Geographical Variability and Network Structure[1]

Carter T. Butts[2]

Department of Sociology and
Institute for Mathematical and Behavioral Sciences
University of California, Irvine

Ryan M. Acton

Department of Sociology
University of Massachusetts Amherst

John R. Hipp

Department of Criminology, Law, and Society
University of California, Irvine

Nicholas N. Nagle

Department of Geography
University of Tennessee, Knoxville

[1] This research was supported by NSF award BCS-0827027 and ONR award N00014-08-1-1015.

[2] To whom correspondence should be addressed: University of California, Irvine; SSPA 2145; Irvine, CA 92697-5100; buttsc@uci.edu

Abstract

In this paper, we explore the potential implications of geographical variability for the structure of social networks. Beginning with some basic simplifying assumptions, we derive a number of ways in which local network structure should be expected to vary across a region whose population is unevenly distributed. To examine the manner in which these effects would be expected to manifest given realistic population distributions, we then perform an exploratory simulation study that examines the features of large-scale interpersonal networks generated using block-level data from the 2000 U.S. Census. Using a stratified sample of micropolitan and metropolitan areas with populations ranging from approximately 1,000 to 1,000,000 persons, we extrapolatively simulate network structure using spatial network models calibrated to two fairly proximate social relations. From this sample of simulated networks, we examine the effect of both within-location and between-location heterogeneity on a variety of structural properties. As we demonstrate, geographical variability produces large and distinctive features in the "social fabric" that overlies it; at the same time, however, many aggregate network properties can be fairly well-predicted from relatively simple spatial demographic variables. The impact of geographical variability is thus predicted to depend substantially on the type of network property being assessed, and on the spatial scale involved.

Keywords: spatially embedded networks, spatial Bernoulli graphs, geographical variability, settlement patterns, graph-level indices

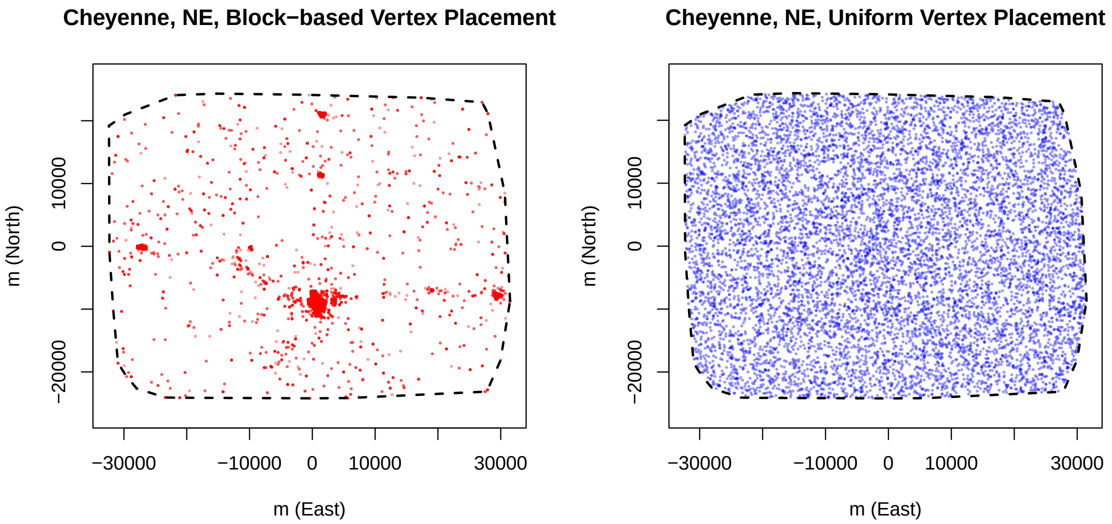


Figure 1: Population distributions for Cheyenne, NE MSA. Points in left panel are placed uniformly by census block; points in right panel are placed uniformly over the convex hull of positions from the census-constrained model. Both contain the same number of points ($N = 9,830$).

A basic observation regarding the distribution of humans across geographical space is that this distribution is extremely heterogeneous. Even leaving aside the contrast between inhabited lands and uninhabited oceans (comprising the majority of Earth's surface area), settlements are typically concentrated in a small set of regions having desirable geological, hydrological, and resource access properties. Within these regions, the resulting settlements are of extremely uneven size, distribution, and structure (Zipf, 1949; Brakman et al., 1999; White et al., 2008). Contrary to the intuition of an evenly inhabited Earth, then, humans are distributed unevenly across a wide range of geographical scales. Moreover, this heterogeneity can be extreme. For example, the left panel of Figure 1 illustrates the population distribution of persons within the Cheyenne, Nebraska metropolitan statistical area, as recorded by the year 2000 Census of the United States.[1] Although the population density in this area is approximately 3 persons per square kilometer (far below the U.S. mean of approximately 32 persons/km^2), this is not reflective of the conditions under which most residents of this region live. Indeed, as the figure indicates, most of the population of this region is concentrated into a small number of dense communities, surrounded by large areas with few residents. The extent of this concentration may be appreciated by comparing the panel on the left with that on the right, which depicts a uniform population distribution over the same area. The difference is stark. Rather than being embedded in a uniform, low-density environment, the median resident of this region faces a local population density of approximately 1,000 persons/km^2 (as computed from block-level data), with densities in some areas being as high as 12,700 persons/km^2 or as low as 0.09. The micro-environments in which individuals form ties may thus differ greatly from a uniform baseline, and these micro-environments may themselves be distributed unevenly across the landscape.

The above observations raise an important question: if human settlement patterns are extremely

[1] This and all maps shown are based on orthographic projections about a central point in the MSA, with distances in meters. Distance and area calculations throughout the paper are based on these projections.

heterogeneous, and if spatial structure influences network structure, then should we not expect that the geographical variability in population distribution will have a substantial impact on the structure of social networks? And, if this is so, what will be the nature of that impact? To investigate these questions, we begin by employing a simple modeling framework which preserves the marginal relationship between distance and tie probability. Using this simplified framework, we then derive a number of ways in which network structure should (or, in some case, should not) be expected to vary based on the underlying population distribution. To explore the ramifications of these results in a realistic geographical context, we perform an extrapolative simulation study of network structure in communities from across the United States, simulating populations ranging from 1,000 to 1,000,000 persons in size over a range of land areas using two relational models. By examining the properties of the resulting networks, we draw some basic conclusions regarding the likely impact of geographical variability on network structure, and identify several concrete targets for empirical research.

1 Spatial Models of Social Networks

It is a well-established result that the marginal probability of a social tie between two persons declines with geographical distance for a wide range of social relations (see, e.g., Bossard, 1932; Zipf, 1949; Festinger et al., 1950; Hägerstrand, 1967; Freeman et al., 1988; Latané et al., 1995; McPherson et al., 2001). While often regarded as a mere curiosity, others have argued that this relationship is a critical determinant of social structure (Mayhew, 1984b). Indeed, Butts (2003) has shown that under fairly weak conditions, spatial structure is adequate to account for the vast majority of network structure (in terms of total entropy) at large geographical scales.

The simplest family of network models to incorporate this notion is the family of spatial Bernoulli graphs, defined by probability mass functions (pmfs) of the form

$$\Pr\left(Y = y \left| D \right.\right) = \prod_{\{i,j\}} B\left(Y_{ij} = y_{ij} \left| \mathcal{F}\left(D_{ij}, \theta\right)\right.\right), \tag{1}$$

where Y is the (random) graph adjacency matrix, D is a matrix of inter-vertex distances, B is the Bernoulli pmf, and $\mathcal{F}$ is a function taking distances into the $[0, 1]$ interval (parameterized by real vector θ). In this context, $\mathcal{F}$ is referred to as a *spatial interaction function* (SIF), and can be interpreted directly as providing the marginal probability of a tie between two randomly selected individuals at some given distance. It can immediately be observed that this family is a special case of the inhomogeneous Bernoulli graphs (w/pmf $\Pr(Y = y|\Phi) = \prod_{ij} B(Y_{ij} = y_{ij}|\Phi_{ij})$), with parameter matrix Φ given by $\Phi_{ij} = \mathcal{F}(D_{ij}, \theta)$. Models of this form have been studied in the context of geographical distances by Butts (2002); Hipp and Perrin (2009); Butts and Acton (2011), and are closely related to the latent space models of Hoff et al. (2002); Handcock et al. (2007). They can also be viewed as special cases of the family of gravity models (Haynes and Fotheringham, 1984), which have been used for several decades in the geographical literature to model interaction between areal units. Butts (2006a) has further shown that the spatial Bernoulli graphs can be written as a special case of a more general curved exponential family of graph distributions. By defining canonical parameters $\eta\left(\theta, d\right) = \text{logit}\mathcal{F}\left(d, \theta\right)$, we may write the pmf for adjacency matrix Y with support $\mathcal{Y}$ as

$$\Pr\left(Y = y \left| D, \theta, \psi \right.\right) \propto \exp\left[\sum_{\{i,j\}} \eta(\theta, D_{ij}) y_{ij} + \psi^T t(y)\right], \tag{2}$$

where $\psi \in \mathbb{R}^p$ and $t : \mathcal{Y} \mapsto \mathbb{R}^p$ are respective vectors of parameters and sufficient statistics. The incorporation of additional statistics (via t) allows for the combination of both spatial and non-spatial effects (e.g., endogenous triangulation, as explored in recent work by Daraganova and Pattison (2007)).

From the perspective of this general family, we re-obtain the spatial Bernoulli graphs in the case for which all non-spatial effects are omitted. In practice, this is naturally an approximation (though possibly a good one for large-scale structure); however, it also has an important theoretical interpretation as the graph distribution having maximum entropy given the marginal distance/tie probability relationship. Thus, if one seeks to understand the "pure" impact of spatial structure on network structure, this model serves as a natural reference point. This model also has the attractive property that it is amenable to large-scale simulation (something not true of models with more complex effects—see, e.g. Snijders (2002) for a discussion of some relevant difficulties), and in some cases to analytical treatment. Given its simple interpretation and theoretical leverage, we employ the spatial Bernoulli framework in the work that follows; we do note, however, that the ability to extend it via the method of Equation 2 provides a ready avenue for further exploration.

2 Effects of Spatial Heterogeneity: Intuitions From First Principles

To reiterate the motivating observation of Figure 1, human beings are heterogeneously distributed over the Earth's surface. If, per section 1, the probability of a social tie between two persons varies systematically with distance, then this heterogeneity in spatial structure will be reflected in the structure of the associated social network. Although some effects of spatial heterogeneity[2] on network structure are difficult to characterize, it is possible to obtain an intuition for some of the relevant mechanisms by examining a few stylized scenarios. In this section, we consider several such scenarios, with the goal of highlighting some ways in which variation in population density would be expected to affect the structure of social networks.

To begin, let us consider some fixed region, A, to which we sequentially add vertices; we assume that the location of each new vertex is chosen by some iid process, and that ties from old to new vertices are governed by the model of Equation 1. For any fixed vertex, v_i, the probability of a tie to a newly added vertex, v_j is given by

$$\begin{aligned}\Pr(Y_{ij} = 1|\ell_i, \theta) &= \int_A p(\ell_j)\mathcal{F}\left(D_{ij}, \theta\right) d\ell_j \\ &\equiv p_Y(\ell_i, \theta)\end{aligned}$$

where ℓ_i, ℓ_j are the locations of the respective vertices and $p(\ell_j)$ is the population distribution over space. Note that, by the iid assumption, the above works for any vertex, v_j, placed after v_i. Moreover, since all vertices are placed independently, we can without loss of generality take $v_i = v_1$ to be the first vertex placed. Then, after k vertices are added to the graph, the conditional "local" expected degree of v_i – i.e., the expected degree within the subgraph induced by membership in A

[2] Throughout this paper, we will use the term "spatial heterogeneity" to refer generically to variation in local population density across space. Other forms of heterogeneity are also possible (e.g., non-stationarity of the SIF), but we focus here on this particular case.

– is simply

$$\begin{aligned}\mathbf{E}\sum_{j=2}^{k} Y_{1j}|\ell_1 &= \sum_{j=2}^{k} \mathbf{E}Y_{1j}|\ell_1 \\ &= \sum_{j=2}^{k} p_Y(\ell_1,\theta) \\ &= kp_Y(\ell_1,\theta).\end{aligned}$$

Marginalizing over ℓ_1 gives the expected degree of v_i,

$$\mathbf{E}\sum_{j=2}^{k} Y_{1j} = \int_A kp_Y(\ell_1,\theta)d\ell_1,$$

which again by the iid assumption reduces to k times a function that depends only on the fixed vertex location pdf and the SIF. Indeed, the above right-hand expression can be rewritten as $k\int_A\int_A p(\ell)p(\ell')\mathcal{F}\left(D_{\ell,\ell'},\theta\right)d\ell d\ell'$, with the double area integral being the marginal probability of an edge between two randomly selected vertices.

This simple exercise leads to an observation that we may call the "in-filling principle": adding vertices to a fixed region in a uniform way leads to a linear increase in expected local (within-region) degree, while holding the expected local density constant. (This last follows immediately from the well-known graph identity $\delta = \bar{d}/(N-1)$, where δ is the density and $\bar{d}$ is the mean degree.) Likewise, given two regions of equivalent shape and area and the same local population distribution, we expect the ratio of their internal mean degrees to be equal to the ratio of their population sizes (with their internal densities again being equal). Where the population gradient is relatively uniform, we thus expect local mean degree to scale linearly with population density.

