Large Scale, Spatial Structure of Interpersonal Networks across the Western United States:
Evidence from a Randomized Survey

Nicholas Nagle – University of Tennessee, Knoxville
Carter Butts, John Hipp, Adam Boessen – University of California – Irvine

Introduction

Relatively little is known about the large scale spatial structure of personal networks. Available information is limited. Some systematic surveys have evaluated the length of personal network ties, and related the extent to individual characteristics of egos and alters. These studies, however, have tended to collect distance, rather than absolute geographic location. Thus, while we know about the relationship between networks and *relative location*, such as geographic distance and proximity, we know relatively little about *absolute location*, and about the physical description of American space and personal networks through that space.

Other studies have explored the absolute location of network, through volunteered or nonspecific network data, such as networks of Facebook users or phone calls. Such data sources are large but incomplete. The demographic and socioeconomic biases of these networks can be large. Furthermore, we know relatively little about the nature of those ties. People's networks are not homogeneous, but they are heterogeneous and multiple.

We present a descriptive analysis of the spatial structure of personal networks across the Western United States. This analysis is based on a unique sample survey that collected information from respondents about the structure of various personal networks, as well as information about the spatial location of themselves and their network alters. This survey used an online format, with a traditional solicitation via mail.

We are able to display geographic space continuously across the Western United States. We show here descriptive analysis based on statistical clustering and visual representation of network, in order to empirically evaluate the structure of personal network through large geographic spaces. Compared to much of the previous network research, which often concentrates on small geographic scales, such as the neighborhood, this sample is better able to analyze network structure over large geographic scales, such as the states and regions within the Western United States.

The rest of this paper is structured as follows. First, we will describe the survey sample used here. We will describe the methods used, including the data preparation, as well as the statistical and visual methods used to characterize network structure. We then describe the empirical results that are obtained from applying these methods to the sampled data.

Data

Data were collected between 2012 and 2013. A fixed sample size was distributed across all census blocks in the Western United States, with inclusion probability proportional to the geographic size of the census block. Sampled individuals were randomly selected from a list of all adults with a mailable address. Individuals were contacted with a personalized letter inviting them to participate in the online survey. This initial mailing contained a $2 incentive, and participants were told that they would receive an additional $10 incentive upon completion of the survey. Participants were sent a reminder postcard one week after the initial mailing, and a reminder letter one month after the initial mailing (Dillman 1991).

Participants were given a unique personal identification number, which they could then use to log in to the online survey. The survey instrument consisted of basic demographic component, combined with name generators for various types networks. For each network name generator, respondents were allowed to nominate as many people as they would like (free response). Once a person was identifies, a check box was placed on the screen, and the person because automatically eligible for nomination in subsequent name generators.

Respondents were then asked to identify the residential addresses for each individual that was identified. Addresses were then geocoded to latitude and longitude by the Google Maps API. A simple map of the location was then displayed to the respondent, and the respondent was presented with the opportunity to alter the location.

After the survey was completed, an automated computer script obfuscated the locations so that they were not available to the researchers. Precise addresses (i.e. those with street and number) were obfuscated to *pairs* of block; while less precise addresses (for example, if a city was provided but not street address) were geocoded to the centroid of the corresponding unit. All study procedures were approved by the appropriate Institutional Review Board.

The final survey included 3,370 responses, for a response rate of 18%. The sampling bias was similar to other forms of mail recruitment, with a median age that is higher than that of the general population.

Methods

Data Preparation

We first divide the Western United States Study region into grid cell of 1 degree latitude by 1 degree longitude and aggregate the egos and alters into these cells. 1 degree latitude and longitude was chosen because it is an easy number quantity to map, and because it was large enough that most grid cells contained at least 10 respondents. Furthermore, many respondents recorded the place name for themselves and their network alters, but not a precise location. The 1 degree grid size is large enough that this locational uncertainty is not problematic for research.

