Title
Predictability of Large-scale Spatially Embedded Networks

Author
Butts, Carter T.

Publication Date
2002-10-01

Predictability of Large-scale Spatially Embedded Networks*†

Carter T. Butts‡

10/23/02

Abstract

Although it is well-known that there is a relationship between socio-physical distance and edge probability in interpersonal networks, the predictive power of such distances for total network structure has not been established. Here, it is shown that upper bounds on the marginal edge probabilities for far-flung dyads can be used to place a lower bound on the predictive power of distance, and one such bound is derived. Application of this bound to the special case of uniformly placed vertices on the plane suggests that only modest constraints are required for distance effects to dominate at large physical scales.

Keywords: social networks, spatial models, Shannon entropy, predictive power, distance

1 Introduction

Numerous studies show a strong relationship between physical distance and social structure (e.g., Merton (1948); Festinger et al. (1950); Caplow and Forman (1950); Blake et al. (1956); Whyte (1957); Sommer (1969); Snow et al. (1981); Latané et al. (1995)); arguably, few other findings in the social sciences can claim such a degree of strength and generality. While this is an interesting and important result in and of itself, it begs a critical question: assuming

*This material is based upon work supported by the National Science Foundation under Grant No. 0100999, as well as an NSF Graduate Research Fellowship. Additional support was provided by the Center for the Computational Analysis of Social and Organizational Systems and the Institute for Complex Engineered Systems at Carnegie Mellon University.

†I would like to thank Kathleen Carley, David Krackhardt, Shelby Stewman, David Rode, and Katie Faust for their comments and input.

‡Institute for Mathematical Behavioral Sciences and Department of Sociology, University of California-Irvine, Irvine, CA 92697, buttsc@uci.edu

that the distance/edge probability relationship is as it appears to be, to what extent can this account for the variability of social structure *writ large*? In his 1984 comment, "Chance and Necessity in Sociological Theory," Bruce Mayhew makes the characteristically bold claim that well over 90% of the variation in social structure is determined by physical space[1]. If Mayhew's assertion is correct, then we would expect for network models based on vertex position to allow us to develop extremely credible predictions of large-scale network structure. Since individual positions can be inferred from population data, such a result should (in principle, at least) allow us to reduce the problem of macrostructural prediction to one of spatial demography. If the assertion is false, by contrast, then other approaches will be required to effectively model the structure of social macrostructure.

While the "Mayhew question" is unlikely to be settled by a single paper, it is shown here that the requirements for predicting network structure from vertex layout are fairly modest. Fairly minimal constraints on the probability of edges between distant alters are sufficient to establish a lower bound on the predictive power of distance, where predictive power is defined in terms of reduction in the Shannon entropy of the total structure. Application of this bound to the special case of vertices in a planar region suggests that the requirements for strong distance effects (e.g., $> 90\%$ uncertainty reduction) are likely to be attainable in practice for moderate to large physical scales, and thus that it is reasonable to expect that large-scale spatially embedded social networks will be readily predictable from vertex position data.

2 Notation and Basic Assumptions

For the results which follow, we will focus exclusively on the case of a loopless undirected graph $G = (V, E)$ with known vertex set V and uncertain (i.e., random) edge set E. (For convenience, we denote the cardinality of V by $N = |V|$.) It is assumed further that G is spatially embedded, in the sense that there exists some space $\mathbb{S}$ and set $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ such that $\mathbf{V} \subset \mathbb{S}$ and $\mathbf{v}_i$ is the *position* of vertex v_i. We assume that there exists some distance function, d, on $\mathbb{S}$, but do not require it to be a metric (i.e., it need not satisfy the triangle inequality). Substantively, the the most obvious interpretation of $\mathbb{S}$ is as a socio-physical space (often called a "Blau" space (Blau, 1977)); such a space may include both physical and demographic dimensions, including gender, age, race, and primary language. For the purposes of our present application, we will focus on a physical space interpretation of $\mathbb{S}$, but it should be emphasized that this is not required for the general result to apply.

