Metrics and Methods for Social Distance

by

Clio Andris

Bachelors of Arts, American Studies, Boston University, 2006
Master of Science, Geography, University of South Carolina, 2008

Submitted to
The Department of Urban Studies and Planning
In partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Urban Information Systems

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2011

Author ____
Department of Urban Studies and Planning
August 11, 2011

Certified by ____
Professor Joseph Ferreira, Jr.
Professor of Urban Planning and Operations Research
Dissertation Supervisor

Accepted by ____
Professor Karen R. Polenkse
Peter de Florez Professor of Regional Political Economy
Head, PhD Committee
Department of Urban Studies and Planning

Metrics and Methods for Social Distance

by
Clio Andris

Submitted to the Department of Urban Studies and Planning
on August 11, 2011 in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy in Urban Information Systems

ABSTRACT

Distance measures are important for scientists because they illustrate the dynamics of geospatial topologies for physical and social processes. Two major types of distance are generally used for this purpose: *Euclidean Distance* measures the geodesic dispersion between fixed locations and *Cost Distance* characterizes the ease of travel between two places. This dissertation suggests that close inter-place ties may be an effect of human decisions and relationships and so embraces a third tier of distance, *Social Distance,* as the conceptual or physical connectivity between two places as measured by the relative or absolute frequency, volume or intensity of agent-based choices to travel, communicate or relate from one distinct place to another.

In the spatial realm, Social Distance measures have not been widely developed, and since the concept is relatively new, Chapter 1 introduces and defines geo-contextual Social Distance, its operationalization, and its novelty. With similar intentions, Chapter 2 outlines the challenges facing the integration of social flow data into the Geographic Information community.

The body of this dissertation consists of three separate case studies in Chapters 3, 4 and 5 whose common theme is the integration of Social Distance as models of social processes in geographic space. Each chapter addresses one aspect of this topic. Chapter 3 looks at a new visualization and classification method, called Weighted Radial Variation, for flow datasets. U.S. Migration data at the county level for 2008 is used for this case study. Chapter 4 discusses a new computational method for predicting geospatial interaction, based on social theory of trip chaining and communication. U.S. Flight, Trip and Migration data for the years 1995-2008 are used in this study. Chapter 5 presents the results of the tandem analysis for social networks and geographic clustering. Roll call vote data for the U.S. House of Representatives in the 111th Congress are used to create a social network, which is then analyzed with regards to the geographic districts of each congressperson.

Each of the case study chapters addresses one or more of the major challenges presented in Chapter 2, which are discussed in Chapter 6 and concluded in Chapter 7.

Thesis Supervisor: Dr. Joseph Ferreira Jr.
Title: Professor of Urban Studies and Planning

TABLE OF CONTENTS

LIST OF FIGURES 6

LIST OF TABLES 9

CHAPTER 1: INTRODUCTION

1.1 A System of Systems 11
1.2 The Social System 11
1.3 Types of Social Flows 13
1.4 Mature Systems 16
1.4.1 Conceptualizing Mature Systems 16
1.4.2 Modeling Mature Systems 19
1.5 Thesis Overview 21

CHAPTER 2: CHALLENGES FOR SOCIAL FLOWS IN GIS

Eight Challenges for Social Flows in GIS 23

CHAPTER 3

Weighted Radial Variation for Visualizing Migration Dynamics 51

CHAPTER 4

Predicting Migration System Dynamics with Conditional and Posterior Probabilities 71

CHAPTER 5

Social and Spatial Patterns in the U.S. House of Representatives 73

CHAPTER 6: DISCUSSION

6.1 Review of System of Systems Concept 125
6.2 Suggestions for GIS and CSS Communities 127
6.3 Experiences with Case Study Chapters 137
6.4 Technical and Computational Improvements with Social Flow Data 142

CHAPTER 7: CONCLUSION

7.1 Theories for Embedded Relationships 153
7.2 Connections and Urban Planning Goals 156
7.3 Ethics and Privacy 163
7.4 Conclusions 167

BIBLIOGRAPHY 171

LIST OF FIGURES

-CHAPTER 2-

Figure 2.1: K-means Clustering of Migration Patterns 29

Figure 2.2: Network Dependency and User Agency Characteristics of Social Flow Transactions 32

Figure 2.3: Three Visualizations of U.S. Migration Flows 33

Figure 2.4: Example of a Prototypical Exploratory Social/Spatial Data Analysis Software Environment 35

Figure 2.5: Image of a Flow from Albany, NY to Akron, OH and Corresponding Table Records in a GIS Environment 39

Figure 2.6: Example Schematic for Components of Flow Data in a GeoDatabase 40

Figures 2.7A and 7B: Examples of Two City Clusters 42

Figure 2.7C Cities Partitioned by Color based on Interaction Patterns 43

Figure 2.7D: Increasingly Complex Interactions between Cities 44

-CHAPTER 3-

Figure 3.1: 2008 Migration Flows in the Continental U.S. 52

Figure 3.2: Rank Size Distribution of Migrants (R) and Rank Size Degree Distribution of Nodes (L) 54

Figure 3.3: Log-Log Distribution of Edge Weights (Migration Flow Volume 55

Figure 3.4: Distribution of Migrant Flows Edge Lengths 55

Figure 3.5: Distribution of Migrant Flows by Edge Angle Values 56

Figure 3.6: Selected Migration "Stars" in the Continental U.S. 56

Figure 3.7: Migration Stars in New York State and 'Unraveled' Signal of Edges 57

Figure 3.8: Results from K-Means Classification of Migration Star Arrays 59

Figure 3.9: Features of 10 K-Means Classes 61

-CHAPTER 4-

Figure 4.1: Metropolitan Statistical Area Boundaries and Available Data for the Contiguous U.S.....77

Figure 4.2: Annual Summary Values for Place-Pair Transition Probabilities........84

-CHAPTER 5-

Figure 5.1: Map of U.S. Congressional Districts, Representative Names and Parties........94

Figure 5.2: Roll Call Vote Responses per Representative........96

Figure 5.3: Frequency Distribution of Pair-wise Agreement Rate........98

Figure 5.4: Social Network Degree Distribution........99

Figure 5.5: Social Network of Congresspersons Force-Directed by Binary Tie........104

Figure 5.6: Core and Periphery Roles of Representatives........106

Figure 5.7: Network Location of High Degree Nodes in the Social Network........108

Figure 5.8: Network Degree (Friend) Values........109

Figure 5.9: Network Degree (Friend) Values Normalized by Party........110

Figure 5.10: Representatives with High and Low Closeness Centrality Measures........111

Figure 5.11: Map of Representative Closeness Centrality Values........112

Figure 5.12: Spectrum of High to Low Betweenness Centrality Values........113

Figure 5.13: Representatives with High and Low Clustering Coefficients........114

Figure 5.14: Location of Representatives with Notable Clustering Coefficient Values........115

Figure 5.15: Pairs with Furthest Maximum Connected Network Distance (4 Hops)........116

Figure 5.16: Social Network of Representative Lynn Jenkins........117

Figure 5.17: Choropleth map of Representatives' Network Distance from Kansas Representative Lynn Jenkins........119

Figure 5.18: Smoothed Network Distance from Kansas Representative Lynn Jenkins........120

-Chapter 6-

Figure 6.1: Network Dependency and User Agency Characteristics of Social Flow Transactions... 144

Figure 6.2: Pairs with Furthest Maximum Connected Network Distance ... 145

Figure 6.3: Table and New York City Region Map View of Congresspersons and Districts ... 149

Figure 6.4: View of a Network Graph Diagram in the Gephi Software Environment ... 150

Figure 6.5: City Clusters and Potential Connections ... 152

-Chapter 7-

Figure 7.1: A Tri-cyclic Model of Variables that Sustain Social Relationships ... 155

LIST OF TABLES

-CHAPTER 3-

Table I: Entity (Star) Statistics by K-means Class ... 60

Table II: Typologies of Common Structures by K-means Class ... 60

-CHAPTER 4-

Table I: Metropolitan Area System Migrants ... 77

Table II: Calibrated Gravity Model Parameters ... 81

Table III: Airline Flights and Gravity Model Correlation Coefficients ... 82

Table IV: Correlation Coefficients for Transition Probabilities at a One-Year Time Step ... 82

Table V: Correlation Coefficients for Expected and Actual Migration Dynamics (Linear Extrapolation ... 83

Table VI: Correlation Coefficients for Expected and Actual Migration Dynamics (10 Year Projection ... 83

Table VII: Temporal Transition Probability Averages and Standard Deviations ... 84

- CHAPTER 5-

Table I: 111th Congress, First Session Vote Frequency ... 95

Table II: Demographics by Party ... 104

Table III: Descriptive Demographic Statistics By Party ... 105

Table IV: Local Moran's I Scores ... 105

Table V: Descriptive Network Statistics by Party ... 107

-CHAPTER 6-

Table I: Cost/Benefit Measures in Transit and Land Use Models 126

Table II: Three Case Studies in "Metrics and Methods for Social Distance" and their Contributions 137

CHAPTER 1: INTRODUCTION

1.1 A SYSTEM OF SYSTEMS

Our great world is driven by increasingly complex systems of systems where (1) humans interact with one another, (2) humans interact with the landscape, and (3) the landscape interacts with itself. These systems are not self-contained or closed; they interact with one another, and across system lines. Thus, our set of multiple systems can be thought of not as a group of systems, but as a system of systems.

This idea is not foreign to some fields of research. For example, in anatomy, the circulatory system interacts with the cardiovascular system, intestine system, muscular and bone structures to produce the system of the human body. These interacting systems are more basic than social systems, as the system components are not driven by human complexities like self-awareness, desires, self-consciousness, and multi-criteria decisions. Therefore, these physical systems are more clearly defined than complex systems in other fields, like social sciences (*anthropology, psychology, sociology, behavioral and cognitive science and organization science*), and spatial sciences (*urban planning, geography, environmental science, and operations research*). Social and spatial systems tend to be individually modeled as closed systems of agents, characterized by the dynamics of the connections between these agents.

Consider a friendship study in a middle school, a network of hyperlinked blog references, or e-mail contacts within a workplace: each network has one type of actor, and the peer-to-peer activity remains within these sets of actors. Consider spatial processes, such as trajectories of metropolitan area commuters travelling from suburbs to central business districts, paths of migrating animals, or a highway route strategy for a fleet of trucks. Like the agent/phenomenon nature of the social processes, in spatial networks, the actors are described not by their choice to connect with other actors, but by their existence in a certain location at a certain time.

When considering social and spatial as standalone systems, there is much evidence for the successful modeling of *social* processes with behavioral, agent based, game-driven or discrete choice models, and *geographic* processes with a proven toolbox of spatial statistics and agent-based processes, such as cellular automata. But to model social and spatial as a system of systems, where social/cognitive choices are represented spatially, our methods and evidence of successful models seem to fall short. The following section introduces the concepts of a geo-located social system.

1.2 THE SOCIAL SYSTEM

We have increasingly reliable evidence that indicates that people relate to one another, and also suggests how and why they relate to one another, under what circumstances, and how frequently. Social network analysis has revolutionized the way social research is conducted. This analysis approach not only allows for the enumeration of a system agent's contacts and "friends", and that the connections between agents can be quantified, but has proven that these social networks exhibit a great amount of regularity. Computational social science models of friendship, communication, or travel focus on autonomous agents making decisions, and indicate that sociologically and psychologically complex micro (agent based) behavior and discrete choices weave large scale (macro) dynamics. As a result, many studies incorporate the work Blau (1986) on power relationships, or Schelling on segregation (1978), Arrow's Impossibility Theorem (1952), and Nash (for example, 1950) on game theory equilibrium convergence, (Kearns et al 2006) Prisoner's Dilemma, belief propagation (Salganik and Watts 2008) or concepts of "homophily", (McPherson et al 2001) "not in my back yard" or "pluralistic ignorance". (Centola and Wilier 2009)

Human decisions drive much of social connectivity and so, social networks configuration can be evidence of segregation, social norms, assimilation, language barriers, and trust. When considering the relative simplicity of Euclidean and Cost Distance network models, the complexity of computational social science ensues: roads do not decide to attach to an intersection; a biological cell doesn't refuse to replicate because he has a fear of being outcast, and so our system of systems proves to be challenging.

In health networks, studies show that friendships and kinships can influence obesity and smoking behavior (Christakis and Fowler 2007, 2008), although there has been recent argument with these findings (Lyons 2011). In a cascade model, individual decisions, based on trust and emulation, viral paths, and their cascades, are very hard to predict. (Watts and Dodds 2007) Social networks are well suited for Power-law degree distributions, network transitivity, and community structure (Girvan and Newman 2002). These systems exhibit *Small World* properties, where the presence of even just a few long edges makes for much easier, fluid, shorter, traversal through branches of agents. (Watts and Strogatz 1998) These systems also exhibit *Scale Free* properties, formed by "preferential attachment", where high-degree nodes are most attractive, and so many nodes have a small degree, and few nodes have high degree values.

Researchers have been successful in correctly predicting "future friendships" (Clauset et al 2008) and showing that unlike an electrical, circuit, or cable networks, social networks are resilient to 'blackout', (Watts 2002) but could fall faster under "attack" than a random network. (Albert et al 2000, Ebel et al 2002) The Strength of Weak Ties asserts that far-fetched friends help the spreading of information across the system. (Granovetter 1973, revisited by Arbesman et al 2008 in an urban context), and maximizing 'social capital' (Burt 1998, Fernandez et al 2000). Without these contacts, tight-knit communities can have trouble spreading information to other communities. (Granovetter 2003)

With the intuition behind why a movement to incorporate social flow metrics into studies of movement and connectivity could be deemed eminent and beneficial, we now further bolster this viewpoint with a review of evidence of the correlation between *social interaction* and *urban or geographic space*. New data and network analysis methods have proven useful in understanding human relationships without needed a geographic or Euclidean distance model. With this in mind we are reminded that physical GIS research and this human relationships research are each successful on stand-alone and so supports the systems of systems approach.

1.3 Types of Social Flows

As mentioned, many human actions, decisions, and behaviors are based on relationships with others. Just as the work of Ben-Akiva and Lerman (1985) and Ben-Akiva and Boccara (1995) has developed the discrete choice models that revolutionized civil engineering and trumped traditional aggregate travel simulations, in geographic systems, strides should be made to model discrete choices in coordination with social relationships.

The occurrence of transactions, either the transmission of people or information, over space may be evidence of the human condition, goals, and emotions, as realized by the ties we forage with friends and kin. Current reasoning for transactions and connectivity do not directly address the social-influential aspect of our decision-making. Social flows are most typically driven by the spreading of a phenomenon between places as quantified by the number of exchanged migrants, commuters, travelers, tourists, phone calls or instant messages, and email. (Limtanakool et al 2007) Some of these data sources are the IRS, Census and Metropolitan Planning Organizations (migration and commuters), tourism bureaus, embassies, intersections and consulates, telecom companies (like Orange, British Telecom, or AT&T), and airline statistics (like the Federal Aviation Administration or a major carrier corporation like Star Alliance). In addition, urban sensing and tracking devices like RFID, Cell Networks and GPS can also be used to describe the human flows between two places by extracting origin and destination. The fusion of these data is a frequent occurrence, as telephone calls and other communication connections and travel are very much intertwined. (Rietveld and Janssen 1990) Moreover, social networks, like those from online networks, UN or political representative networks, neighborhood friend networks at a school, or geocoded dynamics addresses of a church community have led to new social flow analysis. Traditionally, social ties have been volunteered in surveys, but automated communication records, and even cell-phone proximity studies (see the work of Eagle et al (2009) for an example), are more recently used for social analysis.

These quantifiable measures of social connectivity are constantly becoming more bountiful and accessible and may benefit from a rudimentary classification system from which to organize the types of connections and ensure their proper use. Just as sociological fields use a number of different kinds of data sources, like surveys, telephone records, and field observation to uncover structures of interaction and have a set of qualitative (and some quantitative), geographic connectivity draws from multiple data sources as well. We can argue that roads and trains connect places, but in reality, there is no information transfer, analogy, or shared experience between cities or places unless carried by humans. With this newer reasoning, we can formalize newer approaches to measuring functional dispersion in the built environment.

The measurement and quantitative data behind social flows can be categorized into physical movement in transportation (human transfers) and the changing of ideas or the non-physical interacting with others in communication (signaled transfers). With human flows, the connectivity between places is driven by transit and so there is actual, traceable movement. Communication has the same ability to distort the world, and so it is reasonable to say that social connectivity differs from Euclidean proximity and is becoming increasingly important.

PHYSICAL FLOWS

Physical flows, most often realized through transit represent the movement of humans from one place to another. Many years ago, this transfer was easily understood as the ability to travel through the environment—initially by foot, horse, or ship, and then via carriages, trains, automobiles and aircraft. This progress in increasing transportation speed and efficiency can also be understood as the lessening of a traveler interacting with his surroundings as he moves from origin to destination. Evidence of transit includes: Permanent OD Flows [Migrants, Commuters], Traces [Traffic Traces, Pedestrian Traces], Temporary OD Flows [Traffic Counts, Bus Riders, Taxi, Subway, Train, Flights]. These data are collected by: Digital Records [Urban Sensors, Cell Activity, GPS Traces] or by Human Collectors [Surveys, Car Counts, Ridership Volume Observations], or Automated O/D [Subway Cards, Digital Ticket Purchase, Ridership Records).

Cost Distance

Another important factor in physical connections is cost distance, which is the metric most often used for traffic models, trip planning, route assignment, emergency response, and supply chains, for example. Although the 'cost' of something can be complicated, it can generally be computed. Some of these 'complications' are latent variables like a mode of transit's safety for women, or accessibility to the elderly and handicapped, or whether a bus has outlets for one's laptop. These economic variables all have bearing on the absolute attractiveness of transit options along a path from distinct origins to a possible destination. Some cost distance metrics include: Length Of Path [Road Length, Number of Transfers], Time Of Trip [Travel Speed, Time Spend Waiting for Arrival/Departure, Traffic Lights], Monetary Costs [Ticket Prices, Per-Kilometer Fares, Tolls, Gasoline], Temporal Features [Safety, Weather, Traffic, On or Off-Peak Fares, Closings, Construction, Hindering Events, etc.], and Recreational Features [Opportunity for Self-Activity, Privacy, WiFi, Electricity, Cell Service].

Moreover, Economic Geography, Civil Engineering, Transportation Planning and Operations Research communities have rigorously modeled logistical, transportation and some telecommunications flows, leading to important discoveries in shortest path calibration, like solutions for the Traveling Salesman Problem, Chinese Postman Problem, methods like Dijkstra's Algorithm and computationally-friendly solutions like the Christofides Heuristic. (Larson and Odoni 1981)

SIGNALED FLOWS

A signaled transfer can be thought of as the transmission of ideas, information, affect, or personal or mass media. More specifically, communication activity comprises the second part of the social flow family. This activity includes geolocated use of landline telephone calls and physical mail, and digital technologies like SMS (text messaging) and cellular phones, Voice over Internet Protocol (VoIP), e-mails and instant messages. Given the nature of digital transfers, the Internet has driven much of this research through Instant Messages, Social Networking Sites, and sometimes online communities like Mechanical Turk, Support or Health-buddy groups. New technologies have made research in this area more feasible, so now we see our previously invisible communication patterns as large scale evidence of social initiatives to connect. Evidence of the propensity and action to connect with one another over space includes: O/D Communication [Point-To-Point Calls (Landline or Cell), Emails, Instant Messages, SMS, Facsimile] and Interest [Web searches, uploads to Geo-Media]. Communication collection methods include: Web Records [Email, Social Network Sites, IM Records, Web Hits, Geo-Tagged Uploads, Place and Map Searches] Analog Volume [Mail System Counts], Monetary & Spending Records [Tax Records, Personal data from businesses like credit card companies]).

Access Networks

Some types of communication do not require a connection between two agents, but only one agent's desire to access spatial media, like an agent's exposure to advertisements, branding, or broadcast. In analog terms, ordering a newspaper or place-themed publication may indicate a connection between the place of the recipient and the topic of the media. For example, a Washington D.C. transplant from California may still get copies of the *Modesto Bee* magazine in her mailbox. In digital form, geographic web searches can link the geo-located IP address of a user who searches for places and place-related media online. For example, the D.C. migrant may look for flights back to California, could search for a Modesto hotel when visiting, can listen to a local Modesto radio station webcast, or can access the *Modesto Bee* newspaper online. Each of these connections is evidence of a link between Washington, D.C. and Modesto, California.

Interestingly, this intent may be persuaded by better infrastructure or less expensive options, which in turn, may convert intent to a travel event, in an understandably recursive system. If a geolocated user checks a college website often, or a small-town newspaper site, it may be evidence of a future or past migration, which would be evidenced in migrant tables. Some research has already shown that evidence of relationships (telephone calls) often complement traffic, trade, migration, commuting, social visits, and tourism. (Rietveld and Janssen 1990)

Spatializing Social Networks

Communications do not always need to be quantified by volume or intensity. The rise of computational social science, and specifically social networks, seem to uncover interesting patterns in human relations via the existence of a 'friendship' by self-report (ex. an online social network) or happenstance (ex. the boss-worker relationship in a business). Evidence of a landscape of committed agent-to-agent connections can be found on photo upload sites, professional network sites, and sites that host persona pages. These and other social networks can be spatialized by linking a social agent (ex. Governor Brown) to a geographic entity/agent (ex. Texas). Relationships can henceforth be conceptualized at the individual, group, or universal level. For example, there are three 'focal' or 'ego' relationships: Gov. Brown has a relationship with Texas (EgoSocial-Spatial), Texas has a relationship with other states (EgoSpatial-Spatial) Gov. Brown has relationships with other governors (EgoSocial-Social). In the universal system, e.g. without a singular focus, governors have relationships with each other (Social-Social), and their states (Social-Spatial), and the states have topological relationships independent of the social network (Spatial-Spatial).

