its number in  $X$ . Let  $g_{ij}(G)$  be the number of two-paths connecting  $v_i$  and  $v_j$ , then a useful and convenient way to combine all  $t_k(G)$  into a single measure is as follows:

$$T_\lambda(G) = 3t_1(G) - \frac{t_2(G)}{\lambda} + \frac{t_3(G)}{\lambda^2} - \dots + (-1)^{n-3} \frac{t_{n-2}(G)}{\lambda^{n-3}} \quad (14)$$

$$= \lambda \sum_{i < j} x_{ij} \left[ 1 - \left( 1 - \frac{1}{\lambda} \right)^{g_{ij}(G)} \right], \quad (15)$$

where  $\lambda$  is an arbitrary constant. For a detailed discussion of the motivation for this definition, see [45]. A plot of  $\langle T_\lambda(X) \rangle$  for  $\lambda = 2$  against  $p_b$  can be found in Figure 5. There we can see a steady increase in  $T_2(X)$  as  $p_b$  increases. This suggests that graphs have higher tendencies on average to form clumps of triangles — we call these clumps *communities* — for large  $p_b$ . Inspection of instances of graphs when  $p_b$  is large suggests that the delineation of communities are determined by large gaps in the spatial distribution of the nodes. This happens in spite of the fact that the nodes are distributed uniformly randomly in space. This phenomena can be compared with the observation that communities are likely to be separated by large streets, railroad tracks, etc [23].

Now that we have shown that a typical graph in the model is likely to have strong community structures when  $p_b$  is reasonably large. One possible implication of having such kind of structures is that the whole graph can be composed of *disjoint* communities. To detect whether it is the case, we consider the average size of the largest connected component over the ensemble. We define the *component size* containing node  $v_i$ ,  $\gamma_i(X)$ , to be the number of nodes with finite path length from  $v_i$  (including itself), i.e.  $\gamma_i(G) = \sum_{i < j} \delta(l_{ij} < \infty)$ , where  $\delta$  is the normal delta function. Let

$$\Gamma(X) = \max_{v_i} \gamma_{v_i}(X) \quad (16)$$

to be the number of nodes of  $X$  in *largest* component over all  $v_i$ . We have plotted the  $\langle \Gamma(X) \rangle$  against  $p_b$  in Fig. 6. When  $p_b$  is small, a vast majority of nodes is contained in the largest component. As  $p_b$  increases, there is a higher chance that the largest component no longer contains most nodes, leaving a significant proportion of nodes in smaller isolated components. The instance of network in Fig. 2 is an example of such a situation.

### 3.4 Actor degree distributions

Fig. 7 shows the average degree distribution of graphs for various values of  $p_b$ . They are all positively skewed and as  $p_b$  increases the corresponding degree distribution has an increasingly fatter tail. As  $p_b$  becomes very large, the distribution becomes a bimodal distribution. This is due to the boundary effect where nodes near the spatial boundary of the graph are disadvantaged by having less possible neighbours within their neighbourhood radius. This is confirmed by the studying the correlation between the spatial distance of nodes from the centre of a typical graph and the degree of the nodes. A scatterplot for a typical instance can be found in Fig. 8 and in this case it is found that there is a significant correlation between the two ( $p < 0.01$ ).

## 4 An office communication network

We can use our current model to gain insights into the underlying dynamics of a social network. Our example here is a communication network of 33 individuals observed over two days in a single-floor office as part of a large organisation in Australia [6]. We define a communication tie to exist between two individuals if and only if each of the two parties has sought information from each other more than three times over the two-day period. This definition is to avoid brief idiosyncratic encounters that create much noise on top of the regular communication pattern. The network is depicted in Fig. 9. The location of the nodes in the figure are the actual locations of the individual's cubicle or room up to a linear scaling. In particular, the dimension of the office space is scaled so that all relevant locations fits into a  $1 \times 1$  unit square.

The network displays the generic features introduced in Section 1.2. First of all, the density is low ( $\rho = 0.131$ ). The degree distribution is (slightly) positively skewed (skewness statistics = 0.303). The clustering coefficient is high (0.389) while the median geodesic distance is very small (2). Also, in a spatially rearranged layout of the network in Fig. 11, one can easily identify the two distinct communities.

Indeed, one can reasonably conjecture that spatial process is important in this communication network. To verify this, we use our current model and estimate the model parameters, i.e.  $p$ ,  $H$ , and  $p_b$  (see Eq. 2). The empirical edge probability  $p(d)$  is plotted in Fig. 10. It displays the expected big drop at the small values of  $d$ . Based on this, we can fit  $p(d)$  with a simple step function with some neighbourhood size  $H$  such that the sum of squared errors is minimal. In this case,  $H$  is found to be 0.2. As the overall density  $p$  is 0.131, the proximity bias  $p_b$  is estimated to be 0.259, and the correction term  $\Delta$  is estimated to be 0.047<sup>4</sup>. The large bias  $p_b$  indicates that there is a strong spatial component in the underlying social process.

