CO-423

METHODOLOGIES FOR IDENTIFYING SOCIAL NETWORK STRUCTURES ONLINE: UTILIZING GIS FOR CYBERCARTOGRAPHY

STEPHENS M.
University of Arizona, TUCSON, UNITED STATES

INTRODUCTION

Understanding how online social networks connect people in physical space would allow us to better conceptualize the trajectories of the individual and relationships in cyberspace. More specifically, we could gain an understanding of how the internet redefines the role of distance within human relationships. The majority of research in online social networks has been conducted outside of the realm of geography, largely ignoring Cartesian relationships among individuals.

This study examines two different online social networks that connect individuals to social relations in other locations through understanding their relationships in Cartesian space. By emphasizing the structures of internet networks, this research asks the following questions

- Are people bypassing potential acquaintances geographically closer to themselves to connect to those they have more in common with (at a greater distance)?
- How are physical distance and social distance reconfigured by the internet?
- What are the Cartesian spatial dimensions of connectivity within online social networks? Do these adapt to or redefine Tobler's law?

Broadly, this study develops a methodology to examine the structure of social connectivity among individuals in cyberspace through a Cartesian network analysis. I examine whether individuals bypass local opportunities for employment, relationships and friends to focus on the physically more distant, but perhaps more suitable or more similar, opportunities accessible through the internet.

To examine whether spatial proximity is a factor in the selection of social networks, I assume if "near things are more related than distant things" (Tobler, 1970, p. 236), then users should be utilizing virtual social networks as a technological replacement of geographically defined community groups (Wellman & Leighton, 1979). Online social networks only provide a bridge to introducing and connecting people in their local area with somewhat similar interests. If this were not true, then internet social network users would cultivate the local through not developing and maintaining social network ties to more similar users in distant places. I test this through developing a model of distance decay in cyberspace (see methods section below).

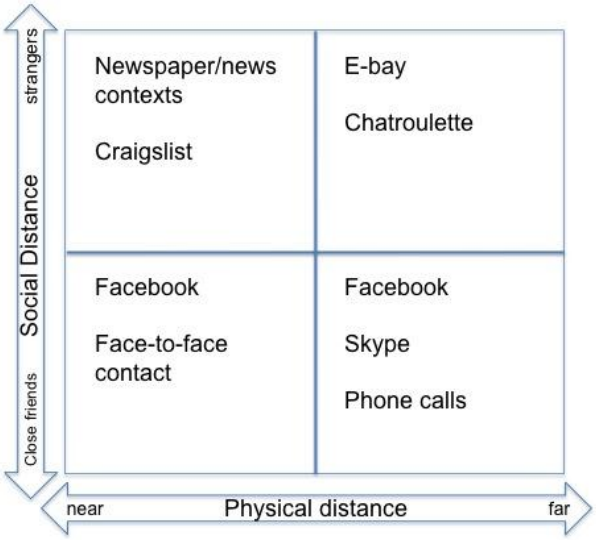

Fig 1. Conceptual internet use over social and physical distance

My methods involve reconstructing a large web-based network in a GIS, through which we can compare all possible matches in the local area with the same number of matches in the national pool of users and determine if persons are more related to those in their local area. This will provide an understanding of the importance of Cartesian distance between nodes (individuals) in the online network. If online networks allow for the bypassing powerful spatial constraints, new and distant connections among individuals are

increasingly possible and hold the potential to reconfigure modern relationships. In short, the internet's potential ability to substitute social distance for physical distance directly challenges Tobler's (1970) law by changing the type of distance that matters.

This paper includes two distinctly different concepts of space. One relates to the Cartesian distance that measures the physical space among individuals. Individuals "near" each other are physically more proximate. The second, conceptualized as "social distance" among nodes in online networks considers "nearness" as shared social commonalities and social relationships.

Since networks frequently apply to global processes, the annihilation of space-time relationships (Harvey, 1989) and the reduction of physical distance as a barrier to inter-personal relationships are frequently cited as two of the primary characteristics of the new global economy. Digital networks have the potential to extend to include remote regions, bringing users together around mutual interests and needs rather than just spatial proximity (Zuckerman, 2008). These networks also enable actors to bypass adjacent opportunities to connect with spaces across the globe (Graham, 1998).