The in-filling principle has a number of further consequences, which may be appreciated by considering the impact of increasing mean degree on other structural properties. As average degree increases, the probability of connectivity rises; for $\bar{d} \approx \ln N$, almost all graphs are connected (see, e.g., West, 1996, pp413–417), and we may thus expect that small regions will become locally connected where $\ln N/N \ll \delta$, with $\delta = \int_A\int_A p(\ell)p(\ell')\mathcal{F}\left(D_{\ell,\ell'},\theta\right)d\ell d\ell'$ being the local expected density. (This threshold is not exact for non-uniform $\mathcal{F}$ due to spatial clustering, but does serve as a lower bound.) This effect is illustrated in the left-hand panel of Figure 2, which shows the probability of a regional network's being connected as a function of N and δ in the uniform case (red and green curves). Increasing N at constant network density also increases the probability of higher order connectivity (e.g., biconnectivity), leading to local populations that are robustly connected. The population threshold for biconnectivity is somewhat higher than that for connectivity (see Figure 2, yellow and cyan curves), but its behavior is otherwise similar.

The practical import of these observations is illustrated schematically in Figure 3. For an arbitrary region of constant shape having area Δ, we can identify some threshold population density $\omega(\Delta)$ such that uniform placement of $\Delta\omega(\Delta)$ vertices within such a region will lead to an expected mean degree exceeding the connectivity threshold for SIF $\mathcal{F}$. "Cutting" the population density surface at $\omega(\Delta)$ will reveal regions expected to be locally connected; specifically, induced subgraphs taken by selecting all vertices belonging to appropriately shaped regions of area Δ in the regions above the threshold are expected to be connected with high probability. Similar subgraphs taken from below-threshold regions are unlikely to be connected. Since the union of two non-disjoint connected subgraphs must itself be connected, the above also implies that large regions exceeding the ω threshold will tend to be globally connected (with connectivity extending to at least the

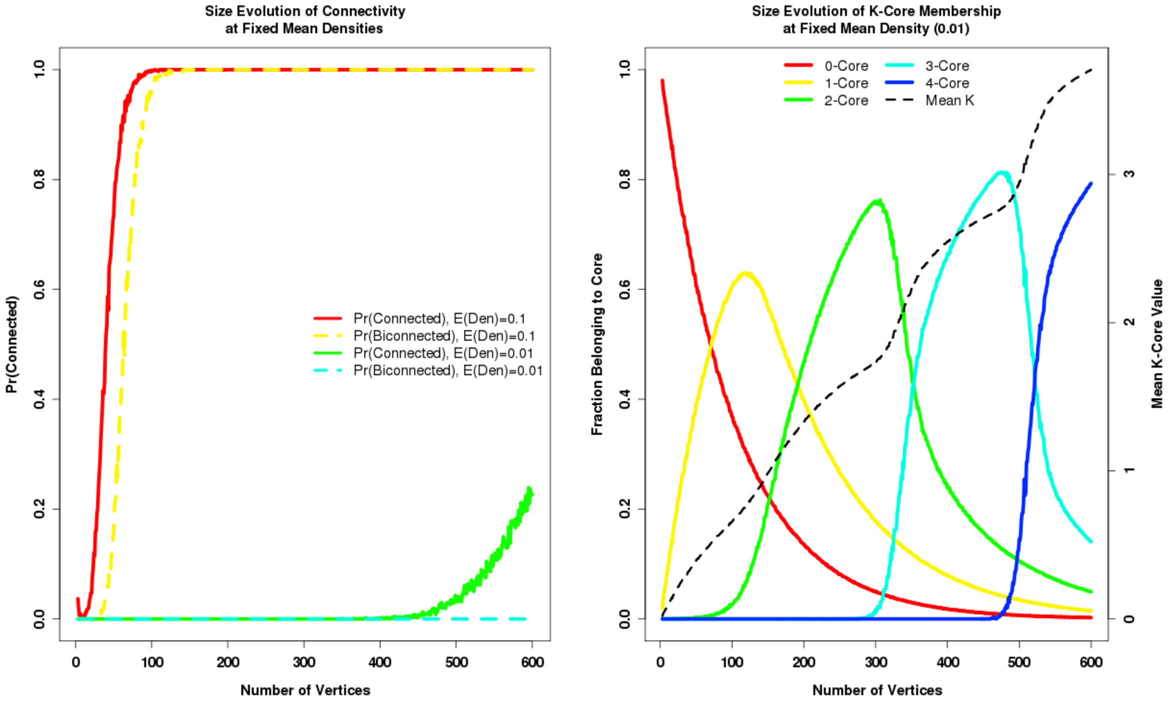


Figure 2: Effects of increasing order on connectivity and cohesion for random graphs of fixed expected density. Left panel shows probability of connectivity and biconnectivity by order and density. Right panel shows fraction belonging to each k-core (and no higher) and mean core number (dotted line) at 1% expected density, by graph order.

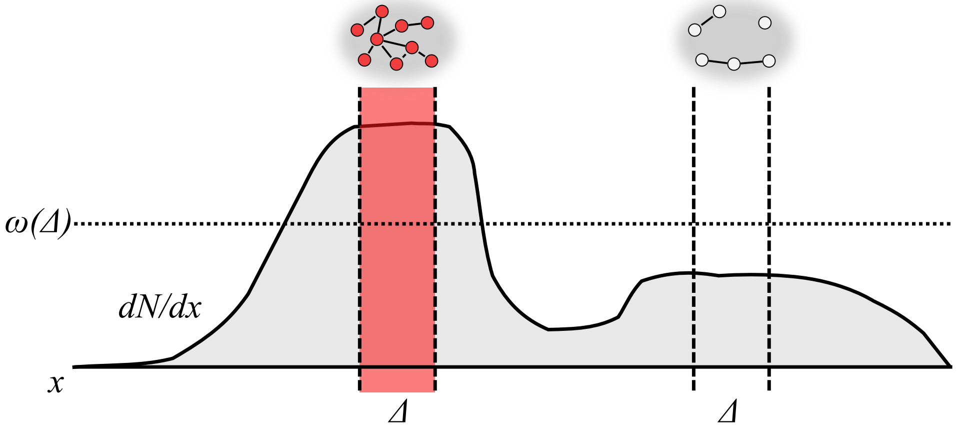


Figure 3: Emergence of local connectivity on an uneven population density surface. Where the threshold population density for an approximately uniform region of area Δ is $\omega(\Delta)$ (such that $\Delta\omega(\Delta) \gg \ln N/N$), local connectivity emerges where dN/dx exceeds the threshold for intervals of appropriate area. Similar intervals of lower population remain locally disconnected.

"edge" of the above-threshold region). Thus, the in-filling principle leads us to expect local social connectivity to vary systematically with the population density surface, and, moreover, for there to be a fairly clear spatial boundary between connected and disconnected areas.

In addition to connectivity per se, another property potentially impacted by in-filling is local cohesion. Many notions of subgroup cohesion exist (see, e.g., Wasserman and Faust, 1994, for a review), with most considering subgroups to be cohesive to the extent that they are locally dense, that members are not easily disconnected from the group, and/or that group members are socially proximate to one another. To take a simple example, the k-core of a graph is defined to be the maximal set of vertices such that all vertices within the set are adjacent to at least k other set members. Although the k-cores are globally cohesive in only a minimal sense—even high-order cores need not be connected, for instance—they can be seen as unions of well-connected groups. In particular, every member of the 1-core belongs to a connected component of size greater than 1, every member of the 2-core belongs to a biconnected component, etc. The number of the highest-order k-core to which a vertex belongs, then, is a convenient indicator of the extent to which it is embedded in a well-connected group (even if the core as a whole is poorly connected). We shall refer to this number henceforth as the "core number" of a vertex. At constant density, the formation of higher-order cores—and hence higher core numbers—is strongly associated with graph order. The right-hand panel of Figure 2 shows the development of core structure in a random graph of constant density as order increases. Initially, vertices transition from the 0-core (i.e., isolates) to the 1-core, with this process slowing as the 2-core begins to emerge. (Core membership as depicted here refers to the highest-order core, or core number.) This process is repeated as N grows, with successive cores emerging and being consumed by higher-order cores at regular intervals. Averaging across the population, however, we see that the mean core number (black line) rises steadily, growing approximately linearly with N. As with connectivity, these behaviors may be altered somewhat by spatial clustering. They provide a useful intuition, however, for the "baseline" impact of in-filling on local structure.

Taking the strong relationship of mean core number with N at constant network density together with the in-filling principle, we would expect that regions of a given shape and area having higher local population densities will exhibit higher mean core numbers than equivalent regions of low population density (again, in the sense of Figure 3. Since the scaling of mean core number with N is roughly linear (in contrast with the threshold behavior of connectivity), we also expect that variation in mean core number across the population surface will be much smoother than variation in local connectivity. At the same time, membership in cores of a particular order is likely to be rare until a particular population density threshold is reached (with that threshold varying depending on the order of the core in question). For phenomena expected to emerge only in subgroups of a particular level of cohesion, then, we may still expect qualitative shifts in behavior between high and low density regions. For phenomena assumed to vary quantitatively with local cohesion, direct proportionality to the local population seems a reasonable first guess.

Although considerable insight can be gleaned from studying stylized situations, the arguments of this section omit several potentially consequential factors. As noted earlier, clustering due to spatial effects should tend to increase the N needed for connectivity, relative to a homogeneous Bernoulli model with the same expected density. We can bound this effect as follows. Let G be the spatial Bernoulli graph on the vertices of region A associated with SIF $\mathcal{F}$. Define $\delta' = \min_{\ell,\ell' \in A} \mathcal{F}(D_{\ell,\ell'})$, and let G' be a homogeneous Bernoulli graph with parameter δ' on the same vertex set. Intuitively, every pair in A is adjacent with probability at least δ', and G' can be thought of as supplying a "lower bound" on the true graph G.[3] Now, define the "residual" Bernoulli graph R, with parameter matrix Φ such that $\Phi_{ij} = 1 - (1 - \mathcal{F}(D_{ij}))/(1 - \delta')$. Let $G' \cup R$ represent the union of G' and R, i.e. the random graph in which an edge appears iff it appears in G', R, or both. Since all edges of $G' \cup R$ are independent and occur with probability $1 - (1 - \delta')(1 - \Phi_{ij}) = \mathcal{F}(D_{ij})$, it follows that $G \sim G' \cup R$.

This "decomposition" of the spatial Bernoulli graph G into a homogeneous Bernoulli graph G' and a residual graph R can provide us with a rigorous tool for understanding the behavior of G in more complex settings. For instance, if some draw g' from G' is connected, then $g' \cup R$ is connected; thus, the probability of connectivity under G' is a lower bound on the probability of connectivity under G. More generally, let z be any graph statistic such that $z(x \cup y) \geq z(x)$ for graphs x and y having the same vertex set. Then clearly $\mathbf{E}z(G) \geq \mathbf{E}z(G')$ and, for any given value z_o of z, $\Pr(z(G) \geq z_o) \geq \Pr(z(G') \geq z_o)$. Since many statistics of interest (e.g., mean core number, probability of k-connectivity) satisfy the condition of z, it follows that we can employ the properties of G' to bound the behavior of G. On the other hand, not all properties are preserved under graph union (e.g., betweenness centralization), and the bound obtained in some cases may be too loose to be useful (e.g., if $\delta' \ll \delta$). Thus, this is an incomplete solution.

Another factor ignored thus far has been the role of non-local ties. Clearly, it is possible to identify some region A whose induced subgraph is disconnected with high probability, while the induced subgraph of some larger region $B \supset A$ is very likely connected—consider, for instance, the case of vertices distributed at unit intervals on the real line, with A being a segment of length 2, B being a segment of arbitrarily long length λ, and $\mathcal{F}$ a constant k such that $\ln \lambda/\lambda \ll k \ll 1$. Although this example is implausible, it points to a real phenomenon: particularly when $\mathcal{F}$ is heavy-tailed, the locally induced subgraph for a small region may not effectively characterize the properties of its members within the larger network. This effect will itself vary across the population surface, to the extent that high-population regions will tend to be embedded within or adjacent to other high-population regions, and low-population regions will likewise be associated with other low-population regions. The exact impact of this effect will vary with the model SIF, graph statistic examined, and population surface, and is difficult to characterize on an a priori basis.

[3] E.g., we can generate G and G' using the same random inputs, such that every edge in G is also in G'; see Butts (2010) for a general treatment of this approach.

Considering simplified scenarios gives us general insights into the first-order effects of spatial heterogeneity on network structure, but does not tell us how these factors will play out in practice. To move beyond these basic intuitions, we must examine the behavior of spatial Bernoulli graphs under realistic conditions: that is, with SIFs based on real data, and vertex positions that are reflective of geographical reality. It is to this problem that we now turn.

3 Bringing Geography Back In: A Simulation Study

While the arguments of the previous section suggest various general ways in which network structure should vary across space, they also underscore the fact that such effects are contingent on the underlying population surface: the same mechanisms can lead to very different networks when applied to populations that are differently distributed. What would we expect to see in real networks, then, given empirically observed settlement patterns? Simulation provides a natural means of addressing this problem. Specifically, we may utilize detailed data on population distributions to simulate draws from spatial Bernoulli graphs with fixed SIFs, and analyze the resulting networks to examine the impact of spatial structure on network properties (holding other factors constant); it is this approach that we employ here. Although the models we employ are empirically realizable (and our simulations based on extrapolations from observed data, rather than first principles alone), it should be borne in mind that our purpose is to explore the theoretical implications of spatial structure for network structure given a minimal set of assumptions, as opposed to making detailed predictions about particular cases. In this, sense, this effort lies at a midpoint on the "intellective" versus "emulative" modeling continuum discussed by Carley (2002), incorporating certain aspects of empirical detail while retaining enough simplifying assumptions to permit a relatively general analysis. At the same time, the generality of the framework in which our models are embedded (i.e., the spatial ERGs of Equation 2) permits subsequent extension and elaboration where appropriate.