---

<sup>4</sup>Note that one can alternatively derive the value of  $\Delta$  using Eq. 4.

## 5 Conclusion

In this study, we have performed Monte Carlo simulations on a class of spatial random graph models. It has been found that the properties of these models differs significantly from the simple Erdős-Rényi random graph model. In particular, for a range of  $p_b$  values, the model displays simultaneously many general features of social networks as discussed in Section 1 while the Erdős-Rényi random graph model fails in many aspects.

We note that the properties of the model are similar to those of the well-known Watts-Strogatz “small world” model [47] but that the current model has the advantage of specifying an explicit probability distribution over the collection of all graphs with a given number of nodes. It is also important to note that the clustering properties of the current model arise entirely from the Poisson process describing node locations and hence from a form of spatial baseline homophily. It is clearly an empirical question whether such models for social networks provide an adequate descriptive account or whether it is necessary to also incorporate inbreeding homophily effects (as in exponential random graph selection models [42]) or endogenous network processes characteristic of the Watts-Strogatz model [47] and more general exponential random graph model specifications [8, 45, 53, 35, 19]. This is a question that has received little attention in the network literature despite its fundamental importance to our understanding of network evolution.

In order to identify models that provide a good match to empirical data, it will be useful to construct a nested family of exponential random graph models that can be used to evaluate the empirical evidence for spatial and other forms of baseline homophily, inbreeding homophily and endogenous clustering effects. Indeed, reference to empirical data immediately raises the possibility of at least two alternative conceptualisations of a spatial model: the first, a geographical space, in which geographical coordinates are associated with each node; and the second, a more abstract “social” space, in which spatial proximities reflect baseline homophily across a broad range of individual attributes. In each case, spatial locations may be observed or unobserved. For the current model and the case of observed locations, it would be necessary to estimate the model parameters  $p$  and  $H$  from the combination of location and network data; in the case of unobserved locations, it would be desirable to use a version of Hoff, Handcock and Raftery’s [19] more general exponential random graph model specifications to estimate model parameters from network data alone. More generally, it would be desirable to construct within the exponential random graph model family, models that also include inbreeding homophily and endogenous clustering effects and associated estimation methods.

The model described in this paper takes a useful first step towards the construction of such a family of models. The results presented here suggest that the fit of models to empirical data will need careful quantitative evaluation (e.g. as in [53, 45]) because of the likely capacity of many models within the family to exhibit in broad terms the commonly observed features of empirical social networks laid out in Section 1.2.

## 6 Acknowledgments

This study is financially supported by a Australian Research Council Discovery Grant and the CSIRO–University of Melbourne Collaboration Supportive Scheme. We thank Sean Bergin and Paul Rogers for sharing with us the communication network data. We express our gratitudes to Paul Walker, Stanley Wasserman and Tom Snijders for their helpful discussions, and the anonymous referee for pointing out the relevant literature and various other suggestions. We would also like to thank Peng Wang for programming the simulation program and the High Performance Computer Facility at the University of Melbourne for allowing us to use the IBM Alfred Cluster for simulations.

## References

- [1] Adamic L., and Adar E., *How to search a social network*, Preprint (2004).
- [2] Albert R., Barabasi A.-L., *Statistical mechanics of complex networks*, Rev. Mod. Phys. **74**, 47-97 (2002).
- [3] Amaral L.A.N., Scala A., Barthélémy M., and Stanley H.E., *Classes of small-world networks*, Proc. Natl. Acad. Sci. USA **97**, 11149-11152 (2000).
- [4] Athanasiou R. and Yoshioka G.A., *The spatial characteristics of friendship formation*, Environment and behavior **5**, 43-66 (1973).
- [5] Barrett A.L. and Campbell K.E., *Neighbor networks of black and white Americans*, in Networks in the global village: life in contemporary communities, Wellman B. (ed.), Westview Press (1999).
- [6] Bergin, S. and Rogers, P., *Private communications*, Commonwealth of Australia (2004).
- [7] Bollobás B., *Random Graphs*, Academic Press, New York (1985).
- [8] Burda Z., Jurkiewicz J., and Krzywicki A., *Network transitivity and matrix models*, Phys. Rev. E **69**, 026106 (2004).
- [9] Caplow T. and Forman R., *Neighbourhood interaction in a homogeneous community*, Am. Socio. Rev. **15**, 357-366 (1950).
- [10] Cox D.R. and Islam V., *Point Processes*, Monographs on Applied Probability and Statistics, Chapman and Hall (1980).
- [11] Dunbara R.I.M., *Neocortex size as a constraint on group size in primates*, J. of Hum. Evolut. **20**, 469-493 (1992).
- [12] Erdős P., and Rényi A., *On random graphs. I.*, Publicationes Mathematicae (Debrecen) **6**, 290-297 (1959).
- [13] Feld S.L., *The focused organization of social ties*, Am. J. Soc. **86**, 1015-1035 (1981).
- [14] Festinger L., Schachter S., and Back K., *Social processes in informal groups*, Stanford Univ. Press (1950).
- [15] Frank O. and Strauss D., *Markov Graphs*, J. Am. Stat. Assoc. **81**, 832-842 (1986).
- [16] Granovetter M.S., *Strength of weak ties*, Am. J. Socio. **78**, 1360-1380 (1973).
- [17] Handcock, M., and Jones, J., *An assessment of preferential attachment as a mechanism for human sexual network formation*, Proceedings of the Royal Society B **270**, 1123-1128 (2003).
- [18] Huberman, B.A., and Adamic L.A., *Information Dynamics in the Networked World*, Lect. Notes Phys. **650**, Springer-Verlag Berlin Heidelberg, 371-398 (2004).

