geographical analysis

Geographical Analysis (2014), 1–23 Geographical Analysis ISSN 0016-7363

Predicting Regional Self-Identification from Spatial Network Models

Zack W. Almquist[1], Carter T. Butts[2]

[1]Department of Sociology, School of Statistics, Minnesota Population Center, University of Minnesota, Minneapolis, MN, USA, [2]Departments of Sociology, Statistics, EECS, and the Institute of Mathematical Behavioral Sciences, University of California, Irvine, CA, USA

Social scientists characterize social life as a hierarchy of environments, from the microlevel of an individual's knowledge and perceptions to the macrolevel of large-scale social networks. In accordance with this typology, individuals are typically thought to reside in micro- and macrolevel structures, composed of multifaceted relations (e.g., acquaintanceship, friendship, and kinship). This article analyzes the effects of social structure on micro outcomes through the case of regional identification. Self-identification occurs in many different domains, one of which is regional; that is, the identification of oneself with a locationally associated group (e.g., a "New Yorker" or "Parisian"). Here, regional self-identification is posited to result from an influence process based on the location of an individual's alters (e.g., friends, kin, or coworkers), such that one tends to identify with regions in which many of his or her alters reside. The structure of this article is laid out as follows: initially, we begin with a discussion of the relevant social science literature for both social networks and identification. This discussion is followed with one about competing mechanisms for regional identification that are motivated first from the social network literature, and second by the social psychological and cognitive literature of decision making and heuristics. Next, the article covers the data and methods employed to test the proposed mechanisms. Finally, the article concludes with a discussion of its findings and further implications for the larger social science literature.

Introduction

Social scientists characterize *social life* as a hierarchy of environments, from the microlevel of an individual's knowledge and perceptions to the macrolevel of large-scale social networks. In accordance with this typology, individuals are typically thought to reside in micro- and

Correspondence: Zack W. Almquist, Department of Sociology, University of Minnesota, 909 Social Sciences Building, 267 19th Avenue South, Minneapolis, MN 55455 or School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street SE, Minneapolis, MN 55455
e-mail: almquist@umn.edu

Submitted: May 23, 2012. Revised version accepted: October 16, 2013.

doi: 10.1111/gean.12045

macro-level structures, composed of multifaceted relations; that is, acquaintanceship, friendship, and kinship (Mayhew and Levinger 1976). In this article, we treat self-identification as occurring when an individual chooses to associate him or herself with a given *label* (e.g., Sam's mother). Self-identified groups occur when each of two or more individuals choose to identify with a label (or category) that exists a priori as a result of a consensus (e.g., "I am black" and/or "I am a mother"). Thus, in this context self-identified groups arise from microlevel processes of individual decision making.

Although this article focuses on the particular case of regional self-identification, self-identification, more broadly, is of special concern to social scientists because it determines racial/ethnic, sexual, gender, class, and other identities (Howard 2000). Self-identified groups also are of particular interest to the subfields of social psychology, social boundaries, and gender relations (see Turner et al. 1987; Howard 2000; Jenkins 2000). Howard (2000) argues that these subfields view identity (self-identification) as a product of modern society and as a core issue, especially when compared with societies with rigidly imposed identities. Specifically, this article proposes that the basic underlying mechanisms for self-identification is a cognitive system, where an individual selects his or her identification from within a set of *salient* items (e.g., cities) and employs a *heuristic*—or set of rules—for choosing among those items.

The main hypothesis of this work, here dubbed the *Social Network Hypothesis of Regional Self-Identification* (SNH), is that individuals choose the region with which they identify based on the salience of the relations of the social networks in which they are embedded (e.g., friends, acquaintances, coworkers, kin; for a visualization of a spatial network see Fig. 1). In other words, individuals choose to identify with the region in which they have the most alters (e.g., friends or kin; see Almquist 2012). We contrast this hypothesis with a series of alternatives that are motivated by, arguably intuitive, *salient* components of modern life (e.g., maps, advertisements, schools, postal codes). For example, one might argue that the region that is most salient to an individual is the one that is most proximal, more so even than the one in which he or she has the most social relations.

To date, social scientists primarily have studied regional identification in the context of national identification (see Gould and White 1986; Tan 2005), with a few studies about urban/rural identification (see Wirth 1938; Fischer 1982). More recently, new developments in online data processing and management allow for larger scale and higher quality geographic data collection by nonprofessionals, what the geographic literature has dubbed *volunteered geographic information* (VGI) (Goodchild 2007). VGI data are detailed geographic data (e.g., latitude and longitude coordinates) collected by nonprofessionals, employing modern *geographic information software* (GIS; e.g., Google maps). One of the more famous of these collection efforts is the *Common Census Internet Project* (Flanagina and Metzger 2008; Baldwin 2010), the data source for this article.

Background: social networks and geography

Spatially embedded social networks have a long history in the geography literature (e.g., gravity models; Phillips, White, and Haynes 1976; Haynes and Fortheringham 1984) and the social network literature (for a review, see Barabási and Frangos 2002; Butts 2002; Butts and Acton 2011). In the geography literature, a historical and recent revival of formal network models has taken place that builds on graph theory, statistics, and machine learning literatures (Tinkler 1972;

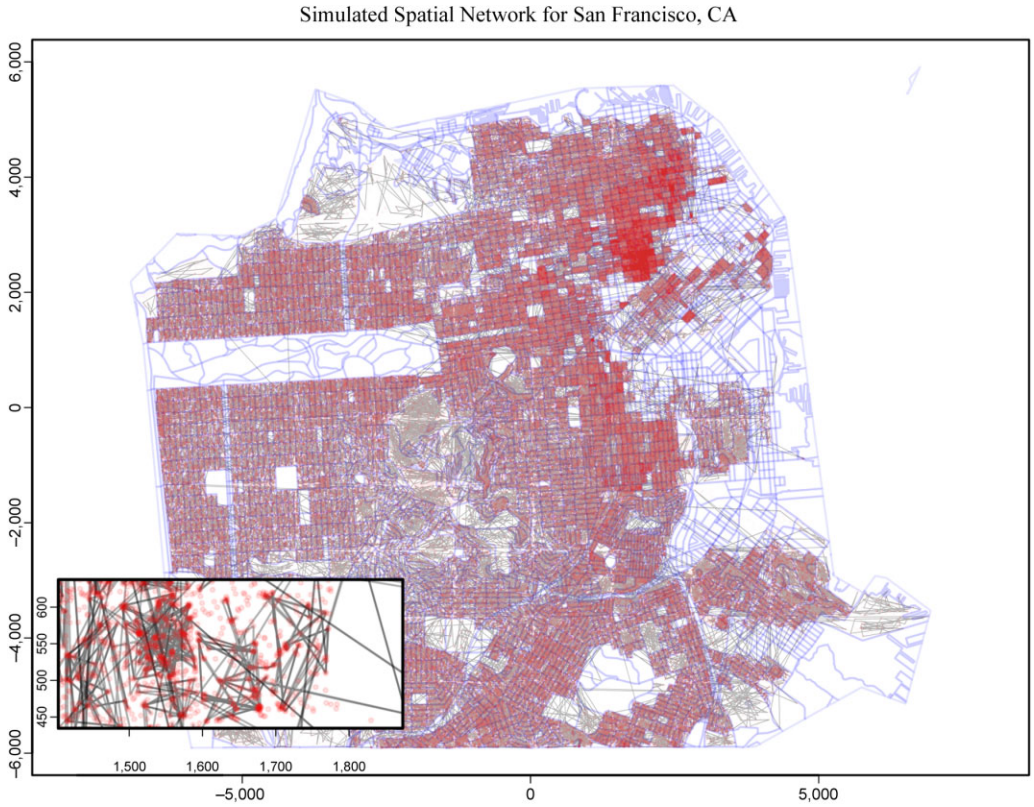

Figure 1. A simulated spatial Bernoulli network for San Francisco, CA. The simulation was performed using the procedure outlined in the Social Network Model (Tie Volume Model) Section using the Facebook SIF in section. The map is as an orthogonal projection around the centroid point in meters. Gray lines represent U.S. Census block lines, dots represent individuals, and black lines represent a social relation (e.g., friendship).

Gopal and Fischer 1996; Rogerson 1997; Gahegan 2000; Griffith 2011). Related extensions within this context include clever optimization and uses of point process models (Boots 1977; Serra and ReVelle 1999; Okabe and Yamada 2001; Schneider 2005; Yamada and Thill 2007; Shiode 2008; Almquist and Butts 2012) and the application and inclusion of network autocorrelation models in the geographic literature (Páez, Scott, and Volz 2008; Farber, Páez, and Volz 2009; Peeters and Thomas 2009; Townsley 2009). Possibly the longest running literature about spatially embedded networks is that for roads (e.g., Hudson 1969; Morley and Thornes 1972; Zemanian 1980; Osleeb and Ratick 1990; Black 1992; Okabe, Yomono, and Kitamura 1995; Peeters, Thisse, and Thomas 1998; Okabe and Yamada 2001; Xie and Levinson 2009; Bentley, Cromley, and Atkinson-Palombo 2013). More recent developments in the network and geography literature include developments concerning the problem of *small worlds* (e.g., Rogerson 1997; Xu and Sui 2009) originally introduced by Travers and Milgram (1969) and Milgram (1967), and later extended by Watts and Strogatz (1998). Other important examples of empirical spatial networks include those for cities (Portugali, Benenson, and Omer 1994; Taylor 2001; Neal 2012), drainage networks (Werner 1972), and t-communities (Grannis 2009; Whalen et al. 2012),

and the use of networks in cognitive models and spatial thinking problems (Morley and Thornes 1972; Mirchandani 1980; Smith, Pellegrino, and Golledge 1982).