[1]This, he quips, being an artificially low estimate due to measurement error.

3 Predictability of Spatially Embedded Networks

What, precisely, is meant by the "predictability" of spatially embedded networks? In the context of this paper, predictability is understood to be the extent to which our initial uncertainty regarding network structure is reduced by the provision of new information. Specifically, we are interested in the extent to which knowledge of vertex positions within $\mathbb{S}$ reduces our uncertainty regarding the edge set of G. A natural measure of uncertainty – and the one which we shall employ here – is the Shannon entropy, which can be interpreted as the expected length of an optimally encoded signal expressing the value of a random variable. Denoting the entropy function by I, we form the R^2-like predictability measure

$$\mathcal{P}(G|\mathbf{V}) \equiv 1 - \frac{I(G|\mathbf{V})}{I(G)} \tag{1}$$

which expresses the extent to which knowledge of the vertex position set, $\mathbf{V}$, can account for the total information content of G. With $\mathcal{P}$ as our notion of predictability, we can state the following general result:

Theorem 1 (Predictability). *Let $G = (V, E)$ be a spatially embedded random graph with vertex position set $\mathbf{V}$ and distance function d, and let G be distributed such that $p\big(\{v_i, v_j\} \in E(G)\big) \leq \epsilon \;\forall\; v_i, v_j : d(\mathbf{v}_i, \mathbf{v}_j) \geq r_c$, for some $\epsilon < 0.5$. Then $\mathcal{P}(G|\mathbf{V}) \geq p\big(d(\mathbf{v}_i, \mathbf{v}_j) \geq r_c\big)\big(1 - I_B(\epsilon)\big)$, and $\lim_{\epsilon \to 0} \mathcal{P}(G|\mathbf{V}) \geq p\big(d(\mathbf{v}_i, \mathbf{v}_j) \geq r_c\big)$, where $\mathcal{P}(G|\mathbf{V}) = 1 - \frac{I(G|\mathbf{V})}{I(G)}$, I is the Shannon entropy, and $I_B(\epsilon) = -\epsilon \log_2 \epsilon - (1-\epsilon)\log_2(1-\epsilon)$.*

Proof. For convenience in notation, let $d_{ij} = d(\mathbf{v}_i, \mathbf{v}_j)$ and let $e_{ij} \equiv \{v_i, v_j\} \in E(G)$. We begin by assuming that, for actors within radius r_c, distance tells us *nothing* regarding edge probability; that is to say, $p(e_{ij} | d_{ij} < r_c) = 0.5$. Then it trivially follows from the definition of the Shannon entropy that $I(e_{ij} | d_{ij} < r_c) = 1$. For the complementary case, we begin by noting that $I(e_{ij} | d_{ij} \geq r_c) = -p(e_{ij} | d_{ij} \geq r_c) \log_2 p(e_{ij} | d_{ij} \geq r_c) - \big(1 - p(e_{ij} | d_{ij} \geq r_c)\big) \log_2\big(1 - p(e_{ij} | d_{ij} \geq r_c)\big)$. The fact that $p(e_{ij} | d_{ij} \geq r_c) \leq \epsilon < 0.5$ then implies that $I(e_{ij} | d_{ij} \geq r_c) \leq -\epsilon \log_2 \epsilon - (1-\epsilon)\log_2(1-\epsilon) = I_B(\epsilon)$.

We now consider the entropy of the entire graph. Using the well-known result that $I(X, Y) \leq I(X) + I(Y)$ for (possibly dependent) random variables X, Y, we can bound the entropy of the graph by the sum of the (independent) edgewise entropies. Therefore we have

$$\begin{aligned}
I\left(G\,|\mathbf{V}\right) &\leq \sum_{\{i,j\}} I\left(e_{ij}\,|\mathbf{V}\right) \\
&\leq \sum_{\{i,j\}:d_{ij}<r_c} 1 + \sum_{\{i,j\}:d_{ij}\geq r_c} I_B(\epsilon) \\
&= \binom{|V\left(G\right)|}{2} p\left(d_{ij}<r_c\right) + \binom{|V\left(G\right)|}{2} p\left(d_{ij}\geq r_c\right) I_B(\epsilon) \\
&= \Big[p\left(d_{ij}<r_c\right) + p\left(d_{ij}\geq r_c\right) I_B(\epsilon)\Big] \binom{|V\left(G\right)|}{2} \\
&= \Big[p\left(d_{ij}<r_c\right) + \big(1-p\left(d_{ij}<r_c\right)\big) I_B(\epsilon)\Big] \binom{|V\left(G\right)|}{2}.
\end{aligned}$$