In the mid-1990s, utopian interpretations of the internet argued that cyberspaces would bring together diverse groups of people to share ideas and knowledge. Barlow (1996) believed that the internet would make factories obsolete as it would allow "whatever the human mind may create [to] be reproduced and distributed infinitely at no cost." The internet's ability to transmit information over large distances would revolutionize the economy by transmitting digital media, aka "bits" (entertainment and news) globally rather than distributing physical goods aka "atoms" (Negroponte, 1995). Negroponte's (1995) idea of the internet as a forum to bring diverse and disparate communities together over common ideas has not been realized to the promised extent.

Aoyama and Sheppard (2003) described the internet as a "key catalyst of digital globalization" which essentially levels the "economic playing field." However the internet does not do this evenly, as both the physical and electronic spaces reconfigure social and economic relationships (Whalley, Williams, 2001). Only some places benefit from the internet highway as others are bypassed entirely for areas of more relevance to the global economy, such as "world cities" (Graham & Aurigi, 1997). It is up for debate whether internet-mediated communications can replace face-to-face communications in "geographic space" (Aoyama, 1999). While some believe distance is becoming less relevant as spatial and temporal boundaries are lowered (Harvey, 1989; Kitchin, 1998), and geospace serves as a surrogate for real-space (M. Dodge & Shiode, 2000). Others claim cyberspace and physical space are intrinsically intertwined (Wellman, 2001), complementing each other while bolstering diversity (Walmsley, 2000). Wellman et al. (2001) described the internet as a way to foster interpersonal relationships as it enables people to autonomously participate in groups with a shared identity while giving individuals more power to communicate.

This paper analyzes the internet as a surrogate for physical space that enables social relationships. The two datasets (www.facebook.com, and www.americansingles.com) utilize the internet to identify individuals for users to connect to. By bringing together a Tobler-ian conception of cost-distance analysis to understand social networks, this methodology merges social network analysis with geospatatial analysis, producing a better understanding of the contrast between social distance and physical (Cartesian) distance within online social networks.

PROBLEM STATEMENT

The internet controls and reconfigures how individuals connect to each other and communicate. It has the potential to modify dating patterns (Holme, et al., 2004), employment prospects (Niles & Hanson, 2003), and day-to-day communication. An abundance of recent research has identified the impact of the internet on society (Boase et al., 2006; Chayko, 2002) and speculated on the causality of this relationship to communities (Chayko, 2008).

It is the structure of the relationships among actors within any social network that determines the impact of an innovation such as the internet (Shirky, 2005). I am particularly interested in how physical distance and social distance are mutually reconfigured by the internet. The associations between Cartesian space and relatedness in a virtual social network are still under examined and hold significant potential for understanding the social structure of communities in cyberspace.

The objective of this paper is to develop methodologies that simultaneously describe and geovisualize the social connections and physical distances among individuals in cyberspace. I will use this method to examine two different networks: an internet dating website (AmericanSingles.com) and a social network (www.facebook.com). Currently geographers do not have the tools or methods to analyze how virtual networks map on Cartesian space. We are limited by the extricate nature of spatial analysis and

visualization, such as econometrics, Visual Interactive Modeling (VIM), Spatial Statistics, and GIS that cannot compare Cartesian distance with social distance, and either are exclusive to visualization of a dataset without analysis or rely heavily on spatial proximity logic (Densham & Armstrong, 1993; Florax & Van der Vlist, 2003). Contemporary network models within the social sciences rarely consider the physical spatial element of 'connectivity' or they weigh connectivity based on Cartesian models without regard to the social and physical space-transcending power of the internet.