Regional identification as a cognitive process

The definition of self-identification employed in this article (i.e., individual *y* identifies with object *x*) is that of a behavior requiring an individual to match him or herself with a label that is drawn from a set of potential labels (or categories) that exist in his or her cultural repertoire (e.g., doctor or Asian). In this sense, a component of self-identification exists that requires a decision from an actor, and which can be further described as a *choice*. This choice, at least at some level, must involve the act of *information processing*, if only for the actor to allocate him or herself to some default option (see Gigerenzer and Todd 1999; Hutchinson and Gigerenzer 2005).

In the case of regional identification, these assumptions imply that an individual has a mechanism for identifying the set of potential geographic categories (e.g., towns, cities, or other culturally recognized places) at the situationally relevant scale, and a way to choose an item from within a given set (e.g., Irvine) with which he or she identifies. This process can be seen in everyday life in a variety of contexts; for example, when an individual proclaims "I am a Persian" or "I am a New Yorker" therefore regional identification can be characterized by the combination of (1) a *choice set* and (2) a *heuristic*. Much of the following discussion is dedicated to describing potential mechanisms that are competing for the "best" (i.e., most predictively accurate) choice set and heuristic in a model of regional identification, including mechanisms involving social structures.

Scale and regional identification

As implied by the proceeding discussion, an individual is potentially able to identify him or herself with a preferred geographical unit at multiple scales; each scale is defined by a culturally relevant set of geographical units (e.g., neighborhoods or local communities, towns or cities, states or provinces, nations), which constitutes the choice set for an identification decision. (Meaningful scales for such identification are themselves culturally defined.) Thus, we may envision the regional identification process as producing for each individual a "cone" of valid identities $x_0 \subseteq x_1 \subseteq \ldots$, each having the property that individual *y* associates more strongly with region x_i than any other region x_i' at the same scale in his or her cultural repertoire. This is depicted schematically in Fig. 2.

Our focus in this article can be viewed as follows: given a uniform "slice" through the cones of regional identities in a population at a given scale, what predicts the units with which each individual will identify? In particular, we here consider identification for local communities among residents of the United States, at a scale that corresponds to "places" designated by the U.S. Census.

Mechanisms of regional identification

We may hypothesize a variety of processes by which regional identification may occur at a given scale. The mechanisms we consider here are divided into two key subgroups: the SNH and the *geography* and *prominence* hypotheses. The first of these hypotheses is based on social structure in which individuals are embedded; the subsequent competing hypotheses are based on particularly salient properties of modern life.

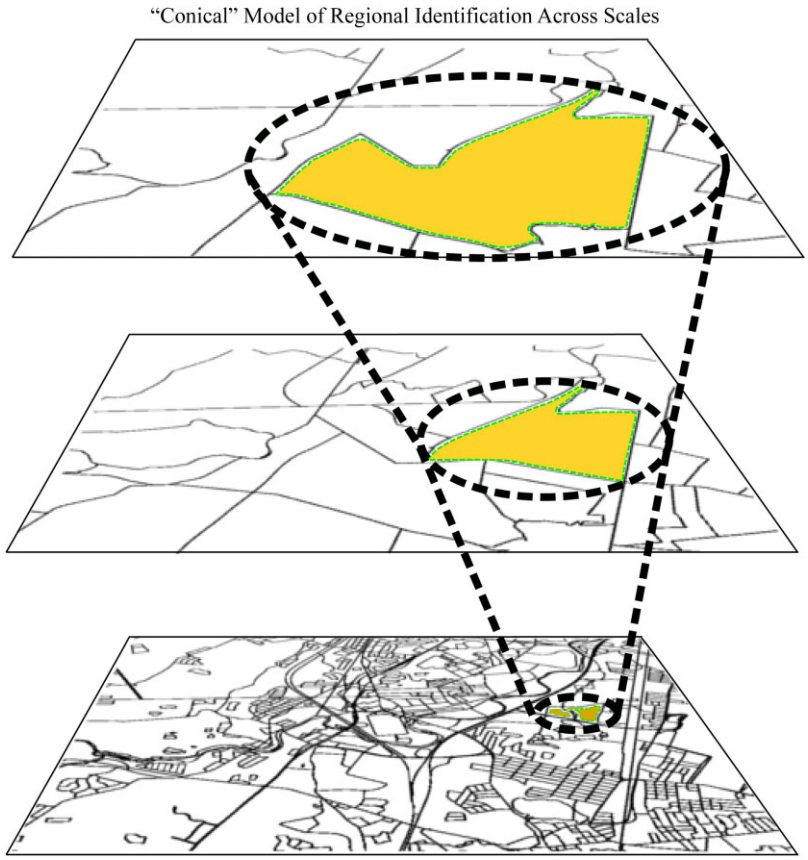

Figure 2. At a given culturally defined scale (planes), an individual most closely identifies with a given geographical unit (circled areas). Identification at any given scale can be conceptualized as eliciting a "slice" through the cone-like structure formed from the union of possible elicitations. "Slicing" at a uniform level allows us to examine identification mechanisms across individuals.

A SNH

One potential mechanism for regional identification is based on the social context in which individuals are embedded (e.g., friendship, coworker, and kinship networks). In such a case, regional identification might be an individual performing a search over his or her personal networks (Dodds, Muhamad, and Watts 2003), and selecting the region that contains within it the *largest* number of alters. Tying this notion back to the concept of salience will be important throughout this article; this hypothesis can be rephrased as an argument that regions containing the maximal number of an individual's alters are the most salient regions to that individual for this type of identification.

SNH. Individuals choose to identify with the region in which they have the largest number of alters.

Different processes (or combinations thereof) could potentially underlie the ultimate mechanism of regional identification and inform the SNH. Intuitively, we might suspect this hypothesis

to be plausible for many reasons. For example, (1) individuals search over their alters and select the region to identify with based on a plurality heuristics; (2) individuals have more exposure to the places they have more alters, and thus such an area is more salient; and (3) individuals mimic their peers and thus choose to identify with the area with which they view the bulk of their peers as identifying. As the data do not allow us to distinguish between these microlevel processes, we view the SNH as representing a *class* of mechanisms (one or more of which may be active at once), which we collectively distinguish from other classes of identification mechanisms.

Geography and prominence hypotheses

The SNH involves one class of mechanisms for regional identification, but others can be entertained. The first alternative hypothesis proposed here is dubbed the *proximity hypothesis*. The proximity hypothesis is based on the intuitive salience of certain geographies to an individual, particularly those who are the closest (most proximal) to an individual (e.g., one lives near Irvine, CA, and identifies with Irvine).

Proximity hypothesis. Individuals choose to identify with the region that is most proximal (closest) to them, given their current geographic location of residence.

An alternative hypothesis—although, one which is related to the proximity hypothesis—is one in which the most salient region is not simply the most proximal, but is a balance of being both the most *prominent* (salient given some characteristics/threshold) and *also* most proximal to an individual's location. In this case, one assumption is that an individual limits his or her choice set to only those regions that meet a particular *prominence characteristic/threshold* (e.g., presence of National Football League team/population threshold; see Gigerenzer and Todd 1999), and subsequently selects the most prominent region within this limited set.

Prominence hypothesis. Individuals choose to identify with the closest *prominent* region to their geographic locations.

One might also propose the reverse of the aforementioned hypothesis, where an individual first limits his or her choice set by the saliency criterion of *distance*, and then chooses a region to identify with based on some prominence characteristic/threshold.

Distance hypothesis. Individuals choose to identify with the most prominent region within a given *distance radius*.

The prominence and distance hypotheses are motivated, first, by the *elimination* heuristics that have been shown to be fast and frugal, as well as accurate in judgement making (Berretty, Todd, and Blythe 1997), and second, by the *vetting models* in the fields of population biology and public health (Handcock and Jones 2004).

Elimination models in the cognitive science literature were conceived for choice tasks; in these models, an object is chosen by repeatedly eliminating subsets of objects from further consideration, thereby whittling down the set of remaining possibilities (Tversky 1972). These heuristics have been extended to include categorization tasks such as length and widths of flower parts by Berretty, Todd, and Blythe (1997). Similarly, the prominence and distance hypotheses may be perceived as a series of elimination heuristics (i.e., limiting the choice set by one criterion after another until a single item remains).

The vetting models were conceived as a two-stage process model for how individuals form sexual partnerships: (1) individuals generate a list of acquaintances and (2) choose their sexual partners (Handcock and Jones 2004). Similarly, the prominence and distance hypotheses may be defined as a two-stage process model in which an individual first limits his or her choice set (e.g., only cities greater than 50,000), and then selects an item in his or her choice set (e.g., the closest remaining city).