Since the uninformative entropy of G is given by $I(G)=\binom{|V(G)|}{2}$, it follows that

$$\begin{aligned}
\mathcal{P}(G|\mathbf{V}) &= 1 - \frac{I\left(G\,|\mathbf{V}\right)}{I\left(G\right)} \\
&\geq 1 - \frac{\Big[p\left(d_{ij}<r_c\right) + \big(1-p\left(d_{ij}<r_c\right)\big) I_B\left(\epsilon\right)\Big]\binom{|V(G)|}{2}}{\binom{|V(G)|}{2}} \\
&= 1 - p\left(d_{ij}<r_c\right) - \big(1-p\left(d_{ij}<r_c\right)\big) I_B\left(\epsilon\right) \\
&= p\left(d_{ij}\geq r_c\right) - p\left(d_{ij}\geq r_c\right) I_B\left(\epsilon\right) \\
&= p\left(d_{ij}\geq r_c\right)\big(1-I_B\left(\epsilon\right)\big)
\end{aligned}$$

which demonstrates the first portion of Theorem 1. To complete the proof, we allow $\epsilon \to 0$ and take the limit:

$$\begin{aligned}
\lim_{\epsilon\to 0} \mathcal{P}(G|\mathbf{V}) &\geq \lim_{\epsilon\to 0} p\left(d_{ij}\geq r_c\right)\big(1-I_B\left(\epsilon\right)\big) \\
&= \lim_{\epsilon\to 0} p\left(d_{ij}\geq r_c\right)\big(1+\epsilon\log_2\epsilon + (1-\epsilon)\log_2(1-\epsilon)\big) \\
&= p\left(d_{ij}\geq r_c\right).
\end{aligned}$$

□

This is a powerful and general result: it tells us that whenever we can place a reasonable upper bound on the marginal edge probability between distant vertices, we can use the quantiles of the distance distribution to place a lower bound on the predictive power of $\mathcal{V}$. Furthermore, when this bound on marginal edge probability becomes small, the predictive power of the position set becomes bounded by the probability that the distance between two randomly selected vertices will exceed the critical threshold. Thus, where the threshold distance is small relative to the overall distribution, we can *guarantee* that the total structure will be easily predicted from vertex position alone.

One important and somewhat counter-intuitive aspect of Theorem 1 is that it does not depend on N: the predictability of the total structure can be bounded by a function which depends only on the geometry of the population layout. Similarly, we did not have to assume dyadic independence to obtain this result (only bounds on the edgewise marginals). These two facts greatly facilitate the application of Theorem 1 in the field, where population distributions and some crude estimates of the distance/edge probability relationship may be all that is available. They also serve to reinforce the argument that the predictive power of distance is robust to varying assumptions about the precise determinants of network structure.

4 Uniform Population Distribution on the Plane

Consider the special case in which a population of arbitrary size is placed uniformly within a square region of size $\ell \times \ell$. Such a model may be thought of as a first approximation to a sparse population distribution in physical space, particularly over large areas. Here, we show the minimum threshold distances necessary to obtain a given level of predictive power for a structure on the plane, as a function of the linear scale (ℓ) of the embedding region. As will be shown, the only modest critical thresholds are required to guarantee high levels of predictability under uniform vertex placement.