LITERATURE REVIEW

I review the bodies of literature that consider the impact of the internet on modern configurations of both social and physical space. More specifically, there are three contending perspectives on the connectivity of the internet. Some believe the Internet simply replaces existing technology (such as phone and mail) by connecting adjoining physical spaces digitally. By this logic, the distance decay theory is employed and the internet decreases in connectivity as distance increases. A second perspective presupposes the total annihilation of space-time relationships believing that everybody connects to individuals universally across the globe instantly. A third position, which this paper subsumes, believes in the internet as both a universal connecting force and a new technology. In this mediated perspective, cyberspace provides "wormholes" which subvert time-space relationships that have changed global interactions (Sheppard, 2002). These perspectives shape this literature review into the following three sections: GIS and Cartesian Networks; a Utopian Cyberspace; and Social Networks in a Mediated Cyberspace.

GIS AND CARTESIAN NETWORKS

Tobler (1970, p. 236) maintained a Cartesian perspective, utilizing the position of x,y coordinates, on Cartesian space with his claim that "near things are more related than distant things." Traditional spatial analysis reverts back to this perspective as Geographic Information Systems (GIS) rely heavily on spatial dependency models. This has been applied to many regional science and economic geography studies such as Econometric models where geographic data maintains two properties: spatial dependence and spatial homogeneity which inherently both apply Tobler's (1970) law (Anselin, 1989). A second example includes time geography, which makes the assumption of a space-path reliant on geographic proximity and relationships in the local domain (Miller, 2005). Social networks online provide a new example of relationships that no longer necessitate physical/geographic proximity.

Geographic research on social networks has largely focused on informal networks, primarily in a metaphorical context. While Saxenian (1992) Leamer and Stroper (2001) and Stroper and Venables (2004) have studied the embedded spatial relationships among places through the lens of the internet; Dodge and Kitchin have mapped the networks of cyberspace (M Dodge & Kitchin, 2001), and others have critically examined the relationships among corporate entities and power dynamics on the internet (Zook, 2000, 2005; Zook & Graham, 2007). The individual as a unit of analysis is often overlooked as studies focus primarily on the larger relationships among spaces connected through the Internet. Individuals are the agents that interact online and are the smallest unit of analysis possible in Geography and individuals are a factor in all spatial units of social analysis. It is the individuals that inform cyberspace and are modifying their socialization pattern through the internet.

While a positivist approach to understanding the internet, Glückler (2007) identified that physical proximity affects social network formation which is frequently geographic and subject to regional agglomeration, however he did not use GIS to analyze any networks. Farber, Páez et al. (2009) compared network analysis to GIS finding they both involve measurement of dependence among observations, while controlling the network topology, they determined the effect of clustering was small but increased connectivity resulted in a decrease of the strength of their models (Farber et al., 2009). I intend to use spatial analysis with GIS as a tool to understand the connectivity among individuals in large social networks online.

A UTOPIAN CYBERSPACE

Assuming Tobler's law--that related things are more related than near things, we can incorporate Negroponte's (1995) belief that the internet would universally connect individuals and create a global community. This liberated community would change our experience of the real world by challenging both the social and the space-based monopolies of the media (Rheingold, 1993) and creating a 'new social space' separate from the offline world (Morley & Robins, 1995).

While Geography is being reconfigured by the internet, and space-time compression is lessening the importance of space (Harvey, 1989), location is still important for utilizing face-to-face social networks, the workforce, and access to physical items (Kitchin, 1998; Zook, 2005). In fact, Geographers have tended to discount the strength of virtual connections by focusing primarily on face-to-face interactions.

Negroponte was optimistic about the possibilities of cyber communities. He assumed that nationality, age, class, race and gender would become obsolete categories as technology allowed for a disembodied, utopian space where communication, participation, andengagement are universally possible (Negroponte, 1995). While lauding the ability to instantly and continuously personalize ones interaction with the world, Negroponte neglected to consider the homogenizing force of personalizing the internet which continues to be segregated as users define their segregated social networks in the digital world (Sunstein, 2001).

SOCIAL NETWORKS IN A MEDIATED CYBERSPACE

Social networks do not require Cartesian space, but require some connectivity among the individuals who comprise the network. At times this may correlate with spatial proximity. With the compression of time and physical space, these networks can extend globally and connect users around mutual interests instead of spatial proximity. However, contemporary research indicates this is not a universal phenomenon but a selective process that bypasses some physically proximate spaces to connect social spaces with more commonalities. This section brings together social networks with the geographies of cyberspace.