The case of community identification

The regional identification processes outlined in the previous section are hypothesized to predict the identification within a scale-induced choice set; to test these hypotheses, it suffices to consider a set of identification decisions (1) made on a common scale, for which (2) a consensus choice set is readily available. In this article, we employ data on identification with local communities collected by the Common Census Project (CCP; Baldwin 2010). As discussed further subsequently, CCP respondents overwhelmingly (> 95%) selected regions of identification that correspond to census-designated places (CDPs) as defined by the year 2000 U.S. Census (US Census Bureau 2001). CDPs are constructed to correspond to towns, cities, or other well-defined local population aggregates with a commonly identified name, and thus serve as an effective operationalization of culturally recognized "communities"; respondents readily selecting CDPs when describing the local community or area with which they identify (despite being given the opportunity to enter alternative labels) further validate the intelligibility of this geographical unit to the study population. Henceforth, we employ CDPs as our geographical unit of interest, using the term *community* as an intuitive shorthand to describe what these units represent.

Our subjects identifying with units at the community scale does not preclude them from identifying with units at other scales. Rather, the community scale serves as a uniform *slice* through a respondents' regional identification cones, giving us a basis for systematic prediction across respondents. We do not, in particular, require that respondents' strength of identification of the community scale be stronger or more salient than, for example, their identification at larger scales. What we do require is that each respondent identify more strongly with the community he or she selects than any other available community, an assumption that is consistent with the nature of the CCP data.

Of particular use to researchers investigating regional identification (of cities or other geographical levels) is the body of spatial and geographic data from the 2000 U.S. Census (US Census Bureau 2001; Almquist 2010) and data from the Common Census Internet project (Baldwin 2010), each of which is readily available, detailed resource of geographic and identification data. Next are a detailed descriptions of the necessary U.S. Census and the Common Census data sets. Before proceeding to a description of our analysis techniques, we provide an overview of these data sets.

U.S. census demographic and geographic data

The 2000 U.S. Census Summary File 1 data consist of population counts and other basic demographics at five geographic resolutions: blocks, block groups, tracts, counties, and states (for detailed definitions, see the US Census Bureau 2001), each of which exhaustively covers the land mass of the United States. The U.S. Census data also contain geographic and demographic data for what it calls *CDPs*, which shall be referred to as *communities* in the remainder of this article.

As noted previously, the U.S. Census Bureau's definition of CDPs closely approximates what most individuals of the United States would consider communities. This linage is reinforced by analysis of the CCP data, for which 96% of respondents' reports of identification are found to coincide with places as they are defined by the U.S. Census Bureau (e.g., a respondent might choose Irvine for his or her identification). This outcome occurs even though respondents were both given the option of choosing items outside the category of places, and provided the option of writing in their own preference. There are a total of 24,670 places ranging in population size from 0 to 8 million (there is no minimum population requirement for a place; US Census Bureau 2001), with most corresponding to towns, cities, or well-defined and commonly named areas within larger urban areas. CDPs can also include military installations or other areas that are well recognized (and which may have a residential population), but that are not captured by conventional definitions of "city," "town," or the like.

Our analysis employs a GIS implementation of the 2000 U.S. Census data by Almquist (2010), implemented in the R statistical computing environment (R Development Core Team 2010). R's spatial tools (Bivand, Pebesma, and Gómez-Rubio 2008) were used for associated data manipulation and analysis.

Common Census Internet project

The Common Census Internet project is a website started in 2005 by Baldwin (2010) to develop a "natural" (perceptual consensus) mapping of the United States such that the borders of/within an area emerge from a consensus among the individuals who reside in that area. In practice, the CCP data are a convenience sample from 2005 to present, which consists of five questions related to one's geography, several of which focus on regional identification. In this work, responses to three of the five questions are used to test the hypotheses proposed in this article. The proceeding analysis also utilizes data from the first of the five questions from the online questionnaire, which elicits a respondent's address and automatically geocodes his or her location (after which these results were anonymized to the census geography of the block).[1]

Given that respondents may answer any of the Common Census questions at idiosyncratic geographic levels, we limit our analyses to those respondents who supply at least one answer at the community (CDP) level. The first[2] and third[3] questions pertain largely to community-level identification, and approximately 96% of respondents supply an answer corresponding to a CDP.

Crucially, both questions request that a respondent ignore any official boundaries and answer only with the region he or she *feels* that he or she identifies with, which should elicit the processes of regional identification this work is interested in, rather than simply a report of the geographic location of individuals.

Responses collected after 2007 were omitted from analysis as a result of the faulty geocoding of respondents' locations; this elimination left a total 51,655 respondents, of which 45,167 answered with a CDP for the second question; this number increases to 49,769 when results of the second and third question are combined. Using a combination of questions two and three, this analysis utilized a sample of 49,769 respondents, which is 96% of the total surveyed population (2005–2007). Fig. 3 visualizes the location of each respondent. The resulting sample includes individuals who identify with 10,325 different places, where approximately 20% of these individuals selected out-of-state places.

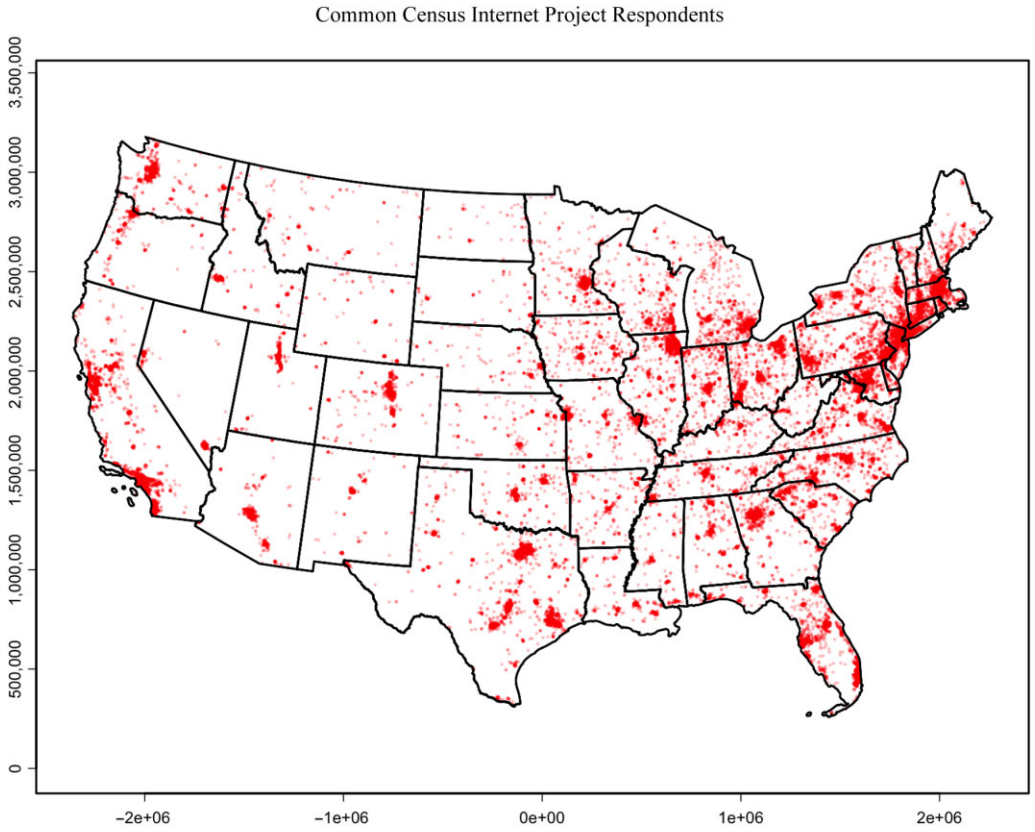

Figure 3. Centroid locations of the Common Census Internet Project respondents in the continental United States, in an Albers Conical Equal Area projection (in meters).

Because the Common Census is an Internet-based self-selected sample, it contain systematic biases.[4] We would expect that these biases would follow those commonly found in Internet surveys (e.g., respondents are younger, wealthier, and more educated people[5]).

Methodology

In order to utilize the Common Census data to evaluate the previously discussed hypotheses, these hypotheses must first be operationalized for a specific level of regional identification (here, the community level). What follows is a series of model proposals, each of which represents one of the aforementioned hypotheses in an analytical framework. The first of these proposals is for a baseline model, here dubbed the *uniform choice model*. This baseline model provides a comparison point for all other models, assuring a reader that the regional identification data of interest here does, in fact, contain structure (is nonrandom).

The uniform choice model

The uniform choice model is a family of parameterized models that, given a respondent and his or her geographic location, map each respondent's location to a randomly chosen item (place) from within the choice set (P). The location of a respondent is coded, using an anonymization

procedure, in terms of the centroid longitude and latitude coordinates of the U.S. Census block in which the respondent resides. Consequently, multiple respondents can have the same coordinates, although they do not live in the same household. The choice set of locations available to individuals for identification is the set of all CDPs in the continental United States (24,670 places).

In effect, the uniform choice model is a mapping of respondents' locations to a CDP randomly drawn from a uniform probability distribution, and is implemented using the following algorithm: (1) map all CDPs onto the natural numbers; (2) select each respondent's place of identification by drawing a random number from a uniform distribution; and (3) map that number back to the corresponding place (e.g., if a respondent with a location [−108.62, 44.97] selects *Ardmor*, *AL*, the model *predicts* this respondent identifies with *Ardmor*, *AL*).