- [19] Hoff P.D., and Raftery A.E., and Handcock M.S., *Latent Space Approaches to Social Network Analysis* J. Am. Stat. Assoc., **97**, 1090-1098 (2002).
- [20] Janson S., Luczak T., and Rucinski A. *Random Graphs*, John Wiley, New York (1999).
- [21] Kingman J.F.C., *Poisson Processes*, Oxford Studies in Probability 3, Clarendon Press, Oxford (1993).
- [22] Lazega E., *The collegial phenomenon. The social mechanisms of co-operation among peers in a corporate law partnership*, Oxford University Press, Oxford (2001).
- [23] Logan J.R., *Growth, Politics, and the Stratification of places*, Am. J. Socio. **84**, 404-416 (1978).
- [24] McPherson M., Smith-Lovin L., and Cook J.M., *Birds of a Feather: Homophily in Social Networks*, Annu. Rev. Sociol **27**, 415-444 (2001).
- [25] Milgram S., *The small world problem*, Psychology Today **2**, 60-67 (1967).
- [26] Morris M., *Sexual networks and HIV*, AIDS **11**, S209-S216 (1997).
- [27] Mok D., Wellman B., and Basu R., *Does distance matter for relationships?*, Presentation at SUNBELT International Social Network Conference, Portoroz, Slovenia (2004).
- [28] Newman M.E.J., *Fast algorithm for detecting community structure in networks*, Phys. Rev. E **69**, 066133 (2004).
- [29] Newman M.E.J., Strogatz S.H., and Watts, D.J., *Random graphs with arbitrary degree distributions and their applications*, Phys. Rev. E **64** 026118 (2001).
- [30] Newman M.E.J., *The structure of scientific collaboration networks*, Proc. Natl. Acad. Sci. USA **98**, 404-409 (2001).
- [31] Newman M.E.J., *The structure and function of complex networks*, SIAM Review **45**, 167-256 (2003).
- [32] Newman M.E.J., *Mixing patterns in networks*, Phys. Rev. E **67** 026126 (2003).
- [33] Newman M.E.J. and Girvan M., *Mixing patterns and community structure in networks*, in Statistical Mechanics of Complex Networks, R. Pastor-Satorras, J. Rubi, and A. Diaz-Guilera (eds.), Springer, Berlin (2003).
- [34] Park J. and Newman M.E.J., *The statistical mechanics of networks*, Phys. Rev. E **70**, 066117 (2004).
- [35] Pattison P.E., and Robins G. L., *Neighbourhood-based models for social networks*, Sociological Methodology **32**, 301-337 (2002).
- [36] Penrose M.D., *On a continuum Percolation Model*, Adv. Appl. Prob. **23**, 536-556 (1991).