This linking of spatial/social networks is a hallmark example of system of systems representations, as using place as an agent's feature is different than using a hair color, gender, education level, or political party as a feature, because a place is not nominal, ordinal, or a ratio or interval value. It is instead part of a complex underlying system of topological relationships. When attaching or fusing two traditionally disjointed data systems we introduce a new level of complexity: Alone, social or geographic configuration each with *n* discrete entities can be configured in n! ways, but given that both interact, this number jumps to n!*n!. This number of combinations is modest and would balloon further when considering dimensions of social or geographic variables (like Census data) and the types of tests applied to each network (like types of centrality measures or spatial autocorrelation). Even when disregarding the seemingly daunting if not unmanageable number of pattern combinations, the baseline implications, orthopraxy, approaches and methodologies for deriving knowledge from these systems of systems are still generally unclear.

After addressing two major types of flows, their sub-types and respective datasets, we now delve further into the social system, and the system in geographic space.

1.4 Mature Systems

1.4.1 Conceptualizing Mature Systems

A geographic information systems model relies mostly on continuous geodesic space, or a discontinuous representation of this space. The core GIS model represents discrete objects in the Euclidean world. Some objects are vector based and can be located in some coordinate in space, while others are raster-based, where each piece of space is characterized depending on its contents. The analysis of geographic data depends mostly on proximity or the combination of multiple layers of co-located information.

The GIS world has grown up to show discrete objects in continuous space, whereas social connectivity has a natural graph theory representation. These configurations lead to a natural layering of geography as a plane that hosts social phenomena that takes place in specific locations. However, GIS models may not serve as the most conducive stage for models that show inter-agent connectivity. Although GIS models have been a fixture in commercial, industrial, military, and civil engineering operations, the spatial functionality and platforms used to construct these models may be fundamentally designed to focus too heavily on entity situation instead of interaction. In the GIS realm, geographic processes are conceptualized and modeled as *static* phenomena, concentrating mostly on the *location* of entities and quantifying geographic dynamics by entity densities or proximities to other entities. When considering the characteristics of communications and transit, notably volume, intensity, scale and frequency of inter-place connections, we see that current GIS platforms can very easily lay the tracks for these transitive processes by holding place-to-place edges in a geodatabase and model or 'solve' cost or Euclidean based optimization problems with operations algorithms. While successful at laying these 'tracks', GIS systems may be improved with increased ability to model activity on these tracks.

More theoretically, from the vantage point of an economist, geographic transactions (flows, connections) are hypothesized to fill certain dynamics because of econometric prophesies of equalizing unemployment rates and income opportunities among a system of places. Some more advanced economic models focus on safety, gender, industry, and housing market affordability. From that of a geographer, these connections are most probably foraged where proximal closeness determines the volume of connectivity. This raison d'être is formalized by Tobler's First Law of Geography, which roughly states that all things are connected, but closer things are more connected. For transportation and operations researchers, "closer" now accounts for travel time and cost, and has evolved into "accessible." The access of friendships, relationships, and social ties is also an important feature of convenience, and perhaps is most often overlooked in GIS endeavors.

Geographic models often do not include the social variables which could explain why a new bridge isn't well used, the wild popularity of an out-of-the-way store, the creation of a new mosque in Nebraska, the emergence of the Boston/Austin Connection, or the reasons why an MIT student cannot find a ride to Worcester, MA but can to Brooklyn, NY. Sometimes we can speculate. But can we empirically know? Currently, as mentioned, our tools generally fall short. In order to measure strength of inter-place connectivity, metrics, methods, and models that determine strength, propensity, and volume of relationships over space are needed.

A model of geographic situation may show that two nations share a border, and therefore we might conclude that these two nations are intertwined such that a change in one nation will affect the neighbor; Historical or political geography factors may render this baseline hypothesis incorrect. (See De Blij et al 2007) Underlying reasons could include different languages, trade embargos, nation-mandated travel sanctions, hostile relations, or lack of transit infrastructure. In a GIS, it is challenging to show a relationship like that of the U.S. and Panama at the turn of the century, modern customer service areas in India, the change of relations after Nixon's 1972 visit to China, or the seminal tearing down of the Berlin Wall, whose divisive aftermath can still be seen today.

The cause of movement and propensity to connect between places can be caused by human trip chaining, where migration and visits to family and friends are chosen because of a friend at the destination. Certainly a decision to visit a friend could be more likely than a decision to visit each place of lower cost distance to ones origin, yet some current models have rarely explored these considerations. The logistics of connectivity also play a role in decisions to frequent places or choose certain paths, as many transit models have described through concepts of trip chaining, activity-based demand modeling, and destination likelihoods. For example, although a grocery store may be close to one's home, and a model may predict that the time-saving customer would patronize the store, he may instead patronize a store in the opposite direction, as it is a convenient stop on the trajectory from his child's school to home. More discreetly, subtle human reactions are increasingly difficult to model: human relationships, our memories, goals and emotions drive us to "return to the scene of the crime", and although this 'return' has been successfully modeled, the human factors that determine these behaviors may be challenging to model. Though seemingly alien to geographic models, these cognitive, self-awareness and affective concepts have been main topics in Computer Science via Artificial Intelligence research for over fifty years.

While GIS models may have matured without computational social science components, the reverse is also true: social networks have grown up outside of the geographic world and could benefit from GIS techniques as well. Geographic concepts of distance are already being utilized in social system analysis. Social space can be conceptualized as a non-Euclidean, but perhaps Cartesian feature space in which scalar values can be used to quantitatively describe the dispersion between agents and entities. Typically, these scalar distance metrics are calculated by enumerating the steps from one entity of interest to another within a social network, or by computing variable/feature distance via linear algebra techniques like matrix covariance or Mahalanobis Distance. These measures are in no direct relation to geography's longitude and latitude grid, but are useful measures of 'distance' for characterizing social topology.

In addition to these measures of dispersion, overlaying social ties on to layers of geographic processes may be able to enrich the prediction power of social behavior models. Euclidean proximity, the main metric of geographic distance for social scientists, is an excellent heuristic for geographic clustering or dispersion, a deeper analysis of geographic layers and processes could also make social models more robust. Specifically, GIS models of the built environment can encourage (1) better variables of cost distance, (2) the possibilities of relationship-enabling infrastructure, (3) the relative convenience of chosen paths compared with the remainder of the landscape and (4) scaling considerations like the Modifiable Areal Unit Problem (MAUP).

Additionally, the push to incorporate geographic identity features can be strengthened when considering that social territoriality is almost a necessity for government policy (school weather closing or liquor laws) as most policies are applied to citizens under the heading of their geographic jurisdiction. Even policies that target a specific income bracket or age, like tax policies or public education, is still applied via geographic area. Furthermore, if a virus was propagating through a social network, a spatial analyst would need to rectify the transmissions with Euclidean space in order to warn others, invoke policy and contain the virus. An effort to contact the 'friends' of an infected or susceptible agent in the epidemiological social network would not only be difficult, but would likely invade serious privacy laws. Social networks can also help us understand how the physical environment is used, spatially and temporally, and how this information relates to the quality of life. For example, if an analyst studies pedestrian connectivity to find common junctions, funnels, and bottlenecks, she could use this information to build places that are more conducive to vis-à-vis connections.

The following section describes some progress in the tandem analysis of social and spatial models.

1.4.2 MODELING MATURE SYSTEMS

One early 'analog' social and spatial experiment is Milgram's Letter Experiment, which traced the paths of friend-to-friend pass-forward 'chain' letters bound for Sharon, Massachusetts from Nebraska and Kansas. Milgram found that 6 hops, on average, are necessary for successful transmission, a finding that Watts and Strogatz (1998) could later confirm with a more complex process and a larger dataset. From a geographic standpoint, this experiment evidenced the Euclidean distance of a 'social reach' of a typical Midwesterner, and surprisingly, that the majority of letter hops were due to fuzzy social ties that could keep letters trapped circulating around the small town of Sharon. (Milgram 1967, Travers and Milgram 1969)

Existing research that uses computational social network methods to analyze geographic processes may indicate that geographers and transportation analysts are receptive and excited about the convergence of social and spatial phenomena. Road Networks have been approached and characterized with formal network science measures and analysis of properties like *Small World* and *Scale Free*. (Limtanakool et al 2009, Xu and Sui 2007) Markedly different in configuration, airline dynamics have been treated with similar measures and tests as non-Euclidean networks: (Xu and Sui 2007) small-world (Guimera et al 2005), rich-club network, triangles with high traffic volumes, with non-stop distance (Xu and Harriss 2008), scale-free properties (DeMontis et al 2010), and relative 'diameter' of air travel flows, and hub and spoke configuration (O'Kelley 1998, Barrat et al. 2004). Additionally, railway systems (Sen et al. 2003), and the subway system (Latora and Marchiori 2002), have undergone the descriptive analysis with 'textbook' network science.

In addition to characterizing geographic space with network system methods, more evidence for impending ties between social and spatial processes can be found through previous work. Making strides in geographic flows, the work of Tobler speaks prominently, providing initiatives for visualization software, theories, and computation. (Tobler 1959, 1978, 1987) The work of Haggett and Chorley (1969) is also a classic for modeling geographic flows. More recently, major steps seem to be taken to advance the agenda of using social networks to enrich information in the built environment (a decidedly feed-forward, reciprocal process). This includes recent advancements in mining trajectories (Giannotti et al 2007, Nanni and Pedreschi 2006, Gonzalez et al 2008, Liu et al 2010), affixing social network topology onto polygonal geographic space (Radil, et al 2010), comparing social network centrality to geographic centrality (Onnela, Arbesman et al. 2011) and modularizing geographic regions based on telephone call patterns (Ratti et al 2010), in addition to older research where telephone data was used to uncover urban-city-town hierarchies (Davies 1979) and draw boundaries between two nearby cities (e.g. is Connecticut New York or Boston-loyal? a la Green 1955).

Furthermore, scientists have shown the relationship between social communication frequency and distance. On a Twitter Network, 80% of friends are separated by over 1,000 kilometers (km), while over 50% of friends on FourSquare are separated by less a< 1 km. (Scellato, et al. 2010) Instant messenger conversations are longer, but less frequent when the pair is distant, (Leskovec and Horvitz 2008). Communication probability decreases with distance in a cell phone network, and friendship probability follows suit in a blog community. (Liben-Nowell et al. 2005) In early efforts to quantify the 'drop off' in connectivity, Lambiotte et al., report that the average duration of phone calls increases with distance, reaching a plateau around 40 km. (2008) Meanwhile, social behavior is further related to geographic space, where researchers find that mere 0.4% of Facebook friendships are those who have never met (Mayer and Puller 2008), baby-naming patterns shows significant locality trends, (Goldenberg and Levy 2009) and that creative innovation is more likely in denser environments. (Bettencourt et al 2007)

Though dialog in the area of merging human intention with geographic space is sparse, the work of Torrens (2010) explicitly discusses the role of computational social science in a geographic context. Additionally, we can look to geographer Matthew Zook, who writes: "While the concepts of strong and weak ties in social action are well established, I feel it is worth revisiting how they relate to space and time. We have moved beyond simply equating physical proximity with strong ties but there is clearly a relationship albeit complex, multifaceted, and with scalar effects." (2010) Encouragement also comes from the theories of Lefebvre, who reflects on the intertwinement of systems, where space is not a scientific object removed from ideology and politics in "Reflections of the Politics of Space."

Others have noticed a lack of convergence of social and spatial fields. (Gastner and Newman 2006, Xu and Harriss 2008) DeMontis et al (2010) are early formalizers of a distance that measures "the magnetism between two places in the digital era", what the authors label as logical distance. Meanwhile, Batty (2005) was an early voice for Complex Network Analysis and showing spatial interactions in GIS. Kwan is an advocate of accounting for "new topologies of spatial interaction" when studying urban travel, and also astutely cites e-shopping and e-banking as examples of how transit networks may evolve. (Kwan 2007) The epidemiological community has also voiced a need for spatial and social methods for health-policy interventions, as conventional approaches underestimate how a place contributes to the risk of disease, and relational views would help identify crucial reciprocal relationships between people and place. (Cummins et al. 2007)

While these inceptions of place and space in social relationships models are encouraging, it seems to be increasingly evident that there may be limitations to the communication of Geographic Information Systems and Social Systems. This prospect leaves us with many questions, some of which are addressed in the following section, but are driven by the overlying question of how we can take advantage of and leverage both. It seems as though social network analysts and computational social scientists have been successful in understanding which kinds of models are needed to model human behavior, but geographers and GIS analysts do not seem to account for these kinds of models in the GIS world. GIS topological models lay a surface which human decisions can be perhaps more accurately modeled, but the existing geographic systems computer infrastructure may be missing something intrinsic from which human behavior models already benefit.

Since social and spatial systems have developed and matured without significant cross-pollination or dialog, we realize that the process of knowledge creation from data, and even the research design process, can be seen as arduous and risky. It is challenging to try to take advantage of worlds that maintain separate models, these models have matured over the past three decades to the point of reduced interoperability, audiences rarely speak both linguae, there is little legacy in terminology or concepts. Often, the meeting of two mature worlds requires a more complex integration of different models, or complex interpretations of current model. Given these challenges, potential benefits include the prospect of added precision to the predictive measurements that are integral for operations, logistics, and planning, like city-to-city transit magnitudes, future road capacity, inventories of infrastructure needs for urban growth, city shrinkage, migration forecasts, site suitability for cooperative meeting places, epidemiological spreading, or containment models, and suggestions for business franchise or advertisement expansion.

We further explore these challenges and shortcomings in the following chapter.

1.5 THESIS OVERVIEW

In our globalizing world, social distance can help us better realize the complex intricacies of the built environment imposed by human decision-making and propensity to connect with one another. Thomas Friedman's *The World is Flat*, Manuel Castells' *Rise of the Network Society*, Bill Mitchell's *City of Bits* and the work of Saskia Sassen (2001, 2007), for example, describe globalization and the effects of a world where a fast food chain can exist anywhere on earth that has sufficient electricity, water, and customers. This corporate ubiquity and pervasiveness is the result of an increasingly frictionless flow system. In stark contrast to past centuries when facile transportation and information transfer were less feasible, today, one's social network is no longer neatly contained in an easily-delineated place because these advancements have made increasingly geo-dispersed social networks sustainable. Our 'digital breadcrumbs' can scatter throughout the world, resulting in complex social ties and ties of ties, that are further complicated by variations of fixed and mutable positions within the constraints of geographic space. With this idea in mind, we approach the concept of distance from a new perspective.

This dissertation presents three case studies that focus on computational and experimental methodologies for implementing and developing social distance metrics. The "how to" guide of our first case study chapter [*Visualizing Migration Dynamics Using Weighted Radial Variation*] gives the researcher a new procedure to apply to his data, in order to classify space, and to visualize complex "haystack" spatial network data. Next, [*Predicting Migration System Dynamics with Conditional and Prior Probabilities*] looks at the predictive power of a probabilistic discrete choice model instead of a traditional distance-based model, to estimate future migration flows. This model is "social" because it accounts for human decisions to move to locales of specific character to locals of another character, and the synergy of city-to-city pairs, instead of the sheer distance and population—not social choices—used in the traditional gravity model. Finally, the spatialization of social network dynamics in [*Social and Spatial Patterns in the U.S. House of Representatives*] is the first known integration of network science metrics into a cartographic environment, or a social network with map linkage. This case study analyses four types of network characteristics, popularity, communities, attachment, and propagation, as they correlate to spatial clusters and socio-demographic constituent make-up.

Since much social network data has only recently been contextualized in geographic space, and because we have yet to apply complex systems metrics to webs of dynamic social flows, there remain a number of challenges to successfully turning these social/spatial (s/s) data into useful knowledge. Before presenting our three case studies, we outline some specific foci and challenges to consider for advancing the agenda in Chapter 2. Following, we present three case studies, and conclude in a final chapter.

CHAPTER 2: CHALLENGES FOR SOCIAL FLOWS

RESTATEMENT OF PROBLEM

The evolution from straight line to cost distance predates the widespread use of computers and telecommunications. Since this movement transacted at least 30 years before the 'digital era' that has afforded us data collection, creation, analysis and storage methods, the newest metrics do not incorporate the capture and use of geolocated social interactions: instant messages, e-mails, telephone calls, web posts, wifi-access points, sensors and real-time transit data, and automated transit records (train, subway, flight volume, etc.).

Since the concept of distance has not been widely revisited since the modern transition into cost distance, a new type of distance may be important to review. Here, we introduce 'social' distance, where two places are spatially separated based on the amount and strength of the human and telecommunications flows between them. This 'strength' of human and telecommunications traces in Cartesian space is measured by the presence and magnitude of 'social flows', or more formally, relationship or human-driven decisions to connect over place and space. Social flows are pragmatically vector data with discrete beginnings and ends that traverse geodesic space, and create a special kind of topology system network that challenges the assumptions of traditional GIS endeavors. When using flow data, the focus is no longer on points, landform lines, and polygons; or adjacency and boundary issues. Better known in the GIS community as 'interaction data', or sometimes 'movement', the nature of interaction datasets has posed numerous problems on computation, prediction, representation, classification, contextualization, exploration, analysis and synthesis of connectivity in geographic space. Perhaps more crucially, a lack of fundamental elements like, data objects their organization within spatial data infrastructure (SDI) and spatial database management systems (SDBMS) constructs, and ontological frameworks have also not been formalized, and thus cannot be implemented, operationalized and taught. Without a reliable GIS platform and tools to host interaction data, the analysis of geospatial connectivity and the geographies of social decisions to meet or communicate have little hope of being realized.

IDENTITY

Research in GIS and has been advanced to the point of elevating the field of geographic information theory, management and computation to a rigorous science, referred to as GIScience. Evidence of this achievement is abound: academic journals, doctoral research, geography and urban planning departments, the solidification of geospatial analysis is the numerous programs and classes in higher education, and the textbooks created to teach geographic information inference and analysis methods with as little bias and as much objectivity as possible. Prominent texts include Longley et al (2005); in remote sensing, Strahler (2007), Jensen (1995) are widespread sources. Although profession GIS advocates may have little desire to be recognized as a 'hard scientists', but with the advancements of Computer Science and Information Science education, the GI sciences seem to fit comfortably with researchers in these categories. Some evidence for standardization in this field is the widespread use of GIS software like ESRI's ArcGIS or Remote Sensing packages created by ERDAS or PCI GeoLeika, and GIS, surveying and Remote Sensing certification and education courses.

When considering the formalization or operationalization of a new measure of distance, it is important to reflect on how to approach impending changes in the GIS and GIScience fields. First, we look at some considerations regarding the field, its progress and its limitations. Next, we look at the standpoints and backgrounds of the end users. Then, we consider how to evaluate whether the changes are good. Following, we give some challenges to incorporating social flows into GIS, and finally, reflect on the implications of these changes.

To define GIS, Cowen (1988) lists four options: the database approach, toolbox approach, application approach, and process-oriented approach. We use this framework to explore the drawbacks and benefits for exploring the addition of social flow formalization, implementation and operationalization in GIS.

The *toolbox* definition of GIS derives from the idea that such a system incorporates a sophisticated set of computer-based procedures and algorithms for handling spatial data, and is organized according to the needs of each process-oriented subsystem (e.g., input, analysis, or output). Functions should work together efficiently to enhance the transfer of a variety of different types of geographical data through the system and ultimately into the hands of the end user. Visualization workbenches (like Processing), have been used to make cartographic visualizations (see the work of Ben Fry, Mauro Martino and Aaron Koblin for examples), but do not provide the overall integration of functions.

A *process-oriented* approach to social flows would look at social flows in terms of more than one integrated subsystems, including procedures for the input, storage, retrieval, analysis, and output of geographic information. A process-oriented definition emphasizes the use of an information product, not the automation aspect of the processes. The *application* approach would put emphasis on the type of information being handled, inventory, planning, and management.

The *database* approach would focus on the data structures behind the social flows, what a flow object consists of, how it is stored, retrieved, updated, compressed, and related to other objects within a transparent yet dynamic storage system. As Cowen warns, however, this approach may result in more concern for how the system is performing instead of its utility. In one case, database discussion questioned raster verses vector systems for their representation, with little attention to than substance.

THE USER

One question to consider when addressing challenges and potential solutions for the following reasons is what kind of user is social flow theory and operations are being tailored to fit. If either the professional or the researcher, but not both, are in mind, the information and processes available will not suit the needs of the universal GIS community and it may be difficult to retrofit the field of social flows once best practices have been established. Accordingly, one challenge is creating infrastructure that helps both the academic or research-focused GIScientist and the professional GIS Analyst.

Since professionals seem to put an emphasis on problem solving will little attention to process (Schon 1983), it should be noted that social flow analysis and representation infrastructure should make process transparent and relatively easy to record, repeat and share; the theory, assumptions and metadata behind the flows accessible; and the ability for social flow infrastructure to produce reliable, communicable results that can be used for decision support and problem solving. Practitioners adhering to specified management systems and goals may have special needs, which are summarized nicely by Tomlinson's (2007) *10 States in the GIS Planning Methodology*: 1) consider strategic purpose 2) plan for the planning 3) conduct a technology seminar 4) describe the information products 5) define the system scope 6) create a data design 7) choose a logical data model 8) determine system requirements 9) consider benefit-cost, migration and risk analysis 10) plan the implementation. While not following these steps perfectly, we can certainly keep them close at hand when addressing challenges into implementing social flows into GIS. For example, metadata must be readily available for a practitioner to 'define the system scope', and different data models must be available for the practitioner to 'choose a logical data model.'