As the connectivity among the actors is the focus of my research, I subscribe to a basic understanding of networks. A network is "A specific set of linkages among a defined set of persons with the additional property that the characteristics of these linkages as a whole may be used to interpret the social behavior of the person involved" (Mitchell 1969, p. 2).

This definition of a "network" transcends both online and offline networks. Prior to the internet, people generally selected demographically, characteristically and behaviorally similar social networks (McPherson, et al 2001), which were also physically more proximate. This preference toward similarity continues to be true with social networks represented on the internet albeit not necessarily with the physically proximate (Lewis, et al 2008). Similarly, some research reports that most people select homogenous internet content that does not conflict with their demographic, situated, viewpoint—in other words, people both online and off-line are largely "xenophobic" (Sunstein, 2001) in their social networks.

Xenophobia decreases as one interacts with diverse body, however, not everybody has the option to interact with a plurality of individuals, especially those living in rural areas. Contemporary research on the spatial limitations of internet communities shows that internet content differs between urban and rural areas. In rural areas, internet-based social networks offer an alternative to the small number of options available through traditional means of meeting people in physical space (Sinai & Waldfogel, 2004). In urban areas there is a greater number of individuals that constitute social networks both online and offline that provide a larger selection of potential social partners. The internet also allows both urban and rural users the option to extend their spatial boundaries and interact with prospective partners farther away. Given the smaller number of social actors in rural areas, this may be seen as particularly advantageous to rural people. This implies that urban and rural internet users have different predispositions in relation to how the internet works with their social networks.

METHODS AND ANALYSIS

The objective of this research is to develop and test a methodology to describe and geovisualize the social connectivity of an online- social network and to analyze the spatial relations within the network to understand if social network users are bypassing physically local social networks for social connections farther away. To do this I use the proxy of an internet dating website (AmericanSingles.com) and a social network (www.facebook.com). In order to achieve this objective, 1.) I develop a model for physical spatial interaction within cyberspace, 2.) I harvest the data to create multi-dimensional geo-spatial and non-spatial databases for each website, and 3.) Then apply the model to query for connectivity, and analyze the results. While this model has other potential applications, I will only be using it to compare virtual connectivity and physical distance.

DATA SOURCES

Facebook.com began in 2004 as a social networking site to connect alumni of a few elite universities, by 2006 the site was available to everybody over the age of 13 (Arthur & Kiss, 2010). In 2010, www.facebook.com surpassed 500 million users to be the most popular social networking site in the world (Zuckerberg, 2010). The social network allows each user to "connect" with "friends" through linkages on their "facebook page." Facebook reports that the average user has 130 friends, or linkages among users within the network, and 50% of users log in to the site daily ("Statistics," 2010). While the Facebook.com network differs substantially from face to face networks, it also serves as an augmentation mechanism for people who know each other already. The connections established by this social network maintain a many-to-many topology that differs from an internet dating network.

AmericanSingles.com is owned by Spark Networks and has declined in popularity in recent years to a membership of roughly 30,000 users. The site is the 84,400th most visited site in the United States and has been surpassed by most other dating sites such as match.com and Okcupid.com (Alexa Social Media, 2010). The site maintains a one-to-one network topology.

MODELING SOCIAL CLOSENESS WITH LSI

Match.com, as well as many other social network sites, use Latent Semantic Indexing (LSI) to search and rank prospective partners (Gould, 2010). LSI does not inherently consider physical distance as a criterion. Through creating a multi-dimensional matrix, in which each attribute a user contributes about herself/himself creates another dimension, LSI indexes all the potential matches. These dimensions are complied and weighted through an unknown algorithm to rank prospective partners by maximum commonalities.