- [37] Penrose M., *Random Geometric Graphs*, Oxford University Press, Oxford (2003).
- [38] Rapoport A., *Spread of information through a population with socio-structural bias: I. assumption of transitivity*, Bull. Math. Biophys. **15**, 523-533 (1953).
- [39] Read K., *Cultures of the central highlands, New Guinea*, Southwestern J. Anthro, **10**, 1-43 (1954).
- [40] Robins G.L., and Pattison P., *Interdependencies and social processes: Generalized dependence structures*, In Carrington, Scott, and Wasserman (Eds) Models and Methods in Social Network Analysis. Cambridge University Press (in press).
- [41] Robins, G.L., and Alexander, M., *Small worlds among interlocking directors: Network structure and distance in bipartite graphs*, Journal of Computational and Mathematical Organization Theory **10**, 69-94 (2004).
- [42] Robins G., Elliott P., Pattison, P., *Network models for social selection processes*, Social Networks **23**, 1-30 (2001).
- [43] Robins G. and Johnston M., *Joint social selection and social influence models for networks: The interplay of ties and attributes*, Presentation at SUNBELT International Social Network Conference, Portoroz, Slovenia (2004).
- [44] Snijders T.A.B., *Markov Chain Monte Carlo Estimation of Exponential Random Graph Models*, J. Soc. Struc. **3**, No. 2 (electronic) (2002).
- [45] Snijders T.A.B., Pattison P., Robins G., and Handcock, M., *New Specification for exponential random graph models*, Socio. Meth., *in press* (2005).
- [46] van Duijn M.A.J., Snijders T.A.B., and Zijlstra B.H.,  *$p_2$ : a random effects model with covariates for directed graphs*, Statistica Neerlandica **58** 234-254 (2004).
- [47] Watts D.J. and Strogatz S.H., *Collective dynamics of 'small-world' networks*, Nature **393**, 440-442 (1998).
- [48] Wellman B., *Are personal communities local? A Dumptarian reconsideration*, Social Network **18**, 347-354 (1996).
- [49] Wellman B., Carrington P., Hall A., *Networks as personal communities*, in Social structures: A network approach, Wellman B. and Berkowitz S.D. (eds.), Cambridge University Press, Cambridge (1988).
- [50] Wellman B., Wortley S., *Different strokes from different folks: Community ties and social support*, American Journal of Sociology **96**, 558-588 (1990).
- [51] Wasserman S., Galaskiewicz J., *Some generalizations of  $p_1$  — external constraints, interactions and non-binary relations*, Social Networks **6**, 177-192 (1984).
- [52] Wasserman S. and Faust K., *Social network analysis: Methods and applications*, Cambridge University Press, Cambridge (1994).
- [53] Wasserman S. and Pattison P., *Logit models and logistic regressions for social networks: I. An introduction to Markov Graphs and  $p^*$* , Psychometrika **61**, 401-425 (1996).

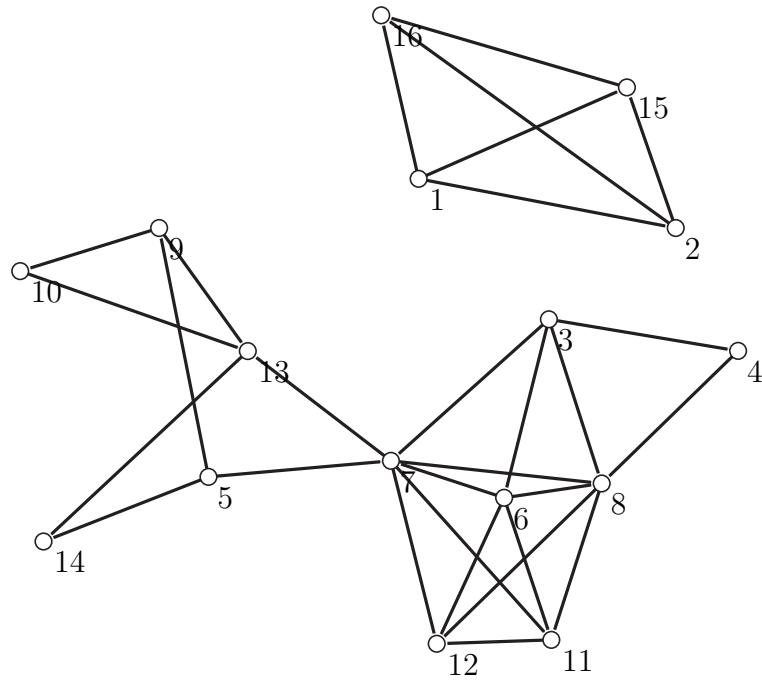


Figure 1: The alliance network in Eastern Central Highlands of New Guinea. Note that the numbers next to the nodes are for illustrative purpose only and the spatial arrangements of the nodes do not reflect the actual tribe locations.

[54] Zipf G.K., *Human behaviour and the principle of least effort*, Addison-Wesley (1949).