For each pair of grid cells, we then calculate an estimated tie volume V_{jk} between the cells. The tie volume is calculated by weighting the observed links between the cells by the probability that the link is in the sample (i.e. by the probability that that one of the two individuals is selected as a respondent).

Spectral Clustering

We use the method of *spectral clustering* in order to empirically identify regions or communities in the Western US. Spectral clustering is increasingly used []. Spectral clustering is based on an eigenvalue (e.g. principal components) decomposition of the graph Laplacian matrix. Graph Laplacians are used to define many properties of network similarity matrices. There are two types of graph Laplacians. Let V be the matrix of tie volumes, and D be the diagonal *degree* matrix, with diagonal values d_{jj} = \sum_j V_{jk}=\sum_k V_{jk}. The un-normalized graph Laplacian is L=(D-V), and the (random walk) normalized graph Laplacian is L=D^{-1}(D-V). The random walk Laplacian derives it's name because the off-diagonal elements L_{lk} can be interpreted as the probabilities that an graph edge originating in region l is destined for region k.

We conduct spectral clustering on the normalized graph Laplacian. Once the normalized Laplacian is calculated, we then extract the smallest k eigenvectors. (The very smallest eigenvector is constant, and it excluded from selection). The smallest eigenvectors correspond to the dominant trends, or *modes,* in interaction. The smallest 12 eigenvectors are plotted in Figure 1.

Hierarchical Bayesian Clustering

We used hierarchical Bayesian clustering in order to cluster the 6 dominant eigenvectors. The model we fit is:

$$
\begin{aligned}
Y_i &\sim \sum_{c \in C} z_{ic} N(\mu_c, \Sigma_c) \\
z_{ic} &= 1 \text{ if } i \text{ in cluster } c; \ 0 \text{ otherwise} \\
z_{ic} &\sim Categorical(\pi_1, \pi_2, \ldots, \pi_C)
\end{aligned}
$$

where C is the number of clusters. A uniform prior over the simplex was placed on the membership probabilities π, and a diffuse Gaussian prior was placed on the cluster centers μ. The covariance matrix Σ was assumed to be constant and known for all clusters. Label switching can be a notorious problem in Bayesian clustering (Jasra 2005). Base don preliminary analysis of the eigenvectors, we specified constraints on the C major cities (i.e. one major city per cluster) to prevent them from switching labels. After implementing this, cluster switching was not a problem in the posterior chains. Thus, our clustering algorithm is a semi-supervised classification problem. Work is continuing to evaluate the robustness of the clusters to these modeling decisions.

The model was fit using the Stan software package (Stan Development Team 2013).

Force-Directed Graph Layout

In contrast to the spectral clustering method, which identifies network communities or clusters based on the similarity of their flows, the force directed graph layout is intended to depict the location of the network in a 2-dimensional, non-geographic space, so that strongly connected nodes are near to each other, and weakly connected nodes are far apart. We use here the Fruchterman-Reingold method, in which the tie volumes create "attractive" forces that bring nodes together, and in which "repulsive" forces keep nodes well separated in space. Force-directed layout methods must be used with some discretion, however, because it is not always possible to suitably represent the distances in a graph as distances in 2-dimensional space. This model was fit in R using the igraph package (Csardi et al 2006).

Results and Discussion

Clustering Analysis

The 12 dominant eigenvectors of the network graph is shown in Figure 1. The graphs are to be read in consecutive order; each graph identifies additional patterns conditional on those identified in the previous graph. The dominant trend in the network is the separation of personal networks between the Pacific Coast and the Rocky Mountain West. The interpretation of this graph is that any person living along the Pacific Coast, is more likely to have a connection elsewhere along the Pacific Coast than in the Rocky Mountain West. The converse holds for those living in the Rocky Mountain West.

This is not simply a factor of geographic separation. If it were a factor of geographic distance, then the largest factor would be a North-South division since that is the largest geographic extent. Thus, there is clear *network* separation between the Pacific Coast and the Rocky Mountain West. Presumably, this is partly caused by the low population densities throughout the intervening inter-mountain West, but this hypothesis is tested here.