3.1 Simulation Design

Our simulation study proceeds as follows. First, we choose a set of locations to examine, selecting them so as to evenly cover the range of observed population sizes and land areas for U.S. micropolitan and metropolitan areas (as well as a number of other, smaller, inhabited locations). For each location in this set, we then simulate population microdistribution using data on block-level population counts and household size distributions. Finally, we simulate networks using two previously calibrated spatial Bernoulli models for each location. By analyzing the resulting set of networks, we are then able to examine the impact of within and across-location geographical variability on network structure.

3.1.1 Location Selection

To compare network structure across regions, one would ideally like to employ units that are both well-defined and socially bounded. While few places in the developed world are fully isolated from one another, it is nevertheless possible to identify regions that correspond to relatively well-defined social units with respect to such processes as daily migration, employment, local commerce, and everyday interaction. Using such criteria, the U.S. Department of the Census divides the populated regions of the United States into *micropolitan* and *metropolitan* statistical areas, each of which contains one or more towns, cities, or other agglomerations together with the immediately surrounding area (to the nearest county or parish boundary, as applicable). A metropolitan area contains at least one city with a population of at least 50,000, whereas a micropolitan area contains

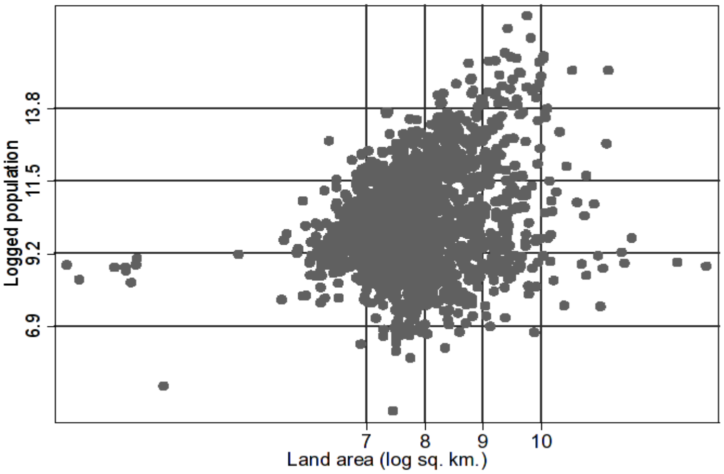


Figure 4: Population size and land area distribution, U.S. micro and metropolitan statistical areas. Vertical and horizontal lines indicate sampling strata for this study; selected locations are those closest to line intersections.

at least one city with a population between 10,000 and 50,000. An "area" in this sense includes the primary city and the surrounding county; it also includes adjacent counties if they are highly socially integrated based on journey to work patterns. At a minimum, micropolitan/metropolitan areas are thus collections of locally dense population surrounded by a low-population buffer, although the Census definition also seeks to avoid "splitting" areas experiencing strong interaction in other respects. The unified set of 922 micropolitan and metropolitan areas, along with the other 1,379 smaller counties in the United States, serve as the population for our study, with the individual area (henceforth "location") being our primary unit of analysis.

In order to cover areas with a wide range of geographical properties, we employ a stratified sample of locations selected by population size and land area (respectively). As location sizes on both dimensions scale roughly logarithmically, we identified four target strata on each dimension based on the overall distribution of locations. (See Figure 4.) These stratum values were 10^3, 10^4, 10^5, and 10^6 for population size, and $e^7 km^2$, $e^8 km^2$, $e^9 km^2$, and $e^{10} km^2$ for land area. For each pair of population size and land area values, the location was identified whose population and area was as close as possible to the target (in a least squares sense), yielding a total of 16 study locations. These are listed in Table 1. As can be seen, the selected locations run the gamut from rural and even near-wilderness locations to long-settled urban environments. By examining the properties in predicted network structure across this range of environments (holding constant other factors), we may assess the extent to which geographic variability would be expected to drive structural properties throughout the population of U.S. micro and metropolitan areas.

3.1.2 Population Microdistribution

For purposes of detailed network simulation, it is necessary to associate each individual within each location with a particular point in space. Such point placement can be immediately constrained by Census counts, which identify total population at the block level (an areal unit roughly analogous to a city block in urban areas, constructed on the basis of physical boundaries, population homogeneity, and population size). Within blocks, most individuals can be further localized to

	Population	Land Area (km^2)	Population Density (Pop/km^2)
Bristol Bay Borough, AK	1,258	1,308	0.96
Golden Valley, MT	1,042	3,044	0.34
Esmeralda, NV	971	9,294	0.10
Yakutat, AK	808	19,815	0.04
Choctaw, MS	9,758	1,085	8.99
Cheyenne, NE	9,830	3,099	3.17
Quay, NM	10,155	7,446	1.36
White Pine, NV	9,181	22,989	0.40
Lawrence, KS	99,962	1,183	84.48
Cookeville, TN	93,417	2,961	31.55
Idaho Falls, ID	101,677	7,676	13.25
Navajo, AZ	97,470	25,779	3.78
Honolulu, HI	876,156	1,553	564.03
Hartford, CT	1,148,618	3,923	292.77
Rochester, NY	1,037,831	7,592	136.69
Salt Lake City, UT	968,858	24,705	39.22

Table 1: Sample Locations, Stratified by Population and Land Area

households, whose size distribution within each block is also known. This leaves the placement of households and the placement of persons within households. In this paper, we refer to the resulting spatial distribution of individuals within blocks as the *population microdistribution.* To simulate the microdistribution, we utilize two competing models representing the extreme cases of a range of potential point processes. The first assumes a maximum entropy solution, in which households (or isolated individuals) are placed uniformly at random subject to known geographical constraints. The second, by contrast, assumes a near-minimal entropy solution, in which households are placed in an extremely even, "grid-like" manner using a low-discrepancy sequence (specifically, a two-dimensional Halton sequence[4]). We refer to these microdistribution models henceforth as "uniform" and "quasi-random" placement, respectively. By examining the impact of these extreme cases on social structure, we can thus infer the likely range of possibilities which could be occupied by processes having intermediate behavior. For both models, we also avoid unrealistic ground-level congestion by means of a simple artificial elevation model, which simulates the effects of multi-story residential structures in densely populated blocks. Specifically, households whose ground position would place them within a 10m radius of k previously placed households are given a vertical elevation of $4k$ meters; thus, intuitively, artificial elevation arises as population density grows, with new households "stacking" on old ones. (Arrival order is treated as random.) Finally, within-household proximity is maintained by requiring household size to satisfy the known marginals within each block, and then placing individuals at their household locations (jittering randomly within a 5m radius to avoid exact overlap).

An example of the result of this microsimulation process is shown in Figure 5. Both panels

[4] A Halton sequence is a deterministic sequence of points that "fills" space in a uniform manner, while also maintaining a high nearest-neighbor distance. The result (sometimes called a "quasi-random" distribution) is similar to a set of draws from the uniform distribution, but substantially more evenly placed; see Gentle (1998) for algorithmic details.

depict the Quay, New Mexico MSA, with lines designating the census block boundaries (2,690 in all). Vertex locations ($N = 10,155$) are indicated by colored dots, with artificial elevation denoted by dot color (ranging from red, at ground level, to violet at 20m). As can be appreciated from the large areas of empty space, most of the population is heavily concentrated into a relatively small fraction of blocks. Thus, population layout is to a great extent constrained by the available data. The difference between uniform (left panel) and quasi-random (right-panel) microdistribution models is seen in the relative evenness with which population is distributed within each block: a process of uniform (random) placement leads to numerous "clumps" in the household distribution, a phenomenon that is minimized under the quasi-random process. An immediate side effect of the latter is that the quasi-random process (approximately) maximizes the distance between households, a property that could be reasonably anticipated to affect network cohesion. Another side effect is a reduction in artificial elevation, stemming from the more efficient use of space under the quasi-random process. This is readily apparent in the Figure 5 insets, which show an expanded view of a 2km square portion of the city of Tucumcari, NM. In this relatively high-density region, both models predict that some degree of artificial elevation will occur; due to the higher rate of "clumping" in the uniform model, however, the later produces a larger number of elevated households. While our intent in this simulation is merely to produce a plausible level of inhibition for ties among individuals in high-density location (rather than to correctly capture the true distribution of residence heights per se), we note that the elevations produced by this process on our 16 locations are broadly consistent with the ranges of residential building heights typically reported in the literature (e.g., Burian et al., 2002).

3.1.3 Spatial Interaction Functions

For marginal models, the effect of geography on social interaction is captured through the spatial interaction function. While many choices of SIF are possible, our interest in this case is in functions that plausibly represent families of relatively proximate relations. To that end, we employ two specific functions obtained in prior work by Butts (2002) based on analyses of existing data sources. The first is from a "social friendship" relation (based on data from (Festinger et al., 1950)), and can be thought of as a locally sparse relation with a fairly long tail (declining as approximately $d^{-2.8}$ for large distances). The second is from a "face-to-face interaction" relation based on data from Freeman et al. (1988), which acts as locally dense relation that attenuates very quickly with distance (apx $d^{-6.4}$). Both are of the general power law form $\mathcal{F}(x) = \theta_1/(1+\theta_2 x)^{\theta_3}$, with parameter vectors $(0.533, .032, 2.788)$ in the case of the social friendship model, and $(0.859, 0.035, 6.437)$ in the case of the model for face-to-face interaction. While we do not assume that all networks—or, indeed, that all types of friendship or interpersonal interaction—follow one of these two forms, we take these as plausible examples of the types of SIFs one is likely to see from proximate relations such as those from which the functions were derived. Similarities and differences in the networks formed by such functions thus provide us with some sense of the range of behaviors one might observe in similar settings.

In computing distances for purposes of the SIF, we employ the Euclidean distances between individual surface positions in the projected geometry, plus a small correction to account for the effects of the built environment (where applicable). Specifically, differences in artificial elevation are added to the base Euclidean distance for those within 25m of one another, while the sums of individuals' artificial elevation distances are added for pairs whose surface positions differ by more than 25m. This simulates a basic feature of travel within the built environment, namely movement within a building for those who are otherwise proximate, versus movement to down ground level, over to the second position, and up for those with distant surface locations. While one could employ

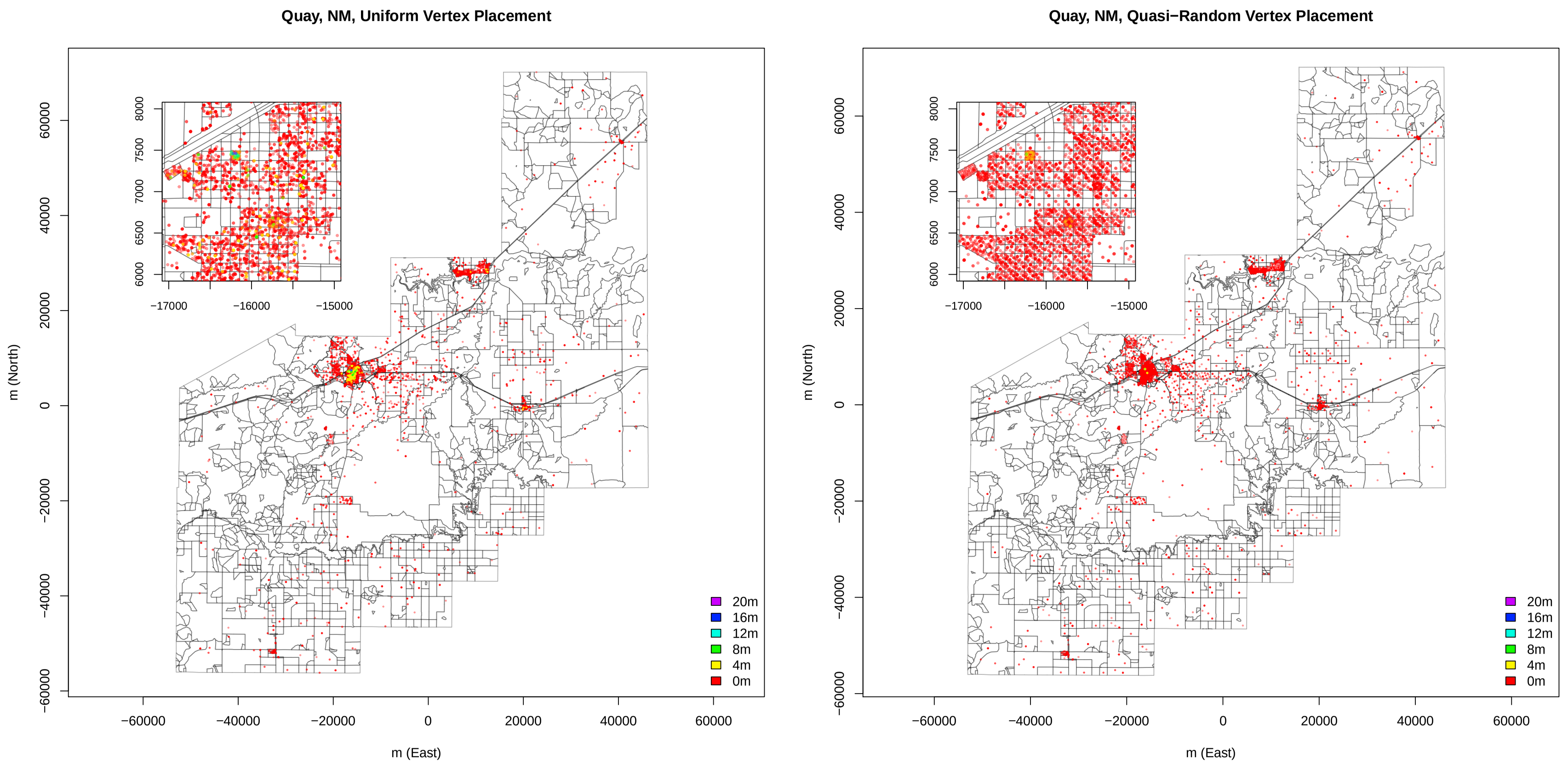


Figure 5: Comparison of uniform and quasi-random vertex placement, Quay County, NM MSA. Lines indicate census block boundaries, with artificial elevation shown via vertex color. Insets provide detail of 2km x 2km portion of Tucumcari, NM.