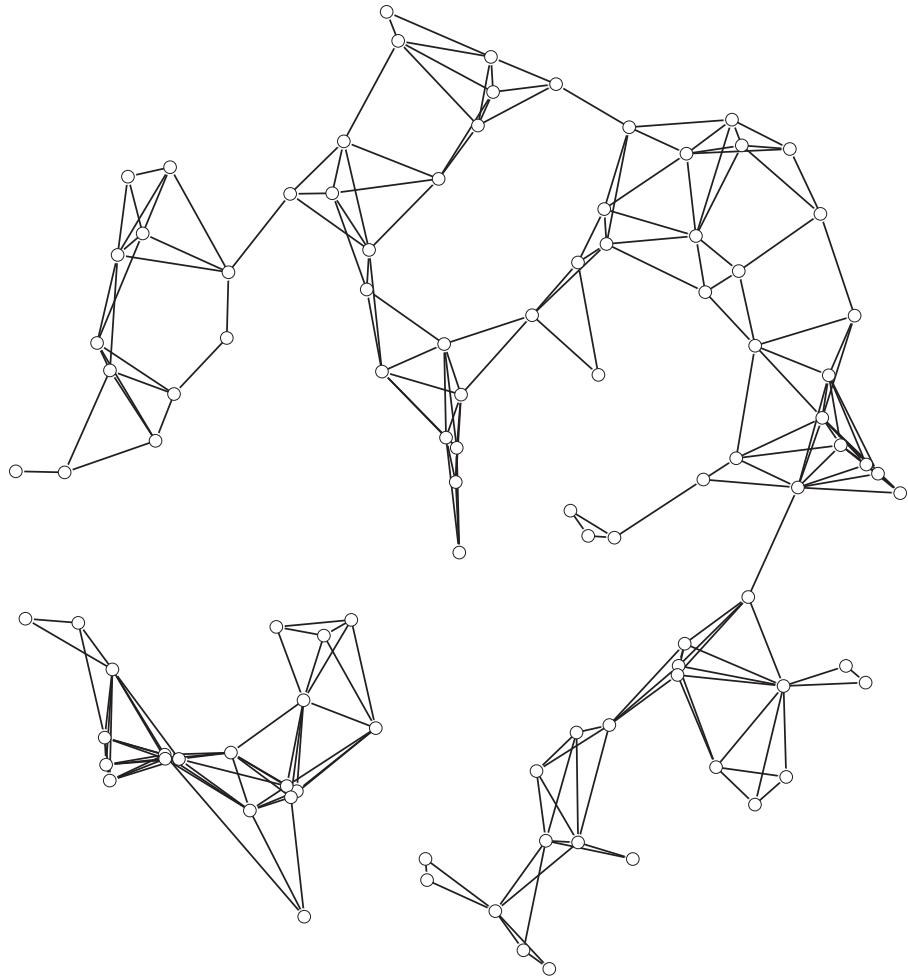


Figure 2: A 100-node graph ( $p = 0.05$ ,  $p_b = 0.95$ , and  $H = 3/2$ ) taken from the Markov Chain Monte Carlo simulation in Section 3. Points are distributed according to a Poisson point process.

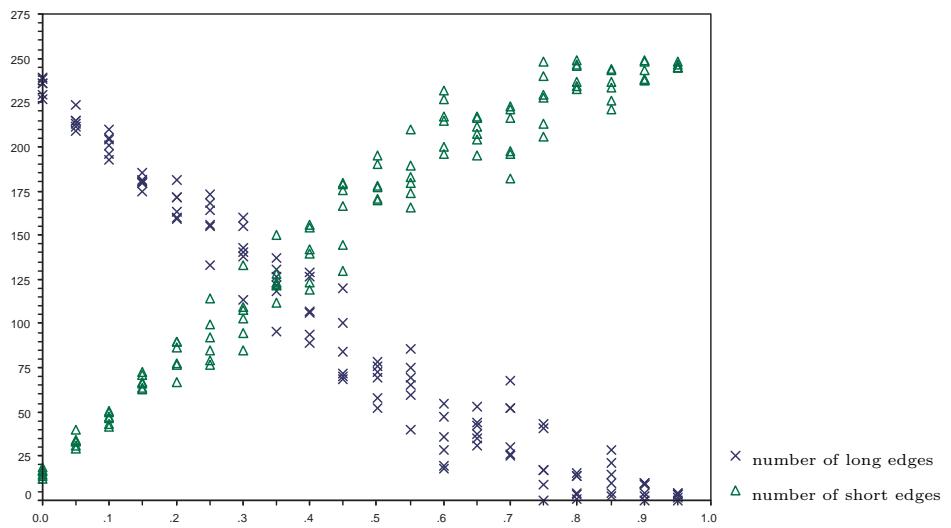


Figure 3: A plot of average number of short edges  $\langle L_{\leq}(x) \rangle$  and average number of long edges  $\langle L_{>}(x) \rangle$  versus  $p_b$ .