Tie volume and the SNH

Interest in large-scale, spatially embedded networks has a long history in the social sciences, stemming from the famous Milgram experiments (Milgram 1967; Travers and Milgram 1969), later repopularized as the "small-world" phenomenon by Watts and Strogatz (1998). Recently, methods for statistical and simulation-based modeling of large-scale spatially embedded networks have been developed by Butts (2003; Butts and Acton 2011; Butts et al. 2012).

Spatial Bernoulli graphs and the spatial interaction function (SIF)

A well-established empirical regularity is that the marginal probability of a social tie between two persons declines with increasing geographical distance for a wide range of social relations (e.g., Bossard 1932; Festinger, Schachter, and Back 1950; Hägerstrand 1966; Freeman, Freeman, and Michaelson 1988; Latané, Nowak, and Liu 1994; McPherson, Smith-Lovin, and Cook 2001). Butts (2003) demonstrates 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. Simple network models based on the distance/tie probabiliy relationship have been shown to produce reasonable distributions for structural features such as degree distributions (Butts et al. 2012) and have been found to have predictive power, for example, crime rates in neighborhoods (Hipp et al. 2013).

The most basic family of such network models is the set of spatial Bernoulli graphs. We define a spatial Bernoulli graph in the manner of Butts and Acton (2011). Consider a set of vertices, V, that are spatially embedded with a distance matrix $D \in [0, 1]^{N \times N}$. Let G be a random graph on V, with stochastic adjacency matrix $Y \in [0, 1]^{N \times N}$. The pmf of G given D is

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

where B is the Bernoulli pmf, and $\mathcal{F}_d : [0, \infty) \rightarrow [0, 1]$ (the SIF). The SIF controls the underlying structure of a network and thus is the key component within this family of models; specifically, the SIF relates distance to the marginal tie probability. Empirically, real-world social networks typically appear to have an SIF, where the marginal tie probability decays with distance (see Butts 2003). Another well-known empirical regularity is that the marginal probability of a tie between two persons declines with geographical distance for a broad range of relationships (e.g., Bossard 1932; Festinger, Schachter, and Back 1950; Hägerstrand 1966; Freeman, Freeman, and

Michaelson 1988; Latané, Nowak, and Liu 1994; McPherson, Smith-Lovin, and Cook 2001; Arentze and Timmermans 2005; Axhausen 2007; Carrasco, Miller, and Wellman 2008). This tendency suggests that the functional form for a social network SIF is some variant of a power law. Here we consider two basic functional forms of an SIF based off of empirical data estimated from two large communication networks (see section Social Network Model [Tie Volume Model]):

$$\mathcal{F}_d(x) = \frac{p_d}{1+(\alpha x)^\gamma}, \quad \text{(attenuated power law)} \tag{2}$$

$$\mathcal{F}_d(x) = \frac{p_d}{(1+\alpha x)^\gamma}, \quad \text{(power law)} \tag{3}$$

where p_d is the baseline tie probability at distance 0, γ is a shape parameter governing the distance effect, and α is a scaling term. For a typical visualization of a network drawn from a model of this type, see Fig. 1.

As the preceding discussion suggests, network structure and geography are intricately linked, and the spatial Bernoulli graphs can be viewed as providing a social structural interpretation of the classical *gravity models* (Haynes and Fortheringham 1984) that are replete in the geographical literature. The gravity models can be viewed as a family of nonlinear regression models for valued relational data, in which the expected degree of interaction between elements is taken to be a product marginal rates (i.e., row/column effects) and an attenuation function dependent on the distance between them. Formally,

$$E[Y_{ij}] \propto P(i)P(j)\mathcal{F}_d(d(i,j)), \tag{4}$$

where $P(x)$ is the *interaction potential* of element x, and $\mathcal{F}_d$ is the SIF. Thus, the spatial Bernoulli graphs can be viewed as a special class of gravity models for dichotomous interactions (although this does not extend to the general class of spatial random graph models; e.g., see Daraganova et al. 2012). Although gravity models are not always motivated by a clear social mechanism, here models of this form (i.e., spatial Bernoulli graphs) are used to capture the expected number of social ties between an individual respondent and all individuals in a given areal unit (based on extrapolative simulation from models fit to network data in prior work). Thus, this article provides an example of the connection between classical geographical techniques and other forms of relational analysis.

Tie volume

Geographically embedded networks have many properties that are jointly related to space and social structure Butts (2003; forthcoming), the most relevant to this work being *tie volume*. The tie volume, $\mathcal{V}(A,B)$ between areal units A and B for graph G is the number of edges (i, j) such that vertex i resides in unit A and vertex j resides in unit B. If we take A_i to be an arbitrarily small region around vertex i (such that A_i contains no other vertices), $\mathcal{V}(A_i,B)$ also can be used to express the total number of ties from vertex i to individuals in areal unit B; we use the shorthand $\mathcal{V}(i,B)$ to denote this special case. When dealing with extrapolatively simulated networks (as in the present context), it is natural to work with the expected tie volume $\mathbf{E}_{\mathcal{F}_d}\mathcal{V}(A,B)$ rather than the

tie volume for an observed graph; in the foregoing, we refer to the expected tie volume simply as the "tie volume" where there is no danger of confusion.

Now, SNH can be operationalized in terms of tie volume in a straightforward manner. Given the calculations of the expected tie volume between two locations (e.g., a respondent's home location and each community in the United States), the SNH predicts that an individual identifies with the community that has the largest expected tie volume with his or her residential location.

Social network model (tie volume model)

To obtain identification predictions from the tie volume model, we proceed as follows: first, calculate the expected tie volume between a respondent's home block to the block groups that make up a given community, dividing by the population of a respondent's block to obtain the expected number of ties from the respondent to residents of each block group in the community. Next, sum the expected tie volumes from a respondent to each block group in the community, providing the expected tie volume between the respondent and the community as a whole. Repeat this procedure for each community in a choice set, then select the community with the *maximum* expected tie volume as the location with which an individual identifies. This procedure may be written explicitly as follows:

(1) Let P be the set of communities, with each P_k consisting of n_k block groups $g_{k1}, \ldots, g_{kn_k}$, with population counts given by $\mathbb{P}$. Let r_i be the census block in which the ith respondent resides.

(2) For each $k \in 1 \ldots, |P|$, calculate $\mathbf{E}\mathcal{V}(i, P_k) = \frac{1}{\mathbb{P}(r_j)} \sum_{j=1}^{n_k} \mathbf{E}\mathcal{V}(r_j, g_{kj})$.

(3) Select $\arg\max_{P_k \in P} \mathbf{E}\mathcal{V}(i, P_k)$; this is the community with which i is predicted to identify.

The expected tie volume between a respondent's location and a given block group depends on both the detailed geometry of the blocks/block groups and the SIF, and is computed via a Monte Carlo quadrature algorithm (Butts forthcoming).[6] In this article, we employ two distinct SIFs. The first is a classic SIF estimated from a large-scale phone network, and the second is a modern example from the social networking site Facebook.[7]

The first SIF used in this article is based on Hägerstrand's data set of phone calls made between regions in rural Sweden in 1950. Butts (2002) computed this SIF from Hägerstrand's (1966), "technologically mediated communication" relation, which acts as a long-tailed example with a slowly decaying distance function (approximately $d^{-2.95}$). The parametric form is an attenuated power law (see equation 2), with parameters (0.937, 0.538, 2.956).

The second SIF used in this article is based on a uniform sample of Facebook users in 2009 collected by Gjoka et al. (2010), where the authors recorded (when identified) a user's university affiliation and his or her alter's affiliation. From this information, Spiro, Almquist, and Butts (2012) computed an SIF for Facebook friendship between university-affiliated individuals. This SIF represents a "modern technologically mediated communication" relation, which acts as a long-tailed example with a slowly decaying distance function (approximately $d^{-6.527}$). Its parametric form is a power law (see equation 3), with parameters (0.627, 0.049, 6.527).

The regional identification proposed in the SNH suggests that a weak interaction SIF such as that from a communication network might be representative of the type of macrolevel structure underlying this phenomenon. The SIFs employed here are two examples of how such a network can scale with distance; by representing a fairly wide range of scaling parameters, they allow us to examine the robustness of the SNH while still employing SIFs based on (previously) observed

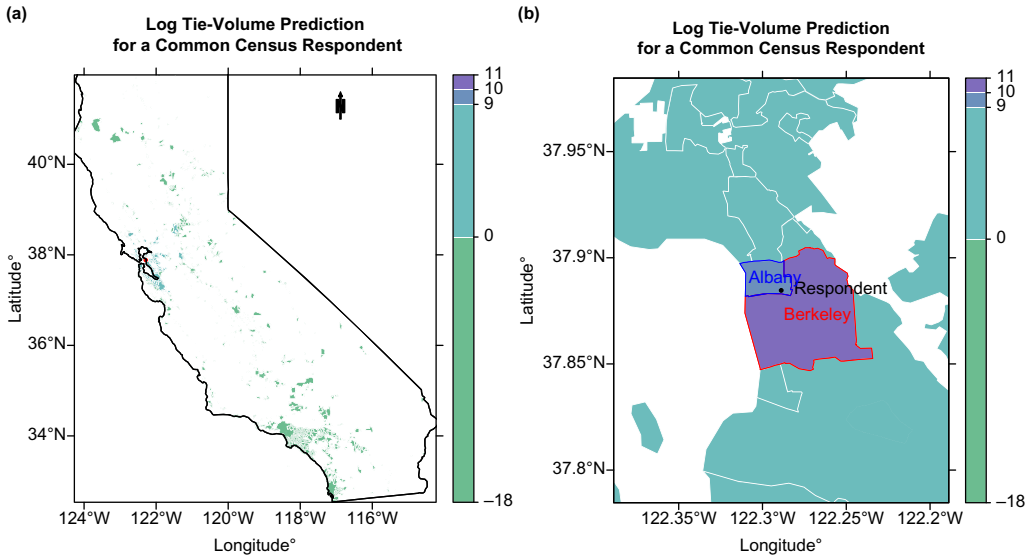

Figure 4. An example of tie volume model for single a respondent living in Albany, CA. Results logged for visualization purposes (log is a rank-preserving transformation and therefore does not change the results). (a) Full state example of the tie volume model for a single respondent living in Albany, CA. (b) A close up of the example of the tie volume model for a single respondent living in Albany, CA.

network structure. Both SIFs were inferred from observed networks in previous studies and were not in any way fit to the CCP (or other regional identification) data. Thus, these models are *zero parameter* with respect to CCP prediction, because they contain no free parameters that are adjusted to improve fit for the regional identification data.