The GIScientist may have other needs to fulfill: for example, how to connect is space to society and how to invoke a politicized social ontology to demonstrate this nexus (Peet 1997); how to show and rationalize the formation of dialectical and homologous relationships between social and spatial structures (Soja 1980). GIScientists and Urban Planning Theorists may also consider the relationship between society and space is further complicated when incorporating the ideologies inherent to socially-created space, as According to Lefebvre, "space is not a scientific object removed from ideology and politics; it has always been political and strategic...Space has been shaped and molded from historical and natural elements, but this has been a political process. Space is political and ideological. It is a product literally filled with ideologies." (Lefebvre and Enders 1976)

A user from another field (like sociology, real estate or psychology) should be kept in mind when creating these systems and addressing the challenges to implementing social flows in GIS. This can mean adhering to widely used Graphical User Interface protocols (Shneiderman 1995), making software intuitive, and visualization easy. It can also lead to ensuring that, as with the professionals, a solution to a social question must be able to be obtained from the system, and like the research community, the system must be a reliable model of the actual happenings in the network of traversal or communication across the built environment. Following the lead of other quantitative techniques in the Social Sciences, like statistics, some data mining endeavors, or existing GIS techniques; social flow infrastructure should be able to guide the user from data to knowledge discovery.

PAST AND FUTURE

The end-result of implementing new theories and systems about social flows can be envisioned by the results and contributions of preexisting GIS advancements. Geographic Information Science has borrowed models that developed in other fields but matriculated into the discipline. These include, but are not limited to, agent based modeling (like Cellular Automata), (Couclelis 1997) Neural Networks, (Woodcock and Gopal 2000), clustering (like K-means) and pattern recognition (like Self Organizing Maps). (Miller and Han 2001), Fractals (Batty 2007), models attached to social physics, catastrophe theory, ecological models of Voterra-Lotka equations, thermodynamics and decreasing entropy, simulated annealing (Goodchild and Janelle 2010) artificial intelligence (Openshaw and Openshaw 1997), and graph theory (Assuncao et al 2006, Guo 2008). Motivation to better model the built environment via social flows and innovative technologies echoes that of previous work in GIS progress for urban planners. Thus, previous successes and failures should be noted for emulation or avoidance when pursuing the integration of cross-cutting methods like social flow modeling—whose naissance can be traced to social networks and network science in physics.

SYSTEM EVALUATION

A challenge of addressing challenges is choosing criteria from which to evaluate the success of a new system, or additions to current systems. Although according to Pickles (1997), GIS has a certain framework of space and place that may have resulted in limitations on the kinds of questions that could be asked to evaluate GIS and a dialog on evaluation techniques is helpful. When looking what to add to a GIS system, French and Wiggins (1990) mention that a cost/ benefit analysis may not be best for evaluating the worth of a GIS system because it considers mostly tangible benefits. Although not explicitly, the authors may imply that the cost/benefit analysis may not account for theoretical, metacognitive and exploratory contributions to a field where researchers use the GIS system tools. Some factors that can be tested after a GIS is implemented and used in a planning or local government setting are: system and information quality, individual and organizational effects information use, user satisfaction, and societal effects.

Nedovic-Budic enumerates distinct intangible factors that can be used in the evaluation of the utility of a GIS system: reducing "maladministration", data standards, better analytical procedures, data security, better information, better access to data, enhanced customer service, rigorous data management, and better visualization. Some added features are staying on the cutting edge of technology (or at least keeping pace), improving public image and professionalism, and finally, enhancing employee pride, job satisfaction, morale and motivation, improving the ability to come with unexpected events, and promoting professional development. (Nedovic-Budic 1999, Budic 1994)

THEORY

During the earlier days of GIS, there is evidence that a theoretical basis for GIS was challenging to address. According to Cowen, "the vague definitions of GIS were doing a great disservice to the field by allowing the label of GIS to be applied to almost any software system that can display a map or map-like image on a computer output device." (1988)

Wright et al (1997) theoretically inquire about social implications of GIS: "the messages it sends, whom it empowers, and the responsibility its developers should bear for its eventual use" (p. 346) In response, Pickles posits that the authors have seemed to transition from their "unreconstructed positivism" platform and towards important issues in "contemporary geography", and asks, "Is GIS merely a tool, is it a tool-making enterprise, or is it a science whose focus is the handling, analysis, and representation of geographic data?" (Pickles 1997)

There is little attempt at generating deep intuitions about how these types of model might relate to what is actually happening in urban systems. (Batty 2002) Instead, there is a lot of evidence that engineering sciences seem to be focused on 'cheaper, better, faster', which may hinder our understanding of the social implications of GIS." (Pickles 1997) Also, GIS is seen as focusing primarily on the domain of problem solving, not the domain of science. Instead, a closer focus on thc *process of discovery* and *the understanding of problems* is suggested. More specifically, GIScicntists and GIS analysts should regard the rules governing (1) the creation of the spatial models, measurement, (2) modeling of error propagation, and (3) proofs or theorems of data structure. (Goodchild and Janelle 2010)

In this context, GIS may represent a new kind of science, one that emphasizes visual expression, collaboration, exploration, and intuition, and the uniqueness of place over more traditional concerns for mathematical rigor, hypothesis testing, and generality.

The following challenges can be divided into the following categories: [A] Characterizing and Managing Components, [B] Lack of Representation Infrastructure and Techniques for Inference and Exploration, [C] Computational Problems, and [D] Issues of Standardization.

EIGHT CHALLENGES FOR SOCIAL FLOWS WITHIN A GIS FRAMEWORK

[A] CHARACTERIZING AND MANAGING COMPONENTS

1. DIFFICULTY CHARACTERIZING SYSTEM NODES

A system node can be characterized by its (1) inherent features, but also (2) by its relationship to other nodes.

Characterizing a system node by its inherent features is not a foreign concept for a geographer or network scientist. In one example, the United Kingdom classifies its census regions using the Index of Multiple Deprivation, where different indicators like Income, Employment, Health, Education, Skills, Housing, Living Environment and Crime are used to typecast places. (Department for

Communities and Local Government 2007) Some less-official typologies are Richard Florida's creative classes and ESRI's over 60 American "lifestyle" types in their Community Tapestry. More broadly, Census data in general can be used to characterize and assign features to geographic system nodes. A system node could be one of five different pre-determined income classes, or could have a certain percentage of citizens reported as a certain race.

The difficulties arise when considering nodes for their participation in a system, or for their connections to other nodes. On a simple scale, two nodes participating in a connection cannot easily be characterized from the metrics above. A connection between a high income node and a low income would be poorly characterized by the average between the two incomes (as specific low-to-high relationship would go unseen). To represent the robust details of this connection, the node set must be marked with a new scheme, perhaps "low to high" or "high to low". Using five different income classes, the number of possibilities for an edge is squared, and grows exponentially with the number of features used to define the connecting nodes.

This 'simple scale' example becomes more complex: Since nodes are typically part of many connections many connected nodes must be considered when trying to characterize the connectivity properties of a node based on the behavior of his neighbors. A node at hand can be categorized as having 6 high income links, 3 middle income links, and so on for every category of every feature in question.

This question is not limited to geographic networks, but geographic networks are further complicated by fixed geometries where each flow radiating from a node has a destination, properties of that destination, flow direction, distance and magnitude. According to Dalton et al, graph theory is not completely suitable for geographers, as graph theory emphasizes the number of edges in a network, rather than distance along road, rail, seaways, flight paths, and other geographic trajectories. The authors give example questions are suited for geographic networks, to paraphrase: *Is the network oriented in any particular direction? What is the pattern of lines-- (radial, random or Manhattan grid)? What sorts of distances are involved? Are the routes straight or winding? How are the routes located in relation to other features? How well does it serve all parts of the country, region or area? Is the traffic continuous (like oil) or discrete (like a train)?*

Accordingly, researchers may put geographic nodes in a separate class than social network nodes.

When an edge weight is high, the two nodes are considered well-connected and likely to 'communicate', host many travelers, or share common features. These networks also exhibit nest-like structures akin to the hot spots geographers use to quantify clusters based on their exhibited values. (Boots and Getis 1988, Getis 1984) The High/Low Clustering measures concentrations of high values or low values. (Getis and Ord 1992) Clusters can change significance with different spatial levels. Similar to Openshaw and Taylor's popular Modifiable Areal Unit Problem (MAUP) (1981), networks also have scaling characteristics, where clusters have certain characteristics at one distance, and others at other distances. In spatial analysis, Multi-Distance Spatial Cluster Analysis, based on Ripley's K-Function, shows spatial dependence (clustering or dispersion) over different study area or neighborhood distances. These measures can also take weights, which often represent incident frequency. (Bailey and Gatrell 1995) Since nodes participate in many flows, it is hard to summarize these variables and group by node without major fidelity loss due to summarization.

To address this challenge, we first suggest that feature reduction, supervised or unsupervised classification methods can be used to classify system nodes by their specific geometric radial configurations, as represented by a series of numeric vectors. Some examples of unsupervised clustering methods include Self-Organizing Maps (Guo 2006), Eigenvector decomposition (Calabrese et al 2009) and K-means (Andris 2011). These methods can manage many attributes, e.g. the features of a node's neighbors, and categorize the node based on its group of chosen connections.

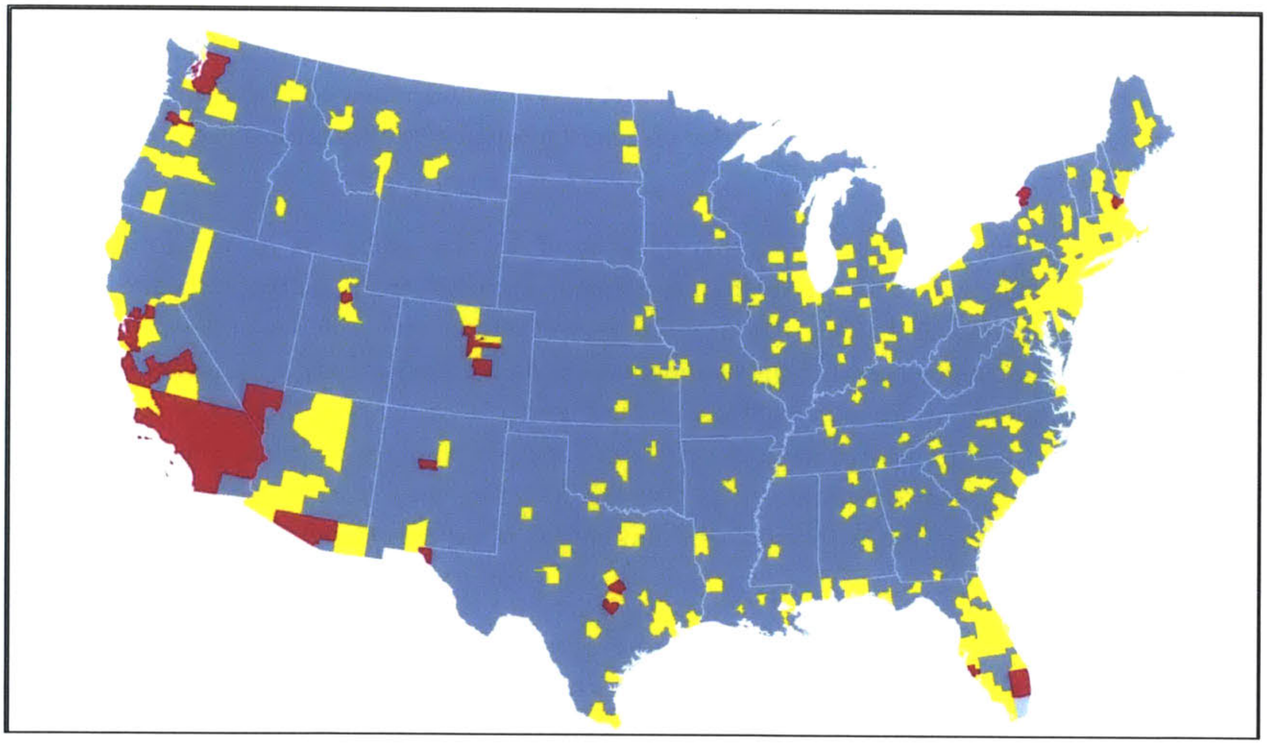


Figure 2.1: K-means Clustering of Migration Pattern **A K-means clustering approach is used to classify counties into one of three categories by their incoming migration flow geometries in 2008. Pink counties have thick radial flows emulating from streamlined directions, yellow counties have flows emulating from more equivocally radial directions, and blue counties show more local, thinner flows from local counties.**

Similar to the work of Skupin's Where do you want to go today (*In Feature Space*)? (2002), another idea for characterizing system nodes is to represent geographic nodes as nominal entities and use a force-directed feature layout (like Frutcherman-Reingold (1991)) to look at similarities among the 'untacked' nodes. This contort the traditional map topology to show a more natural, spring-like network--where edges don't just represent distance, as they do in the tacked network, making traditional social network analysis techniques more applicable to the geo-independent network.

The implication of understanding system nodes in non-Euclidean space perhaps poses a new interpretation Tobler's First Rule of Geography, which generally states that all things are connected to each other but things that are geographically closer are more connected, but does not go against its literal meaning. If "close" is re-engineered to include social and physical interaction, perhaps we can hold this rule in esteem for our purposes. But however much we seemingly ignore cartographic adjacency and distance, the spatial configuration topology will be reflected in the interactions, as geographic proximity affects the number of interactions between places.

2. UNCONSTRUCTED THEORIES OF EDGE ASSIGNMENT

Transportation flows represent the movement of humans from one place to another, but the level of interaction between the flow and the built environment is not always easy to discern from digitized flows. How can be express edges that overlay on maps, in order to better inform the user as to the relevance of the trace?

Many years ago, this transfer was easily understood as the ability to travel through the environment—initially by foot, horse, or ship, and then to carriages, trains, automobiles and aircraft. This progress in increasing transportation speed and efficiency can also be understood as the lessening of a traveler interacting with his surroundings as he moves from origin to destination. With each advancement, humans are immersed less in the experience of the path, yielding a spectrum of total immersion (a walk) to nearly total evasion (an airplane ride) with the path between the travelers target location. When science fiction writers tout teleportation, they describe traversal immersion levels that approach 0, as the path would be untouched in an episode of instantaneous human transfer.

With this spectrum of varying levels of immersed experience on a path from origin to destination, we have another dimension that seems to correlate with simple to technologically advanced transit: a spectrum of increased streamlining and decreased tortuosity. Otherwise stated, the more efficient and modern the mode of transport, the increased likelihood of flows to be pre-determined and straight. For example, a train or an airplane can only transport people through certain geographies, on direct routes, whereas the automobile has increased autonomy, and the pedestrian has virtually unconstrained choice of her flow from A to B.

We should emphasize how much a traveler interacts with his environment because the fundamentals, or unwritten 'rules' of geographic flows, and geography itself apply in some cases, but not others, and it is important to differentiate between the 'cooperating' and 'deviant' types of flows. The representations of how we interact with space in a flow system or a network, especially in visualization cases, do not indicate whether these network edges are meaningful to a geographer or urban planner, leaving many questions unanswered: If we see significant Oi/Dj flows, should we widen this space to allow for more transit? Should we eliminate traffic between two places? Should we make a route more pedestrian friendly? Would this space be suitable for a certain feature, like a hospital, rest stop, or university, based on the observed and potential new flows that could access the area? Is this area good for railroad tracks?

Or, are the travellers impartial to the land between Oi and Dj, as even the most saturated airline route between to two places does not signal that this flow edge should be singled out for urban renewal. A San Francisco-bound Washingtonian's chosen route should not be of primary concern to a planner in Nebraska, although in a GIS, this trace may span the state. Similarly, a long rural highway connecting homes to a work environment does not necessarily demand pedestrian infrastructure. These questions seem straightforward to someone who knows that the primary modes of transit between the two aforementioned place pairs are air and road transit, respectively, but what if the only data about these flows was in matrix form, as it usually is? We know little about whether the *physical edges* between origins and destinations are significant to those looking to examine and improve the built environment—as some muddle through urban space (like a worker in a central business district). Of course urban sensing and tracking devices like RFID, Cell Networks and GPS can also be used to describe the flows between two places. Surely there is a correlation between distance travelled and mode—a bicoastal flow leaves little reason to inspect the mode of travel, as would a walk from the office to lunch—but the origin/destination flow data are still incomplete.

Libertarian and Deliberative Infrastructure

Libertarian infrastructure assumes that the user has some agency in the network traversal process. The agent can make route choices, and at the most libertarian case, is not subject to predetermined infrastructure. At the next level, when physical infrastructure is implemented, the user still has choices as to turns, legs, speed and other cost factors. This infrastructure is usually marked with more tortuosity, and indirect paths. In the most libertarian system, the user has maximal decision-making ability. This ability to decide when to stop, start, and how to get from an origin to a chosen destination is relatively flexible for pedestrians, bikers and car drivers, but wanes as a car driver is not in an area with small side roads, but only straight highways. Additionally, although a passenger is riding in a car on the same infrastructure as he could drive himself, the user loses agency when subject to a taxi driver's chosen route.

The agency decreases steadily until we see increasingly deterministic systems, and finally physical transfers that have little or no interaction with the physical network. (Figure 2.2)

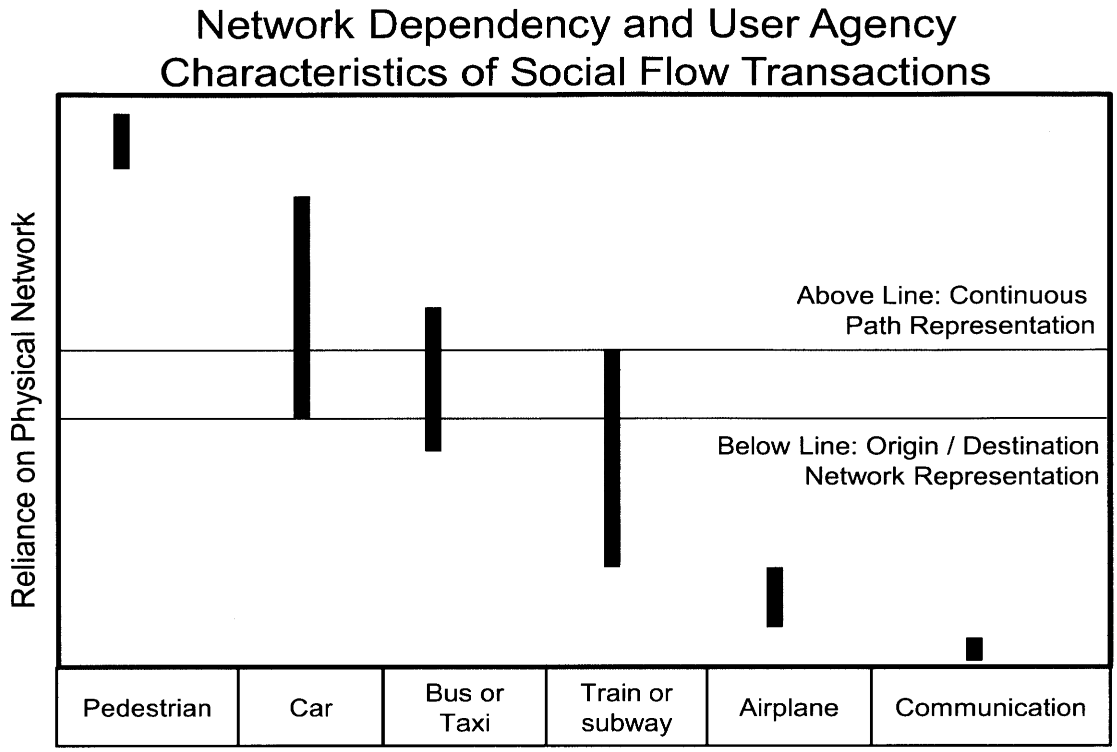


Figure 2.2: Network Dependency and User Agency Characteristics of Social Flow Transactions **A system for classifying different geodesic traces informs a user of the appropriate representation of a trace by its reliance on the physical network.**

Deliberative (below line in Figure 2.2) infrastructure can be thought to begin with the exchange of agency from the cab driver to the cab driver's passenger. Deliberative infrastructure can also be seen as the transfer of a biker choosing his route to one constrained to a single path. Users on a bus are subject to a predetermined path, where the system becomes increasingly deliberative with the fewer possible stops choices the bus provides for its riders; Consider a local bus verses an express bus. With the bus perhaps marks the end of libertarian infrastructure, and the beginning of deliberative infrastructure. Increasingly predetermined are subway and train structures, where agents do not typically determine the route between two places. These structures also mark the beginning of immutable physical proof of route systems, as their tracks provide proof of route choice. With these cases also comes the solidifying of origin/destination representation, as decisions as to how to travel from origin to destination, and the physical obstacles en route, become obsolete. For example, consider the debilitating effects of a bridgeless river on a pedestrian, in comparison to a subway rider, or a cumbersome street parade's effect on a driver contrasted with a train-rider. These point-to-point deliberative systems for human transit perhaps conclude with airplane ridership, the ultimate human travel for outmaneuvering the physical obstacles that hinder ground-bound systems. Agents patronize transit systems for social interaction in the majority of cases, and recreation in a small minority. Thus, we consider transit, and the volumes and dynamics on these systems to be social connectors. From a network perspective, in these increasingly deterministic systems intersections become rarer, hubs have more connections, and system dynamics begin to resemble those of scale free and small world networks, along with themes of preferential attachment.