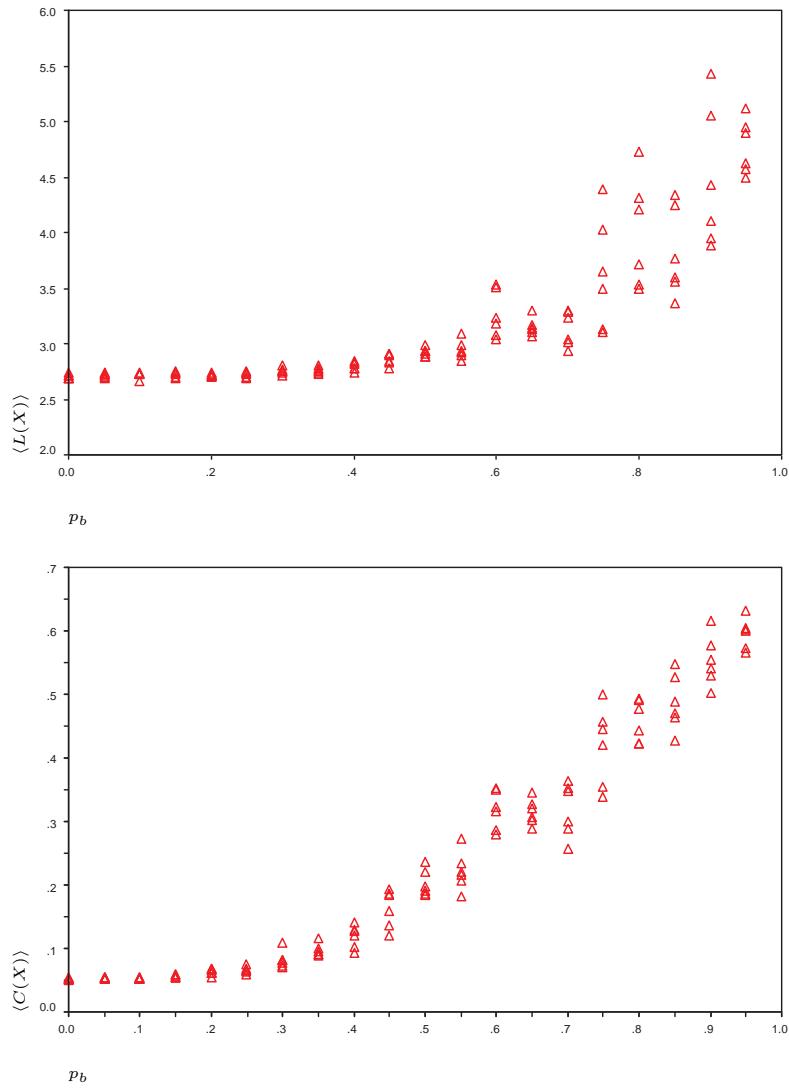


Figure 4: A plot of average geodesic distance  $\langle L(X) \rangle$  (upper graph) and average clustering coefficient  $\langle C(X) \rangle$  (lower graph) against  $p_b$ .

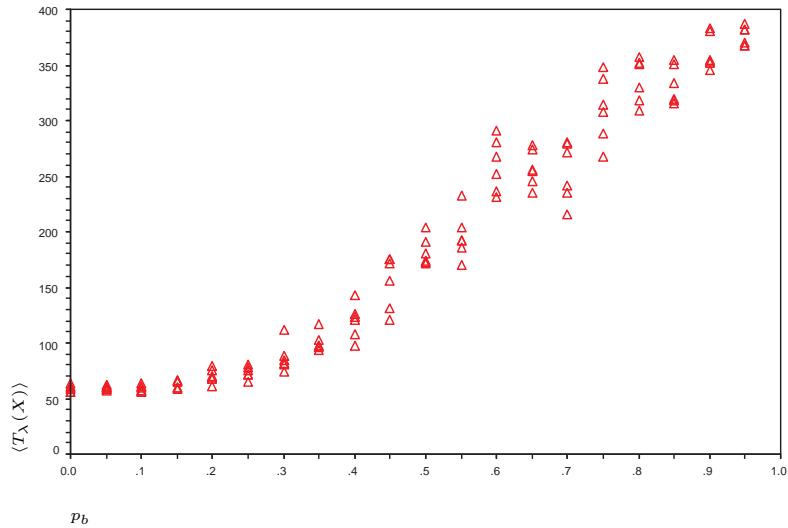


Figure 5: A plot of average  $k$ -triangle statistics  $\langle T_\lambda(X) \rangle$  [45] for  $\lambda = 2$  against  $p_b$ .