4.1 Distribution of Inter-point Distances

In order to apply Theorem 1, we must first know the distribution of inter-point distances for square planar regions. Under the assumption that d is the euclidean distance, we derive this distribution in the following lemma:

Lemma 1. *Let $\mathbf{v}_i, \mathbf{v}_j$ designate two randomly selected points on a two-dimensional plane, with each coordinate being IID $U(0, \ell)$. Then the density function of $d(\mathbf{v}_i, \mathbf{v}_j)$ is given by*

$$f(d) = \begin{cases} 2d\left[\frac{\pi}{\ell^2} - \frac{4d}{\ell^3} + \frac{d^2}{\ell^4}\right] & 0 \leq d \leq \ell \\ 2d\left[\frac{2}{\ell^2}\sin^{-1}\left(\frac{2\ell^2-d^2}{d^2}\right) - \frac{4}{\ell^3}\left(\ell - \sqrt{d^2-\ell^2}\right) + \frac{2\ell^2-d^2}{\ell^4}\right] & \ell < d \leq \sqrt{2}\ell \\ 0 & \textit{otherwise} \end{cases},$$

with associated distribution function

$$F(d) = \begin{cases} 0 & d < 0 \\ 2\left[\frac{\pi d^2}{2\ell^2} - \frac{d^3}{3\ell^3} + \frac{d^4}{4\ell^4}\right] & 0 \le d \le \ell \\ \frac{1}{3} + \frac{2d^2}{\ell^2}\left[1 - 2\left(\frac{\ell^2 d^2 - \ell^4}{d^4}\right) + \sin^{-1}\left(\frac{d^2 - 2\ell^2}{d^2}\right)\right] + \frac{8\left(d^2-\ell^2\right)^{\frac{3}{2}}}{\ell^3} - \frac{d^4}{\ell^4} & \ell < d \le \sqrt{2}\ell \\ 1 & d > \sqrt{2}\ell \end{cases}.$$

Proof. For the two-dimensional case, we may write the euclidean distance in terms of coordinate differences:

$$d\left(\mathbf{v}_i, \mathbf{v}_j\right) = \sqrt{\left(\left(\mathbf{v}_i\right)_1 - \left(\mathbf{v}_j\right)_1\right)^2 + \left(\left(\mathbf{v}_i\right)_2 - \left(\mathbf{v}_j\right)_2\right)^2}$$

By assumption, these coordinates are uniformly distributed on $[0, \ell]$. It can easily be shown that the difference between two such uniform deviates is distributed Triangular with lower bound $-\ell$, upper bound ℓ, and mode 0. Thus, we may simplify the distribution of d as follows:

$$\begin{aligned} d\left(\mathbf{v}_i, \mathbf{v}_j\right) &\sim \sqrt{\left(U\left(0,\ell\right) - U\left(0,\ell\right)\right)^2 + \left(U\left(0,\ell\right) - U\left(0,\ell\right)\right)^2} \\ &\sim \sqrt{\left(T\left(-\ell,\ell,0\right)\right)^2 + \left(T\left(-\ell,\ell,0\right)\right)^2} \end{aligned}$$

Note that $T\left(-\ell,\ell,0\right)$ is symmetric about the origin; thus, it is a standard result that $F_{T^2}(x) = 2F_T\left(\sqrt{x}\right) - 1$, where F_{T^2} and F_T are the cumulative distribution functions of variates T^2 and T, respectively (Evans et al., 2000). For the distribution function of a Triangular deviate with lower bound a, upper bound b, and mode c we have

$$F_T(x) = \begin{cases} 0 & x < 0 \\ \frac{(x-a)^2}{(b-a)(c-a)} & 0 \le x < c \\ 1 - \frac{(b-x)^2}{(b-a)(b-c)} & c \le x \le b \\ 1 & x > b \end{cases},$$

and hence

$$\begin{aligned} F_{T^2}(x) &= 2F_T\left(\sqrt{x}\right) - 1 \\ &= \begin{cases} 0 & \sqrt{x} < 0 \\ 2\frac{\left(\sqrt{x}-a\right)^2}{(b-a)(c-a)} - 1 & 0 \le \sqrt{x} < c \\ 1 - 2\frac{\left(b-\sqrt{x}\right)^2}{(b-a)(b-c)} & c \le \sqrt{x} \le b \\ 1 & \sqrt{x} > b \end{cases} \end{aligned}$$

which, by symmetry, can be collapsed to

$$\begin{aligned}
&= 1 - \frac{2\left(b-\sqrt{x}\right)^2}{2bb} \\
&= 1 - \frac{\left(b-\sqrt{x}\right)^2}{b^2}
\end{aligned}$$

(for $0 \leq x \leq b^2$).