Furthermore, the human transit portion of the spectrum may be complete, information transfers, as communication flows, lie on the metaphorical ultra-violet end of this spectrum. Edges that represent telecommunications have little interaction with the space between calling agents, Adams (2010) while in comparison, edges between a pedestrian's origin and destination represent an embeddedness in the area between start and stop points. We find difficulty expressing edges that overlay on maps, as they do not inform the user as to the relevance of the trace. A framework for representing these connections is presented as a spectrum that ranges from fine-grained traced paths to a more tabular origin/destination representation. (Figure 2.2)

[B] Lack of Representation Infrastructure & Techniques for Inference and Exploration

3. Muddled visualization

Since flows connect two places in absolute, discrete space, the edges that connect these places are poorly suited for large-scale visualization because the number of edges in a typical dataset is too dense for the constraints of 2D space. The work of Waldo Tobler speaks prominently in this field. The rendering of interpolation for surface analysis first formally explored in Tobler's Automated Cartography (1959), and so flows were streamlined to a continuous topological surface, where the eye can easily see values in between-point data (like a weather map, from weather stations). Similarly Tobler's work on movement mapping (1987) expanded and created a new outlet for visualizing spatial data through vector surfaces. For visualizing complex interactions in space, as trajectories tend to overlap, FlowMapper (CSISS 2005) has allowed for the visualization of spatial flow trends, by way of aggregating flows and querying for specific parameters. Additionally, Tobler's use of flow methodology with migrant datasets, and has yielded exciting findings about the nature and patterns of human migration. (Figure 2.3)

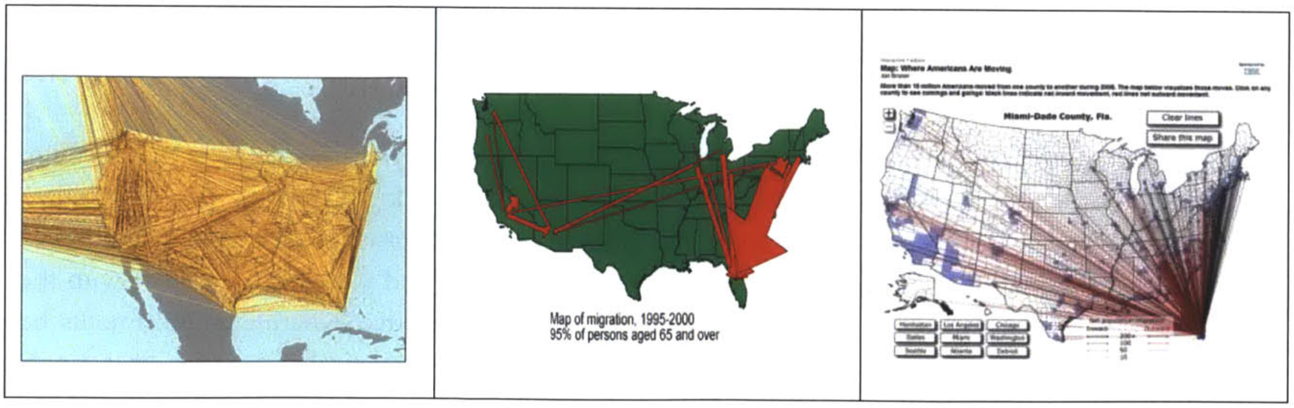


Figure 2.3: Three Visualizations of U.S. Migration Flows **Three views of migration systems in the U.S. show (L-R), muddled visualization, aggregate and queried flows (Tobler 2005), and interactive partial flows (Bruner 2010)**

Cross-disciplinary psychology and cognition research should be considered when determining how maps are understood for wayfinding (Hirtle 1985, Golledge 1999), perception of distance (Montello 1997) distortion (Battersby and Montello 2009), symbologies, (Slocum 1999) color ramps and patch series (Harrower and Brewer 2003), and web-based GIS (Plewe 1997) .

For this, we can suggest some of the classification methods described in (1), as typecasting places that radiate many flows would synthesize a 'haystack' of data into a one-variable cartographic representation. In addition we suggest querying, filtering and automating visual and tabular ordinal hierarchies. Since the implementation of these functions turn renders previously-static representations interactive and dynamic, a new problem arises which addresses the prospects of realizing this solution, and is discussed below.

4. LACK OF VISUAL-ANALYTIC SYSTEMS

There is currently a lack of ESDA and software platform systems for exploring (a) the relationship between a social network and geographic space and (b) spatial relationships between non-adjacent entities. Although guided by researchers who rigorously study Human Computer Interaction and Graphical User Interfaces (Scheniderman and Plaisant 1998), interactivity for data visualization is a widespread challenge. When introducing a new system of variables, this challenge becomes more complex. First, we revisit our concepts of knowledge representations.

According to Davis et al (1993), a knowledge representation fills two roles that can be addressed under the heading of a lack of visual-analytic systems. First, malleable and intuitive infrastructure is needed to support 'thinking' instead of 'action'. Flow datasets may be well-suited for conceptualization as a model in which applying changes to the system can yield different scenarios from which predictions can be made without irreversible results. Adding or eliminating a flow's role in a network can certainly change propagation dynamics in a social system, and could also alter the configuration of interconnections in Cartesian space, if, for example, a lack of origin or destination activity pushes agents to travel and communicate elsewhere. Thus, visual-analytical systems should include direct manipulation of discreet variables in the social and spatial datasets.

Davis et al's criteria also includes the role of knowledge representation of flow datasets as a medium for efficient computation should be addressed by building and evaluating platforms from which data can be organized and the user can access this data to find patterns, test scenarios and infer knowledge. Following the work of Takatsuka and Gahegan (2002) and Anselin et al (2006) with the GeoVISTA Studio and GeoDA, respectively, we suggest that interactive dynamic environments be created for users to explore social and spatial configurations, and statistical properties, in a single view. (Figure 2.4)

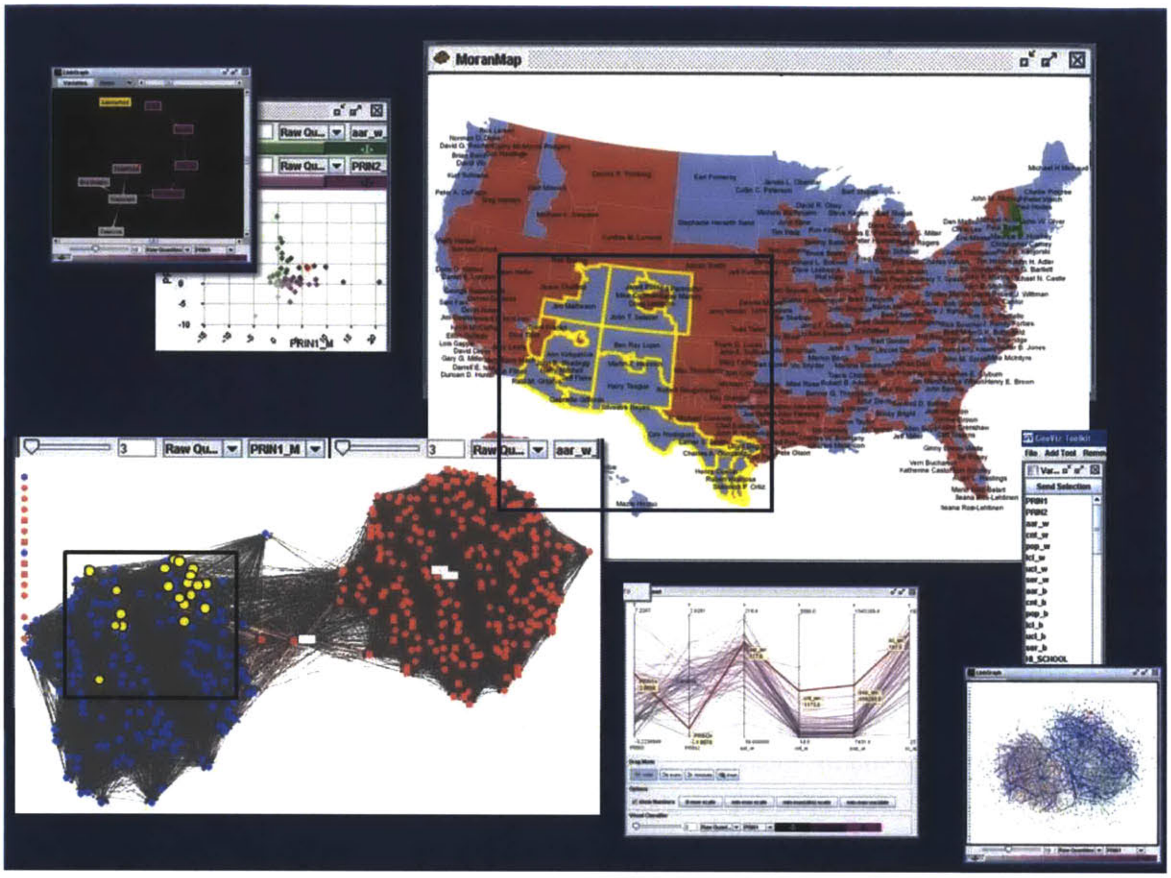


Figure 2.4: Example of a Prototypical Exploratory Social/Spatial Data Analysis Software Environment **A prototype for an interactive social/spatial geo-visualization environment allows the user to manipulate entities in a force-directed social network in tandem with a linked map. In this example, yellow entities in the network are political figures, and correspond to their representative districts, as also highlighted on the map.**

In general, visual data mining (Keim 2002) and Exploratory Data Analysis (EDA) (Tukey 1980) drives much of quantitative analytical visualization, and the principals of these fields of study should be applied to the EDA of social flow analysis. Moreover, MacEachren and Kraak (1997) put forth the needs of ESDA, and MacEachren et al (2004) functions of geovisualization are to "explore, analyze, synthesize and present," while Guo et al (2006) add that exploring data should be intuitive, visual and able to support decision making. Other interactive features like zooming, panning, scrolling, focusing, interactive selecting and "linking and brushing" (MacEachren and Taylor 1994) help users find patterns and anomalies. (Plaisant et al 1999)

Dynamic query-based visualization is a helpful tool in exploring spatial or aspatial data to find correlations and handle many variables, dubbed the "curse of dimensionality." Thus query options, via singular, group, union and intersect functions are essential for these systems. A dynamic query based gives the opportunity to select certain traces with certain features—like all traces that start or end in a Central Business District between 8AM and 9AM. Recent work by Guo (2010) on a flow mapping environment that combines unsupervised classification and automatic selection with flow data, and may solve a few challenges above. This system makes a significant contribution, as it undeniably facilitates innovative interactive ESDA for flow data, but does not consider space for the integration of a social network.

A social/spatial system would be especially useful for a 'representative' network, where a node in a force-directed social network (e.g. linked by friendship or other recorded association) also corresponds (either one to one, many to one, one to many, or many to many) to an entity in projected geographic space (e.g. a county on a map). As we view social flows as a "system of systems", interactive selection of visual patterns can relate one system to another. For example, selecting a certain group of agents in the feature space of a force-directed network can highlight their corresponding locations on a map (Figure 2.4). Conversely, selections of regions on a map could yield a pattern of clustering or dispersion within the force-directed network.

Following the capabilities of packages in the research community like UCINet (Borgatti et al 2002) and PajeK (Batagelj and Mrvar 2003), we suggest that the tools, operations and computational abilities, should be paired with spatial statistics for use in one environment. Statistical components should be readily available in these systems, if not calculated beforehand. From the network science community, computational methods of *community detection* such as cliques, hierarchical clusters, modularity measures; *popularity* measures such as degree centrality, betweenness centrality, flow betweenness, and hubs & authorities; *spreading processes* such as time-step propagation, resistance to percolation, cascades and rule-based diffusion; and *attachment behavior* such as homophily and clustering coefficients. (Jackson 2008) To compute the statistical significance of spatial patterns, the spatial position of competing or cooperating districts, their geographic adjacency, proximity and propensity to form cohesive regions, cluster detection measured by *Moran's I* (Moran 1950) or *Geary's C* (Geary 1954), statistics for interaction (Ord 1975), Hot & Cold Spot Detection (Cressie 1992), and LISA (Local Indicators of Spatial Autocorrelation) (Anselin 1995). In addition to measuring proximal regions, demographic (feature) clustering shows the correlation between certain social features and U.S. Census information like of Income, Urban Areas, Racial Percentages, so social network agents can sometimes be associated with demographic features, with precaution for assuming ecological fallacy and inaccurate variable correlations.

We have just addressed broad visualization issues for flow datasets, and the lack of exploratory data environments for social and spatial flows, as they each limit the ability for a user to infer patterns by sight and statistical exploration. On a more technical level, the rendering, storage, retrieval and manipulation of flow datasets also lack capabilities that would enable and facilitate better social/spatial pattern recognition. We explore issues regarding the design and structure of geospatial flow data in a GIS environment, with an emphasis on the recent work of Glennon (2010).

5. Lack of GIS infrastructure for flow manipulation

The phenomenon of a spatial flow incorporates the features at its origin and its destination, as well as features of the flow. In the field of Database Management Systems (DBMS), the theory and implementation of Spatial Data Infrastructure (SDI) is not yet developed to treat two unique points and a connecting edge as a unit of analysis, making selection, operations and manipulation difficult. Perhaps it is for these reasons that flow data has not been as prominent of a fixture in Geographic Information Systems, in terms of software, computation, statistical endeavors, and academic and professional use. Solutions to these challenges are manifold, but early advancements suggest combining graphs for easier querying. (Doytsher et al 2009) We suggest creating a new object (called not point, line or polygon, but 'flow' etc.) that packages nodes and edges into a single entity. This entity can then be queried and interactively selected by its named components (e.g. the origin, destination and trace).

It is surprising that software packages for GIS do not cater to those looking to analyze spatial interaction data. Others have noted this lack of integration, saying that interaction data is not generally supported by GIS, and that the functionalities within contemporary GIS generally lack standard functionality in contemporary GIS (Tobler 2004, Goodchild and Glennon 2008; Cova and Goodchild 2002).

Since complexity is an ever-growing challenge for research that combines social and spatial systems, we can expect that the storage and software capabilities for datasets to be complex as well. According to Unwin and Unwin, optimal software for flow analysis may include: direct linking, spatial data structures, graphics, prototyping, statistical programming and interactive interfaces. According to Ferreira (1990) we may think of our data design and management system as a "mechanism for accumulating knowledge in ways that would otherwise be too cumbersome, time consuming or expensive." Moreover, a data system designer who is considering flow data may want to consider the following questions: Can the data be feasibly stored in a single table? Is the data static or dynamic, private or shared, and following these questions, should the database be amenable to interactivity? Also, does the data need to be fused with other sources or reformatted for output? (Ferreira 1990)

More specifically, Glennon (2010) lists four requirements that a flow model should fulfill: "(1) it must have structural **locations** to hold all data required to recreate each use case; (2) its structure must be able to logically describe the **linkages** and components of the case; (3) the model must provide adequate logical infrastructure for exemplar associated **queries** to be successfully performed; and (4) **redundancies** and **dependencies** should be minimized." (p. 32). Furthermore, the author is sensitive to the type of queries that the system could handle, and provides three example queries: 'Which direction is the flow moving; how large is the flow at a location; or what is the difference in magnitude at one location compared to another?' These concerns specifically echo those addressed in this section, but we keep all the aforementioned definitions and considerations in mind when addressing the following limitations in current mainstream GIS design.

FUNDAMENTAL REPRESENTATION

It seems that most proponents of advancing infrastructure for geographic flows have implicitly accepted node-edge (graph) models as the standard for modeling interaction and connectivity. However, this model is not explicitly accepted as the best fundamental structure for representing flows. In order to actively accept the graph structure, it may be advantageous to enumerate the benefits of this model for comparison against other models and for full realization of the functionality of graph structure representation. The following is a list of benefits: 1) Community structure detectors can show how clusters can be understood in terms of how well entities they work as a unit (like a "portfolio problem"), not simply that they are densely collocated (like the ice cream parlor and theater). Euclidean distance may not capture social closeness—as defined by high social connectivity between places. 2) Node measures, like 'degree', can be an instant heuristic for comparing the relative power of place. 3) Network system measures like degree distribution, edge weight distribution, diameter (when appropriate) and clustering coefficients are standardized metrics that can be used for both comparing and applying ordinal structures to groups of systems. 4) Different GIS data can be fused without layering techniques, by embedding one network inside another. Poly-flow models combining multi-modal transit flows, migration, social network contacts and communication magnitudes can be used to better quantify interaction between places. 5) Interaction data is well-suited for real-time simulation, as edge travel is easier to visualize and yields itself to greater precision, as the travelling entity is contained to an edge, than the uncertainty of polygon-to-polygon travel. 6) These systems are also well-suited for modeling connectivity where topological adjacency, or even closeness, does not necessarily correlate with the volume of social exchanges between two places. Whether realized or unrealized, these benefits of the link-node structure are important fundamental concepts for those modeling interaction, but their implementation yields significant limitations in functionality and utility.

SPECIFIC FUNCTIONALITY LIMITATIONS

First, at one root of the problem is the general acceptance of input data, namely sparse matrices that seem foreign as GIS dat. Origin/Destination data (and social network data) often are stored as grid matrices, where many, if not most of the values are null (indicating no connection between two row/column entities). (Figure 2.5) In *Flowmapper*, Tobler has allowed users to form graphic flows from tabular interaction matrix data, although this example is rare. Software packages, MATLAB in particular, are typically used to analyze matrices for knowledge discovery and critical statistics. One exception to these constraints is software for transportation modeling and logistics, some prominent examples of which are TransCAD and ESRI's ArcGIS Network Analyst package. However helpful for traffic modeling scenarios and supply chain management, the functionality of these packages caters almost exclusively to terrain networks as measured via 'cost distance' topology, without incorporating 'social distance' factors like communications, migrants, and representative flows. The systems set up for operations and civil engineering-type tasks also exhibit the theoretical basis that a node A (intersection) is typically a joining factor of two of the node's neighbors, B and C, making system transitivity a useful endeavor. In social networks, a node A can be a joining factor between B and C, but in a model where nodes represent a collection of people (e.g. a place), nodes B and C may have little in common by each having connections to A.

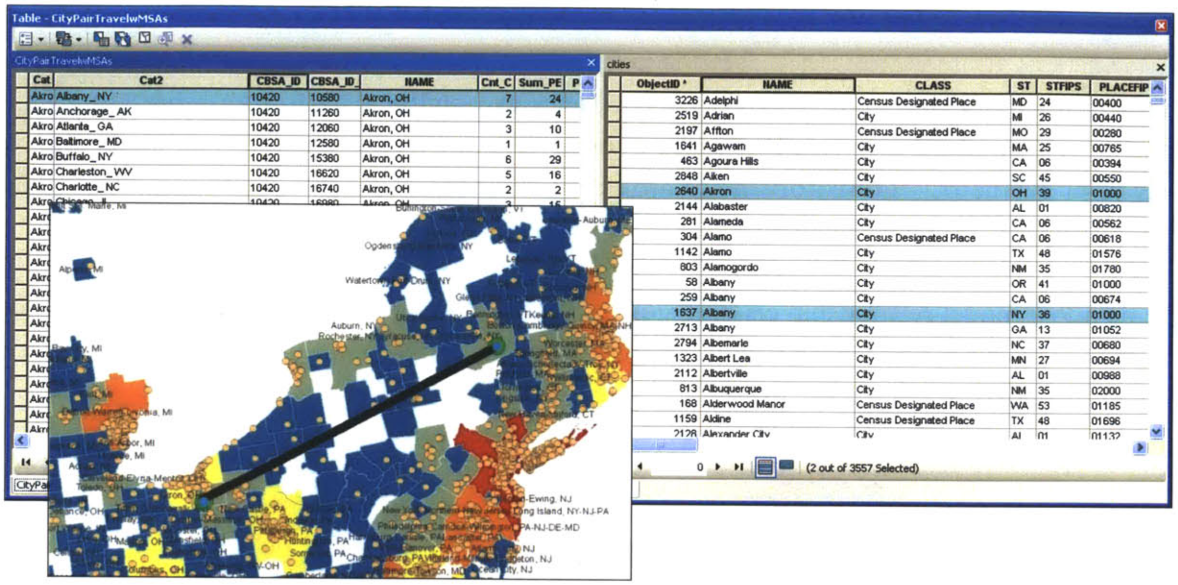


Table - CityPairTravelwMSAs

CityPairTravelwMSAs

Cat	Cat2	CBSA_ID	CBSA_ID_	NAME	Cnt_C	Sum_PE	P
Akro	Albany_NY	10420	10580	Akron, OH	7	24	
Akro	Anchorage_AK	10420	11260	Akron, OH	2	4	
Akro	Atlanta_GA	10420	12060	Akron, OH	3	10	
Akro	Baltimore_MD	10420	12580	Akron, OH	1	1	
Akro	Buffalo_NY	10420	15380	Akron, OH	6	29	
Akro	Charleston_WV	10420	16620	Akron, OH	5	16	
Akro	Charlotte_NC	10420	16740	Akron, OH	2	2	
Akro	Chicago_IL	10420	16980	Akron, OH	3	16	

cities

ObjectID *	NAME	CLASS	ST	STFIPS	PLACEFIP
3226	Adelphi	Census Designated Place	MD	24	00400
2519	Adrian	City	MI	26	00440
2197	Affton	Census Designated Place	MO	29	00280
1641	Agawam	City	MA	25	00765
463	Agoura Hills	City	CA	06	00394
2848	Aiken	City	SC	45	00550
2640	Akron	City	OH	39	01000
2144	Alabaster	City	AL	01	00820
281	Alameda	City	CA	06	00562
304	Alamo	Census Designated Place	CA	06	00618
1142	Alamo	City	TX	48	01576
803	Alamogordo	City	NM	35	01780
58	Albany	City	OR	41	01000
259	Albany	City	CA	06	00674
1637	Albany	City	NY	36	01000
2713	Albany	City	GA	13	01052
2794	Albemarle	City	NC	37	00680
1323	Albert Lea	City	MN	27	00694
2112	Albertville	City	AL	01	00988
813	Albuquerque	City	NM	35	02000
168	Alderwood Manor	Census Designated Place	WA	53	01185
1159	Aldine	Census Designated Place	TX	48	01696
2128	Alexander City	City	AL	01	01132

(2 out of 3557 Selected)

Figure 2.5: Image of a Flow from Albany, NY to Akron, OH and Corresponding Table Records in a GIS Environment **An example flow in the ArcGIS software environment, and its corresponding tables illustrates the geographic entity selection framework. Here, even though flows and origin/destinations are in two separate tables, origins cannot be distinguished from destinations without geographic redundancy (e.g. Albany has two geographic points in the GIS, one as an origin, another as a destination, but these cannot be distinguished when selecting interactively).**

Secondly, there are few summary statistics for flow data, although some, like degree and sum of edge weights, can be derived by combining attribute queries with group by, sum by and count table functions. Summary statistics of interest would be: the second degree of a node, the weighted degree of nodes, or number of edges in participant network. Fourthly, as addressed in briefly in when discussing a lack of visual-analytic systems, GIS functionality does not have the capability to perform any prediction, modulation or partitioning methods. Like the work of Guo (2006) and Assuncao et al (2005) on algorithms *REDCAP* and *ICEAGE*, respectively, graph partitioning methods have proven helpful for spatial data. The authors use minimum spanning tree partitioning methods to better understand how a social system graph's natural breaks can be directly translated to the underlying geographic topology, resulting in more socially-integrated regions.