more complex schemes (including explicit adjustment for roadways, local obstructions, etc.), this would require more detailed knowledge of household position, built environment, and other aspects of local geography than were available for our test regions. As a practical matter, experimentation with reasonable alternatives to the approach employed here did not produce substantively different results. By way of explanation, it should be noted that different notions of distance tend to be very strongly correlated, even at fairly small scales. For instance, a comparison of Euclidean distances for the cases used here with Manhattan distances (the so-called "city block" metric sometimes suggested as an alternative for urban environments) produced a median correlation of approximately 0.99 under both uniform and quasi-random microdistribution models; even considering only very proximate points (within 100m of each other in Euclidean space), the median correlation is still approximately 0.98 for both microdistribution models.[5] While the impact of alternative distance models on network structure at smaller spatial scales is an interesting and potentially productive target for further research, our experiments suggest that the results reported here are not sensitive to reasonable variations in how distance is defined.

3.1.4 Network Simulation

Given the above, simulation proceeds as follows.[6] For every location, we generate a population microdistribution using each placement model (uniform and quasi-random), subsequently generating 25 networks from each microdistribution using the two spatial models. The resulting set of 1600 networks is the primary basis for our subsequent comparative analysis. In addition to this larger sample, an additional single network was drawn in each condition. This smaller set of 64 networks was retained for within-network analysis.

3.2 Results

Our analysis of the simulated networks includes an examination of both within and between location heterogeneity. Here, we discuss our major findings in each area, beginning with cross-locational comparisons.

3.2.1 Graph-Level Properties Across Locations

Given our widely dispersed set of study locations, and the high level of spatial heterogeneity within each location, it seems plausible that few systematic patterns will be found that span the sample of simulated graphs. If, on the other hand, we find that there are aggregate properties that remain stable – or that change in a predictable way – across locations, then we may provisionally conclude that these properties will be robust enough to justify empirical investigation (for networks with similar SIFs to those considered here, at least). In this section we consider several contexts in which graph-level relationships can occur: comparisons of spatially conditioned and uniform random graphs; associations of graph-level properties with aggregate geographical features; and correlations among multiple global properties on the same networks.

Comparison with Random Baselines

As we have noted, spatial structure affects network structure by adding heterogeneity to edge

[5] Since Manhattan distance is affected by the choice of coordinates, we also considered the correlation of Euclidean distance against Manhattan distance on a randomly rotated axis set. The resulting median correlations were nearly identical (apx 0.99 in the unconstrained case, and 0.97 within 100m).

[6] Simulation and analysis was performed using the `statnet` and `sna` libraries for R (Handcock et al., 2008; Butts, 2008) and the R spatial tools (Bivand et al., 2008), along with additional functions created by the authors.

probabilities, and by creating correlations among those probabilities (an effect which is distinct from conditional dependence among edges, as in Pattison and Robins (2002)). Nevertheless, the possibility exists that such changes will have only a limited impact on the global properties of the resulting networks. It has long been known that many global network properties are sharply constrained by basic factors such as size and density (Mayhew and Levinger, 1976; Anderson et al., 1999; Faust, 2007) and analyses such as those of Watts and Strogatz (1998) show that highly structured networks can behave much like random graphs in certain respects. Apart from their substantive adequacy, simple random graph models are also useful as baselines (Mayhew, 1984a) against which to compare the behavior of more complex models. To compare the behavior of networks generated under the spatial Bernoulli model with their homogeneous counterparts, we therefore constructed a paired comparison sample to our 1,600 spatially-conditioned networks. For each of our spatially-conditioned networks, we drew a single conditional uniform graph (CUG) with identical size and density to that in the original sample. This resulted in a sample of equal size to the original, whose corresponding members had the same size and densities (and, by extension, mean degrees) as the original networks, but which were free of spatial structure. To assess the ways in which space distorts global structure in the present context (above and beyond density), we compare the distributions of several graph-level indices (GLIs) on the spatial and CUG samples.

Figure 6 summarizes the relationships between global properties of the spatial and uniform networks, as captured by several standard graph-level indices. Each panel shows case pairs, with vertical and horizontal coordinates respectively indicating GLI values for the spatially-conditioned and CUG networks. While all of the selected GLIs show clear differences between the models, the nature and extent of the deviations vary. The top left panel of Figure 6, for instance, shows the standard deviation of the degree distribution for each simulated pair. Although the mean degrees in each case are constrained to be equal, their variations are not: as can clearly be seen, virtually all of the spatial networks are well above the 45-degree line, indicating that the degree distributions under the spatial models are substantially more variable (and, in practice, more right-skewed) than their random counterparts. Interestingly, this amplification of variability appears to be quite systematic, with degree standard deviation in the spatial model scaling as approximately the 5/3 power of the random baseline ($R^2 = 0.96$). Moreover, we see that this relationship appears to be generally homogeneous with respect both to the choice of SIF and to the micropopulation model. This would seem to imply that the form of the variance increase stems from a relatively robust property of the spatial model, rather than from any region-specific geographical features.

While degree variance is substantially and systematically different in the spatial and uniform cases, mean core number provides an example of a GLI that behaves fairly similarly in both cases. As the top-right panel shows, mean core number in the spatial sample is closely and linearly related to mean core number in the uniform sample ($R^2 > 0.99$), with the observations deviating only slightly from the 45-degree line. Although there is clearly a systematic difference—the spatial models have slightly higher mean core numbers when the baseline is low, and slightly lower mean core numbers when the baseline is high[7]—this difference is small compared to the overall variation in the statistic. Intuitively, we may understand strong correlation as arising from the fact that mean core number tends to be robustly related to degree, but that spatial correlation makes it slightly easier to assemble low-order cores (by grouping edges together in space) while also making it slightly more difficult to assemble high-order cores (by reducing connectivity between distant groups). As with variation in degree, this relationship appears quite fundamental, and shows little evidence of depending upon the particularities of the location at hand.

[7] Note that this is not a regression artifact; as the high R^2 suggests, the result does not reverse when one regresses the CUG score on the spatial model score.

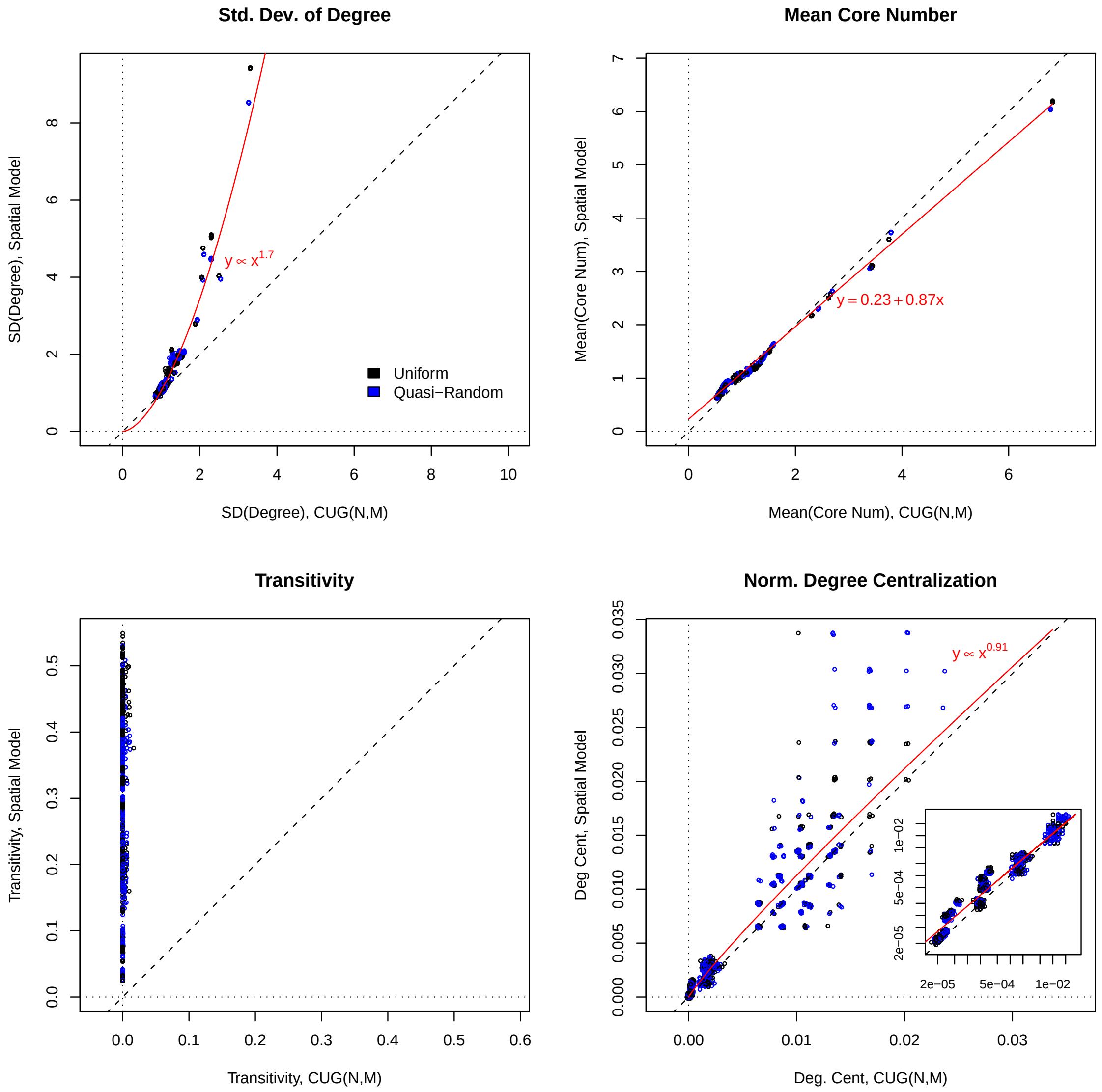


Figure 6: Graph-Level properties, spatial models versus paired CUG baselines. Each point represents a single simulation outcome (location by SIF by microdistribution), with color indicating choice of microdistribution. 45 degree line indicates equality; red lines, where applicable, show least squares prediction of spatial from baseline GLI values. Inset in last panel shows variables in log-log scale.

A strong contrast with the two previous cases is provided by transitivity, i.e., the fraction of completed two-paths. As the bottom left panel of Figure 6 shows, there is no clear relationship between transitivity in the CUG baselines and transitivity in the spatial networks, and the correlation between the two, while significant, is extremely small ($\rho = 0.085, p = 0.0006$). Indeed, this is one of the most substantial areas of divergence between the models, with transitivity being almost entirely absent in the baseline networks (median=0, IQR=3.0e-5) and very substantial under the spatial model (median=0.28, IQR=0.32). That the spatial model produces a higher rate of triadic clustering is easily understood (indeed, this was a major motivation for the latent space model of Hoff et al. (2002)), but the quantitative extent of the difference is nevertheless dramatic. Here, we thus have an example of a network property for which there is not only great divergence from the baseline but also for which the baseline model appears to have little predictive value.

Finally, we consider degree centralization, a case which lies between the previous extremes. The measure we employ here is the size and density normalized degree centralization of Butts (2006b), which reduces the impact of scale and mean degree factors on the centralization score. Since high levels of degree centralization in large networks require the assembly of extremely large stars, it is not surprising that levels are low overall; nevertheless, we see that centralization is on average slightly higher in the spatial versus non-spatial networks, a tendency which reduces as the baseline centralization grows. The relationship here is clearly weaker than in the first two cases (R^2 of 0.96 in log-log scale, but only 0.88 on the original scale), suggesting a greater role for idiosyncratic effects. On the other hand, the overall relationship is still strong enough to indicate a close connection between the baseline behavior and the behavior of the spatial model.

To summarize, we find overall that there are fairly systematic relationships between the overall behavior of the spatial model and the behavior of the corresponding baselines—implying that one of the more important properties of the spatial model in any given context is its expected density, a point to which we will later return. This does not imply that the features of networks drawn from the spatial model are necessarily similar to those of random graphs with equivalent dyadic characteristics (although they are somewhat in the cases of mean core number and degree centralization), but that the spatial model deviates from the baseline model in a quantitatively predictable way. That such relationships would arise across locations of such highly variable structure is an interesting finding, and suggests an avenue for gaining further theoretical leverage. On the other hand, it is also true that some network properties differ greatly between the spatial and uniform cases, with little linkage between the two. Transitivity seems to be a clear example of this, having behavior in the spatial context that is essentially decoupled from the size/density controlled baseline. As we shall see, however, this does not mean that transitivity is wholly unpredictable. Rather, the way in which it arises in spatial networks is rather different from its behavior in uniformly mixed graphs.