Because optimal prediction from the tie volume model requires that one have either a priori knowledge of the *exact* SIF governing identification-relevant relationships or infer an SIF from the data (to guarantee the best fitting model), computing the expected tie volume in the manner implemented here is a more *stringent* test of the SNH than for example fitting the observed data to a gravity model. If the tie volume model outperforms competing models in predicting regional identification, the extent of this superior performance would only increase in a better-fitted model. In effect, the aspects of sub-optimality of this model make it a stronger test of the effects of large-scale social networks on regional identification.

To demonstrate the tie volume model, we consider an illustrative case within California. Respondent A lives within a census block in Albany, CA. First, we compute the expected tie volume of respondent A, given their home location within the city of Albany to all other communities in California under the aforementioned SIF (see Fig. 4). We then rank the results and select the community with the highest expected tie volume between respondent A and all communities within California. In this case, respondent A lives in Albany but identifies with Berkeley, as the tie volume model predicts (see Fig. 4).

The proximity model

The *proximity model* is a family of parameterized models that map the location of each respondent to the nearest item (community) in the choice set (P), where *nearest* here is defined as the

item with minimum distance between itself and the respondent's location. Given the notation of the The Uniform Choice Model section, in combination with a distance function $d(\cdot, \cdot)$, the algorithm first calculates a respondent's distance from his or her location to every item in the choice set, and then selects the item with the smallest corresponding distance.

In order to best approximate the actual physical distance between a respondent and each place in the continental United States, the *great circle* distance[8] is calculated between the longitude and latitude of each respondent and the center point of each community. For example, a respondent with a location (–107.53, 41.03) would be predicted to identify with Dixon, WY, as a result of the respondent having a distance of zero between his or her location and the location of *Dixon*, *WY*, and greater than zero distance for all other places.

Vetting models

Given the notation in section and the distance function of The Proximity Model section, two distinct families of single-parameter vetting models are proposed: the first of which is a *distance-based vetting model*, here dubbed the distance vetting model, and the second of which is a *prominence-based vetting model*, here called the population vetting model. All vetting models are named according to the initial *rule* an individual uses to first limit his or her choice set.

Each vetting model may be viewed as a two-stage process (Handcock and Jones 2004) in which an individual first limits his or her personal choice set with a *decision rule*, and subsequently selects a final choice based on a different decision rule. This procedure follows the same basic logic as the elimination heuristics in the cognitive science literature (Tversky 1972; Berretty, Todd, and Blythe 1997) and involves the following three basic steps:

Step 1) Select a rule to limit the choice set (e.g., individuals contemplate only communities within 50 miles of where they live).
Step 2) Select a rule to pick from among the limited choice set (e.g., individuals choose the highest population community within the resulting choice set).
Step 3) Apply the conjunction of steps 1 and 2.

Step 1 constrains a choice set using a decision rule, motivated by the hypotheses in the Mechanisms of Regional Identification section, which is operationalized as a parameter constraint, θ, and relation operator, R (e.g., a binary relation R usually is defined as an ordered triple $[X, Y, G]$ where X and Y are arbitrary sets, and G is a subset of the Cartesian product $X \times Y$; this is commonly written xRy). For example, in the case of the distance vetting model, a choice set is limited to only those communities less than θ distance from a respondent.

In step 2, another decision rule is chosen, again motivated by the hypotheses in the Mechanisms of Regional Identification section. In this article, two decision rules are proposed: CLOSEST (C) and LARGEST (L). CLOSEST is where an actor chooses the community nearest to where he or she lives that is contained within his or her limited choice set. LARGEST is where an actor chooses the most salient item in terms of population size (e.g., largest) within the limited choice set. The LARGEST decision rule requires a monotonicity assumption for a choice set, which can be accomplished by listing cities in descending order based on population size.

The distance vetting model

The distance vetting model assumes that, in the process of regional identification, an individual considers only those regions within some maximum distance of where he or she lives. This initial

limitation is achieved by narrowing the choice set to only those cities that are less than or equal to θ distance from an individual (i.e., R is the ≤ operator). Subsequently, an individual makes his or her ultimate choice by selecting the most prominent community from within that radius (θ).

For example, a respondent with a location (−86.816, 33.272) and a θ = 106.58 km has an initial, limited choice set of 171 communities. The largest three communities within this radius (in descending order) are Birmingham, AL (population 242,820); Tuscaloosa, AL (population 77,906); and Hoover, AL (population 62,742). Thus, this respondent is predicted to identify with Birmingham, AL.

The population vetting model

According to the population vetting model, an individual considers only communities whose population is greater than or equal to θ (e.g., population ≥50,000, for θ = 50,000). This individual then makes his or her final choice by selecting the closest community from within this choice set.[9]

For example, a respondent with a location (−86.816, 33.272) and θ = 289,315.4 has a resulting initial choice set of 57 communities. The closest three communities from within this set are (in ascending order of distance) Atlanta, GA (228.9 km); Nashville, TN (322.3 km); and Memphis, TN (356.6 km). Thus, this respondent is predicted to identify with Atlanta, GA.

Computational considerations

Each of the aforementioned algorithms employed in this article are implemented in the R statistical programing environment (R Development Core Team 2010). The uniform choice model and proximity model are implemented exactly as discussed, as is the tie volume model, including estimation of the expected tie volume between a respondent's block and the block groups of a given city using the spatialNetwork package (implemented in R). The population vetting model and distance vetting model employ modern techniques of optimization (specifically, we employ the optimization function provided in the R base code; R Development Core Team 2010) to obtain their parameter estimates (all code written in the R statistical programing language).

For parametric models, estimates of model standard errors and confidence intervals are performed using a nonparametric bootstrap (10,000 replications), allowing for tests of statistically significant differences in performance among the proposed models (Dwass 1957).

Analysis and results

Uniform choice model and baseline models

To test the hypotheses proposed in the Mechanisms of Regional Identification section, each model discussed in the Methodology section has been applied to the Common Census data set, the results of which (at national-level estimates) are summarized in Table 1. Currently, computation of the tie volume model solutions for the entire national data set is not feasible; rather, it has been applied on a state-by-state basis to the contiguous United States. This implementation means that any individual who resides in one state, but selects a city in another state, counts against the model for its predictive analysis (e.g., if a respondent lives in New Jersey and selects New York City as the city with which he or she identifies, then the model cannot predict it and is penalized).

Table 1 shows model prediction for community-level regional identification. These results illustrate the poor performance of the uniform choice model, which predicts only 0.02% of the

Table 1 National Comparison of Each Model; Standard Errors and 95% CI Calculated Using a Nonparametric Bootstrap with 10,000 Replications

	Proportion Correct	95% CI
Uniform choice model	0.00020	(0.00007, 0.00033)
Tie volume model	N/A	N/A
Proximity model	0.62968	(0.62028, 0.63909)
Distance vetting model	0.34939	(0.34002, 0.35874)
Population vetting model	0.62956	(0.62003, 0.63908)

CI, confidence interval; N/A, not applicable.

data. This performance is interpreted as evidence of the presence of underlying structure in the data set (i.e., individuals do not choose to identify with a place at random). Of the four baseline models proposed, the proximity model performs the best. The distance vetting model performs quite poorly ($\hat{\theta} = 53.13$ kilometers), which may be a result of heterogeneity of human settlements (an assumption not accounted for in the single-parameter distance vetting model). Although including additional parameters in any of the models, including the distance vetting model, improves overall performance, the relative performance of the distance vetting model most likely would not change given its initial shortcomings in the single-parameter version (e.g., 20% reduction in accurate predictions compared with the proximity model). The proximity model and population vetting model $\hat{\theta} = 1{,}480$ people) are statistically indistinguishable, and the limit of the population vetting model *is* the proximity model (if $\theta = 0$ the population vetting model is identical to the proximity model). Overall, from the national results presented in Table 1 imply (1) regional identification is *not* a random process and (2) of the baseline models proposed, the proximity hypothesis is the most likely mechanism for regional identification at the community scale.