Next, in the current framework, nodes and edges are not currently able to be stored as singular entities or objects in a GIS DBMS, which can pose challenges to entity manipulation. Hence, an object is hard to select and define because it currently must relate to other tables to be realized as a fully-descriptive flow. Glennon (2010) divides flows into two categories: (A) Steady flows are aggregate movements that have been completed or exist in a stable state. (B) Transitory flows exist in a state of unfolding motion, requiring the handling of time. An example of the former is a model of intra-state migration for one year, and an example of the latter is a parade marching through a town's streets. Similarly, the data model also allows two conditions, which echo the 'edge assignment' theory posited previously: (1) Flow along a known route and (2) flow where origin and destinations are known, but the exact path geometry is not certain.

In the data flow model feature class, Glennon defines as **polyline** as a feature representing a line containing one or more line segments, and a **node** is a single coordinate point containing one (usually x, y) geometry. **Transitory** flows are denoted by the pink track, and **singular** flows are denoted by the yellow track. (Figure 2.6)

.In the dataflow model relationship class, (Figure 2.6) a 'flow' is a relationship that holds the quantity of the flow. This class's association with a network or link provides a flow's length and direction. The structure allows for a method 'magnitude' to either be invoked or set as a constant.

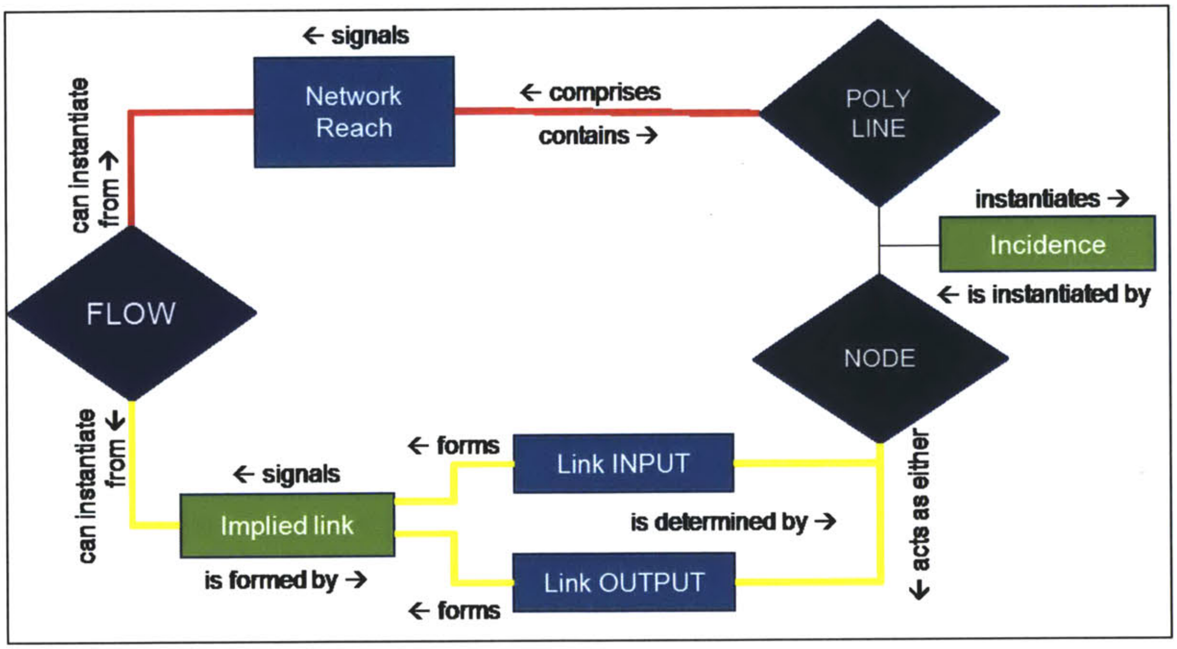


Figure 2.6: Example Schematic for Components of Flow Data in a GeoDatabase **A re-engineered version of Glennon's (2010) flow system design includes perceived actions as well as design components and objects. Objects and methods are denoted as navy blue diamonds; descriptive components are blue rectangles and intention components are green rectangles.**

The benefit of this design is that a flow can be a singular entity in a database, which would also indicate that a single table can be used as well. Glennon illustrates the robustness and benefits of this system with three case studies: U.S. migration flows, Napoleon's famous march and a stream system in Kentucky, U.S.A.

After looking at visualization, system and framework design issues, we've addressed some important challenges in human-computer interaction, exploratory data analysis, visual pattern recognition and inference. However, for statistical and computational endeavors we must address additional problems in the flow system framework. We move next to problems and challenges with computation, where the first issue is a pragmatic stop in applying traditional spatial statistics to flow data (given flow data's transitive nature), and the second issue is a theoretical concern regarding the application of social variables to prediction methods.

[C] COMPUTATIONAL ISSUES

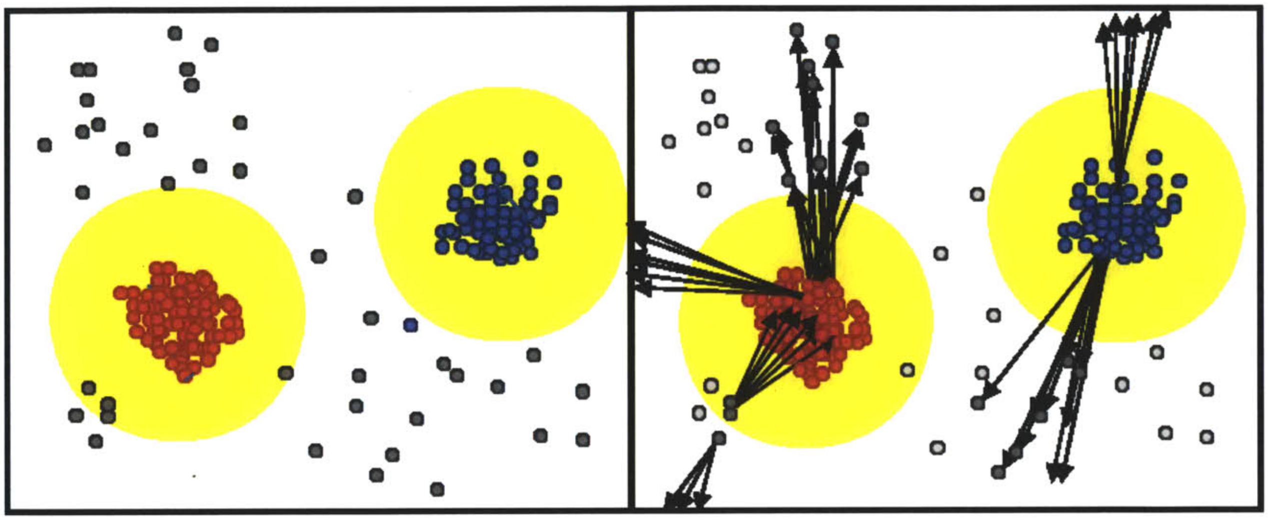
6. INABILITY TO EMPLOY SPATIAL ANALYTIC METHODS

Going against some fundamentals of geographic relationship theory, social flow data may be the only data found in GIS where a coincidence at an absolute location can be completely unrelated, or more related to a process 10000 kilometers away than a flow 0 kilometers away. This makes statistics that rely on absolute location helpless, unless a temporal snapshot, or a summary of origin and destination statistics are the input data for the test.

Additionally, these interaction data cannot be analyzed with spatial statistics, point-pattern analysis, and spatial operations (like a clip or spatial join) as these statistics are fit for single points, polygons, based on each entity's proximity or adjacency to one another. Since a flow connects two places that are generally not adjacent, the measure of proximity between two places is a misleading metric for their relationship. Spatial statistics, however, rely on this metric, with the same underlying assumptions that closer things are more related. To illustrate the different challenges in flow (transient, movement) data than static, single geometry data, consider two statistically significant clusters—say two cities (Fig 7A). These clusters (or polygonal centroids) can be characterized by statistics like Moran's I, LISA (Local Indicators of Spatial Autocorrelation), Getis-Ord Hot Spot Detection, Ripley's k-function or Geary's C.

However, when flow data is considered, it may be discovered that entities that seemed close in Euclidean space actually had 'designs' on entities that would not have been considered closely connected. (Figure 2.7B) Perhaps an outlying node exhibits behavior that can surely be considered a contributing factor to the city. Conversely, a neighborhood in the city may connect with only entities that are out of scope, or with a certain part of the suburbs. Similar to the overarching idea behind a k-means algorithm or a modularity algorithm, the cities are re-grouped according to the modules of interact with one another. The result of this example leaves three clusters of note in the left cluster and two clusters of note in the rightmost cluster. Totaling five clusters, the social topology shows a more fractured pattern of population center clusters. Since these 'social' clusters better reflect the connectivity patterns of landscape, it can be inferred that population travels mostly within the confines of the cluster, and that travel time may either be naturally shorter (spurring the connectivity) or should be improved (to support the connectivity), within these interacting centers. Other benefits and potential assumptions are abundant.

Although we offered modularity and k-means as potential effective models for the re-configuring of conceptual social clusters, this scenario remains in the 'challenges' section because these interlocking patterns are usually more complex than the linearly separable (or 'fence' separable, since our medium is land) connectors. The connections usually form triangles, or even a clique-like webs, where if 100 population centers were present, it would not be unreasonable to expect up to 9,900 inter-node directed connections. Figure 2.7D shows that just a small fraction of these connections can already illustrate the computational problems for a spatial analyst looking to characterize space.



Figures 7A and 7B Examples of Two City Clusters **show two cities, as formed by clusters of population. Figure 2.7A shows the cluster configuration. Figure 2.7B (left) shows the underlying social relationship connections (e.g. commuter or email volume) between the two cities and their surrounding population centers.**

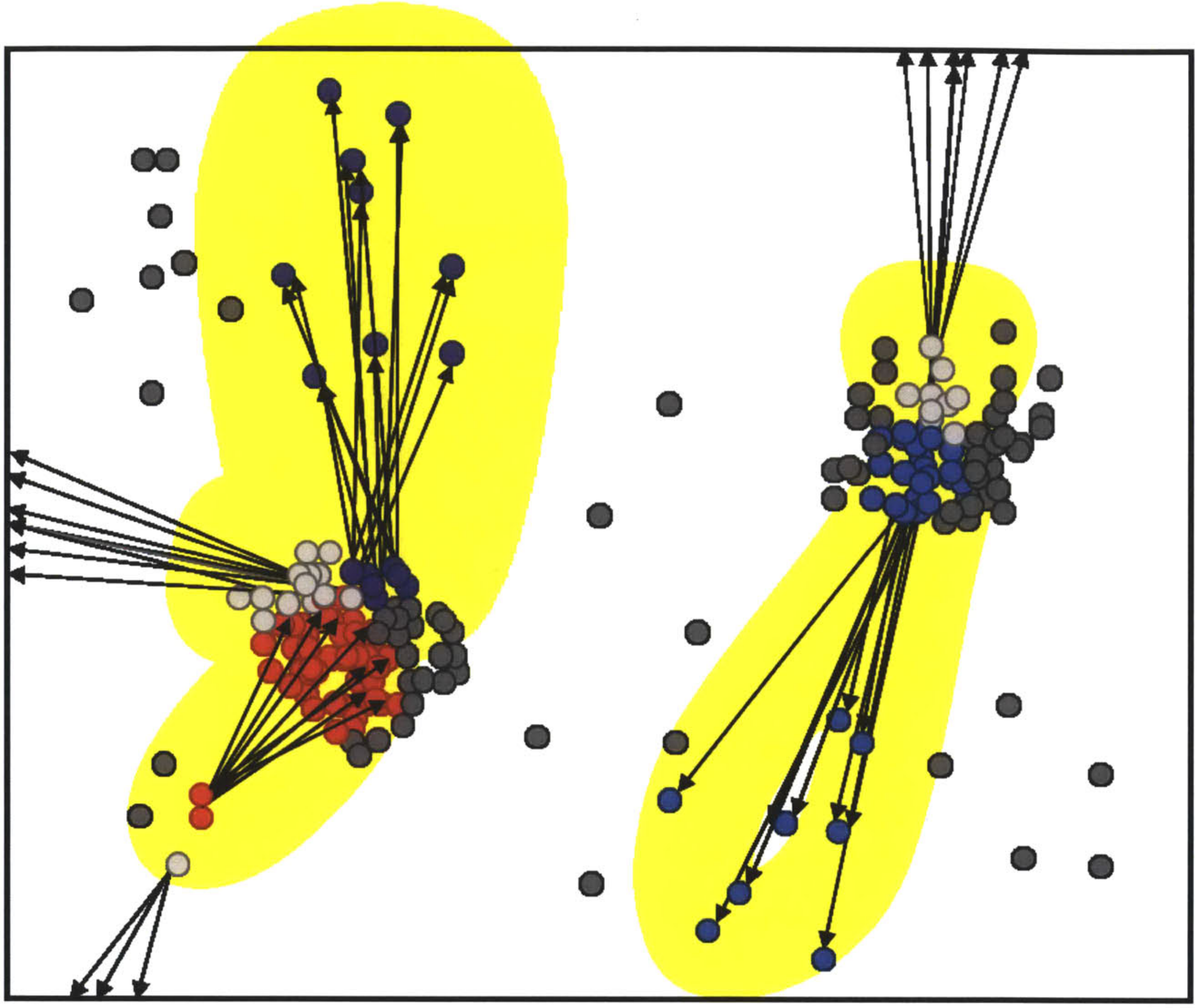


Figure 2.7C Cities Partitioned by Color based on Interaction Patterns **shows the reconfiguration of modules, according to the new 'families' of connected entities. The left cluster engulfed two exurban centers, and the northern neighborhood exhibited connections with a number of population centers in the periphery directly north. Western (leftmost) population centers asserted their apparent disassociation with the urban cluster, as evidenced by their communication with a far-away population. The rightmost cluster exhibits these behaviors as well.**

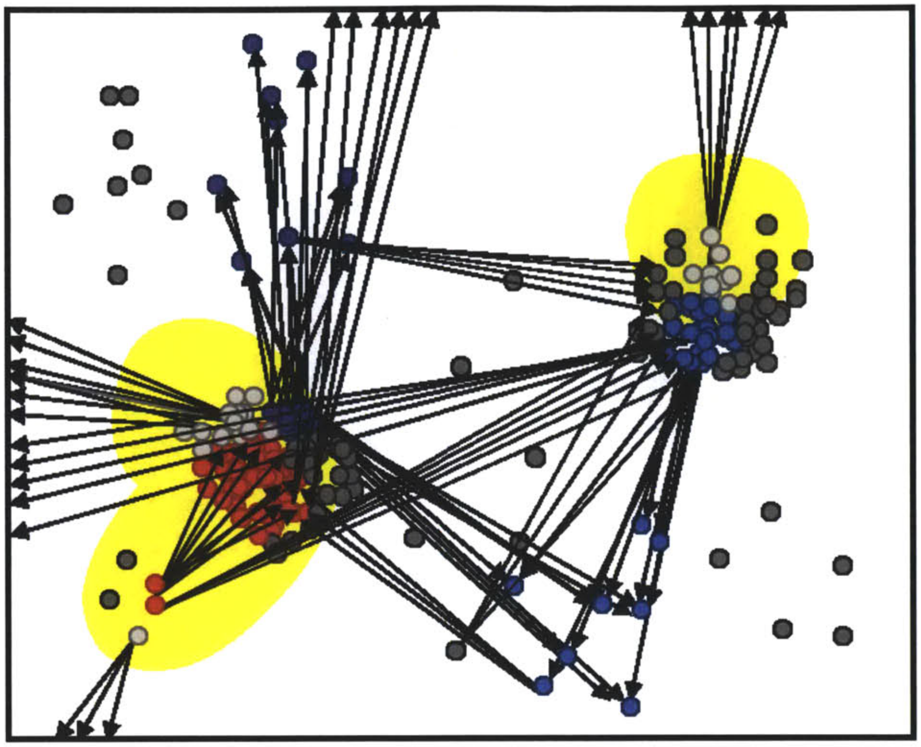


Figure 2.7D: Increasingly Complex Interactions between Cities **depicts a slightly more realistic situation of inter-urban communities. Unlike the linearly-separable connections in Figure 2.7C, the cross cutting, and multi-node connections in this figure portray the complexities of interaction data that make it difficult to characterize nodes and their interactions with Cartesian space.**

Theoretically, geographic data show spatial dependence, are subject to scaling issues, and are unlike statistical data in that stationarity is hard to assume. (Unwin and Unwin 1998) With flow data, the stationarity issue is exacerbated, and the spatial dependence issue is very reliant on the nature of the flow. We can assess homogeneity in flow behavior, by aggregating like vectors, but there is currently a lack of methods for assessing the heterogeneity or homogeneity of absolute space. Since flow data typically links separate entities, the spatial distribution of a single flow object cannot be easily assigned to a space, therefore many flows cannot be easily assigned to a distribution that characterizes, or illustrates a finding about the space beneath the flow. Central features and mean center, directional mean or distribution of the points shifts in a different way than expected. Additionally, the directional skew (angle/orientation) of the data may change, and may correlate with new or different features in the built environment (like a shift to a technology corridor instead of a historical trail). Distance decay functions or dispersion of data (at different scales—related to the MAUP problem) can be altered as well.

The relationships between entities are difficult to discern: Colinearity, regression or dependence processes for finding variable correlation are also under developed, and can benefit from more attention to how to manage dependent and independent variables found at the origins, destinations and on the flow itself. Processes like Geographically Weighted Regression (GWR) are difficult to employ via two separate locations (an origin and destination).

In terms of density, event and point pattern analysis, it is difficult to count events in a single space: some flows are completely inside, some origins are inside, and some destinations are inside. These three statistics lengthen the process of finding critical statistics and make it difficult to characterize a phenomenon with a convenient single value. Comparison of spaces also poses problems: Holding edge weights constant, is an area with 10 complete flows and 4 incomplete origins more socially dense than an area with 8 complete flows, 4 incomplete origins and 4 destinations, because activity is contained more contained in the first example? Or since there are more initiator and receptor points in the second space, is this space more socially active?

When calculating the distance to the nearest neighbor, does a neighbor count when the focal entity does not connect with this space? They are closest in Cartesian space, but in the flow system, they are disconnected (directly) spatial autocorrelation is also hard to calculate because flows do not interact with other flows—the entities themselves do not interact, like points or polygonal regions.

Another notable drawback of flow systems is the inability to interpolate data points for a smooth surface. Interpolation methods, like Inverse Distance Weighted (IDW), Kriging or Nearest Neighbor, helps to estimate values of interest in places where data is not present, by predicting the value based on neighboring values. Since the intersections of multiple flows, and flow centroids are not usually meaningful, only summary characteristics of origins and destinations can be interpolated. Except in some efforts to define a good 'meeting place,' the interpolation of the connections regarding flows is typically irrelevant. The geometry of flows does not lend itself to parallelisms, meaning that the instance of two flows over a certain x, y may be completely unrelated, whereas a measure of rainfall at a certain x,y and a nearby jittered x,y in any direction, is very closely related.

As mentioned in (1), it is difficult to characterize nodes, as their connections are not characterized as a single connection, but an array of many connections, and proper representation would account for characteristics of all tangential places. Similarly, assigning demographic (or other) variables to flows raises problems for flow to flow correlation of variables. Instead of comparing points by their location and a feature (for example, incidences of cancer), we must decide whether to sum, average, count, find the standard deviation, etc. of participants at the origin and at the destination. Thus, we may try to apply proximal clustering and statistical methods to origin/destination nodes, or assigning statistical properties to the edges themselves. Without sensitivity to how the flow interacts with the ground below it, the entity's 'existence' in geographic space can have various meanings. We can face these challenges, but we must be careful about the question we are asking, and ensure that our research design is thoughtfully configured to address the research question.

Some computational methods can still be used—namely trace mining and data mining. One property of flows that is consistent with some spatial statistics theory is Stevens' Levels of Measurement: Nominal, Ordinal, Interval, Ratio and its newer types, Directional and Fuzzy. Each of these data label types can still be used to define flows, and are beneficial for use of supervised classification and machine learning applications like Decision trees, Bayesian Nets, Markov Chains or Support Vector Machines and unsupervised clustering methods like K-Means, Neural Networks or Self-Organizing Maps (SOM). (Miller and Han 2001) Ordinal data is also helpful for ranking flows by criteria of interest, while nominal data can be used for extracting flows with certain criteria. Although the work of Tobler has explored some of these issues, perhaps the use of movement aggregation methods from other disciplines, like physics (stressing magnetism), hydrology, or mechanical engineering (stressing fluid dynamics) may be useful.