Aggregate Geography and GLIs

As the baseline comparison above suggests, many aspects of our simulated networks seem to be well-ordered, and their global behavior surprisingly predictable from baseline properties. Given this, we may be inclined to ask whether we can predict some aspects of global network structure from aggregate properties of the regional geography on which they are based (without resorting to detailed knowledge of the population surface). As it happens, there are a number of GLIs which can be at least roughly predicted from basic locational characteristics, several of which we summarize here.

Across scenarios, mean degree is strongly related to the mean (spatial) nearest-neighbor distance. Indeed, a simple power law model based on this alone (with an interaction for choice of SIF) explains approximately 79% of the variance in mean degree. This may be understood by a

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
(Intercept)	0.47	0.035	13.39	<2e-16	$***$
FriendSIF	2.37	0.050	47.26	<2e-16	$***$
log(PopDen)	0.11	0.004	30.78	<2e-16	$***$
log(FFunMed)	0.13	0.006	22.45	<2e-16	$***$
FriendSIF*log(PopDen)	0.28	0.005	55.74	<2e-16	$***$
FriendSIF*log(FFunMed)	0.25	0.008	30.56	<2e-16	$***$
$***p < 0.001$, $**p < 0.01$, $*p < 0.05$					
Residual standard error: 0.1717 on 1594 degrees of freedom					
Multiple R^2: 0.9314, Adjusted R^2: 0.9311					
F-statistic: 4326 on 5 and 1594 DF, p-value: $<$ 2.2e-16					

Table 2: Regression of logged mean degree on population density (PopDen) and median empty space function (FFunMed), with an interaction for model type.

"sparse/local" approximation, in which it is assumed that the probability of being tied to higher-order neighbors is small compared to the probability of being tied to the nearest neighbor; in the limit, mean degree under this model becomes determined by the nearest-neighbor distance (and takes an approximate power law form because of the choice of SIF). An alternative approach is to assume that degree will scale with the local population density, and that this will in turn depend upon the degree of global clustering (in the sense that the larger the regions of "empty space," the higher the population density must be in the remaining areas). One natural measure of the latter is the median of the F (or "empty space") function (Ripley, 1988), which is the median radius that one must go from a randomly selected coordinate within the sample area until one encounters a vertex. A power law model incorporating population density, the median empty space function, and an interaction for SIF choice explains 96% of the mean degree variance (93% on log scale), while still being quite parsimonious. This model is shown in Table 2. As we saw earlier that mean core number generally grows linearly with degree (a theoretical result which we will revisit below), we may expect this model to predict the former fairly well; in fact, using identical predictors on log mean core number (not shown) yields an R^2 of 0.93. Thus, both local cohesion and degree are fairly well explained by a combination of macro-clustering and population density.

Given the ability to explain mean degree from basic geographical characteristics, density is explained even more directly. First, we note from the earlier mentioned identity that density must be equal to $\bar{d}/(N-1)$; since population varies here by three orders of magnitude, while mean degree varies by less than a single order of magnitude, the simple relation $\delta \propto 1/N$ is expected to be fairly predictive. Indeed, this is the case: the correlation between density and the inverse of population is approximately 0.96 ($R^2 = 0.92$), and we can thus immediately account for most of the variation in density by population alone. This prediction can be improved upon by using what we have already discovered regarding mean degree. Combining the predictors of Table 2 with the logged population count yields an adjusted R^2 of over 0.99 for logged density, accounting for over 99% of the variance in the statistic in both the logarithmic and original scale. (Because of the transparent similarity of this model to the degree model, we omit it here.)

We saw earlier that transitivity was a graph property whose behavior in the spatial case was very far from—and generally unrelated to—the conditional uniform baseline. This does not mean, however, that transitivity is unpredictable. Intuitively, we should expect transitivity to be increasing in overall population density, and decreasing with nearest neighbor distance. A simple

	Estimate	Std. Error	t value	Pr(> $\lvert t \rvert$)	
(Intercept)	-2.75	0.141	-19.56	<2e-16	$***$
log(NNDist)	0.84	0.060	14.00	<2e-16	$***$
log(PopDen)	-0.09	0.013	-6.99	4.13e-12	$***$
FriendSIF	-5.11	0.199	-25.71	<2e-16	$***$
log(NNDist)*log(PopDen)	0.05	0.004	12.53	<2e-16	$***$
log(NNDist)*FriendSIF	1.35	0.085	15.89	<2e-16	$***$
log(PopDen)*FriendSIF	-0.18	0.019	-9.53	<2e-16	$***$
log(NNDist)*log(PopDen)*FriendSIF	0.07	0.006	12.31	<2e-16	$***$
$***p < 0.001$, $**p < 0.01$, $*p < 0.05$					
Residual standard error: 0.2433 on 1592 degrees of freedom					
Multiple R^2: 0.9345, Adjusted R^2: 0.9343					
F-statistic: 3247 on 7 and 1592 DF, p-value: < 2.2e-16					

Table 3: Regression of logged transitivity on population density (PopDen), mean nearest neighbor distance (NNDist), and SIF, with all pairwise interactions.

log-transformed model with these terms (and an interaction for SIF) explains just over 89% of the variance in (log) transitivity, but the effects are not exactly as one would expect: while population density does correlate positively with transitivity, nearest neighbor distance is positive as well! The solution to this apparent puzzle appears to lie in the fact that graphs with short mean nearest neighbor distances provide more opportunities for creating bridging ties, thereby generating intransitive two-paths. (This intuition is reinforced by the observation that transitivity is lower under the Social Friendship SIF, precisely because this more spatially diffuse model allows ego to connect to relatively distant alters that are thereby unlikely to be tied to one another.) Allowing an interaction term between density and nearest neighbor distance boosts the R^2 to 93% (92% on original scale) without adding an excessive number of parameters (though interpretation of individual parameters is difficult due to the multiplicative effects). We provide this more complex (but conceptually similar) model in Table 3.

When networks are embedded in space, we may note that they acquire additional properties that are due neither to network structure nor space alone, but rather to their superposition (Butts, 2002). One simple example of such a property is the mean edge length, meaning the average spatial distance between endpoints of a randomly selected edge. Intuitively, we may expect mean edge length to scale with the characteristic length of each location (i.e., the square root of the location area), and a simple log-scale model with this term and an SIF interaction accounts for 86% of the variance in the log edge length statistic. Interestingly, population provides an even better model, leading to an R^2 of just under 0.95. (See Table 4) This latter effect may be due to the fact that population tends to fill space unequally, and that one tends to get substantial numbers of actors at long distances (and hence the opportunity for long edges) for MSAs with high population counts.

To summarize the forgoing, many global properties of the networks formed under the spatial model appear to be predictable from basic characteristics of the underlying geography. Choice of SIF is also of critical importance, here, unlike in the case of spatial/baseline comparisons, as it governs the nature of the interaction between the geographic variables and the properties of the resulting network. On the other hand, choice of population micro-distribution does not appear in any of the above models, having been found to add little or no explanatory power in any of the above cases. This is consistent with the notion that, while low-level geographical details may affect

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
(Intercept)	0.57	0.051	10.97	<2e-16	$***$
log(Pop)	0.10	0.005	20.72	<2e-16	$***$
FriendSIF	0.83	0.072	11.60	<2e-16	$***$
log(Pop)*FriendSIF	0.18	0.007	26.98	<2e-16	$***$
$***p < 0.001$, $**p < 0.01$, $*p < 0.05$					
Residual standard error: 0.3461 on 1596 degrees of freedom					
Multiple R^2: 0.9468, Adjusted R^2: 0.9497					
F-statistic: 9461 on 3 and 1596 DF, p-value: $< 2.2\text{e-}16$					

Table 4: Regression of logged transitivity on population density (PopDen), mean nearest neighbor distance (NNDist), and SIF, with all pairwise interactions.

those in particular parts of the network, the global features of regional scale networks are driven more by general properties of the population surface than by its intimate details.

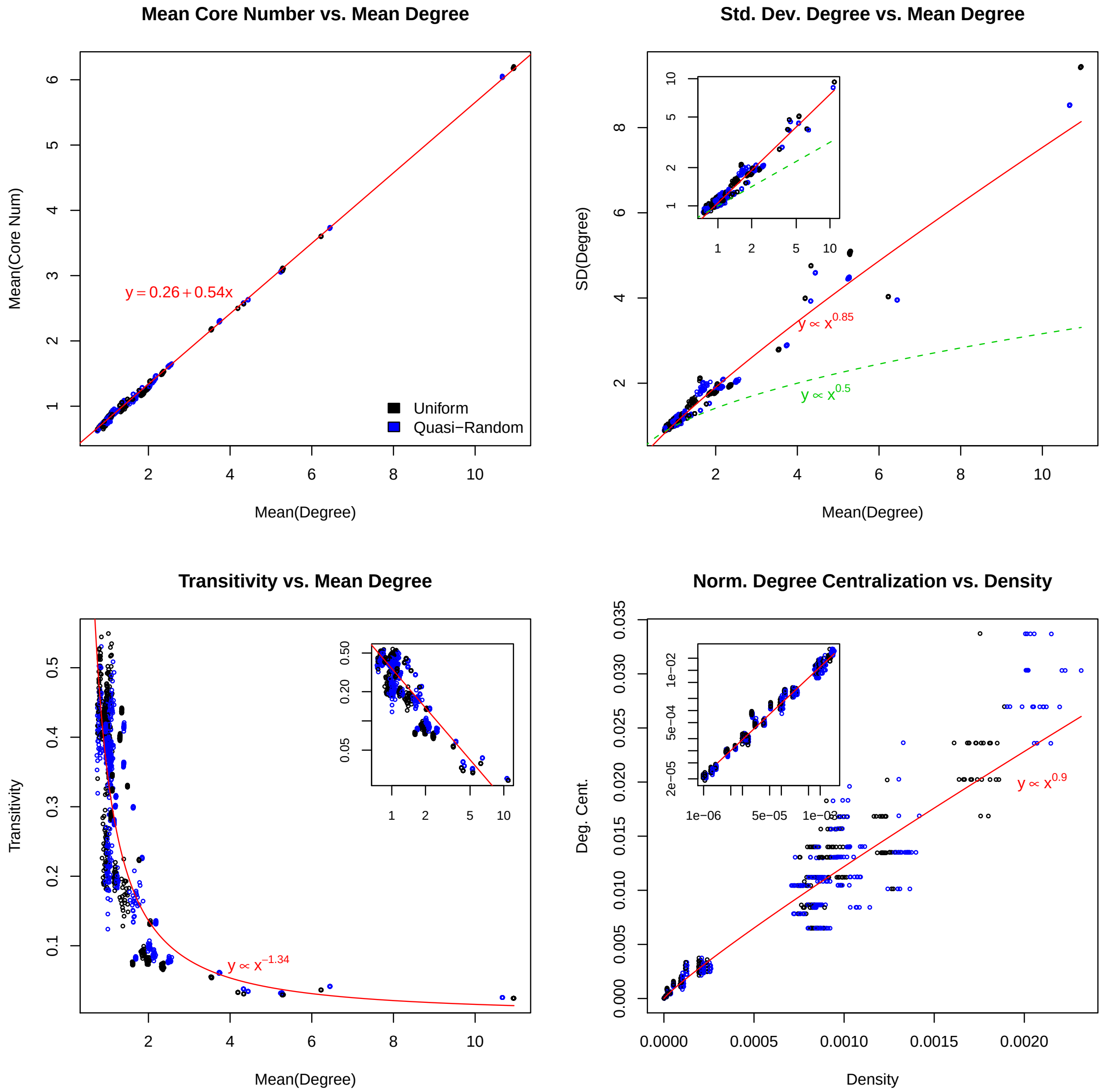
Internal Relationships Among GLIs

While there is a direct relationship between geographical variables and GLIs, it is also important to note that GLIs are also related to each other—this can in some cases prove to be a useful predictive device in its own right (i.e., constraining one graph feature allows one to predict another), and it can also provide insight into the factors that drive the creation of network structure. Before proceeding to a consideration of within-location heterogeneity, then, we summarize some particularly noteworthy relations among global properties for networks in our simulation sample.

We begin by considering mean degree. As with other (non-spatial) networks, we find that mean degree is itself a powerful predictor of other structural properties. Notably, the mean core number correlates with mean degree at greater than 0.99, with the relationship being strongly linear. (See Figure 7, top left panel.) Likewise, the marginal probability of belonging to the 2-core (and thus to a bicomponent) is essentially determined by mean degree (the logit of the 2-core membership probability correlates with the log mean degree at greater than 0.99). These effects are qualitatively similar to what one observes in homogeneous random graph models, despite the dramatically inhomogeneous structure induced by geography. We interpret this as being consistent with the notion that, at large levels of aggregation, much of the impact of geography is on its determination of baseline network parameters; although spatial structure clearly induces many other "non-random" features (e.g., triadic closure), the baseline effects are often powerful enough to constrain many aspects of network structure. (See also similar arguments by Mayhew (1984a); Butts (2001); Faust (2007), and related results shown above.)