Tie volume model versus the proximity model: a state-by-state analysis

Because it was not possible to utilize the tie volume model nationally, this article presents a state-by-state comparison of the proportion correctly predicted by the tie volume model and the proximity model (where both models have been provided a limited choice set such that only the cities within a state are considered; both models are penalized by individuals who select communities out of state; Table 2). The proximity model is chosen for this comparison because it is the best performing model for the baseline hypotheses.[10,11]

Inspecting the pattern of results presented in Table 2, the tie volume model consistently outperforms the proximity model, sometimes by as much as 27% in the Hägerstrand SIF case and by as much as 45% in the case of Facebook SIF. The tie volume model performs *significantly* better than the proximity model for 31 of the 48 continental states and DC in the case of the Hägerstrand SIF, and 47 of the 48 continental states and DC in the case of the Facebook SIF. In other words, the tie volume model has significantly better prediction of regional identification for almost all of the states analyzed, and a greater raw number of correct predictions for *all* but one state analyzed (and this case is not significant) in the case of the Hägerstrand SIF, and all states in the case of Facebook SIF. If one takes the aggregation of the tie volume model applied to each state individually as an estimate for the full contiguous United States and then compares this estimate to that for proximity model, one again finds a highly significant result for both SIFs

Table 2 State-by-State Comparison of the Tie Volume Model versus the Proximity Model; Tie Volume Model and Proximity Model Proportions Predicted Correctly Where Both Models Have Been Provided a Limited Choice Set Such That Only the Cities within a State Are Considered

	Hägerstrand SIF				Facebook SIF			
	TV	Prox.	Diff.	*P*-value	TV	Prox.	Diff.	*P*-value
Alabama	0.7107	0.6311	0.0796	0.0046*	0.7767	0.6311	0.1462	0.0000*
Arizona	0.7120	0.4839	0.2281	0.0000*	0.7414	0.4839	0.2540	0.0000*
Arkansas	0.8320	0.7033	0.1286	0.0000*	0.8594	0.7019	0.1618	0.0000*
California	0.6384	0.5302	0.1081	0.0000*	0.6763	0.5302	0.1476	0.0000*
Colorado	0.7684	0.4433	0.3250	0.0000*	0.8067	0.4429	0.3626	0.0000*
Connecticut	0.3984	0.3938	0.0047	0.8642	0.5611	0.3938	0.1738	0.0000*
DC	0.5510	0.5510	0.0000	1.0000	0.5510	0.5510	0.0000	1.0000
Delaware	0.5852	0.5057	0.0795	0.1351	0.6640	0.5057	0.1619	0.0000*
Florida	0.6435	0.4597	0.1838	0.0000*	0.6925	0.4488	0.2426	0.0000*
Georgia	0.5430	0.4547	0.0883	0.0000*	0.5803	0.4547	0.1252	0.0000*
Idaho	0.4081	0.3969	0.0112	0.7318	0.4271	0.3969	0.0334	0.1334
Illinois	0.8121	0.6002	0.2120	0.0000*	0.8768	0.6002	0.2761	0.0000*
Indiana	0.6151	0.5666	0.0485	0.0294*	0.6659	0.5666	0.1003	0.0000*
Iowa	0.8291	0.7450	0.0840	0.0001*	0.8794	0.7450	0.1341	0.0000*
Kansas	0.7930	0.6356	0.1573	0.0000*	0.8805	0.6343	0.2455	0.0000*
Kentucky	0.4859	0.4382	0.0477	0.1067	0.5181	0.4374	0.0797	0.0000*
Louisiana	0.7340	0.4601	0.2739	0.0000*	0.7778	0.4601	0.3190	0.0000*
Maine	0.5447	0.5000	0.0447	0.3218	0.7688	0.5000	0.2754	0.0000*
Maryland	0.5244	0.4810	0.0434	0.0252*	0.5703	0.4781	0.0897	0.0000*
Massachusetts	0.4604	0.4490	0.0113	0.4667	0.5824	0.4490	0.1338	0.0000*
Michigan	0.6858	0.6585	0.0273	0.1174	0.8081	0.6585	0.1484	0.0000*
Minnesota	0.8231	0.6248	0.1983	0.0000*	0.8559	0.6224	0.2334	0.0000*
Mississippi	0.7429	0.6381	0.1048	0.0192*	0.8430	0.6351	0.2089	0.0000*
Missouri	0.7093	0.6166	0.0927	0.0000*	0.7794	0.6166	0.1598	0.0000*
Montana	0.7228	0.7065	0.0163	0.7278	0.7727	0.7065	0.0721	0.0201*
Nebraska	0.8000	0.7323	0.0677	0.0422*	0.8543	0.7278	0.1295	0.0000*
Nevada	0.6842	0.3454	0.3388	0.0000*	0.6574	0.3454	0.3109	0.0000*
New Hampshire	0.5535	0.5203	0.0332	0.4385	0.7636	0.5203	0.2429	0.0000*
New Jersey	0.5105	0.5252	–0.0147	0.4331	0.6851	0.5252	0.1614	0.0000*
New Mexico	0.8182	0.5273	0.2909	0.0000*	0.8374	0.5273	0.3130	0.0000*
New York	0.4409	0.4337	0.0072	0.5943	0.5171	0.4336	0.0830	0.0000*
North Carolina	0.7440	0.5245	0.2195	0.0000*	0.8225	0.5245	0.2981	0.0000*
North Dakota	0.8288	0.6937	0.1351	0.0162*	0.8700	0.6937	0.1798	0.0001*
Ohio	0.6341	0.5592	0.0749	0.0000*	0.7034	0.5595	0.1450	0.0000*
Oklahoma	0.7896	0.4239	0.3657	0.0000*	0.8171	0.4239	0.3939	0.0000*
Oregon	0.6785	0.5302	0.1483	0.0000*	0.6924	0.5302	0.1594	0.0000*
Pennsylvania	0.5421	0.4730	0.0691	0.0000*	0.6299	0.4730	0.1560	0.0000*
Rhode Island	0.6080	0.5227	0.0852	0.1047	0.6923	0.5169	0.1757	0.0000*
South Carolina	0.6524	0.5025	0.1499	0.0000*	0.7145	0.5025	0.2106	0.0000*
South Dakota	0.8226	0.7661	0.0565	0.2700	0.8981	0.7661	0.1242	0.0007*
Tennessee	0.5272	0.4877	0.0395	0.1348	0.5618	0.4871	0.0772	0.0001*
Texas	0.7317	0.5213	0.2104	0.0000*	0.7832	0.5214	0.2616	0.0000*
Utah	0.6831	0.6399	0.0432	0.1525	0.7315	0.6399	0.0895	0.0001*
Vermont	0.4533	0.3667	0.0867	0.1233	0.8243	0.3667	0.4551	0.0000*
Virginia	0.5623	0.5192	0.0430	0.0044*	0.6313	0.5192	0.1092	0.0000*
Washington	0.4770	0.3730	0.1040	0.0000*	0.4830	0.3730	0.1108	0.0000*
West Virginia	0.6270	0.5164	0.1107	0.0133*	0.7015	0.5164	0.1952	0.0000*
Wisconsin	0.7799	0.5619	0.2181	0.0000*	0.8876	0.5602	0.3282	0.0000*
Wyoming	0.8601	0.8042	0.0559	0.2060	0.8958	0.8042	0.0902	0.0060*
Pooled	0.6330	0.5199	0.1131	0.0000*	0.7010	0.5190	0.1814	0.0000*

The difference of the two proportions compared using an unpooled z-test and bootstrap estimated standard errors.
*Denotes significant at 0.05 alpha level.

(Table 2; over 11% for the Hägerstrand SIF and over 18% in the case of Facebook SIF). The Facebook SIF performs significantly better than the national estimates of all the baseline models with a 7% improvement over the best performing model (Tables 1 and 2). As a cautionary note, the pooled tie volume model only moderately outperforms the unconstrained proximity model in the case of the Hägerstrand SIF (which would not be statistically significant); this outcome is moderated by the Facebook SIF results, which statistically outperform all the baseline models.

For the states that do not exhibit a statistically significant difference in the performance of the tie volume versus proximity models (20 of 49, < 50% for the Hägerstrand SIF and 2 of 49; < 5% for the Facebook SIF), at least some of these cases may be due to power constraints (e.g., a state like Delaware, which has only 176 respondents, may lack the requisite statistical power for such a comparison). A closer look at several of the worst performance states (from the perspective of the tie volume model) reveals that several of these cases are ones furnishing arguably fertile ground for out-of-state identification based on the size of the state, as well as the size of nearby (yet not in-state) cities (e.g., Connecticut or Maryland, which are near New York and Washington, DC, respectively).

Discussion and conclusion

This article outlines a cognitive representation of *self-identification*, and further makes a case for regional self-identification as a particularly interesting case study of self-identification. Further, it summarizes an evaluation of six competing hypotheses where we find the social influence model performs the best. The social network model performs the best without fitting to the data (i.e., it is a zero parameter model), whereas the other five models are optimized to the data, thus providing a stronger result. The superior performance of the SNH-based model affirms the theory that regional identification is both a social and a geographical process.

The application of comparable models (and hypotheses of social structure) to other forms of identification (e.g., gender, racial/ethnicity, urban/rural, and national) may possibly shed light on many different areas of identification. For example, the large-scale social network methods can be used to accurately predict even difficult cases of identification (e.g., boundary cases of ethnic identity).