In conclusion, as mentioned in (1), it is difficult to characterize nodes, as their connections are not characterized as a single connection, but an array of many connections, and proper representation would account for characteristics of all tangential places. Similarly, assigning demographic (or other) variables to flows raises problems for flow to flow correlation of variables. Instead of comparing points by their location and a feature (for example, incidences of cancer), we must decide whether to sum, average, count, find the standard deviation, etc. of participants at the origin and at the destination. Thus, we may try to apply proximal clustering and statistical methods to origin/destination nodes, or assigning statistical properties to the edges themselves. Without sensitivity to how the flow interacts with the ground below it, the entity's 'existence' in geographic space can have various meanings. We can face these challenges, but we must be careful about the question we are asking, and ensure that our research design is thoughtfully configured to address the research question.

After reviewing some issues that separate social flow data from the kinds of spatial data that have been successfully analyzed with spatial statistics, we now turn to the type of data that is being modeled. The next section suggests that social flow data may be helpful in predicting future interactions.

7. FEW SOCIALLY-DRIVEN FLOW PREDICTION METHODS

It's rarely contested that humans rely on their families, friends and neighbors for support, that there may be decreased efficiency when traveling somewhere unknown without friend recommendations than to a familiar place, or that human emotion and affect feed into decisions to connect and travel. Now that micro-behaviors are able to be quantified and enumerated we may benefit from their predictive capabilities, as human social needs are a more reliable artifact of society than politics, economics or infrastructure.

Pragmatically, geographically-dependent decisions to meet, migrate, travel or communicate are comprised of more than distance factors, but social factors, yet models for flow prediction do not usually reflect evidence of interpersonal relationships. We often assume that the shape of the environment determines the strength and patterns of social ties, like the fundamental distance-dependent theories from Losch (1944) that says a political border produces an effect identical to that of increasing the distance between two nearby areas. Yet we rarely consider that the environment form is an effect of social ties.

Spatial flow prediction methods have been very reliant on aggregate gravity models, and econometric forecasts the envelop disparities between income, land value, socio-demographic, distance, population, density and other variables. Conversely, independent variables are found and tested for correlation with the flow magnitude. While the combination of variables and parameterization methods may change, universally, empirical models and deterministic calculations are typically used to predict how much interaction two places would exhibit in the future.

Economic theory has not completely avoided human considerations in its models. Alfred Weber's classic [Least Cost] "Theory of the Location of Industries" states that the process of locating an industry should account for the optimal transportation of all objects from raw materials to the finished product, as well as the availability and travel cost of skilled and unskilled labor and the ease of agglomeration and deglomeration of firms. (1909) William Reilly's Law of Retail Gravitation (1953) states that people are willing to travel further to connect to larger cities. This phenomenon can be approached from a cost/benefit perspective, where citizens must absorb a high cost to access the amenities of a bigger city, and measures of elasticity between amenity scarcity and travel can be modeled and tested. Research that more closely represents social parameters can be found as early as 1922, where Gras suggests a variety of ways for delineating markets: a region could be where insiders are drawn to homogenous social (museums, galleries, fashions, etc.) infrastructure. (echoed by Andrews 1954 for city delineation) After modern computing capabilities improved computation, Ben-Akiva and Lerman's (1985) demonstrate the ungrouping of mass movement patterns via the individualistic nature of discrete choice analysis. This advancement allows for personal behavior choices (like stopping at a store) and latent variables (like weather) to better predict the magnitude, demographics and spatial dynamics of trip generation, trip distribution, mode choice and route assignment.

Case-based research has discovered meaningful patterns from human transit data. For example, travel and commuting patterns have played a major role in defining urban areas, Karlsson and Olsson (2006) provide a good overview of how the data have been used for city and regional partitioning. Similarly, Limtanakool et al. (2009) use work and leisure flows to reveal a complex sub-national systems and polycentric urban structure that may have otherwise gone undetected. Also with regard to aggregate human behavior in geographic space, Davies (1979) and Holmes (1983) use aggregate landline calling – between areas of Montana and Australia, respectively – to try to quantify inter-urban connectivity and compare it to more established measures of interaction.

While these research studies are evidence for the integration of social interaction data (via transit and communication). In one experiment, an author notes that cost distance of trade routes lacked the variables necessary to predict current city power structures.

Predicting the present economic network based on a network of medieval trade routes qualitatively showed that Moscow exhibited high connectivity, but not always the highest for some accessibility measures. (Straffin 1980) The author asks: "Does Moscow indeed have the higher accessibility in this network, or must other factors have modified the forces of 'situational determinism'?" Although not explicit, Straffin may have been indicating that travel magnitudes and temporal trace patterns, not just routes, could have better predicted the Russia's current metropolitan system dynamics.

In sum, there is a notable lack of models that incorporate human interaction data. The proliferation of the specific 'digital breadcrumbs' that link people to one another should be quantitatively harnessed in order to model the propensity of future communication, economic flow and travel, from a bottom-up approach. One burgeoning example is the study of remittances, which follow the theoretical phenomenon known as 'chaining' (Castells 2000), where future migrants can be predicted by communication back to a place of origin. We expect these models to be better predictors of the future, because evidence is compiled not by income or distance variables, but by evidence of human nature and agent-based decisions.

[D] ISSUES OF STANDARDIZATION

8. DIFFICULTY DEVELOPING A LINGUA FRANCA AND TAXONOMIES FOR THE FIELD

There seems to be little research on ontologies, nomenclatures, typologies and frameworks for flow data. Considering that ontology can be defined as "a conceptual scheme that enumerates and describes the relationships and rules for a particular domain" (Worboys and Duckham 2004) , linguae seem to be successfully implemented for the separate 'domains' of Complex Network Analysis and Geographic Information Systems but not for the coordination of these domains. Ontologically, semantics and communication play a theoretical, standardizing and functional role, without which, a flow system would have little staying power in the GIS field. (Davis et al 1993)

According to Foncesca et al (2000), an 'urban ontology' contains: **objects** (like a country or a school) **relations** (like a country trades with another country, or a school is part of a larger district), **events** (like a closed bridge or a parade) and **processes** (such as pedestrian flow or weather). This schema could be used to merge social and spatial lingua under the same dialog. Refining these typologies may enable progress in the synergistic field, and its ability to be communicated, improved, and understood. Initiative for this streamlining should come from multiple GIS entities: (1) software corporations, (2) GIS textbooks, (3) Higher Education terminology in labs and lectures, and (4) peer-reviewed literature.

CONCLUSION

Given that geographic analysis can benefit from better manipulation of social flows and connectivity over space, we consider eight pragmatic challenges and possible solutions for use of social flows and social distance in the digital era. Our questions are still both theoretical and practical: What is the first 'jump'? On what should we invest resources? How can social flows best serve analysis for operations, logistics, and planning, in tasks such as city-to-city airline passenger magnitudes, future road capacity, inventories of infrastructure needs for urban growth, city shrinkage, migration forecasts, site suitability for cooperative meeting places, epidemiological spreading or containment models, and suggestions for business franchise or advertisement expansion?

This chapter has outlined obstacles to the development of social/spatial research under the assumption that flow (interaction, connectivity, or network) data connects two places either by the transfer of humans or information. We first addressed a framework for theoretically approaching, implementing, integrating, and learning from these "systems of systems", or more specifically, the intersection of social networks and geographic space. The facets of this framework included considerations for ethics and the user, a peripheral look at the landscape of GIS and GIScience in terms of identity, growth theory, technology, past and future, and finally, suggested rubrics and parameters for system evaluation. Following, we listed and discussed eight pragmatic issues for the use of these place-to-place connectivity measures in a spatial context.

As mentioned, not only are the final payoffs of integrating social flows into GIS are unclear, but how to gauge the payoffs of addressing these challenges is also unclear. In thc following three sections, we present three different case studies for using social flows in context, and attempting to address the challenges listed here.

The aftermath of which challenges each case study addressed will be discussed in the conclusion, but in summary, the first case study chapter presents a new method for visualizing flow data, the second case study chapter tests the predictive power of social flow data, and the third case study chapter addresses the lack of interactive environments for social network and spatial network research.

ACKNOWLEDGEMENT

The author gratefully acknowledges Joseph Ferreira Jr. and Frank Hardisty for input and supervision for the introductory and challenges chapter as well as Leon Andris and Vanessa Andris for editing and proofreading.

CHAPTER 3
Visualizing Migration Dynamics Using Weighted Radial Variation

ABSTRACT

Directional flows created from an origin/destination matrix have been traditionally difficult to visualize because of the number of flows to be rendered in a small cartographic space. Because visualizing geographic flow dynamics are useful for understanding the complex dynamics of human and information flow that connect non-adjacent space, techniques that allow for visual data mining or static representations of system dynamics are a growing field of research. Here, we use a Weighted Radial Variation (WRV) technique to classify places based on their group's radially-emanating vector flows. Each entity's vector is syncopated in terms of cardinality, direction, length, and flow magnitude. The WRV process unravels each star-like entity's individual flow vectors on a 0-360° spectrum, to form a unique signal whose distribution depends on the flow presence at each step around the entity, and is further characterized by flow distance and magnitude. The signals are processed with a supervised classification method that clusters entities with similar signatures or trajectories in order to learn about types and geographic distribution of flow dynamics. We use U.S. county-to-county human incoming and outgoing migration data to test our method.

Keywords: movement, flow mapping, geovisualization, feature reduction, graph structures

INTRODUCTION

We present a novel way to classify geographic entities (origins and destinations) in a flow system while preserving the individual characteristics (flow magnitude, weight and direction). We call this method Weighted Radial Variation, or WRV. Using these classifications, we can better visualize, and thus better understand the nature and dynamics of large complex geographic flow systems.

This research is driven by the increasing availability of large flow datasets: Data from cell phone traces, traffic sensors, flight schedules and telephone records, and government digital collections are now becoming more common sources for analysis and for fields such as transportation, logistics and operations, geography and civil engineering.

Computational methods for matrix datasets have already helped researchers in these fields learn more about human and communication transactions across the built environment. The dynamics of large, multi-scale flow systems are often measured with summary factors—like a node's degree (number of neighbors), or a hub's centrality in a whole flow system (O'Kelly 1998, Xu and Harriss 2008, Jaing and Claramunt 2004).

While these methods continue to inform spatial system dynamics and benefit from cutting-edge complex network analysis (CNA) techniques, their progress has been largely unaccompanied by spatial visualization techniques. The need for good characteristic-reducing techniques for multi-featured spatial systems, like complex flows, are an important component of the static and dynamic, user-interactive, geovisualization tools that currently support visual spatial data mining. (Anselin 2005, MacEachren and Kraak 1997) A recent focus towards visualizing flows (Marble 1997 and Guo 2009) and object movement (N. Andrienko et al 2000, G. Andrenko et al 2007) and space-time dynamics (Dykes and Mountain 2003) has demonstrated the benefits of these kinds of methods.

Though rendering spatial systems is a growing topic of interest, the nature of complex geographic networks' geographic flow intersections and overlaps, present a natural visualization problem, as point-to-point datasets can have many links, yielding a 'haystack' of links—from which little analysis can be performed. (Figure 3.1)

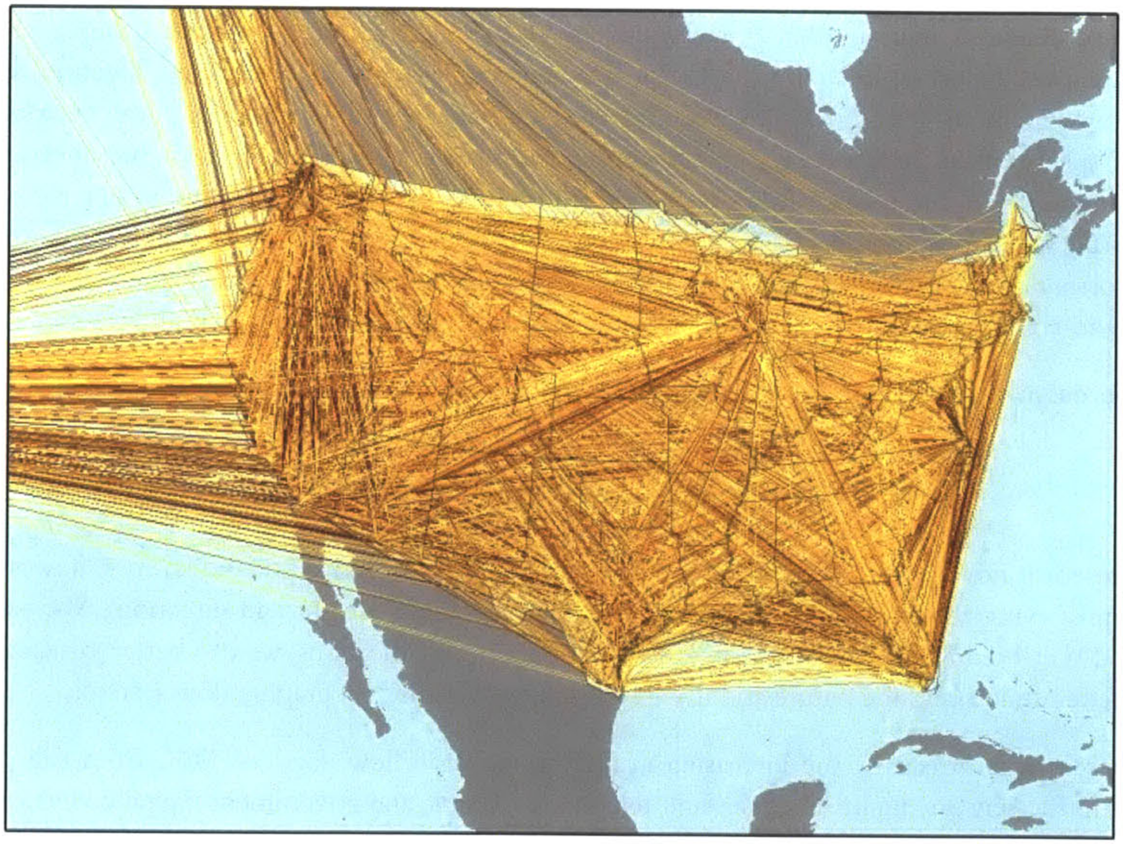


Figure 3.1: 2008 Migration Flows in the Continental U.S. **Visualizing flow lines often results in a 'haystack' view of overlapping entities.**

Early efforts to spatialize flow dynamics include Tobler's computer mapping, where flows were rendered as aggregate arrows in order to fit in a cartographic space, and also woven into vector surfaces that resembled magnetic fields. (Tobler 1959, 1978, 1987) More recently, Andrienko and Andrienko (2010) echo the benefits of aggregating flows, adding that multi-scale analysis is now possible. Similarly, Woods et al take a unique approach to flow aggregation by assigning characteristics of OD vectors to cells, instead of the more traditional line summaries. (Wood et al 2010) One method that fixes the haystack problem and the aggregation problem (sidesteps summarizing or averaging values, distances or flow direction.) is an interactive system where nodes selection, so that certain links are shown instead of all links. This provides a clearer picture of small-scale behavior, but at the expense of losing the 'all data in one view' advantage, so pre-selected views must be stored and retrieved by memory.

WRV was developed to allow analysts to better visualize and understand network data. In the remainder of this study, we introduce an example problem: understanding migration between U.S. counties. We use this problem to explain the method and illustrate how it may be used to extract novel characteristics about the network. We first introduce the data set and note some of its salient characteristics. Next, we explain the WRV method in detail. We then give a detailed characterization of the clusters of types of migration patterns identified by WRV. We then conclude with possible future applications of WRV.

DATASET

We use data on U.S. county-to-county migration, collected by the IRS, to simulate a node/edge flow system in geographic space. We start with a matrix of 3140 x 3140 entries, where each column and row represents a U.S. county centroid. Our dataset then is whittled to 3061 nodes, as some of the 3140 counties do not post any significant migration. Unlike some network datasets, we do not count self-nodes, when movers choose a new home within the same county, but note that these values are typically high for each place. We add a distance matrix, listing the distance in kilometers between each county centroid and a matrix of angles between each node pair, where the vector head is at the origin and tail at the migrant's destination. (These can, of course, be reversed, if the destination is the node in question, but here, we concentrate on one type of flows only.)

Migrants are recorded for county-to-county flows over 10 people. The distribution of migrants per node shows an S-curve, where most counties have between 100 and 1000 migrants, relatively few have 10 to 100, while a few counties have over 10,000 or 100,000 migrants. (Figure 3.2) This indicates regularity among counties for over 70% of the dataset. In total, the migrant dataset contains around 5,500,000 migrants. The total migrants per county node are the summation of migrants to a county from different places of origin. The degree distribution, or the number of origins sending migrants to a destination-node, shows a similar s configuration as the distribution of migrants per node. Each node has at least one flow radiating from its center and at most 620 flows. (This maximum is found in the Phoenix, Arizona area.) Here, nearly half of the counties in the system have 10 or fewer county origins from which migrants arrive. Close to 1/3 of the number of counties in the dataset have about 10 to 100 different origins, and the remainder of the counties each has over 100. Note that the upward draw of this s curve is slower than that of the migrant distribution, indicating that there are fewer anomalies or outliers in the rank-size degree distribution and so groups of counties can be considered for their participation in classes of flow edge cardinality. These distributions are important to consider because cardinality can be a heuristic for worldliness, as it can be imagined that a geographic variety of migrants may bring more diversity to a place.

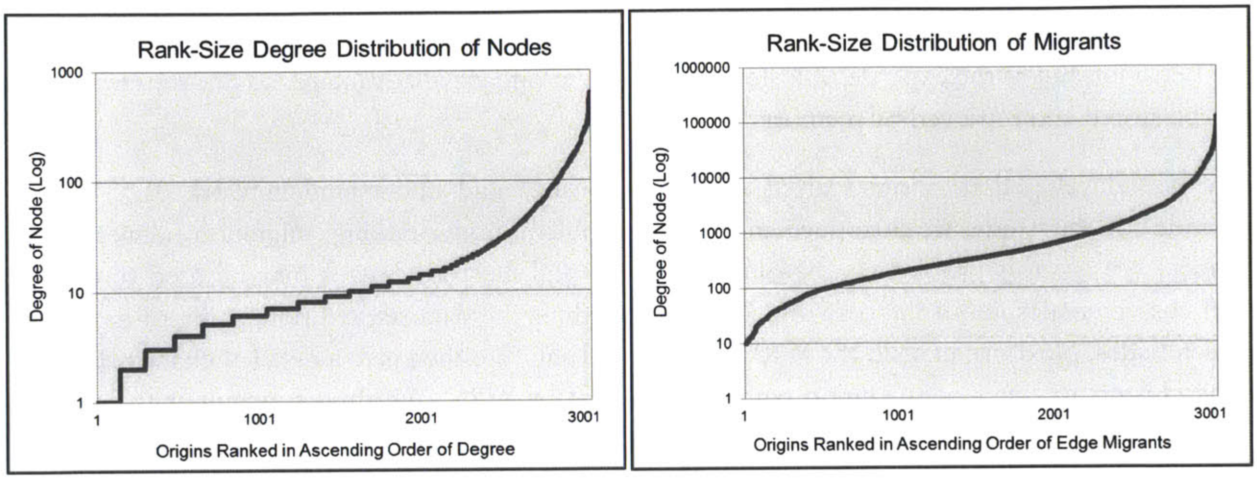


Figure 3.2: Rank Size Distribution of Migrants (R) and Rank Size Degree Distribution of Nodes (L) **each exhibit S curves, indicating regularity towards the log-median values.**

Each individual edge flow in the 'haystack' has three properties: a migrant weight (how many migrants travel on the edge), a length (the distance that the migrant travels) and an angle (the direction that the migrant travels). We explore the distribution of these statistics from an disaggregate perspective below.

We first explore the distribution of migrants per edge and find that the edge weight of migrants decreases quickly, meaning that streamlined channels are not as prevalent as edges that carry between 10 and 100 migrants. (Figure 3.3) The distribution of edge lengths shows that, surprisingly, 34 km is the most frequent distance travelled to migrate (541 edges participate), with a drop off at 50 km (a distance where 334 edges participate) and a final drop at around 100 km (where 150 edges participate). (Figure 3.4) This distribution shows that local migration is a frequent occurrence. Finally, distribution of edge angle values shows a preference for east-west migration, where edges most frequently surround peaks of 90 and 270 degrees, due east and west, respectively. (Figure 3.5) Interestingly, the symmetry of this distribution shows a lack of preference for a certain target migration channel.

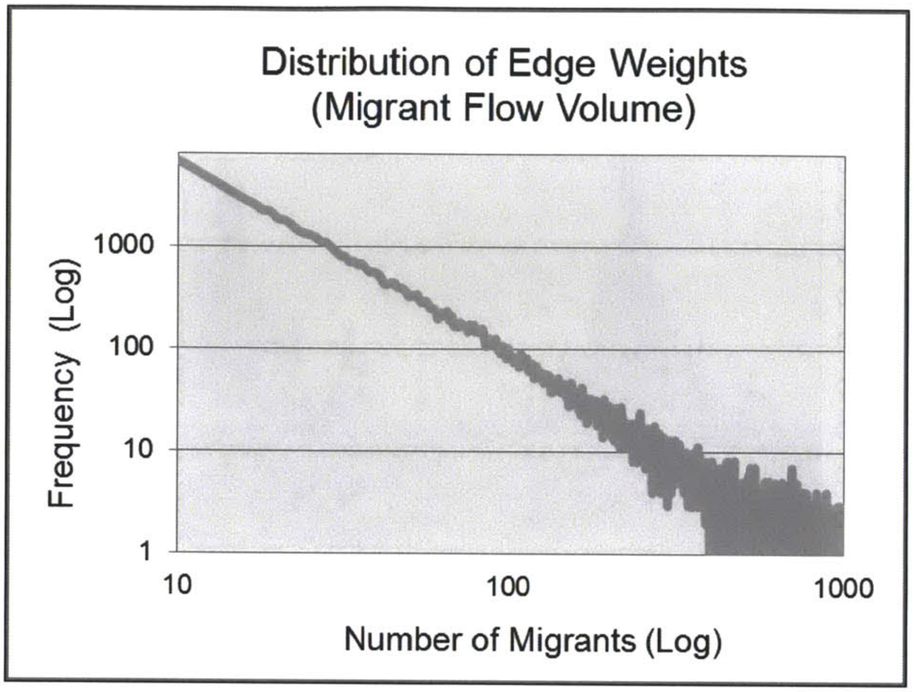


Figure 3.3: Log-Log Distribution of Edge Weights (Migration Flow Volume) The distribution of edge weights for the entire system shows that a single edge rarely has more than nearly 200 migrants.