The remaining maps identify successively less dominant trends in the network structure, apparently selecting out differences between, in succession, Colorado, New Mexico, Utah, and the Pacific Northwest, and Arizona and Northern California. The higher eigenvectors are increasingly dominated by local deviations, which my be due outliers and noise. We choose to concentrate on clustering analysis of the first 6 eigenvectors.

Results from the Bayesian Gaussian Mixture Model are shown in Figure 2. We show here the result using 6 eigenvector dimensions and C=7 clusters, but results from lower dimensions and clusters are available on request. The cluster analysis shows relatively well defined clusters, apart from regions in Nevada and Idaho, which are not robustly identified with a cluster. The analysis with fewer eigenvectors and clusters suggests that the division between 4 clusters in the Mountain West clusters

explains more of the network structure than the division between the 3 Pacific Coast Clusters.

The clear division along some state lines is interesting. For example, the space along the Arizona – New Mexico is very sparsely settled. Yet, our samples include grid cells on either side of the border, in which we have at least ten respondents in each grid, and, which appear to be robustly assigned to a cluster within their state. The persistent effect of some state borders needs to be assessed more fully. Future research will explore simulating populations with the correct population density, but with the null hypothesis of no border effects, in order to more robustly test this finding (see, for example, Butts and Acton 2011; Butts et al 2012)

Graph Layout

The force directed graph layout is shown in Figure 3. The force-directed algorithm is an iterative method that attempts to place the locations on a surface, so that the places with strong ties are near to each other, and those with weak ties are far apart. We have labeled the primary city in each of the 7 clusters. In general, however, we see that there is pretty good visual separation of the 7 clusters, even though this method is not a clustering technique. Los Angeles and the Southern California cluster points are near the center of the graph. Each of the other cluster is located roughly like a flower petal away from the center. The dominant cities in each cluster are more centrally located, with a position toward the center of the graph. The smaller places are not labeled, however, the periphery of the graph is dominated by small places.

Conclusion

We have created a representative sample of personal networks throughout the Western United States. A distinguishing characteristic of this sample is the oversampling of rural places, allowing us to conduct a thorough spatial analysis of the network throughout the study region, without restriction to the more populous places. We are thus able to conduct a systematic analysis of regionalization and connection across space. We believe that these results largely support to "folk wisdom" and lend empirical evidence to our understanding of space and American Society. Specifically, we find that the nearer places are more connected, but so are larger places. Thus, we see that large places are near to each other based on the number of connections between them, but that small places are more connected to their local neighborhoods.

Future research will explore whether the spatial characteristics of the American social network vary significantly with demographic and social characteristics, such as age and education.