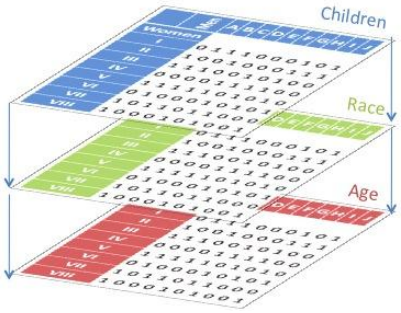

Fig 2. A simplification of LSI

These internet sites use many advanced database queries to suggest potential matches for each user. I am unable to replicate this exact function because each matching algorithm is proprietary to the company concerned. For this reason, I use Simulink's MatLab software to query each database for all potential connections within a user's local area and then use this number of matches (*m*) to pull the same number (*m*) of random matches within the national database. For this study, a 'potential connection' is a dyadic relationship between users with a matched specified attribute, for example a male user may seek a female user, or users may share a potential connection through their stated interest in sushi. This creates two matrices of potential "matches" for each user, one based on physical proximity and a randomly selected sample of the entire database. These two matrices will then be compared using LSI to identify which users have more commonalities, the random users selected from the national pool (*nca*) or the local users (*lca*). I believe the comparison of these aggregates will indicate the homogeneity of local people over the national pool.

This will develop as follows:

Let *I* be the set of integers

$I = \{1,2,...,m\}$

Each local user will be identified by a vector u_i of length *P* (the number attributes about a user) where *i* is an integer in the index set *I*. Each national user will be identified by a vector u'_i of length *P* as well as where *I* is an integer in the index set $J = \{1,2,...m-1\}$

The latent semantic index of two users *u, v,* is a function on the vectors *u, v*, returning a real number and written as:

LSI(u, v).

The *local compatibility aggregate* for the user u_i, written as $lca(u_i)$, is the mean of the set $\left\{LSI(\vec{u}_i,\vec{u}_j)\right\}_{j\in I\setminus\{i\}}$. That is,

$$lca(\vec{u}_i)=\frac{1}{m-1}\sum_{\substack{j=1\\ j\neq i}}^{m} LSI(\vec{u}_i,\vec{u}_j)$$

The *national compatibility aggregate* for the user u_i, written as $nca(u_i)$, is the mean of the set $\left\{LSI(\vec{u}_i,\vec{u}'_j)\right\}_{j\in J}$. That is,

$$nca(\vec{u}_i)=\frac{1}{m-1}\sum_{j=1}^{m-1} LSI(\vec{u}_i,\vec{u}'_j)$$

.

This is done as a bulk operation with a binary variable attributed to each user consolidated into a singular database for each network.

I then apply this database to a GIS, in which each user has been geocoded to an explicit spatial coordinate based on information contained in the dataset. Users are geocoded at the zip code level with the zip-code aggregated to the common name (city, state) areas. For example, Ann Arbor, Michigan would be geocoded as one spatial area although it is the agglomerate of nine different zip codes. This is done to avoid the Modifiable Area Unit Problem (MAUP) where the network patters formed by individuals' internet use are compared at a different geometry geometric scale than they are located within. This forms three databases for the remainder of this analysis.

I will also run multiple tests to understand the equilibrium point between nca and lca in the following sense; Consider a randomly selected collection of n users ($n \geq m$) from the national database, denoted by R_n. Let $nca(u_i,R_n)$ be the national compatibility aggregate for the user u_i with respect to the users in R_n. As n increases, we hypothesize that the difference between $nca(u_i,R_n)$ and lca will decrease until $nca(u_i,R_n)$ is greater.

Using a summation of the LSI matrices a final matrix is created. In this matrix each potential partner has a score based on the number of variables for which each dyadic pair match.

Men / Women	A	B	C	D	E	F	G	H	I	J
I	10	20	12	5	15	2	6	8	10	4
II	4	8	7	2	20	5	3	9	11	3
III	1	0	18	5	3	6	8	14	10	9
IV	10	0	4	9	2	1	20	1	0	7
V	1	1	16	1	9	0	1	0	4	0
VI	2	4	0	20	2	10	4	12	0	7
VII	1	9	1	11	9	1	16	11	3	0
VIII	16	7	9	9	4	0	12	4	8	1

Fig 3. An example LSI summation matrix

For each user these scores are then ranked. Two databases are created from this matrix with each user and their ten best-matches in the both the local and national sample. Ten was chosen as a threshold as match quality declines rapidly and ten maintains the integrity of the dataset. This database is exported into the GIS.