We saw earlier that conditioning on mean degree, the standard deviation of degree increased at an accelerating rate with the expected baseline standard deviation. The top right panel of Figure 7 shows that the standard deviation of degree likewise grows systematically (but sublinearly) with the mean degree itself ($R^2 = 0.96$). While the degree standard deviation grows more slowly in these models than the mean, it nevertheless grows more rapidly than the square root of the mean (dotted line), and hence is inconsistent with a Poisson distribution. We will return to this issue in the next section, when we directly evaluate the degree distributions for the full set of networks.

We saw previously that transitivity was related to many of the same geographical properties that predict mean degree, and the relationship between the two GLIs is clarified in the lower left panel of Figure 7. Transitivity *falls* smoothly with mean degree ($R^2 = 0.86$), an effect related (as noted



Mean Core Number vs. Mean Degree

Std. Dev. Degree vs. Mean Degree

Transitivity vs. Mean Degree

Norm. Degree Centralization vs. Density

Figure 7: Relationships among various graph-level properties, simulated spatial networks. Each point represents a single simulation outcome (location by SIF by microdistribution), with color indicating choice of microdistribution. Red lines, where applicable, show least squares prediction of vertical axis GLI from horizontal axis GLI. Insets show variables in log-log scale.

earlier) to the fact that the same circumstances giving rise to high degree also give rise to an increase in local bridges, and thus to intransitivity. The transitivity/mean degree relationship helps to clarify the fact that, for these networks, global transitivity is a feature of highly fragmented, extremely local structures in which only small, spatially proximate clusters tend to be well-connected.

Finally, the upper right panel of Figure 7 shows that a moderately strong relationship exists between normalized degree centralization and density ($R^2 = 0.98$ on log scale). It is noteworthy in this regard that a similar relationship does *not* hold for mean degree ($R^2 = 0.13$ on log scale), bearing in mind that density is more heavily driven by population size than by mean degree in this sample. It should also be borne in mind that the density measure employed here is already normalized for density as well as graph size, and hence the observed density relationship is more subtle than it might appear. We note in this regard that the highest density graphs are simultaneously those that are small and that have higher mean degree. These graphs will be more likely to, on the one hand, produce high-degree outliers (since they have relatively higher degree variance), and, on the other, to have maximum raw centralization scores (with respect to which the measure is normalized) that are small enough that the normalized measure is not damped heavily towards zero. Complex measures such as centralization thus require careful interpretation (even when "adjusted" for baseline effects), no less so for these than for other network models.

In summary, strong and systematic relationships exist between GLIs within this sample. Some appear to be driven by baseline effects; their presence here reinforces the observation that even strong underlying heterogeneity does not immediately overcome combinatorial considerations. Other relationships, however (like that between mean degree and the degree variance) are clearly the result of spatial effects per se. It is interesting to note that neither SIF choice nor microdistribution model were found to be important mediators of the relationships among GLIs for these networks, even for those with an obviously spatial character; this suggests at the very least that the relationships are reasonably robust to detailed modeling assumptions, and perhaps that they are reflective of very general effects arising in a range of settings (spatial and otherwise). While these simple—often linear or multiplicative—relationships among global network properties thus demonstrate an order arising from a heterogeneous foundation, they do not tell the whole story. As we shall now show, the detailed structure of these networks clearly shows the impact of geographical factors.

3.2.2 Within-Location Heterogeneity

Having shown consistent behavior in the cross-location context, we now look within each network (using the smaller sample of 64 example structures) to examine the influence of geographical heterogeneity on local network structure. Consistent with the arguments of Section 2, inspection of the 64 networks reveals that local properties such as degree and core number vary systematically across space. As Figures 8 and 9 illustrate, the frequency of persons with high degree and high-order k-core members increases as one approaches regions of high population, with those in relatively sparsely populated areas tending to have both lower degree and lower levels of local cohesion. These relationships are as we would expect from first principles. While there is a strong spatial relationship with both degree and core number (both tracking the population surface), it should be noted that the highly clustered nature of population at the micro-level (in part associated with family structure) helps to ensure that some persons of at least moderate degree can be found even in outlying areas (and some of low degree in areas of high population). In practice, this mixing would mean that spatial degree variation could easily go unperceived by residents of a given area, even where the mean differences between dense and sparse regions are fairly substantial. Core number is even less likely to be perceived, since this would require extensive knowledge of the local network (and the ability to carry out the necessary calculations). While these effects may be invisible to

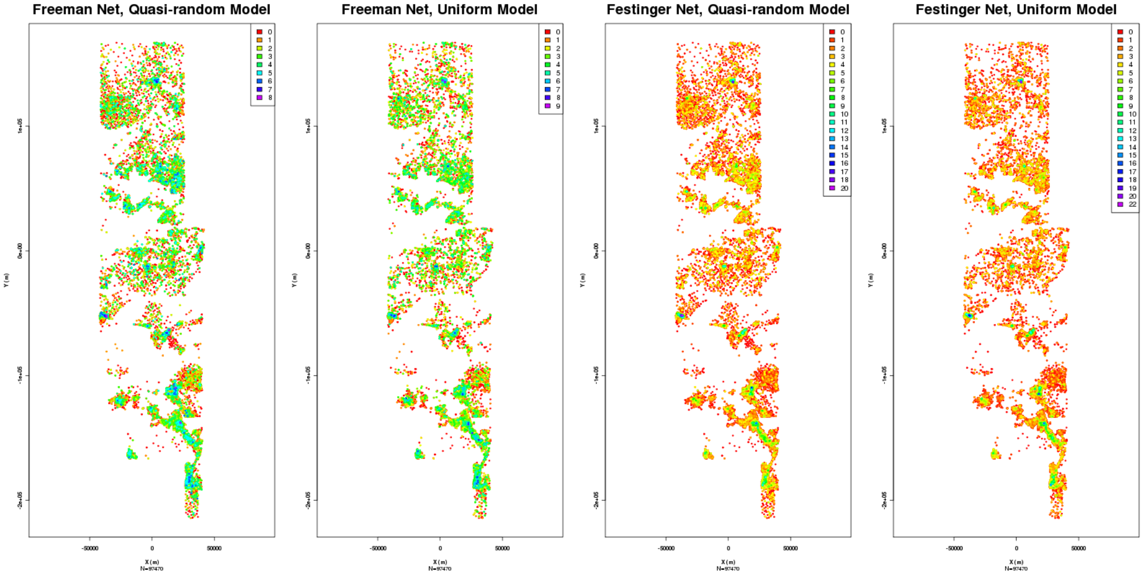


Figure 8: Spatial variation in predicted Degree (by SIF and layout model), Navajo County, AZ. Vertex color indicates degree, with bluer colors indicating higher greater numbers of ties.

residents, they could nonetheless be verified by appropriate empirical studies (though the authors know of no extant data set that speaks to this issue).

Also in line with the arguments of Section 2, the wildly unequal distribution of population across space leads to dramatic differences in local connectivity and tie volume. This is graphically illustrated in Figure 10, which shows ties among individuals in blocks near the center of Cookeville, TN (quasi-random distribution, Friendship SIF). While activity is present throughout the region, the intense clustering of persons in blocks like that near the center of the figure creates a corresponding social cluster whose members have both higher mean degree and who are on average more cohesively connected than those in nearby blocks. Even at scales on the order of 1km, we thus expect to see substantial heterogeneity in structural characteristics that are driven in part by geographical variation.

As the above example suggests, subgraphs in a spatial context have a dual existence: they can be considered on the one hand in terms of their network properties, and on the other in terms of the spatial positions of their members. In this light, we note that the convex hull of the set of vertices in a particular subgraph provides an intuitive notion of the region "covered" by that subgraph, as illustrated in Figure 11. The Figure graphically illustrates the emergence of large cohesive subgroups (here, members of high-order cores who are themselves biconnected) within the Cookeville, TN case. As argued in Section 2, such groups should develop relatively suddenly when a sufficiently large area exceeds the requisite threshold density; the location of large cores "covering" the high-density regions of the map is consistent with this behavior. Such spatially large cohesive sets are of potential interest for theories such as those of Sampson et al. (1997), which relate to the ability of social groups to monitor and control activities within a given area. Models of the kind studied here suggest a relatively sharp boundary between the conditions under which such cohesion is feasible, and those under which it is not. Such boundaries may account in part for the frequently voiced sense of qualitative difference between social interactions in cities and those in sparsely populated environments.

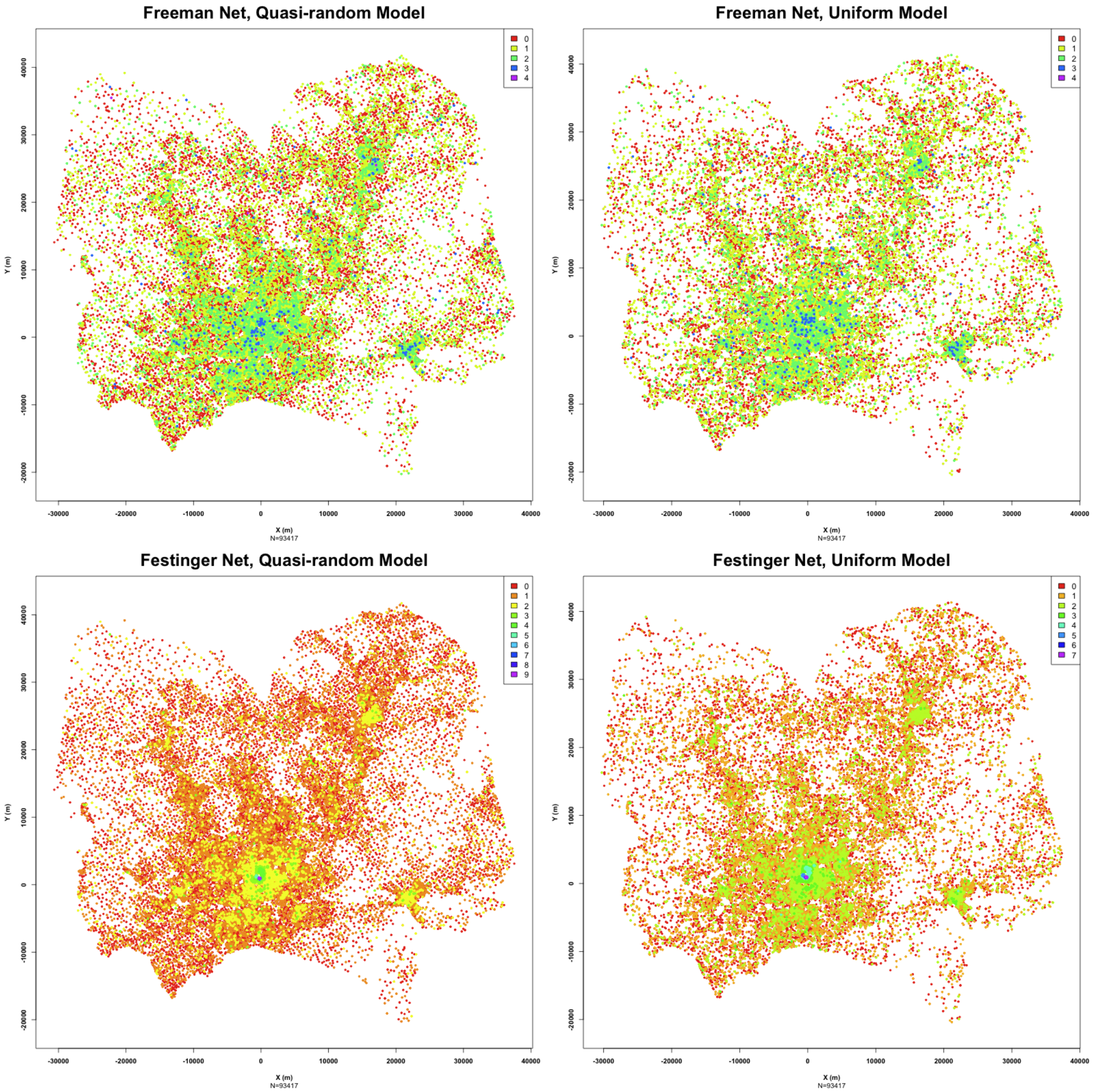

Freeman Net, Quasi-random Model
Freeman Net, Uniform Model
Festinger Net, Quasi-random Model
Festinger Net, Uniform Model
0
1
2
3
4
5
6
7
8
9
Y (m)
X (m)
N=93417
-30000
-20000
-10000
0
10000
20000
30000
40000

Figure 9: Spatial variation in predicted core number (by SIF and layout model), Cookeville, TN MSA. Vertex color indicates maximum core membership, with bluer colors indicating membership in higher-order cores.