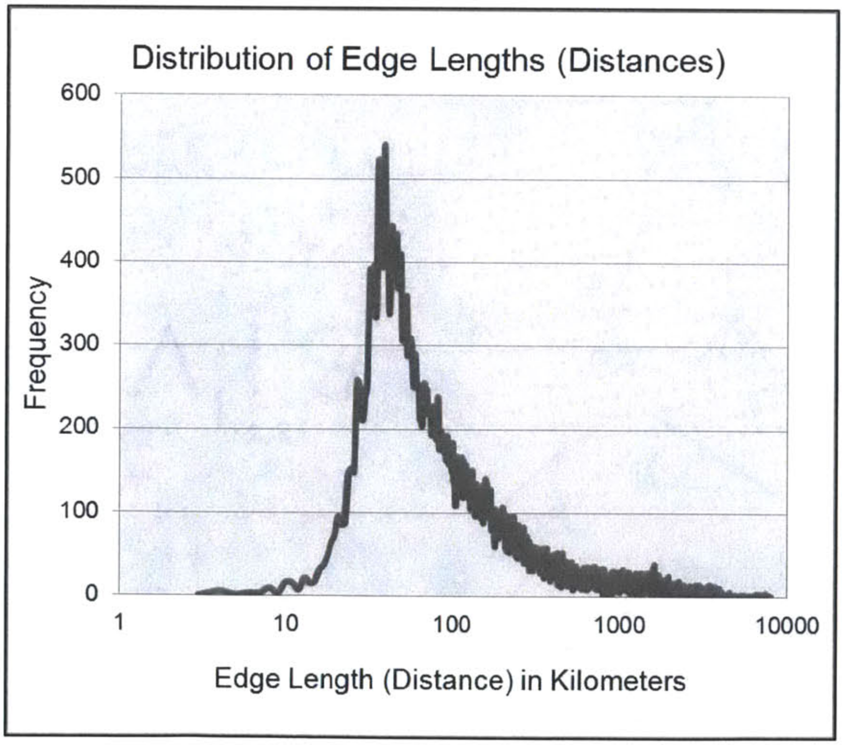


Figure 3.4: Distribution of Migrant Flows Edge Lengths The distribution of edge lengths shows a generally normal configuration with a peak around 36 km.

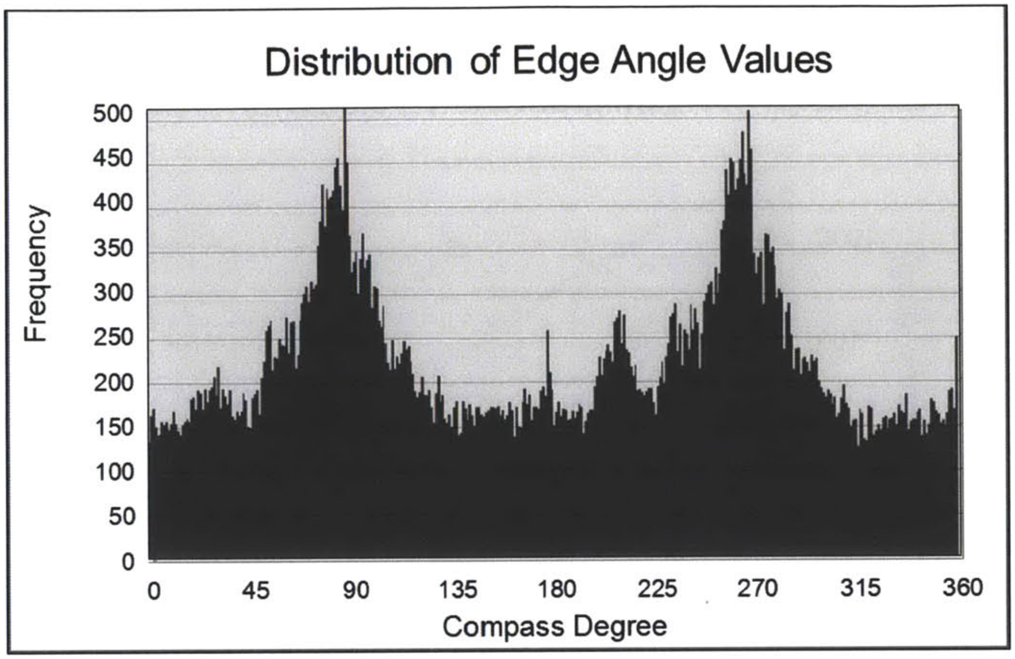


Figure 3.5: Distribution of Migrant Flows by Edge Angle Values **The distribution of edges radiating from a single node shows that the majority of moves go from east to west. The graph assumes that 0 degrees is due north.**

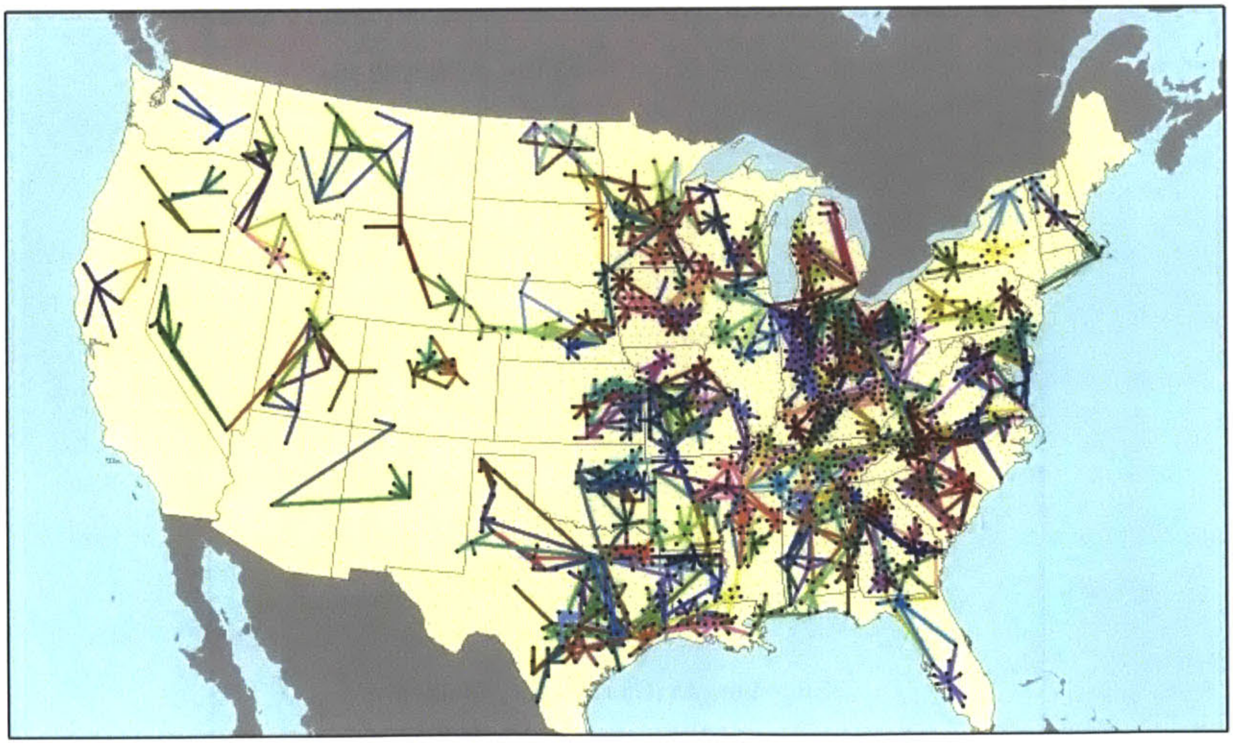


Figure 3.6: Selected Migration "Stars" in the Continental U.S. **Selected star plots in the Continental U.S. visually represent the larger migration network system.**

METHODOLOGY

Our goal is to characterize individual places by their out migration features. For each county, we extract 'stars,' where the node in focus is the center, and the flows leaving the node are attached to the central county centroid, and treated as part of the star. (Figure 3.6)

The star method has been used before for multidimensional data visualization, where each spoke from the center has a length equivalent to a specified quantitative feature of the entity. (Noirhomme-Fraiture 2002, Kandogan 2001) Others have taken the tool a step further, stressing interactivity (Teoh and Ma 2003) and evaluating the effectiveness of different star symbologies (Klippel et al 2009). The difference between this glyph-type entity visualization technique and our geographic case is that each vector in the radial system represents three (or more) characteristics instead of a single characteristic, as we have measures of (1) distance, (2) direction and (3) magnitude, for each. Also, our vectors are "tacked" to geographic space, meaning that an arc's radial direction is a variable with syncopated occurrences around the 0 – 360° radius, instead of an evenly-spread series of a pre-determined number of spokes in non-spatial star glyphs.

Noting that geographic stars (like graph structures) have a nearly infinite number of possible configurations, our probability of having a certain graph occur could be calculated by the convolution of 4 continuous variables: ray cardinality, magnitude, distance, and angle. To manage and group these e use an 'unraveling' technique to characterize the radial dynamics of each county's graph structure, to measure Weighted Radial Variation. For each county, we create a signature vector comprised of an edge weight (number of migrants) and distance value for each angle circling around the node from 0 – 360°. This signature vector is laid out as a signal over the radial steps to show similarities and differences between counties. (Figure 3.7)

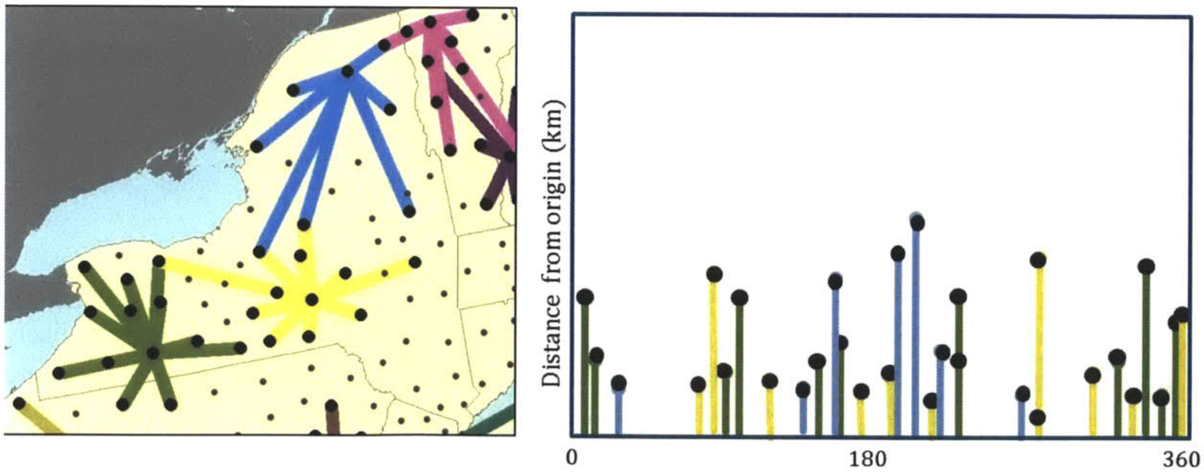


Figure 3.7: Migration Stars in New York State and 'Unraveled' Signal of Edges **Three migration stars in Upstate New York (L) are 'unraveled' at right to show three individual 'signals' whose**

components post different distances (on the y axis) as the signal progresses radially from 0 to 360. These branches are also weighted by number of migrants per edge.

Vector Instantiation

We create ***N*** vectors, where each individual node $\boldsymbol{N_k}$ is comprised of ***n*** concatenated tuples ***t***. In our system ***N*** is equivalent to the number of counties, 3061. The number (***n***) of tuples is equivalent to the node's degree (Figure 3.2), and ranges, as mentioned, from 1 to over 600. Additionally, each tuple ***t*** is comprised of only 3 values: Differential Angle, Distance (KM), and Number of Migrants. This value sequence is chained ***n*** times to form a vector, so that our shortest vector has three values (many examples), and our longest vector (representing the Phoenix area) has 620 * 3, or 1860 values.

Implementation

Angles and distance for each edge are calculated in the ArcGIS environment to preserve geodesic distance of vector components. The weighted signals are then clustered using an unsupervised K-means clustering algorithm in the Rapid Miner Environment (Mierswa et al 2006). From these classes, a typology is created for each cluster, where the number of desired resultant cluster classes can be chosen. (Jain et al 1999) This type of signal processing and pattern recognition analysis has been successful in a geographic context for classifying and understanding space. (Reades et al 2009) We choose 10 classes for differentiating between place types.

RESULTS

When these typologies of county graph types are visualized in geographic space, we are able to compare which counties are similar in their migration behavior, look at regional variation, and join demographic information. We also find that this method is robust with respect to including many edge weights per flow instead of a single measure of total migrants from county i to county j. For example, we can use our method to cluster stars where each flow has a measure of female migrants and male migrants, or migrants by age group. These typologies are then visualized via a single-variable cartographic representation that still represents the anisotropic 'spread/reach' of people migrating from different kinds of locales. From this representation, we can answer questions that would be difficult to answer otherwise, for example: how far flows travel, to what geographic direction are the flows travelling, and the magnitude of movers from each locale.

Overall, the results differentiate cities from rural areas, and show almost no clustering or autocorrelation. In fact, unless counties are of class 2, two adjacent counties are surprisingly unlikely to fall into the same class. These results lead us to believe that hierarchical drivers are in play, where k-means classes have less to do with regionalization, but with metropolitan area designation, population and population density. At first glance, we may guess that these classes are divided by such variables, as cities are certainly highlighted in Figure 3.8. However, population, number of migrants and total distance travelled are not factors that can be used as singular discriminants for splitting these data (Figure 3.9) leading us to believe that underlying graph structure, and the weighted dynamics of these graph structures, are significant components of the K-means clustering

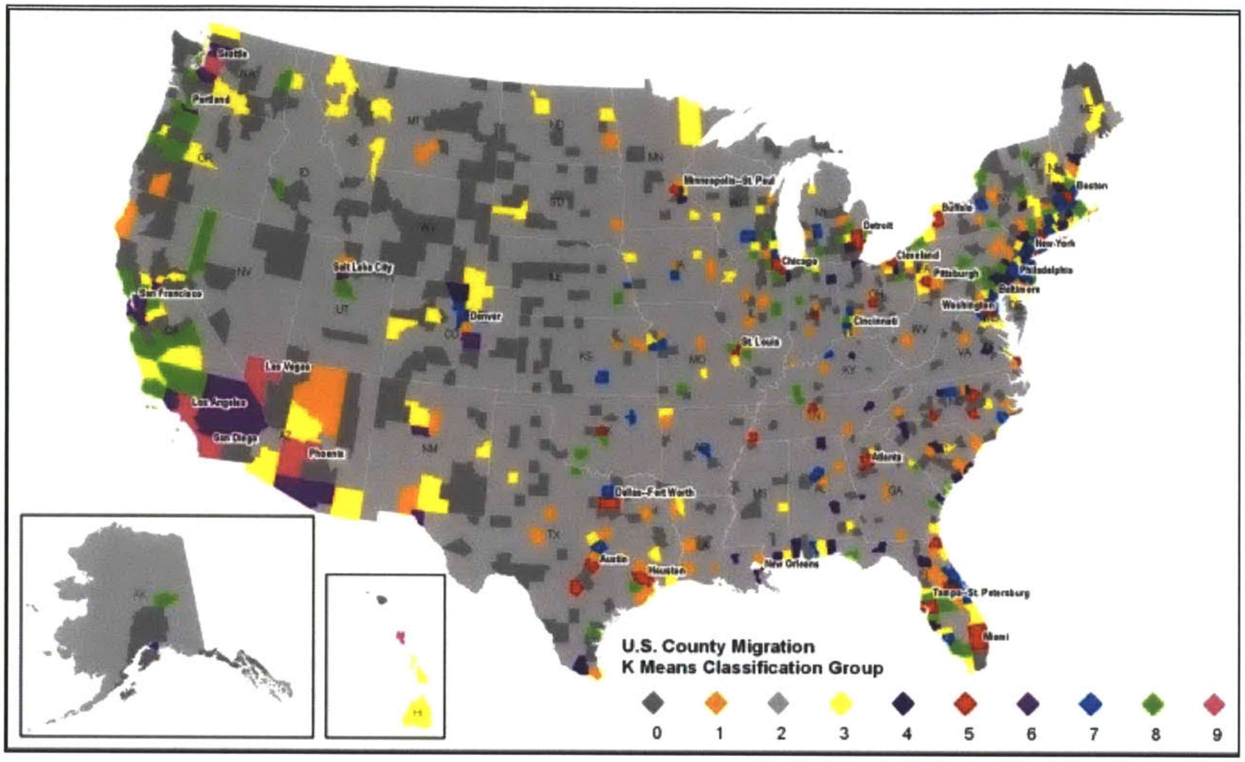
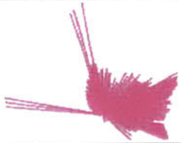
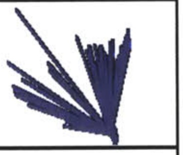
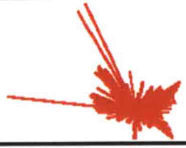
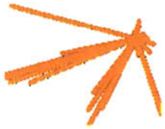
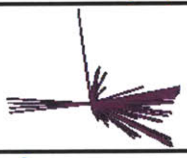
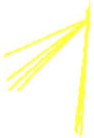
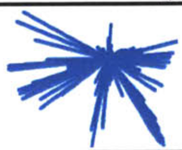
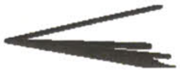
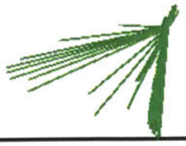
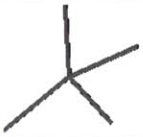


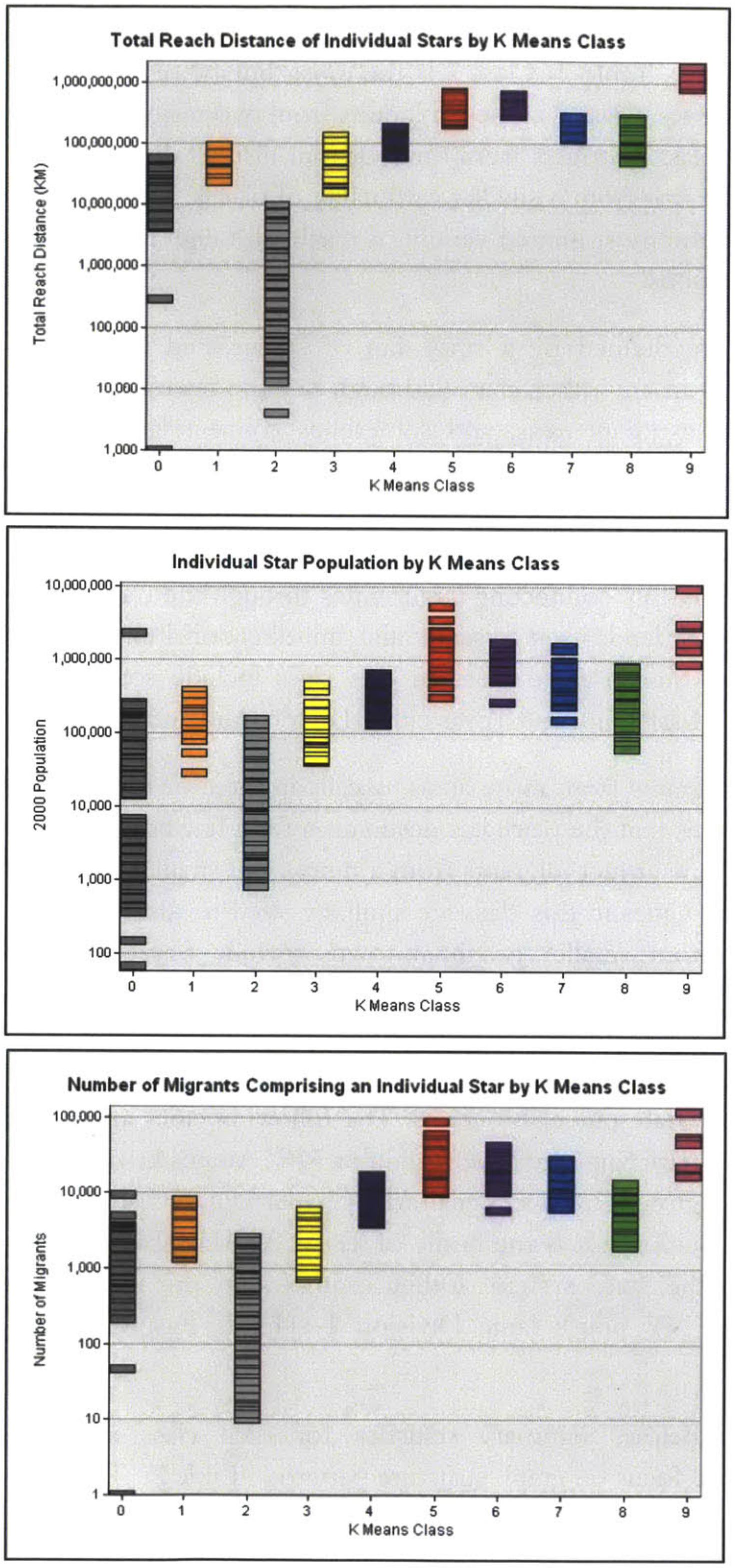
Figure 3.8: Results from K-Means Classification of Migration Star Arrays **A map of each county by K-means class shows that geographic distribution of migration classes is dispersed.**

TABLE I
ENTITY (STAR) STATISTICS BY K-MEANS CLASS

K-means Class	Number in Class	Avg. Number of Edges	Avg. Migrants Per Star	Avg. Total Distance	Avg. County	Average Age
0	394	21	1115	8,500	66	37.1
1	96	64	3384	46,400	178	34.6
2	2,297	8	320	900	24	37.8
3	105	43	2321	35,600	132	36.0
4	56	127	8031	120,300	352	34.9
5	43	301	27900	324,700	1,191	34.7
6	19	218	21065	408,000	888	33.8
7	49	181	13599	187,000	597	35.5
8	75	97	5718	101,300	303	34.6
9	7	465	51080	1,191,000	3,124	33.9
National Average					*(89)*	*(37.3)*

TABLE II
TYPOLOGIES OF COMMON STRUCTURES BY K-MEANS CLASS

Class	Common Structure	Class	Common Structure
Class 9		Class 4	
Class 5		Class 1	
Class 6		Class 3	
Class 7		Class 0	
Class 8		Class 2	



Total Reach Distance of Individual Stars by K Means Class

Individual Star Population by K Means Class

Number of Migrants Comprising an Individual Star by K Means Class

Figure 3.9: Features of 10 K-means Classes **(From top to bottom) Plots of total distance, population, and number of migrants at each individual star node per K-means class show that a single variable is not responsible for class splits.**

At first glance, we see a few different typologies of stars: Class 5, 9 and 6 draw from the widest variety of places. (Table 1, Table 2) Class 7 is the most 'bursty' of the classes, and exhibits evenly distributed radial properties. Class 8 attracts migrants from two major directions, while class 4 shows a fuller 'fan' of reach. Class 1 attracts from three or four distinct directions. Class 3 is often a two-pronged structure, attracting from a smaller cardinality of places. Class 0 is the most uni-directional class, while class 2 has the most limited variety of reach, although its configuration lends itself to a variety of cardinal directions.