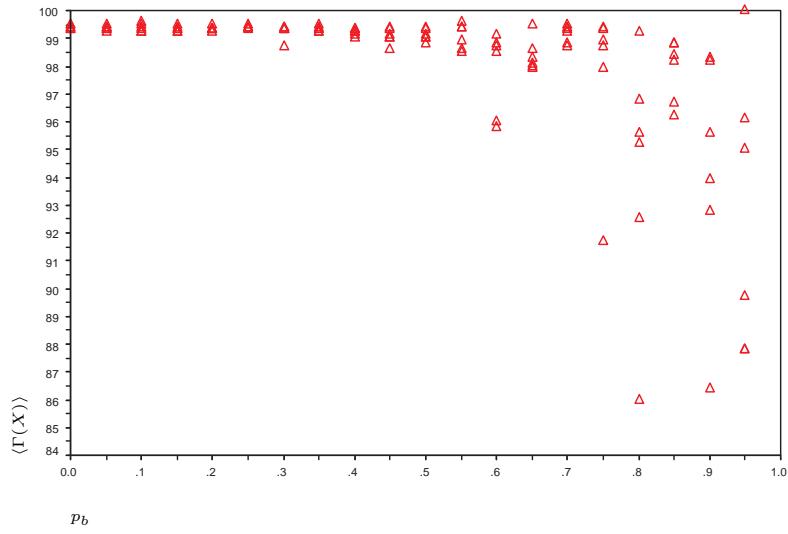


Figure 6: A plot of average largest component size  $\langle \Gamma(X) \rangle$  against  $p_b$ .

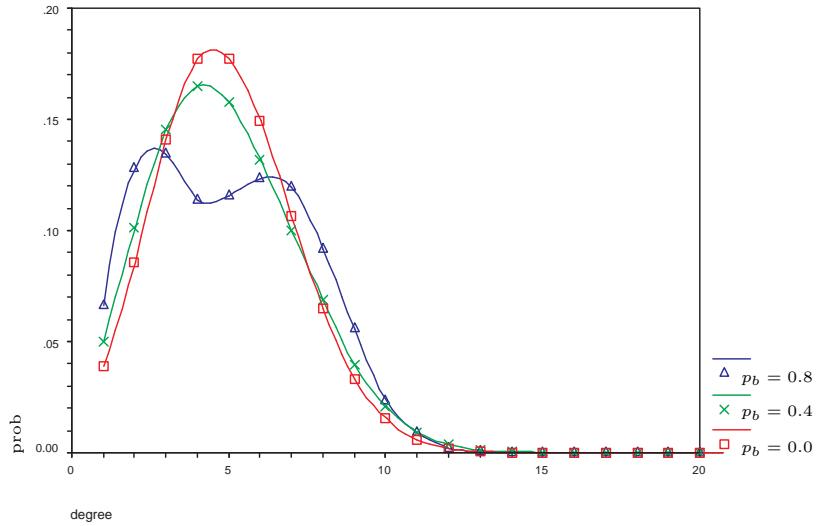


Figure 7: Average degree distribution for various values of  $p_b$ .

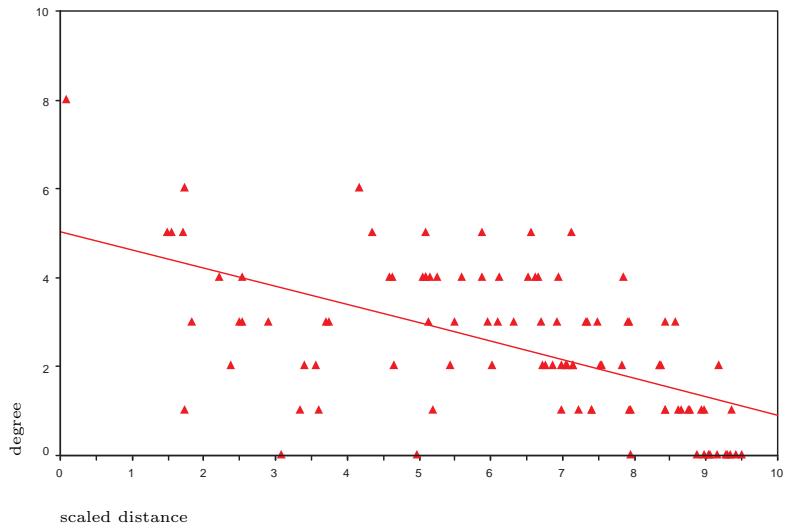


Figure 8: Scatterplot for the degree of nodes versus distance from the centre of the graph in one typical instance of the model ( $N = 100$ ,  $p = 0.05$ ,  $p_b = 0.8$ ). The line is the least square fit line.