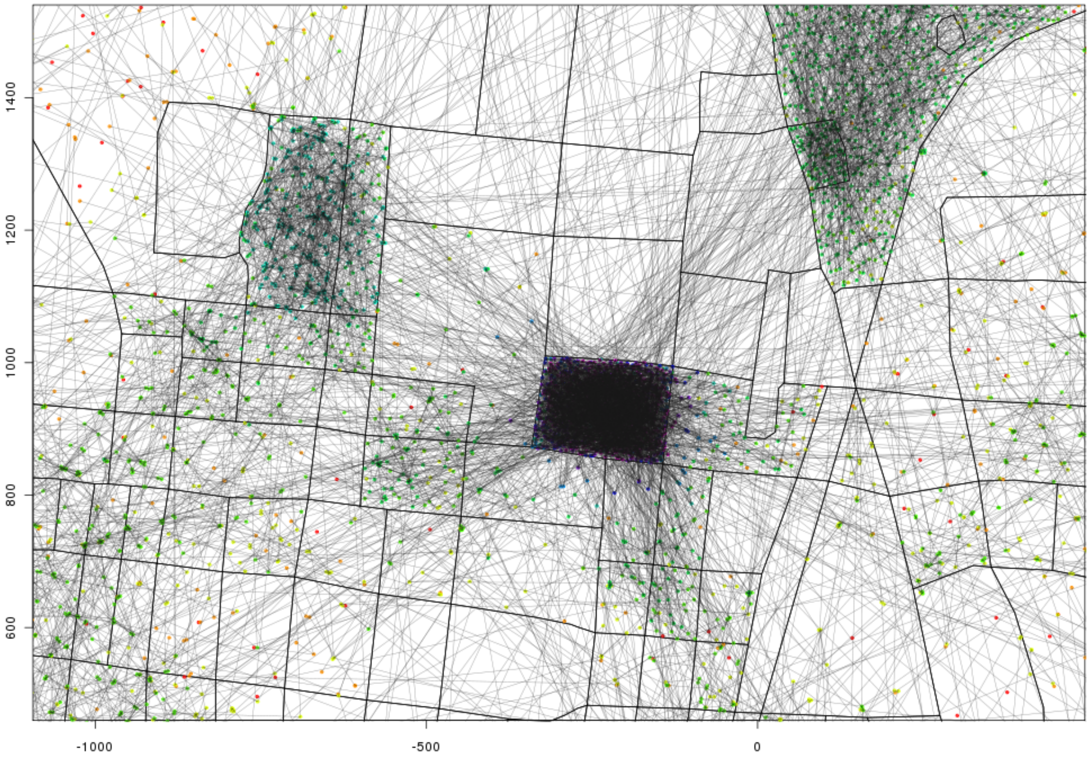


Figure 10: Detail of edge structure (quasi-random placement, Friendship SIF) for a portion of the Cookeville, TN MSA. Vertices are shaded by core number, from red (minimum) to violet (maximum); dark lines indicate census block boundaries.

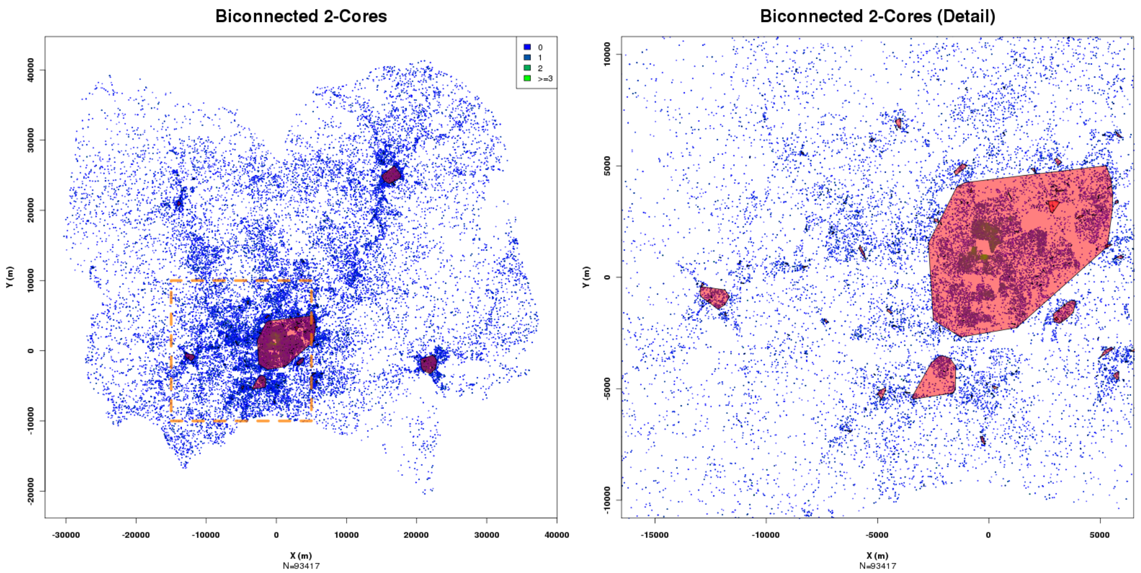


Figure 11: Spatial structure of cohesive components, Cookeville TN MSA (uniform placement, Friendship SIF). Shaded regions indicate convex hulls of membership locations for biconnected sets of k-core members, with pink shading indicating 2-cores, and green indicating 3-cores. Right-hand panel shows detail of dotted area.

While the detailed structure of the simulated networks reveals substantial variation in both positional and group characteristics, there are nevertheless strong similarities across cases. Figure 12 shows the marginal degree distributions for all 64 networks, each representing the aggregate effect of the interaction of population surface and SIF across vertices. While the profiles of Figure 12 obviously differ in some respects, there are also some similarities (not the least of which being a relatively long upper tail). Are these similarities only a matter of appearance, or do they indicate a common underlying functional form? To assess this, we attempted to fit geometric, negative binomial, Poisson, Waring, and Yule distributions to each of the depicted degree distributions (models fit via `degreenet` (Handcock, 2003) using maximum likelihood). After fitting all five distributions to each network, we select the model that best approximates the observed distribution (in the sense of minimizing the expected Kullbach-Leibler distance) using the AICC statistic. The resulting classification of network by model is provided in Table 5. The outcome of this analysis is overwhelmingly clear: in 59 out of 64 cases the preferred model is the negative binomial, with the long-tailed Waring distribution preferred in four and the Poisson preferred in only 1. The Waring cases seem to be associated with the Hartford, CT and Cookeville, TN MSAs under the Friendship SIF, and may reflect particularly high levels of heterogeneity in these locations. These possible exceptions aside, the vast majority of cases can be seen to be well-approximated by distributions of the same form, despite differing in population size and land area by several orders of magnitude.

Turning to core number, we note in the marginal distributions of Figure 13 the same combination of family resemblance and difference in detail seen earlier in Figure 12. As before, we attempt to assess the presence of a common underlying distributional form by fitting models to each distribution, selecting that chosen by the AICC. The results of this process are shown in Table 6. Once again, we find that the negative binomial is overwhelmingly preferred, with the geometric distribution (itself a special case of the negative binomial) favored in two cases (both Hartford, CT

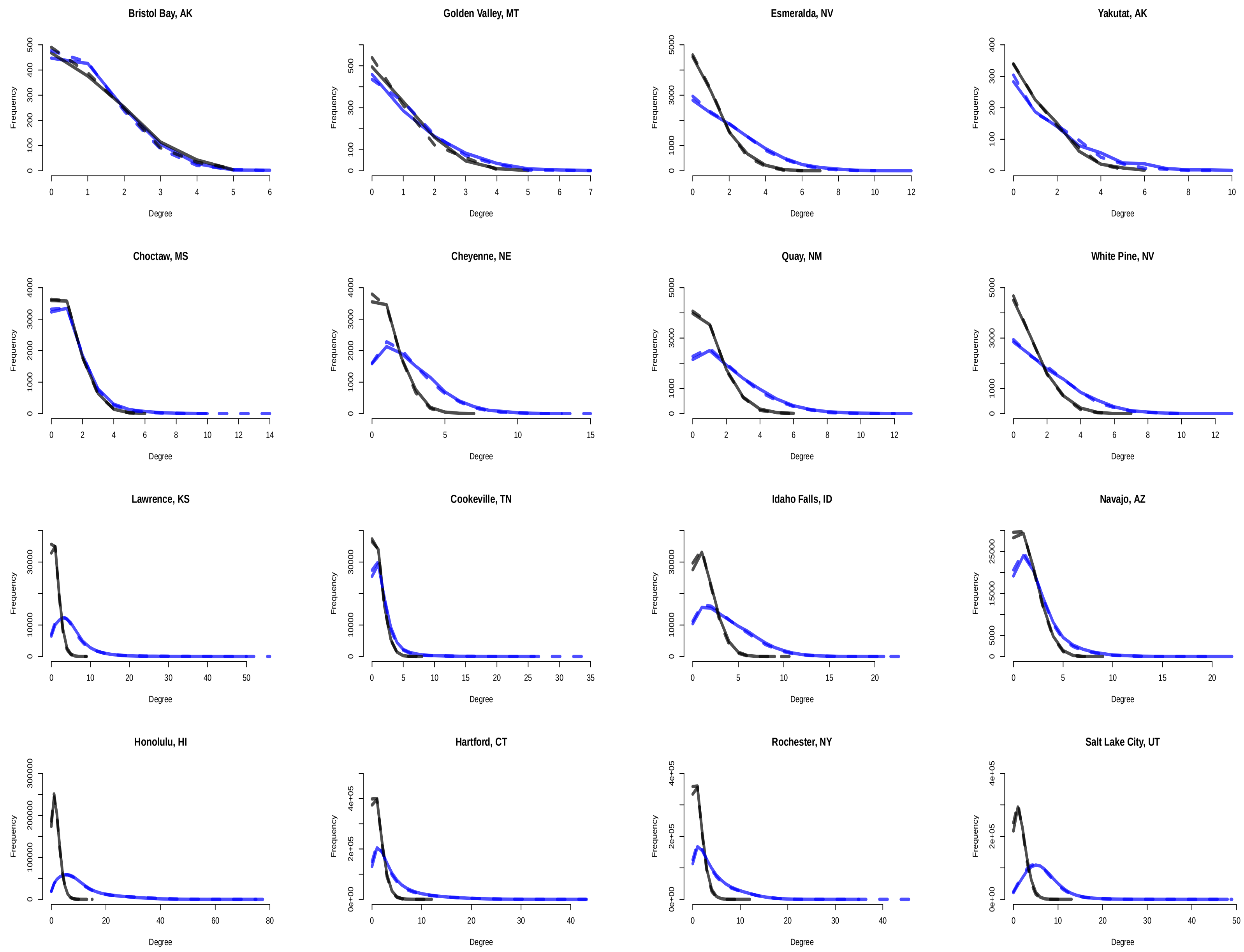


Figure 12: Marginal degree distributions by location, SIF, and placement model. Friendship model distributions are shown in blue, interaction model distributions in black; solid lines indicate uniform placement, with quasi-random placement in dotted lines.

	Freeman/Uniform	Festinger/Uniform	Freeman/Quasi	Festinger/Quasi
Bristol Bay, AK	NB	NB	NB	NB
Golden Valley, MT	NB	NB	NB	P
Esmeralda, NV	NB	NB	NB	NB
Yakutat, AK	NB	NB	NB	NB
Choctaw, MS	NB	NB	NB	NB
Cheyenne, NE	NB	NB	NB	NB
Quay, NM	NB	NB	NB	NB
White Pine, NV	NB	NB	NB	NB
Lawrence, KS	NB	NB	NB	NB
Cookeville, TN	W	NB	W	NB
Idaho Falls, ID	NB	NB	NB	NB
Navajo, AZ	NB	NB	NB	NB
Honolulu, HI	NB	NB	NB	NB
Hartford, CT	W	NB	W	NB
Rochester, NY	NB	NB	NB	NB
Salt Lake City, UT	NB	NB	NB	NB
G=Geometric, NB=Negative Binomial, P=Poisson, W=Waring, Y=Yule				

Table 5: AICC selected models for degree distribution, by location, SIF, and placement model.

Interaction SIF models) and the Waring favored in one. For core number, as for degree, then, the unique pattern of variation in each individual population surface nevertheless combines to generate a consistent family of marginal distributions.

To summarize our findings regarding within-region/within-network variation, the results from our simulated networks closely follow our a priori expectations. Degree and core number vary with the population surface, and cohesively connected subgroups appear over regions with systematically above-threshold density. While all of this suggests that many local structural properties will vary greatly both within and across regions, this variation is also systematically structured. In addition to the above patterns, we find that both degree and core number for these networks are well-modeled by negative binomial distributions (though the parameters of those distributions clearly vary by location and SIF). Although it is not clear how robust this result is to the imposition of other (currently unmodeled) factors, its prevalence here leads to the speculation that the negative binomial degree and core number distributions may serve as easily falsifiable signatures for a spatial Bernoulli process. It is interesting in that light to note that the negative binomial has been found to provide a reasonable fit to at least some empirical data sets (see, e.g. Hamilton et al., 2008), suggesting that this pattern is not beyond the bounds of plausibility.