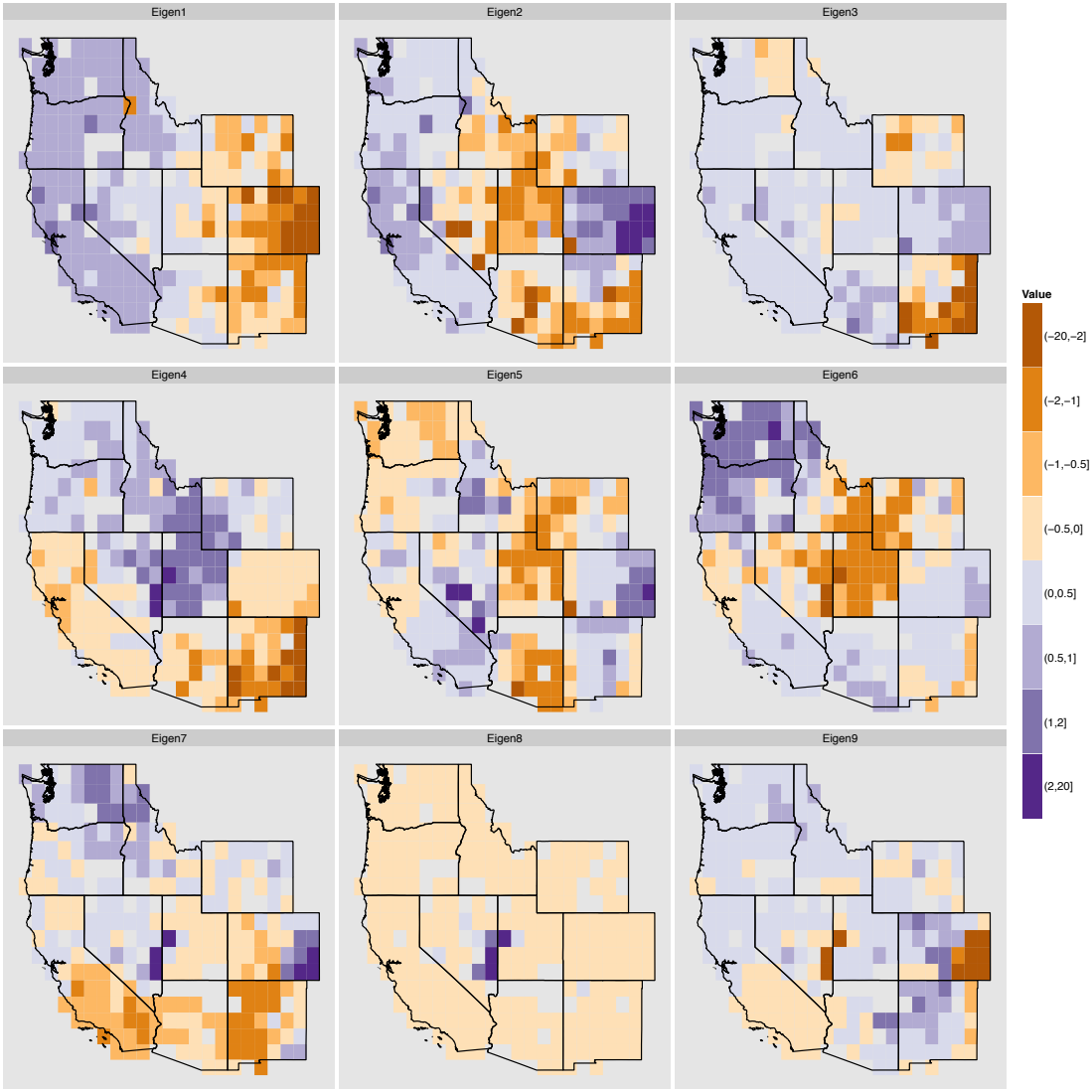

Figure 1. The dominant eigenvectors of the network graph for the Western United States.