To obtain the associated density function, $f_{T^2}(x)$, we simply differentiate:

$$\begin{aligned}
f_{T^2}(x) &= \frac{\partial}{\partial x} F_{T^2}(x) \\
&= \frac{\partial}{\partial x}\left(1 - \frac{\left(b-\sqrt{x}\right)^2}{b^2}\right) \\
&= \begin{cases} \frac{1}{b\sqrt{x}} - \frac{1}{b^2} & 0 \leq x \leq b^2 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

Having derived the density of a single T^2 variate, we now must consider the sum of two such (IID) variates. Since the variates in question are independent, their joint density is simply the product of their individual densities, and we may obtain the density of the sum via convolution. In this case, however, we note that the domain of $T^2 + T^2$ depends on the individual variates, and hence some care is needed in choosing the limits of integration. We divide the density at $x = b^2$ (the point beyond which both variates must be greater than 0), taking the lower region first:

$$\begin{aligned}
f_{T^2+T^2}\left(x \,\middle|\, 0 \leq x \leq b^2\right) &= \int_0^x f_{T^2}(y)\, f_{T^2}(x-y)\,\mathrm{d}y \\
&= \int_0^x \left(\frac{1}{b\sqrt{y}} - \frac{1}{b^2}\right)\left(\frac{1}{b\sqrt{x-y}} - \frac{1}{b^2}\right)\mathrm{d}y \\
&= \frac{1}{b^2}\sin^{-1}\left(\frac{2y-x}{x}\right)\Big|_0^x - \frac{2\sqrt{y}}{b^3}\Big|_0^x + \frac{2\sqrt{x-y}}{b^3}\Big|_0^x + \frac{y}{b^4}\Big|_0^x \\
&= \frac{1}{b^2}\left[\sin^{-1}(1) - \sin^{-1}(-1)\right] - \frac{2\sqrt{x}}{b^3} - \frac{2\sqrt{x}}{b^3} + \frac{x}{b^4} \\
&= \frac{\pi}{b^2} - \frac{4\sqrt{x}}{b^3} + \frac{x}{b^4}
\end{aligned}$$

Now, we consider the upper region:

$$
\begin{aligned}
f_{T^2+T^2}\left(x \,\middle|\, b^2 < x \le 2b^2\right) &= \int_{x-b^2}^{b^2} f_{T^2}(y)\, f_{T^2}(x-y)\,\mathrm{d}y \\
&= \int_{x-b^2}^{b^2} \left(\frac{1}{b\sqrt{y}} - \frac{1}{b^2}\right)\left(\frac{1}{b\sqrt{x-y}} - \frac{1}{b^2}\right)\mathrm{d}y \\
&= \frac{1}{b^2}\sin^{-1}\left(\frac{2y-x}{x}\right)\Bigg|_{x-b^2}^{b^2} - \frac{2\sqrt{y}}{b^3}\Bigg|_{x-b^2}^{b^2} + \frac{2\sqrt{x-y}}{b^3}\Bigg|_{x-b^2}^{b^2} + \frac{y}{b^4}\Bigg|_{x-b^2}^{b^2} \\
&= \frac{1}{b^2}\left[\sin^{-1}\left(\frac{2b^2-x}{x}\right) - \sin^{-1}\left(\frac{x-2b^2}{x}\right)\right] - \frac{2}{b^3}\left(b - \sqrt{x-b^2}\right) \\
&+ \frac{2}{b^3}\left(\sqrt{x-b^2} - b\right) + \frac{2b^2-x}{b^4} \\
&= \frac{2}{b^2}\sin^{-1}\left(\frac{2b^2-x}{x}\right) - \frac{4}{b^3}\left(b - \sqrt{x-b^2}\right) + \frac{2b^2-x}{b^4}
\end{aligned}
$$