4 Discussion and Conclusion

Assuming a simple, maximum entropy graph distribution constrained by empirically obtained marginal distance/tie probability relationships for two relations, we find that spatial variability should indeed exert substantial influence on network structure at the settlement level. The highly uneven density of population within our sample areas results in “lumpy” networks that are characterized by regions of differential local connectivity, spatially correlated gradients of expected degree and core number, and other such properties. At small spatial scales, then, we predict that the

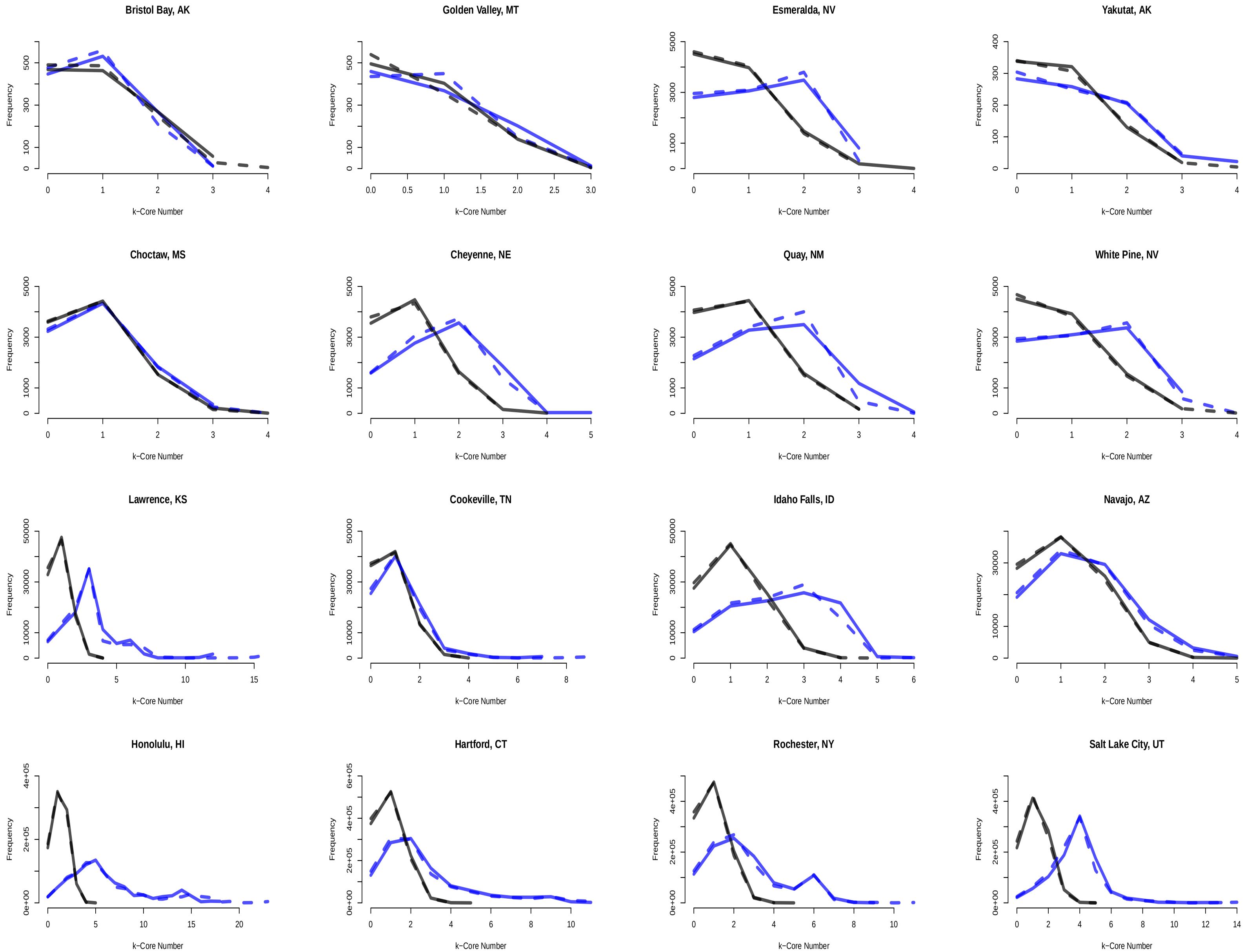


Figure 13: Marginal core number distributions by location, SIF, and placement model. Friendship model distributions are shown in blue, interaction model distributions in black; solid lines indicate uniform placement, with quasi-random placement in dotted lines.

	Freeman/Uniform	Festinger/Uniform	Freeman/Quasi	Festinger/Quasi
Bristol Bay, AK	NB	NB	NB	NB
Golden Valley, MT	NB	NB	NB	NB
Esmeralda, NV	NB	NB	NB	NB
Yakutat, AK	NB	NB	NB	NB
Choctaw, MS	NB	NB	NB	NB
Cheyenne, NE	NB	NB	NB	NB
Quay, NM	NB	NB	NB	NB
White Pine, NV	NB	NB	NB	NB
Lawrence, KS	NB	NB	NB	NB
Cookeville, TN	NB	NB	W	NB
Idaho Falls, ID	NB	NB	NB	NB
Navajo, AZ	NB	NB	NB	NB
Honolulu, HI	NB	NB	NB	NB
Hartford, CT	G	NB	G	NB
Rochester, NY	NB	NB	NB	NB
Salt Lake City, UT	NB	NB	NB	NB
G=Geometric, NB=Negative Binomial, P=Poisson, W=Waring, Y=Yule				

Table 6: AICC selected models for core number distribution, by location, SIF, and placement model.

character of the local structural environment will—for these types of relations—depend heavily on local population distribution.

While spatial heterogeneity does induce substantial within-network heterogeneity, we also observe that geography drives many aggregate network properties in a predictable way. For relations like those modeled here, aggregate mean degree, edge length, and local clustering can be well-predicted by properties such as the mean nearest-neighbor distance, together with SIF-specific factors. The nearest-neighbor distance is itself driven in large part by land area and total population size, though microdistributional factors do play a role. Taken together, this implies that, for these sorts of relations, it should be possible to predict differences in a number of aggregate structural properties from fairly basic features of the underlying social geography.

Looking forward, we note that while the present investigation leaves out many factors—e.g., endogenous clustering, differential mixing due to age or race, and degree constraints, to name a few—it nevertheless includes a far more detailed consideration of spatial factors than has been pursued in past studies. In that light, it is interesting to observe that many features taken as generically indicative of complex endogenous processes (including long-tailed degree distributions, differential mixing, and triadic clustering) arise here in natural way without recourse to more complex assumptions. While the existence of such processes is well-documented and uncontroversial, their relative importance in determining the properties of large-scale networks is less clearly established. To that end, it would be productive to undertake more systematic empirical research to identify the limits of simple properties such as spatial marginals for structural prediction. The spatial Bernoulli graphs are obviously an approximation to a more complex reality, but may still be adequate for many purposes. If so, their simplicity and tractability (theoretical, computational, and inferential) have much to recommend them.

In like vein, we would suggest that in settings for which spatial Bernoulli graphs are inadequate, augmenting the base model (per Equation 2) with covariate effects preserving Bernoulli structure (e.g., age or race mixing) rather than sources of dependence among edges should be considered as an initial strategy (at least for modeling of large-scale networks). Although more complex than purely spatial models, such hybrids retain many of the scalability and theoretical tractability advantages exploited here. One natural avenue for such expansion is via the use of Blau-space models (McPherson, 1983; 2004) that employ a notion of distance over a combined socio-physical space. (Some initial steps in this direction have been taken by Hipp and Perrin (2009).) One major obstacle to current progress in this area is a lack of high-quality network data sets containing sufficient information on both geography and demographic characteristics to permit estimation of a joint socio-physical SIF. As such data becomes available for multiple relational types, it will be possible to substantially expand the range of questions that can be asked within the spatial network paradigm.

Another obstacle to further theoretical progress is a lack of detailed population data on the co-evolution of social ties and residential mobility. In this paper, we have focused on the problem of instantaneous prediction: given the (current) distance structure, predict the contemporaneous properties of an associated social network. In so doing, we neither make nor require assumptions regarding the dynamic processes that produce this joint socio-geographic structure—so long as we know how space is instantaneously related to social relationships, we can predict the latter from the former. Many interesting questions, however, relate to these underlying processes, and a detailed understanding of them would advance current knowledge in many respects. For instance, we here take the SIF as given (estimated from prior data), but the marginal relationship it describes clearly arises from a combination of (possibly tie-influenced) movement of persons in space, and (possibly spatially-influenced) formation and dissolution of social ties. A richer understanding of these mechanisms could potentially allow for the prediction of spatial interaction functions themselves, as well as prediction of the circumstances in which such functions might vary or change. Likewise, sudden perturbations to normal mobility patterns (e.g., displacement following wars or natural disasters) may produce short-term changes in socio-geographic structure that are poorly predicted by comparative statics (i.e., equilibrium structure before and significantly after the event). Modeling such shocks requires knowledge of how rapidly ties decay following relocation, the rate at which new ties form in response to this relocation (and to whom), and selective influences of tie acquisition and loss on secondary mobility. These are complex phenomena, which in our opinion require a much deeper understanding of social dynamics than is currently available. However, the first step in developing this understanding will clearly be the design of studies that are sensitive to both spatial and temporal concerns.

Finally, we close by reiterating some simple predictions that, seeming to be robustly present in the cases studied here, lend themselves to empirical evaluation. First, we predict a positive correlation between local population density and both mean degree and core number over scales that are at least comparable to the relevant SIF. Second, we predict transitivity far in excess of CUG baselines, declining globally in mean degree, increasing in nearest-neighbor distance, and declining in SIF tail weight. Third, we predict sudden changes in the formation of large, cohesively connected subgroups with population density, with such groups transitioning from small and rare to large and common as a threshold function of local population. Finally, we predict that internal density for areas of equivalent geometry and population distribution will be approximately constant in total population (with the solution to the degree bound problem of Mayhew and Levinger (1976) tending to be resolved through changes in population distribution, including artificial elevation, rather than through density change). None of these predictions seems to us self-evidently true, and all could be far from the mark. Systematic evaluation of each, however, for a range of networks, could tell us much about how geographical variability relates to network structure.

5 References

Anderson, B. S., Butts, C. T., and Carley, K. M. (1999). The interaction of size and density with graph-level indices. *Social Networks*, 21(3):239–267.

Bivand, R. S., Pebesma, E. J., and Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R.* Springer, New York.

Bossard, J. H. S. (1932). Residential propinquity as a factor in marriage selection. *American Journal of Sociology*, 38:219–244.

Brakman, S., Garretsen, H., Van Marrewijk, C., and Van Den Berg, M. (1999). The return of Zipf: Towards a further understanding of the rank-size distribution. *Journal of Regional Science*, 39(1):183–213.

Burian, S. J., Brown, M. J., and Velugubantla, S. P. (2002). Building height characteristics in three U.S. cities. In *Fourth Symposium on Urban Environment*, pages 129–130. American Meteorological Society, Norfolk, VA.

Butts, C. T. (2001). The complexity of social networks: Theoretical and empirical findings. *Social Networks*, 23(1):31–71.

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

Butts, C. T. (2003). Predictability of large-scale spatially embedded networks. In Breiger, R., Carley, K. M., and Pattison, P., editors, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers.* National Academies Press, Washington, D.C.

Butts, C. T. (2006a). Curved exponential family parameterizations for spatial network models. Presentation to the 26th Sunbelt Network Conference (INSNA).

Butts, C. T. (2006b). Exact bounds for degree centralization. *Social Networks*, 28(4):283–296.

Butts, C. T. (2008). Social network analysis with sna. *Journal of Statistical Software*, 24(6).

Butts, C. T. (2010). Bernoulli graph bounds for general random graphs. Technical Report MBS 10-07, Irvine, CA.

Butts, C. T. and Acton, R. M. (2011). Spatial modeling of social networks. In Nyerges, T., Couclelis, H., and McMaster, R., editors, *The SAGE Handbook of GIS and Society*, chapter 12. SAGE Publications.

Carley, K. M. (2002). Computational organizational science and organizational engineering. *Simulation Modeling Practice and Theory*, 10:253–269.

Daraganova, G. and Pattison, P. (2007). Social networks and space. Presentation to the 2007 International Workshop on Social Space and Geographical Space.

Faust, K. (2007). Very local structure in social networks. *Sociological Methodology*, 37(1):209–256.

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

Freeman, L. C., Freeman, S. C., and Michaelson, A. G. (1988). On human social intelligence. *Journal of Social and Biological Structure*, 11:415–425.

Gentle, J. E. (1998). *Random Number Generation and Monte Carlo Methods.* Springer, New York.

Hägerstrand, T. (1967). *Innovation Diffusion as a Spatial Process.* University of Chicago Press, Chicago.

Hamilton, D., Handcock, M. S., and Morris, M. (2008). Degree distributions in sexual networks: A framework for evaluating evidence. *Sexually Transmitted Diseases*, 35(1):30–40.

Handcock, M. S. (2003). *degreenet: Models for Skewed Count Distributions Relevant to Networks.* Seattle, WA. Version 1.0.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1).

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model based clustering for social networks. *Journal of the Royal Statistical Society, Series A*, 170:301–354.

Haynes, K. E. and Fotheringham, A. S. (1984). *Gravity and Spatial Interaction Models.* Sage, Beverly Hills, CA.

Hipp, J. R. and Perrin, A. J. (2009). The simultaneous effect of social distance and physical distance on the formation of neighborhood ties. *City and Community*, 8(1):5–25.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

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. (1984a). Baseline models of sociological phenomena. *Journal of Mathematical Sociology*, 9:259–281.

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

Mayhew, B. H. and Levinger, R. L. (1976). Size and density of interaction in human aggregates. *American Journal of Sociology*, 82:86–110.

McPherson, J. M. (1983). An ecology of affiliation. *American Sociological Review*, 48:519–532.

McPherson, J. M. (2004). A blau space primer: Prolegomenon to an ecology of affiliations. *Industrial and Corporate Change*, 13:263–280.

McPherson, J. M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444.

Pattison, P. E. and Robins, G. L. (2002). Neighborhood-based models for social networks. *Sociological Methodology*, 32:301–337.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes.* Cambridge University Press, Cambridge.

Sampson, R. J., Raudenbush, S. W., and Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277:918–923.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2).

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.

West, D. B. (1996). *Introduction to Graph Theory*. Prentice Hall, Upper Saddle River, NJ.

White, D. R., Tambayong, L., and Kejzar, N. (2008). Oscillatory dynamics of city-size distributions in world historical systems. In Modelski, G., Devezas, T., and Thompson, W. R., editors, *Globalization as an Evolutionary Process: Modeling Global Change*. Routledge, London.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Hafner, New York.