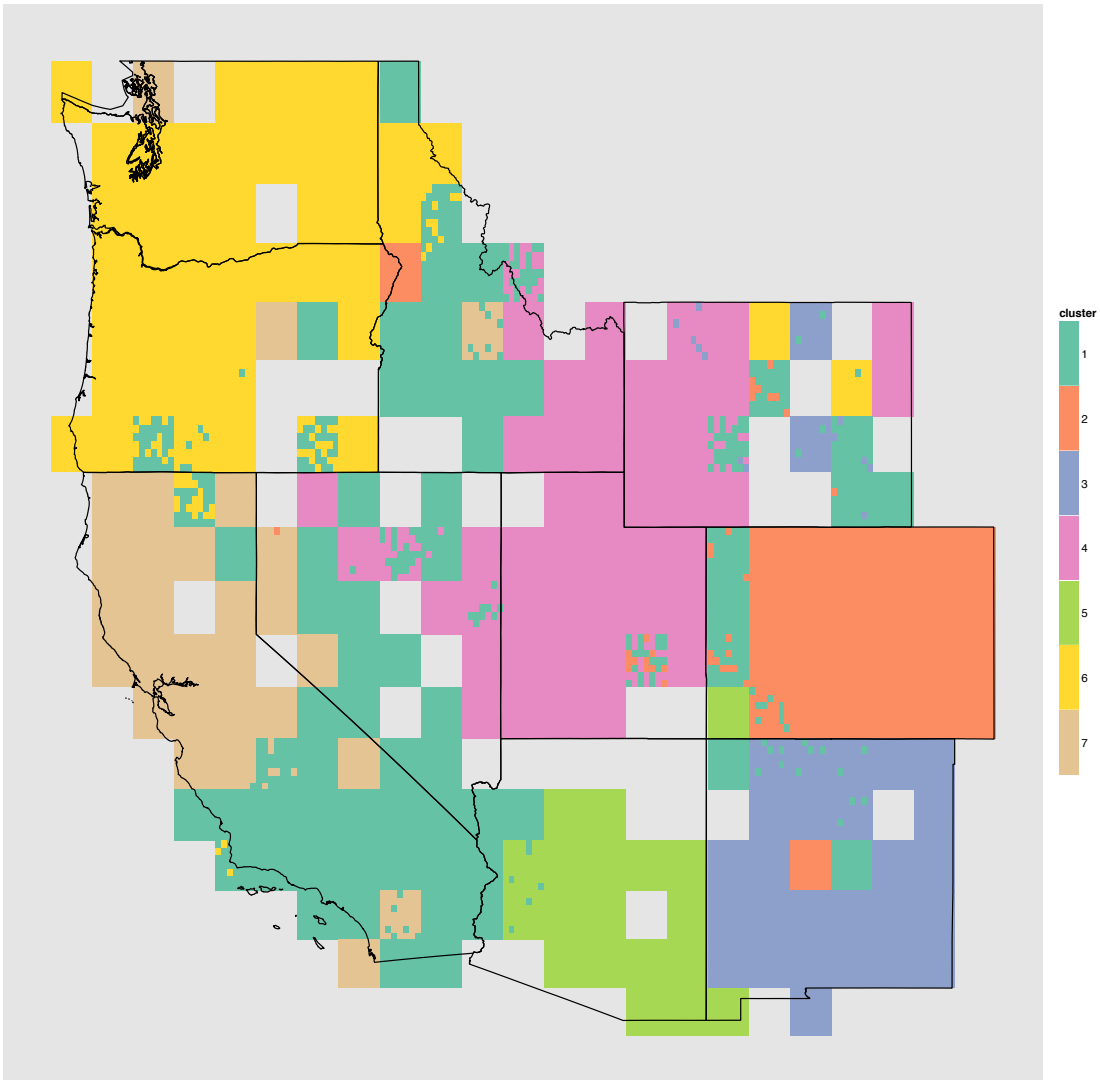

Figure 2. Gaussian Mixture Model analysis of the Western US network graph, with C=7 clusters.