Finally, the successful application of large-scale social network models to the regional identification problem provides further validation for geographical factors as critical drivers of social process (Mayhew 1984). Even very simple spatial network models, incorporating marginal distance effects, are able to predict a complex social psychological process. Applications of such models to other social processes would seem to be a fruitful direction for further research.

Acknowledgements

This work was supported in part by an Office of Naval Research (ONR) award (# N00014-08-1-1015), National Science Foundation (NSF) awards (# BCS-0827027), (# SES-1260798) and (# OIA-1028394), and National Institute of Health (NIH)/National Institute of Child Health & Human Development (NICHD) award (# 1R01HD068395-01). The authors also want to thank Katherine Faust, Yen-Sheng Chiang, Jessamy Norton-Ford/Almquist, and the anonymous reviewers for all their helpful comments.

Notes

1 As a volunteered, self-reported/administered survey, the CCP is necessarily limited by both the design of the instrument and informant inaccuracy. Although error from such sources can never be ruled out, we did not observe evidence suggestive of data quality problems, and we note that our findings are robust to fairly large perturbations in the data set.

2 Question two states: many Americans have addresses that say they live in one town or neighborhood, have a government or police force of another name, and fall in the school district of yet another area. For this question, forget about what all "official" sources have told you and answer whatever you feel you identify with most. What do you consider to be your local community? Do not confuse this with your whole local area; that will be in the next step. This is about the single local community you most feel you live in.

3 Question three states: this step asks for you to identify with a slightly larger area. Again, please ignore all "official" boundaries like counties, telephone area codes, or zip codes, and answer what you feel you identify most with. We need a way to identify your local area—the local community you just specified, together with the local communities that immediately surround it. So, please choose the name of the local community that you feel is the natural cultural and economic center within your local area. Or, if you feel a general name (i.e. "Hope Valley," "Pleasant Lake Area," or "Midway-Fairview Area") is more descriptive of your local area culturally than the name of a single central community, then please give what you feel to be the best commonly accepted name for your local area.

4 It is not obvious that this should be a problem for this article as the mechanisms proposed should be largely universal; however, one might be concerned that the geography/population of the respondents could be systematically different than most "Americans." This however does not appear to be the case as far as can be tested with the anonymized data. We used the block-level data of each individual's home as proxy for their neighborhood and calculated the racial composition of each respondents neighbored as compared with the city, county, and state over series of common demographics and detected only minor variations from what would be expected from a random sample of individuals.

5 The Pew Internet & American Life Survey, December 2010, http://www.pewinternet.org.

6 These algorithms have been implemented in the spatialNetwork in the R statistical environment (R Development Core Team 2010; Butts and Almquist 2013). The spatialNetwork software package requires the user choose a particular parametric form of the SIF.

7 Facebook, an online social networking site, offers a rich context in which to study social relations. Further, it has attracted researchers from many different fields (Lewis et al. 2008; Tufekci 2008; Wimmer and Lewis 2010). Users of the website build detailed personal profiles, including information about demographics, interests, and activities. Beyond personal characteristics, Facebook allows users to publicly declare "friendships" with other users (so called Facebook friendship). Declared friendships must be confirmed by both parties involved, and therefore constitute mutual relationship acknowledged by both individuals. Although much debate exists over the nature of Facebook friendships evidence suggests that Facebook users maintain a significant degree of online/offline integration (Lampe, Ellison, and Steinfield 2006; Wimmer and Lewis 2010). That is, individuals primarily use the service to "friend" others whom they met in an offline context, rather than search out friends with whom they have had no offline interaction. The popularity and global penetration of Facebook makes it extremely attractive to researchers as a source for rich population-level social interaction data. It is one of the most prominent sources for large-scale social network data. Given its extremely high membership (and daily usage) rates, Facebook users have access to a extremely large, diverse (both spatially and demographically) population of potential social contacts.

8 $d(v_i, v_j) = C_r \cos^{-1} [\cos(v_i)_2 \cos(v_j)_2 + \cos((v_i)_1 - (v_j)_1) \sin(v_i)_2 \sin(v_j)_2]$, where C_r is the spherical radius (approximately 6,371 km in the case of the earth).

9 Note that the population vetting model reproduces the proximity model when $\theta = 0$, resulting in a "limited" choice set that is, in fact, the entire choice set. This effect also may be observed in cases in which θ is sufficiently small.

10 One might worry that performing a state-by-state comparison unfairly limits the choice set for the tie volume model. Although this might be the case, we have no evidence that this should be an issue. To this effect we took a moderately sized state (Nebraska) and performed our procedure giving the choice set as all contiguous states with Nebraska and itself (i.e., Wyoming, South Dakota, Iowa, Missouri, Kansas,

Colorado) and the model performed approximately identically with the sole state constraint. Notice that many of the adjoining states have large nearby cities that might influence the prediction, for example, Denver, CO.

11 Results for the baseline models maintain their rank order when given the more limited choice set with a linear decrease in prediction.

References

Almquist, Z. W. (2010). "US Census Spatial and Demographic Data in R: The UScensus2000 Suite of Packages." *Journal of Statistical Software* 37(6), 1–31.

Almquist, Z. W. (2012). "Random Errors in Egocentric Networks." *Social Networks* 34(4), 493–505.

Almquist, Z. W., and C. T. Butts. (2012). "Point Process Models for Household Distributions within Small Areal Units." *Demographic Research* 26(12), 593–632.

Arentze, T. A., and H. J. Timmermans. (2005). "Representing Mental Maps and Cognitive Learning in Micro-Simulation Models of Activity-Travel Choice Dynamics." *Transportation* 32(4), 321–40.

Axhausen, K. W. (2007). "Activity Spaces, Biographies, Social Networks and Their Welfare Gains and Externalities: Some Hypotheses and Empirical Results." *Mobilities* 2(1), 15–36.

Baldwin, M. (2010). "The Common Census Internet Project." http://www.commoncensus.org (Accessed October, 2010).

Barabási, A.-L., and J. Frangos. (2002). *The New Science of Networks*. New York, NY: Perseus Books Group.

Bentley, G. C., R. G. Cromley, and C. Atkinson-Palombo. (2013). "The Network Interpolation of Population for Flow Modeling Using Dasymetric Mapping." *Geographical Analysis* 45, 307–23.

Berretty, P. M., P. M. Todd, and P. W. Blythe. (1997). "Categorization by Elimination: A Fast and Frugal Approach to Categorization." *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, 43–48.

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

Black, W. R. (1992). "Network Autocorrelation in Transport Network and Flow Systems." *Geographical Analysis* 39, 268–92.

Boots, B. N. (1977). "Contact Number Properties in the Study of Cellular Networks." *Geographical Analysis* 9, 379–87.

Bossard, J. H. S. (1932). "Residential Propinquity as a Factor in Marriage Selection." *American Journal of Sociology* 38(2), 219–24.

Butts, C. T. (2002). Spatial Models of Large-Scale Interpersonal Networks. Doctoral Dissertation in the Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA.

Butts, C. T. (2003). Predictability of large-scale spatially embedded networks. In Breiger, R. L., Carley, K. M., and Pattison, P., editors, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, 313–323. National Academies Press, DC.

Butts, C. T. (forthcoming). *Space and Structure: Models and Methods for Large-Scale Interpersonal Networks*. New York: Springer.

Butts, C. T., and R. M. Acton. (2011). "Spatial Modeling of Social Networks." In *The Sage Handbook of GIS and Society Research*, 222–50, edited by T. Nyerges and R. M. Helen Couclelis. Thousand Oaks, CA: Sage Publications.

Butts, C. T., and Z. W. Almquist. (2013). networkSpatial: Tools for the Generation and Analysis of Spatially-Embedded Networks. R package version 1.0.

Butts, C. T., R. M. Acton, J. R. Hipp, and N. N. Nagle. (2012). "Geographical Variability and Network Structure." *Social Networks* 34, 82–100.

Carrasco, J. A., E. J. Miller, and B. Wellman. (2008). "How Far and with Whom Do People Socialize?: Empirical Evidence about Distance between Social Network Members." *Transportation Research Record: Journal of the Transportation Research Board* 2076(1), 114–22.

Daraganova, G., P. Pattison, J. Koskinen, B. Mitchell, A. Bill, M. Watts, and S. Baum. (2012). "Networks and Geography: Modelling Community Network Structures as the Outcome of Both Spatial and Network Processes." *Social Networks* 34(1), 6–17.

Dodds, P. S., R. Muhamad, and D. J. Watts. (2003). "An Experimental Study of Search in Global Social Networks." *Science* 301, 827–9.

Dwass, M. (1957). "Modified Randomization Tests for Nonparametric Hypotheses." *The Annals of Mathematical Statistics* 28, 181–7.

Farber, S., A. Páez, and E. Volz. (2009). "Opology and Dependency Tests in Spatial and Network Autoregressive Models." *Geographical Analysis* 41, 158–80.

Festinger, L., S. Schachter, and K. Back. (1950). *Social Pressures in Informal Groups: A Study of Human Factors in Housing*. Palo Alto, CA: Stanford University Press.

Fischer, C. S. (1982). *To Dwell Among Friends—Personal Networks in Town and City*. Chicago, IL: The University of Chicago Press.

Flanagina, A. J., and M. J. Metzger. (2008). "The Credibility of Volunteered Geographic Information." *GeoJournal* 72, 137–48.