Finally, putting this together, we obtain the complete density for a sum of two squared triangular variates:

$$
f_{T^2+T^2}(x) = \begin{cases} \frac{\pi}{b^2} - \frac{4\sqrt{x}}{b^3} + \frac{x}{b^4} & 0 \le x \le b^2 \\ \frac{2}{b^2}\sin^{-1}\left(\frac{2b^2-x}{x}\right) - \frac{4}{b^3}\left(b - \sqrt{x-b^2}\right) + \frac{2b^2-x}{b^4} & b^2 < x \le 2b^2 \\ 0 & \text{otherwise} \end{cases}
$$

From the above, we may now derive the density of d. The last step in this process involves applying a positive square root transformation to the sum of squared triangular variates. This is a monotonically increasing one-to-one transformation, and we can thus derive the new density by a simple change of variables:

$$
f_{\sqrt{T^2+T^2}}(x) = f_{T^2+T^2}\left(x^2\right)|J|
$$

(where $|J|$ is the Jacobian determinant of the transformation)

$$
\begin{aligned}
&= f_{T^2+T^2}\left(x^2\right) 2x \\
&= \begin{cases} 2x\left[\frac{\pi}{b^2} - \frac{4x}{b^3} + \frac{x^2}{b^4}\right] & 0 \le x \le b \\ 2x\left[\frac{2}{b^2}\sin^{-1}\left(\frac{2b^2-x^2}{x^2}\right) - \frac{4}{b^3}\left(b - \sqrt{x^2-b^2}\right) + \frac{2b^2-x^2}{b^4}\right] & b < x \le \sqrt{2}b \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

To arrive at the distribution function of d, we need merely integrate $f_{\sqrt{T^2+T^2}}$ over the desired

Scale		$\mathcal{P}(G\|\mathbf{V})$		
ℓxℓ		90%	95%	99%
0.1x0.1 km	$r_c \leq$	0.019	0.013	0.006
1x1 km	$r_c \leq$	0.192	0.133	0.058
10x10 km	$r_c \leq$	1.924	1.331	0.578
100x100 km	$r_c \leq$	19.236	13.311	5.923
All r_c values given in kilometers.				

Table 1: Maximum Critical Radius as a Function of Scale and Uncertainty Reduction, Uniform Vertex Placement

range.

$$F_{\sqrt{T^2+T^2}}(x) = \begin{cases} 0 & x < 0 \\ \int_0^x 2x\left[\frac{\pi}{b^2} - \frac{4x}{b^3} + \frac{x^2}{b^4}\right] \mathrm{d}y & 0 \leq x \leq b \\ \int_0^b 2x\left[\frac{\pi}{b^2} - \frac{4x}{b^3} + \frac{x^2}{b^4}\right] \mathrm{d}y & \\ + \int_b^x 2x\left[\frac{2}{b^2}\sin^{-1}\left(\frac{2b^2-x^2}{x^2}\right)\right. & \\ \left. -\frac{4}{b^3}\left(b - \sqrt{x^2-b^2}\right) + \frac{2b^2-x^2}{b^4}\right] \mathrm{d}y & b < x \leq \sqrt{2}b \\ 1 & x > \sqrt{2}b \end{cases}$$

$$= \begin{cases} 0 & x < 0 \\ 2\left[\frac{\pi x^2}{2b^2} - \frac{4x^3}{3b^3} + \frac{x^4}{4b^4}\right] & 0 \leq x \leq b \\ \frac{1}{3} + \frac{2x^2}{b^2}\left[1 - 2\left(\frac{b^2x^2-b^4}{x^4}\right) + \sin^{-1}\left(\frac{x^2-2b^2}{x^2}\right)\right] & \\ + \frac{8\left(x^2-b^2\right)^{\frac{3}{2}}}{b^3} - \frac{x^4}{b^4} & b < x \leq \sqrt{2}b \\ 1 & x > \sqrt{2}b \end{cases}$$