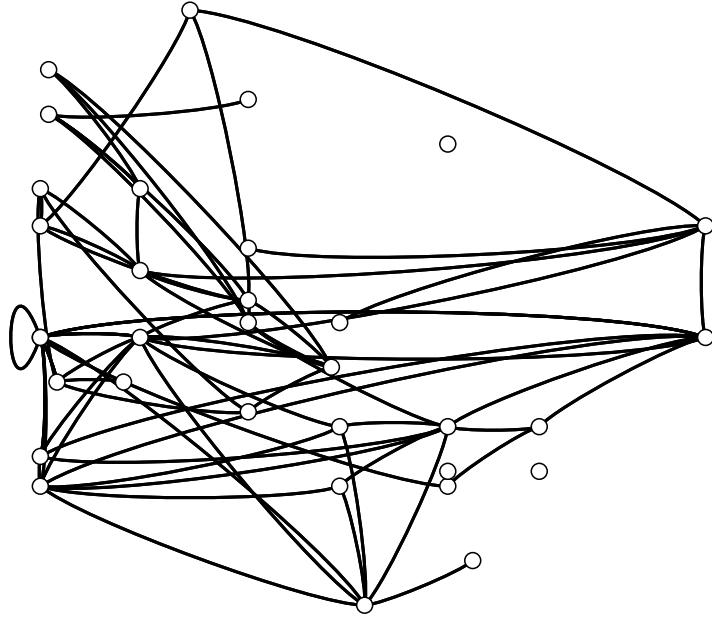


Figure 9: An office communication network of 33 individuals [6]. The location of the nodes reflects the cubicle or room locations. Note that the “self-loop” in fact indicates a link between two individuals who share the same office space.

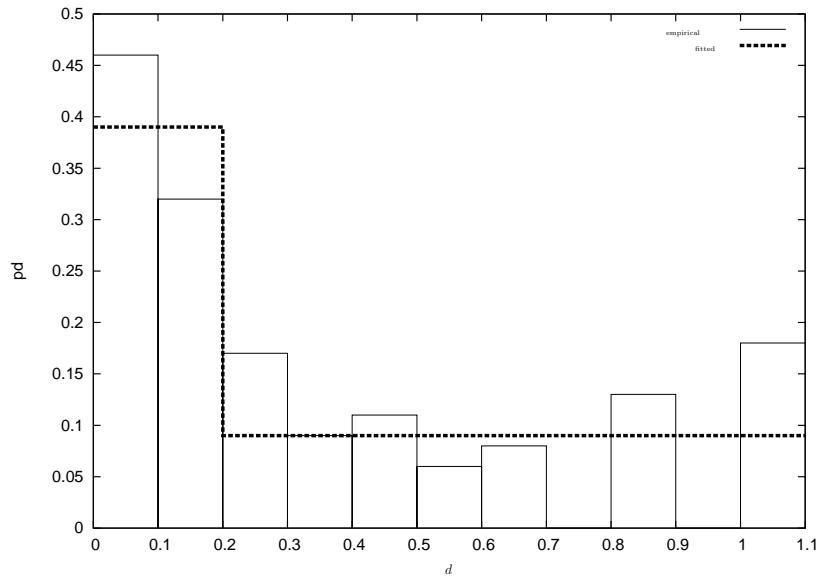


Figure 10: A plot of edge probability  $p(d)$  against distance  $d$  in the communication network. The solid bars are the empirical edge probability and the dashed line is the fitted step function with neighbourhood radius set at  $H = 0.2$ . Note that the distance has been scaled so that all node fits into a  $1 \times 1$  square. All pairs of nodes are less than distance 1.1 apart.

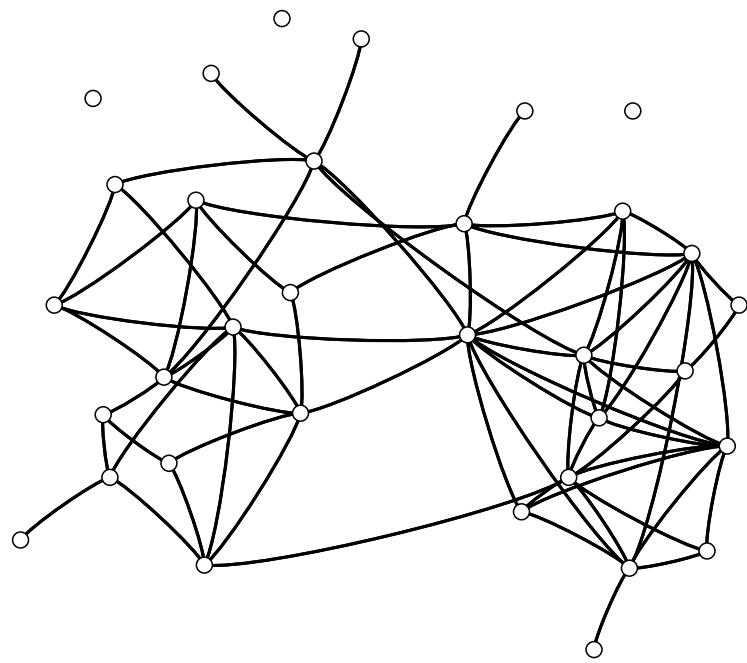


Figure 11: A spatial re-arrangement of the office communication network.