Freeman, L. C., S. C. Freeman, and A. G. Michaelson. (1988). "On Human Social Intelligence." *Journal of Social Biological Structure* 11, 415–25.

Gahegan, M. (2000). "On the Application of Inductive Machine Learning Tools to Geographical Analysis." *Geographical Analysis* 32, 113–39.

Gigerenzer, G., and P. M. Todd (eds.) (1999). *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Gjoka, M., M. Kurant, C. T. Butts, and A. Markopoulou. (2010). Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *Proceedings of IEEE INFOCOM '10*, San Diego, CA.

Goodchild, M. F. (2007). "Citizens as Sensors: The World of Volunteered Geography." *Geo-Journal* 69, 211–21.

Gopal, S., and M. M. Fischer. (1996). "Learning in Single Hidden-Layer Feedforward Network Models: Backpropagation in a Spatial Interaction Modeling Context." *Geographical Analysis* 28, 38–55.

Gould, P., and R. White. (1986). *Mental Maps*, 2nd ed. London: Routledge.

Grannis, R. (2009). *From the Ground Up: Translating Geography into Community through Neighbor Networks*. Princeton, NJ: Princeton University Press.

Griffith, D. (2011). "Geography, Graph Theory, and the New Network Science." *Geographical Analysis* 43, 345–6.

Hägerstrand, T. (1966). "Aspects of the Spatial Structure of Social Communication and the Diffusion of Information." *Papers in Regional Science* 16(1), 27–42.

Handcock, M. S., and J. H. Jones. (2004). "Likelihood-Based Inference for Stochastic Models of Sexual Network Formation." *Theoretical Population Biology* 65, 413–22.

Haynes, K. E., and A. S. Fortheringham. (1984). *Gravity and Spatial Interaction Models, Volume 2 of Scientific Geography*. Beverly Hills, CA: Sage Publications.

Hipp, J. R., C. T. Butts, R. M. Acton, N. N. Nagle, and A. Boessen. (2013). "Extrapolative Simulation of Neighborhood Networks Based on Population Spatial Distribution: Do They Predict Crime?" *Social Networks* 35, 614–25.

Howard, J. A. (2000). "Social Psychology of Identities." *Annual Review of Sociology* 26, 367–93.

Hudson, J. C. (1969). "A Model of Spatial Relations." *Geographical Analysis* 1, 260–71.

Hutchinson, J. M., and G. Gigerenzer. (2005). "Simple Heuristics and Rules of Thumb: Where Psychologists and Behavioral Biologists Might Meet." *Behavioral Process* 69, 97–124.

Jenkins, R. (2000). "Categorization: Identity, Social Process and Epistemology." *Current Sociology* 48(7), 7–25.

Lampe, C., N. Ellison, and C. Steinfield. (2006). A Face (book) in the Crowd: Social Searching vs. Social Browsing. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, 167–70. ACM.

Latané, B., A. Nowak, and J. H. Liu. (1994). "Measuring Emergent Social Phenomena: Dynamism, Polarization, and Clustering as Order Parameters of Social Systems." *Behavioral Science* 39, 1–24.

Lewis, K., J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. (2008). "Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.Com." *Social Networks* 30(4), 330–42.

Mayhew, B. H. (1984). "Chance and Necessity in Sociological Theory." *Journal of Mathematical Sociology* 9, 305–39.

Mayhew, B. H., and R. L. Levinger. (1976). "Size and the Density of Interaction in Human Aggregates." *The American Journal of Sociology* 83(1), 86–110.
McPherson, M., L. Smith-Lovin, and J. M. Cook. (2001). "Birds of a Feather: Homophily in Social Networks." *Annual Review Sociology* 27, 415–44.
Milgram, S. (1967). "The Small-World Problem." *Psychology Today* 1(1), 61–7.
Mirchandani, P. B. (1980). "Locational Decisions on Stochastic Networks." *Geographical Analysis* 12, 172–83.
Morley, C. D., and J. B. Thornes. (1972). "A Markov Decision Model for Network Flows." *Geographical Analysis* 4, 180–93.
Neal, Z. (2012). "Structural Determinism in the Interlocking World City Network." *Geographical Analysis* 44, 162–70.
Okabe, A., and I. Yamada. (2001). "He K-Function Method on a Network and Its Computational Implementation." *Geographical Analysis* 33, 271–90.
Okabe, A., H. Yomono, and M. Kitamura. (1995). "Statistical Analysis of the Distribution of Points on a Network." *Geographical Analysis* 27, 152–75.
Osleeb, J. P., and S. J. Ratick. (1990). "A Dynamic Location-Allocation Model for Evaluating the Spatial Impacts for Just-in-Time Planning." *Geographical Analysis* 22, 50–69.
Páez, A., D. M. Scott, and E. Volz. (2008). "Weight Matrices for Social Influence Analysis: An Investigation of Measurement Errors and Their Effect on Model Identification and Estimation Quality." *Social Networks* 30(4), 309–17.
Peeters, D., and I. Thomas. (2009). "Network Autocorrelation." *Geographical Analysis* 41, 436–43.
Peeters, D., J.-F. Thisse, and I. Thomas. (1998). "Transportation Networks and the Location of Human Activities." *Geographical Analysis* 30, 355–71.
Phillips, F., G. M. White, and K. E. Haynes. (1976). "Extremal Approaches to Estimating Spatial Interaction." *Geographical Analysis* 8, 185–200.
Portugali, J., I. Benenson, and I. Omer. (1994). "Sociospatial Residential Dynamics: Stability and Instability within a Self-Organizing City." *Geographical Analysis* 26, 321–40.
R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rogerson, P. A. (1997). "Estimating the Size of Social Networks." *Geographical Analysis* 29, 50–63.
Schneider, B. (2005). "Extraction of Hierarchical Surface Networks from Bilinear Surface Patches." *Geographical Analysis* 37, 244–63.
Serra, D., and C. ReVelle. (1999). "Competitive Location and Pricing on Networks." *Geographical Analysis* 31, 109–29.
Shiode, S. (2008). "Analysis of a Distribution of Point Events Using the Network-Based Quadrat Method." *Geographical Analysis* 20, 122–39.
Smith, T. R., J. W. Pellegrino, and R. G. Golledge. (1982). "Computational Process Modeling of Spatial Cognition and Behavior." *Geographical Analysis* 14, 305–25.
Spiro, E. S., Z. W. Almquist, and C. T. Butts. (2012). The Persistence of Division: Geography, Institutions, and Online Friendship Ties. Working Paper, Department of Sociology, University of California, Irvine.
Tan, S. (2005). *Challenging Citizenship: Group Membership and Cultural Identity in a Global Age*. Burlington, VT: Ashgate Publishing Limited.
Taylor, P. J. (2001). "Specification of the World City Network." *Geographical Analysis* 33, 181–94.
Tinkler, K. J. (1972). "Bounded Planar Networks: A Theory of Radial Structures." *Geographical Analysis* 4, 5–33.
Townsley, M. (2009). "Spatial Autocorrelation and Impacts on Criminology." *Geographical Analysis* 41, 452–61.
Travers, J., and S. Milgram. (1969). "An Experimental Study of the Small World Problem." *Sociometry* 32(4), 425–43.
Tufekci, Z. (2008). "Grooming, Gossip, Facebook and Myspace." *Information, Communication & Society* 11(4), 544–64.
Turner, J. C., M. A. Hogg, P. J. Oakes, S. D. Reicher, and M. S. Wetherell. (1987). *Rediscovering the Social Group: A Self-Categorization Theory*. New York: Basil Blackwell Ltd.

Tversky, A. (1972). "Elimination by Aspects: A Theory of Choice." *Psychological Review* 79(4), 281–99.

US Census Bureau. (2001). *Census 2000 Summary File 1 United States/Prepared by the U.S. Census Bureau. Technical Report*. US Census Bureau.

Watts, D. J., and S. H. Strogatz. (1998). "Collective Dynamics of 'Small-World' Networks." *Nature* 393(6684), 440–2.

Werner, C. (1972). "Patterns of Drainage Areas with Random Topology." *Geographical Analysis* 4, 119–33.

Whalen, K. E., A. Páez, C. Bhat, M. Moniruzzaman, and R. Paleti. (2012). "T-Communities and Sense of Community in a University Town: Evidence from a Student Sample Using a Spatial Ordered Response Model." *Urban Studies* 49, 1357–76.

Wimmer, A., and K. Lewis. (2010). "Beyond and below Racial Homophily: Erg Models of a Friendship Network Documented on Facebook." *American Journal of Sociology* 116(2), 583–642.

Wirth, L. (1938). "Urbanism as a Way of Life." *The American Journal of Sociology* 44, 1–24.

Xie, F., and D. Levinson. (2009). "Effect of Small-World Networks on Epidemic Propagation and Intervention." *Geographical Analysis* 41, 263–82.

Xu, Z., and D. Z. Sui. (2009). "Effect of Small-World Networks on Epidemic Propagation and Intervention." *Geographical Analysis* 41, 263–82.

Yamada, I., and J.-C. Thill. (2007). "Local Indicators of Network-Constrained Clusters in Spatial Point Patterns." *Geographical Analysis* 39, 268–92.

Zemanian, A. H. (1980). "Two-Level Periodic Marketing Networks Wherein Traders Store Goods." *Geographical Analysis* 12, 353–72.