A GIS layers data for analysis purposes, however rather than using textual or binary layers like LSI, a GIS utilizes spatially referenced layers. To achieve this it is necessary to establish a spatially referenced database by geocoding each member of the website to an area that agglomerates zip-codes with the same name. The database created from the LSI process will then be joined with the geocoded layer of users.

A GIS analysis of networks must be based largely on the quantitative techniques of network analysis. This type of analysis provides the tools to study the structure of the entities and the interactions through mapping nodes to edges (Wasserman & Faust, 1994). Actors in the social network will be considered "nodes" while "edges" will represent theconnectivity among them and will be analyzed for the distance between each dyadic pair.

A spider mapping algorithm is used to draw directional lines (network edges) representing connectivity between the users (nodes). When compiled, this represents the spatial-social topography of a large internet network. Edges are drawn to connect each node to

their top ten ranked suitable matches. This allows for conclusions to be drawn about the scale and dimensions of this network. For example, a user with short edges would have a physically constrained network while long edges suggest space-time compression.

This analysis is twofold—first I employ the methodology laid out above to create a model for geovisualizing large networks. Second, I create a spider-diagram from the dataset, from this we can derive the physical distance required to establish a "connection" among users in the database. From the summation score derived in the LSI process we determine what percentage of each users' social network can be met and at what physical distance. Essentially this is a mapping of availability. Through visually inspecting the network we can understand the level at which physical distance and social distance reach equilibrium.

If "near things are more related than distant things" (Tobler, 1970, p. 236), then users should be more attracted to people in their local area. As the entire process is repeated with a larger sample of national

users, when does the summation matrix for the local domain become smaller than the national? The relationship between the weight given to location and the score of potential matches is a distance decay of cyber space. I imagine this relationship would be in the form of a weighted distance decay such as:

$S=1/d^2$

Where 'S' is equal to the average score of all the potential matches derived from the LSI summation matrix and 'd' is the distance. This relationship will be calculated and presented as part of this project as will a statistical analysis of the distance decay of cyber-space (formula above). Additionally, I will attempt to map a mathematical function for distance decay to the observed pattern.

To conduct the demographic analysis I will compare US Census data to the demographics harvested from each social network for each geographic area using a one-tailed T-test. This will determine if any demographic variable (race, education, income) is over represented in the sample derived from the social network. This will help determine if results from this study are applicable to the larger population.

Each network will differ slightly in the topology derived from the network. The Facebook.com network, with a many-to-many topology shape, will be an undirected network with more edges than vertices (approximately 150 to 1), a small clustering coefficient, and a high degree correlation coefficient. The AmericanSingles.com network with a one-to-one network topology will be a directed network with a small number of vertices roughly equivalent to the small number of edges, an insignificant clustering coefficient and potentially a negative degree correlation (Newman, 2003). Both networks would presumably have a "social distance" as individuals associate based on occupation, location, interest, etc… (Newman, 2003; Watts, et al., 2002). This study will determine a Cartesian measure for "social distance" within both social networks. Lastly, for a comparative measure, I will overlay the network maps to evaluate the differences between the network structures.

Geographic research on the internet has largely focused on the production of internet content (Zook, 2000), virtual social communities created by the internet (Fortin and Sanderson 1999), discrepancies in access to the internet (Grubesic & Murray, 2002), and the role of the internet in modifying spatial-economic relationships (E. Leamer & M. Storper, 2001). This research will contribute an added dimension to the literature surrounding internet geography through adding an analysis of spatial variation within social networks.

Through combining and expanding upon these studies I hope to contribute a methodology that can be useful to conduct future GIS- based analyses of the Internet. Through challenging the necessity of spatial dependency within these social networks, I will suppose that related things are more related than near things.