Substitution of d for x and ℓ for b completes the proof. □

4.2 Predictability

Using Lemma 1 together with Theorem 1, we may determine the maximum r_c value needed to guarantee that a given fraction of the uncertainty in G can be accounted for by the euclidean distances between vertex positions; these threshold values are shown in Table 1. As the table indicates, a critical radius at or below approximately %20 of the linear scale of the region is adequate for a %90 uncertainty reduction, with a reduction of %99 possible for radii of 0.06ℓ or less.

How do these threshold values compare to empirical assessments of the distance/edge probability relationship? Fits of dyadic edge models to existing data sets suggest that a low-probability threshold is attainable, but also clearly indicate that thresholds will depend upon relational content (Butts, 2002). Recalling that $\mathcal{P}(G|\mathbf{V}) \geq p\big(d\left(\mathbf{v}_i, \mathbf{v}_j\right) \geq r_c\big)\big(1 - I_B\left(\epsilon\right)\big)$, we can express the adequacy of the threshold approximation in terms of $I_B\left(\epsilon\right)$. In this regard, $\epsilon \leq 0.001$ is sufficient to bring the predictability bound within approximately 1% of the limit; such a bound is not hard to achieve in practice. Based on the Butts models, thresholds of as little as 0.05km may be reasonable approximations for face-to-face contact, with larger thresholds of approximately 0.25km and 18km for social friendship and telephone contacts, respectively. Although it may be possible to obtain *more* predictive power using these or other models than Table 1 would suggest, the lower bounds alone indicate that physical layout has the potential to account for the overwhelming majority of network structure at even modest spatial scales.

5 Conclusion

To summarize, then, it would seem that even a *very* modest null model based on physical distance (the threshold model) must account for the vast majority of network structure in large-scale networks, under quite minimal assumptions. Since fitted models have the capacity to be much more informative than the null model, they are expected to provide even more information about network macrostructure at even smaller scales. Thus, not only might one reasonably speculate that distance could account for most of the uncertainty in large-scale interpersonal networks, it almost *has* to do so. This would seem to vindicate the intuition of theorists such as Mayhew, who perceived that physical space was a critical structuring force, but who did not demonstrate the extent of that result.

References

Blake, R. R., Rhead, C., Wedge, B., and Monton, J. (1956). Housing architecture and social interaction. *Sociometry*, 19:133–139.

Blau, P. M. (1977). *Inequality and Heterogeneity.* Free Press, New York.

Butts, C. T. (2002). *Spatial Models of Large-scale Interpersonal Networks.* Doctoral Dissertation, Carnegie Mellon University.

Caplow, T. and Forman, R. (1950). Neighborhood interaction in a homogeneous community. *American Sociological Review*, 15:357–366.

Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical Distributions.* John Wiley and Sons, New York, third edition.

Festinger, L., Schachter, S., and Back, K. (1950). *Social Pressures in Informal Groups*. Stanford University Press, Stanford, California.

Latané, B., Liu, J. H., Nowak, A., Bonevento, M., and Zheng, L. (1995). Distance matters: Physical space and social impact. *Personality and Social Psychology Bulletin*, 21(8):795–805.

Mayhew, B. H. (1984). Chance and necessity in sociological theory. *Journal of Mathematical Sociology*, 9:305–339.

Merton, R. (1948). The social psychology of housing. In Dennis, W., editor, *Current Trends in Social Psychology*, pages 163–217. University of Pittsburgh Press, Pittsburgh, PA.

Snow, D. A., Leahy, P. J., and Schwab, W. A. (1981). Social interaction in a heterogeneous apartment: An investigation of the effects of environment upon behavior. *Sociological Focus*, 14(4):309–319.

Sommer, R. (1969). *Personal Space*. Prentice-Hall, Englewood Cliffs, NJ.

Whyte, W. H. (1957). *The Organization Man*. Doubleday, Garden City, NY.