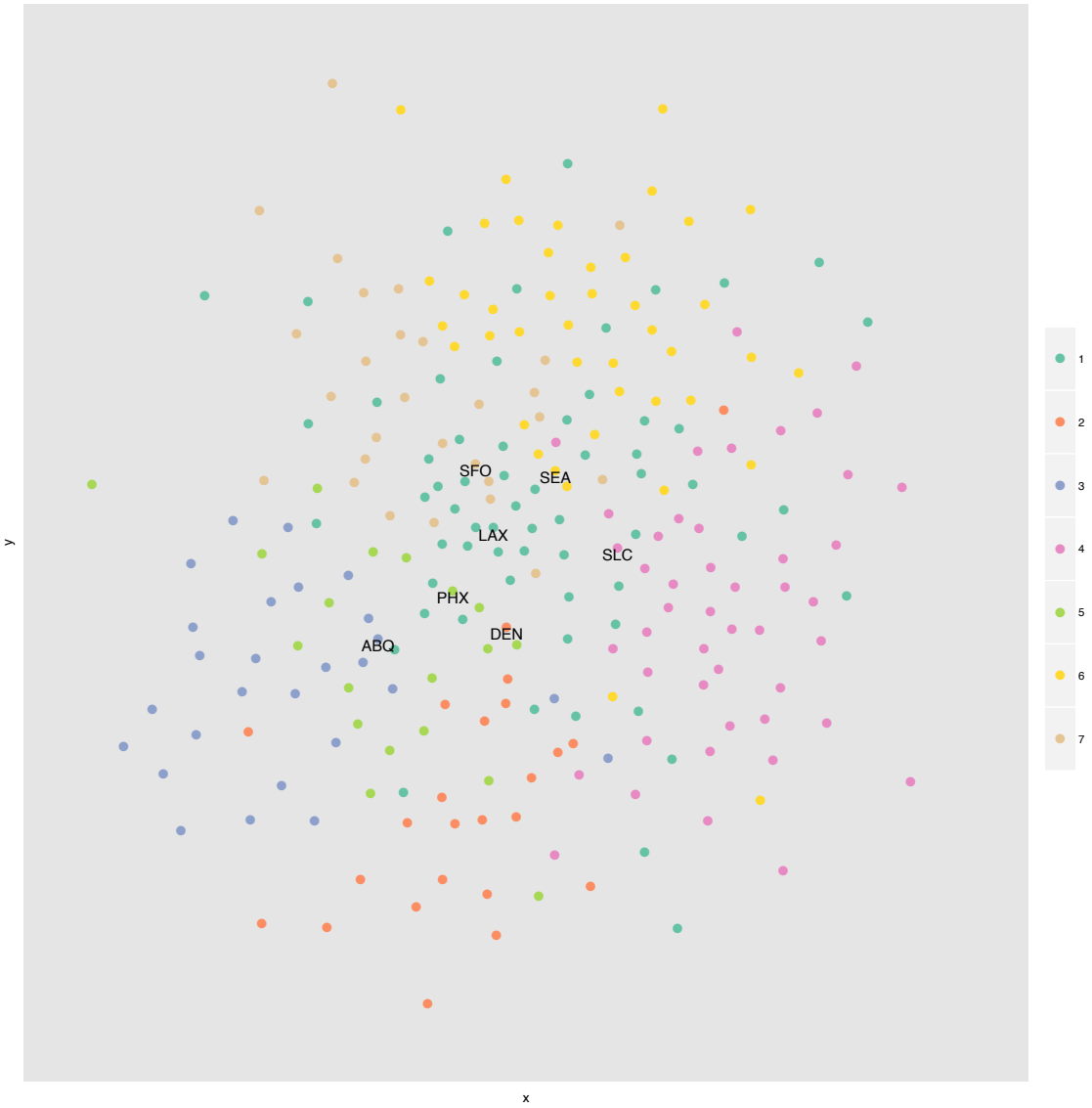

Figure 3. Force Directed Graph Layout of the Western US Network. The coloring is based on the clusters depicted in Figure 2. The primary cities in each cluster are labeled by their 3 letter airport code, but this in no way suggests a graph based on air travel.

References

Butts, Carter T and Ryan A Acton. Spatial Modeling of Social Networks. In The SAGE handbook of GIS and society. 2011.

Butts, Carter T, Ryan A. Acton, John R Hipp and Nicholas N. Nagle. Geographic Variability and Network Structure. *Social Networks*. 34(1): 82-100. 2012.

Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.sf.net

Fruchterman, T. M. J., & Reingold, E. M. "Graph Drawing by Force-Directed Placement." Software: Practice and Experience, 21(11). 1991.

Jasra, A. C.C. Holmes and D.A. Stephens. "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modelng." *Statistical Science*. 20(1): 50-67. 2005.

Stan Development Team. 2013. Stan: A C++ Library for Probability and Sampling, Version 1.3. http://mc-stan.org/.