Counties in Class 0 are defined by a small fan of streamlined inflows, and have the smallest populations of counties in any other class (although not the fewest migrants). (Figure 3.9) These counties are found in exurban areas, and sometimes connect larger cities together by way of adjacency. Planners in these cities can expect to draw migrants from few directions, possibly due to linear infrastructure (like highways) or lack of compass options for migration. For example, counties in these classes mushroom along a chain of highways skirting the Appalachian Mountains from Virginia to Mississippi, visibly connecting larger cities through the Carolinas, Atlanta and Alabama. This intermittent behavior lends itself to sequential, uni-directional migrant transfer along roads that enable low cost travel times. Large cities in this class include southern cities like Tuscaloosa, Alabama and Jackson, Mississippi, and Texas cities, Laredo, Midland, and Amarillo.

Class 1's cities draw migrants from more areas, usually in three or four main directions. This 'claw' type of structure indicates that the place is a destination for a few nearby pockets of places. In many cases counties in class one attract migrants from a number of smaller areas that are likely to be on a similar highway system. Cities in this class are similarly sized to those in class 0. Anchor cities that likely draw migrants from smaller proximal towns include Shreveport LA, Chattanooga TN, Hampton VA, Green Bay WI, Billings MT, Topeka KS. Unlike cities found in suburban outskirts and sprawl, these cities are markedly not part of urban agglomerates, nor as synergistic cities with a nearby partner, and range in population from 100,000 to 200,000. One interesting finding in this group is the number of state university towns. The following cities are the primary campus of the state-wide university system: Santa Fe NM, Columbia MO, Athens GA, Lawrence KS, Iowa City IO, Medford OR, Bloomington IN, Fayetteville AR, Chapel Hill NC. Fargo ND is home to North Dakota State, College Station TX is the home of Texas A&M; Ashville NC, Lafayette LA, to large satellite campuses of the state system. Other examples in this category are Wilmington NC, Bethlehem PA, Ithaca NY house large Division 1 schools East Carolina, Cornell and Lehigh Universities.

assignment schedule. Below, summary statistics for each class also do not offer clear-cut discriminants, and so we focus on graph structure patterns. (Table 2)

The vast majority of counties in the U.S. fall into Class 2, although the populations of these counties are generally smaller than any other group. (Table 1) The graph structure of counties in this group are the sparest and simplest of the system (Table 2), with few spokes in a line, angle, three-tier, cross or simple star configuration. Also, entities in this class have markedly lower distances and migrants that the remainder of the classes (Figure 3.9). Class cities include Midwestern Capitals Frankfort KY and Jefferson City MO. Class 2 also includes large cities in the Deep South, Vicksburg MS, Tupelo MS, Decatur AL, in West Virginia: Beckley WV, Wheeling WV and Virginia: Roanoke VA, Lynchburg VA, Suffolk VA, Fredericksburg VA, Fairfax VA, Manassas VA. Although Fairfax, Manassas and Fredericksburg, are all considered part of the Washington D.C. metropolitan area, with notable commuter streams, they exhibit the same migration patterns as the plethora of small cities and sparsely populated counties. With the exception of the capital cities, the aforementioned entities have historical preservation components to their cities, and small tourist industries. Eastern VA and West Virginia are known for elderly residents; this category's counties have the oldest average age of any category. This category's cities are also generally devoid of major transportation infrastructure like airports, ethnically diversity, higher education and international draw.

Classes 3 and 4 show distinct fanning, with class 4 exhibiting a fuller range of cities from which migrants arrive. (Table 1) Class 3 typically draws from cities in on two or three distinct directions, also indicating that road infrastructure might be in play. Cities include Lubbock TX, Kansas City KS, Greeley CO, Rochester MN, Duluth MN, Sioux City SD, Racine WI, Oshkosh WI, Kenosha WI in the Midwest, West Cheyenne WY, Missoula MT, and small cities Charlottesville VA, Dover DE, Warwick RI, Niagara Falls NY in the East. Each of the aforementioned cities is marked with a steady, stable population, and a generally homogenous citizen body. The mountain region sees a lot of class three, indicating that this type may be characteristic of sparsely populated regions. Class three also notably fills in space between coastal counties with fuller migrant draw patterns.

California cities are typically considered diverse and far-reaching, especially with respect to Asian and Hispanic populations, but Modesto CA, Davis CA, San Louis Obispo CA fall into this category of low cardinality, and near-reaching pull. Their existence in this class may indicate that populations drawn to these cities are more standard, local and homogenous, and perhaps that the offerings of these cities are less colorful, and better-suited for attracting local community.

Cities in class 4 are marked with a wider fan of migrant origins, and are situated almost exclusively west of the Mississippi River. Class 4 counties dot the Deep South and Gulf, as well as the Northeast. Class 4 counties include larger cities in the industrial north, beginning with Baltimore MD, westward to Toledo OH, Akron OH, Lexington KY, Fort Wayne IN and St. Paul MN. St. Paul MN is often touted as the "Last City in the East", a dictum that is evidenced by its participation in this eastward-pulled class. In the Deep South and Gulf, Norfolk VA joins cities Charleston SC, Savannah GA, Huntsville AL, Montgomery AL, Gulfport, Biloxi MS, Mobile AL, New Orleans LA, in this category. Each of these cities contain older infrastructure, historical influence and less progressive policy—like relatively lax tobacco usage. As discussed by Hulten and Schwab (1984) these areas have been 'slowed by an aging public infrastructure, deteriorating urban environment an obsolete capital stock and institutional sclerosis.' (Pg. 152)

Those looking to economically invest in the temperate, cultural and aesthetically-pleasing cities like Savannah, Charleston and New Orleans may note their presence in a category with the same limited migrant magnetism as aging cities.

As an aside, this class also includes a number of large college towns: Tallahassee FL (Florida State), Gainesville FL (Florida), Baton Rouge LA (Louisiana State), Fort Collins CO (Colorado State), Trenton NJ (Rutgers), Wilmington DE (Delaware), Ann Arbor MI (Michigan), Knoxville TN (Tennessee) and Durham NC (Duke), Winston Salem NC (Wake Forest), Providence RI (Brown).

In class 5, we see the strongest pulls of any class until class 9. These cities are found exclusively in the Midwest and east, and include migrant-magnet New York NY. In concordance with each class, cities in class 5 are rarely, if ever nearby one another, but act as regional anchors throughout the Sunbelt, Snowbelt. Class 5 typically draws from the entire U.S., and therefore it is not surprising that regional "Command and Control Centers" (Frey 1990) such as Atlanta and Minneapolis-St. Paul, draw diverse populations to their "world class" influential centers. The cities distinctly dot 3 regions: the Midwestern Snowbelt region, including Buffalo and Rochester NY, Pittsburgh, Columbus OH, Cleveland, Indianapolis, St. Louis, Detroit to metropolitan Chicago; Sunbelt cities Memphis TN, Nashville TN, Charlotte NC, Raleigh NC, Tampa FL, Jacksonville FL and Texas powerhouses, Dallas, Austin and Houston, and Northeaster corridor. These designations as coherent regions are founded by previous research. For example, Rodgers (1952) defines industrial inertia in terms of The Pittsburgh-Cleveland-Buffalo triangle, marked by steel production, and its expansion to St. Louis, Indiana, Detroit and Chicago-Gary. In comparison to the cities in the 'Snow belt' region in class 4, these Snow Belt cities may be sidestepping economic stagnation, as they differentiate themselves with the poor-cardinality draw of class 4 cities like Fort Wayne IN. The inclusion of these classical industrial regions in the same category as warmer climates, especially the large Atlanta and Charlotte, a favorite for Fortune 500 Companies, is evidence that this 'rusty' region still has a strong similar pull, despite some discussion of its current or impending dilapidation.

Thirdly, although Boston and Washington D.C. are not explicitly in this category, neighborhoods indicate that their gravitational pull is relevant. In the Boston area, areas rich with medical, technical, pharmaceutical, research and firms like Cambridge MA, Somerville, and Arlington, as well as D.C.'s Bethesda, and Gaithersburg MD, home of the National Institutes of Health (NIH) and National Institute of Standards and Technology (NIST) can understandably fall into this category because of the national pull for skilled workers in an area dense with higher education. Additionally, migrants to these more suburban areas are likely drawn to downtown amenities and 'worldliness', a migrant's place of residence in one of these cities can indicate that he works or reaps the benefits of the larger surrounding city.

This class is especially important because it illustrates the relative powerlessness of intervening opportunities (Stouffer 1966, Galle and Taeuber 1966, Bright and Thomas 1941) when migrants are considering these cities. Because cities in this class exhibit a strong national draw, we can insinuate that migrants cannot be satisfied with closer locales, and thus these cities have a special magnetism. Importantly, the 'rust belt' has been cited recently as an area of economic decline. However, their participation in this group illustrates that they are still in the same class as cities that are seen as prospering.

Class 6 cities draw a variety of migrants for their relative sizes. Their social reach is diverse, with steady currents to many major cities. This class includes cities exclusively in the West: Pacific Northwest cities Tacoma WA, Everett WA, which flank Seattle, home to mega corporations Microsoft and Boeing, and Portland OR, known for its progressive urban planning and sustainability initiatives, join equally progressive Tucson AZ, and Denver CO. Most notably, California cities San Francisco, San Jose, San Bernardino, Sacramento, Oakland, Riverside CA, Fremont CA, Santa Clara, Berkeley, Oxnard, Thousand Oaks, San Mateo, Mountain View, Redlands, Redwood City, Citrus Heights, Palo Alto, Laguna, Menlo Park, and Saratoga and others fall into this group. These Silicon Valley, and satellite silicon cities are marked with high literacy and education rates, high incomes, and diverse lifestyles. This area is marked with up-and-coming research and industry, including Google, Adobe Systems, Disney Pixar, Facebook, Oracle Corporation, eBay, Apple Inc., Cisco Systems, Hewlett-Packard, Intel, Yahoo!, Sony, TiVo, Nokia, YouTube, VeriSign and Netflix. (Sturgeon 2000) This category succinctly extracts places that have the opportunity to attract and evidence of attracting from a variety of places, accordingly the average age of this class is the lowest at 33.8 (Table 1). Surprisingly, other cities in this category are Salt Lake City UT, Albuquerque NM, El Paso TX, Colorado Springs, CO and Anchorage AK, indicating that new technology firms may be sprouting here, or that site suitability models for new firms should also take these cities into account.

Class 7 cities include metropolises Philadelphia and Washington DC and a steady chain of coagulate satellites within the gravitational pull (Huff 1964) of New York City, such as Newark NJ, New Haven CT, Hartford CT, Hackensack NJ, New Brunswick NJ, White Plains NY, Levittown PA, Hicksville NY. Similarly, cities and suburbs of Boston, MA like Quincy, Lynn, Haverell, Peabody, Revere, Needham, Gloucester, Methuen, Leominster, Fitchburg, Worcester MA, are also in this class. With the exception of Needham, these proximal places, nested in the Massachusetts Bay periphery are historically known for industry, especially manufacturing.

In the Midwest, Milwaukee WI, and Cincinnati OH anchor industrial mid-western cities Grand Rapids MI, Kansas City MO, Dayton OH, and Louisville KY. Similarly-sized Birmingham AL, and Little Rock AR also join from the same longitude in the Deep South. The class continues westward to include large Great Plains cities Tulsa OK, Omaha NE, Wichita KS, and Des Moines IA.

Boulder CO, Madison WI, Annapolis MD and Syracuse NY, in this category, may attract a similar array of researchers due to Universities, research facilities and military organizations like the Naval Academy. Boulder and Annapolis are also proximal to Denver and Washington D.C., respectively, giving potential migrants the benefits of efficient air travel and access to diverse goods and services. Florida havens Pensacola, Fort Meyers and Daytona Beach are known to be warm-weather havens and may draw similar channels of aging migrants.

Class 8 migration structure is limited for the size of its populations. The structure consists of draws from two major directions, and exhibits fanning in these directions. These counties are geographically situated in the southwest and west with cities Fresno CA, Bakersfield CA, Stockton CA, and Provo UT Reno NV and Boise ID. These cities do not have the draw of larger California cities on the Pacific Coast, although interestingly, acclaimed Santa Barbara CA, considered a wealthy haven, is included in this class. Provo and Reno do not draw the same diverse migrants as their respective same state anchors Salt Lake City and Las Vegas. In a similar fashion, small Pacific Northwest cities Salem OR, Eugene OR, Vancouver WA, and Spokane WA draw a smaller, less diverse set of migrants than their local anchors, Seattle and Portland. Midwestern and Plains cities Lincoln NE, St Louis MO, Springfield MO, Lansing MI, Champaign IL are also included. Of these, St Louis is the largest city. As state capitals, Lansing and Lincoln, along with Harrisburg PA and Richmond VA, may draw intra-state citizens, like legislators and politicians. Lincoln and Champaign, along with South Bend IN, Norman OK, Binghamton NY boast large universities—three of which are comprised mostly of in-state students.

In a more temperate climate, Augusta GA, Hilton Head SC, Land O'Lakes FL, Naples FL, Holiday FL draw "snow birds" from the Midwest more than the Northeast. These cities offer amenities for the retired, especially notable golfing infrastructure.

Class 9 cities exhibit a heavy eastward draw. The opposite of class 2, this group has few counties, but each of these few counties (9) has a high cardinality of flows. Following suit, this group also draws the most migrants per city. (Table 1) The Phoenix area draws the maximum flow edges at over 600. This class also includes, Las Vegas, Los Angeles and their respective large surrounding areas, as the counties in which these southwestern metropolises are situated are notably large in area. This class also includes San Diego, whose population is comprised of a Hispanic proportion similar to the aforementioned cities. Seattle and Honolulu HI, are also included here. Geographically it could have been expected that Portland and San Francisco would have been included in this class, but their structures better matched those of class 6. San Diego could have been included in class 6, but perhaps because of a lack of progressive technology, and a more traditional military hold with Naval influences, San Diego remains in the same class as its larger, sprawling cities.

Our results lead us to question what it means to have community typologies, and what can we learn from this. Even at first examination, we could tell heuristically from the cardinality of flows that some communities exhibit much extroverted outreach qualities within the flow system while others stretch less but these categories give a more colorful illustration of community typologies. However, we are still unsure how to completely interpret this behavior from a local perspective, some potential interpretations can come from the field of economics.

In the next section, we discuss insights and suggest applications for this technique.

APPLICATIONS AND CONCLUSION

(1) Infrastructure Planning

There are many potential applications for this method since this technique can take in a wide variety of flow data. Relevant data types include phone calls, emails, air transportation and commuting. Given that commuters want the most efficient way to get to work, with challenges of traffic, commuters often fight for funding for new or better infrastructure—types of infrastructure like rail, widened roads, or better airport access.

The timing of migration, and channels through which it flows, means that a model with declining moving costs fits better than a neoclassical model of free or fixed-cost mobility.

(2) Economic Assessments
Using WRV, it should be possible to find out what the advantages of certain features are, or the impact of those features on the local network structure. It can be expected that a city with a world-class facility may exhibit different flow properties than its neighbors. Since the town with the special facility general holds the same characteristics as the places nearby. For example a college, or in the United States, the Mayo Clinic in Minnesota, Santa Fe Institute in New Mexico, Oak Ridge National Laboratory in Tennessee, and Cape Canaveral NASA Station in Florida are examples of facilities that draw workers and visitors from a wider variety of places, than other cities in their region.

(3) Delineation
Although our results are not clustered, a clustering of entities of homogenous typologies could allow for the delineation and partitioning of similar places. These places can then be used for functional regions, planning collations or policy instantiation.

(4) Prediction
A temporal assessment of flow types could inform the changing nature of places by their flow dynamics. An entity that exhibits a syncopated temporal trajectory with another entity can use the latter entity's current state as a probabilistic preview of future flow behavior. In our case, analyzing the temporal graph structure change in Las Vegas since 1950, and matching increments of this change to other cities, might yield

(5) Tobler's First Law for Flows
The amalgamation of flow data characteristics does not seem to yield a conclusion that nearby places exhibit similar behavior. Perhaps when including in and out flow to characterize a place, Tobler's classic First Law of Geography may benefit from parameterization or reconstruction.

In conclusion, our aim for this new visualization technique was to preserve individual, disaggregate characteristics of flow data while allowing the information to be explored in a single view. By extracting and clustering different geographically-tacked graph configurations, we are better able to understand the distribution of human movement patterns in space, while using disaggregated data. This method is not limited to migration patterns, but can be used for other datasets where origin "reach" is a metric of interest like commuting flows, phone call volume, or temporal vacation/leisure flows.

We foresee many possible applications for WRV, both scientific and applied.

ACKNOWLEDGEMENT

This work is supported in part by the ASEE National Defense Science and Engineering Fellowship Program, the VACCINE Center (Visual Analytics for Command, Control, and Interoperability Environments, Grant 2009-ST-061-CI0001), and the Contextual influences on the category construction of geographic-scale movement patterns (ConCat) grant from the National Science Foundation (Grant 0924534). The Vaccine Modeling Initiative grant from the Bill & Melinda Gates Foundation (Grant 49279) is gratefully acknowledged, however no endorsement is implied.

REFERENCES

[1] O'Kelly, M., A Geographer's analysis of hub-and-spoke networks *Journal of Transport Geography,* **1998,** *2,* 171—186

[2] Xu, Z. & Harriss, R. Exploring the structure of the US intercity passenger air transportation network: A weighted complex network approach *GeoJournal,* **2008**, *73*, 87—102

[3] Jiang, B. & Claramunt, C. Topological analysis of urban street networks *Environment and Planning B: Planning and Design,* **2004**, *31*, 162

[4] MacEachren, A. & Kraak, M. J. Exploratory cartographic visualization: Advancing the agenda *Computers and Geosciences,* **1997**, *23*, 335—343

[5] Anselin, L. Exploring spatial data with GeoDA: A workbook spatial analysis laboratory *Department of Geography, University of Illinois at Urbana-Champaign, Center for Spatially Integrated Social Science,* **2005**

[6] Marble, D.; Gou, Z.; Liu, L. & Saunders, J. Recent advances in the exploratory analysis of interregional flows in space and time *Innovations in GIS 4: selected papers from the Fourth National Conference on GIS Research UK (GISRUK), CRC* **1997** *75*

[7] Guo, D. Flow mapping and multivariate visualization of large spatial interaction data *IEEE Transactions on Visualization and Computer Graphics,* **2009**, *15*, 1041—1048

[8] Andrienko, N.; Andrienko, G. & Gatalsky, P. Supporting visual exploration of object movement *Proceedings of the Working Conference on Advanced Visual Interfaces (Palermo, Italy),* V. Di Gesù, S. Levialdi, L. Tarantino eds., *ACM Press*, **2000**, 217—220

[9] Andrienko, G.; Andrienko, N. & Wrobel, S. Visual analytics tools for analysis of movement data *ACM SIGKDD Explorations Newsletter,* **2007**, *9*, 38—46

[10] Dykes, J. & Mountain, D. Seeking structure in records of spatio-temporal behaviour: Visualization issues, efforts and applications *Computational Statistics & Data Analysis*, **2003,** *43,* 581—603

[11] Tobler, W. Automation and cartography *Geographical Review,* **1959**, 526—534

[12] Tobler, W. Migration fields *Population,* **1978**, 215—232

[13] Tobler, W. Experiments in migration mapping by computer *Cartography and Geographic Information Science,* **1987**, *14*, 155—163

[14] Andrienko, N. & Andrienko, G. Spatial generalization and aggregation of massive movement data *IEEE Transactions on Visualization and Computer Graphics,* **2010**, *16*

[15] Wood, J.; Dykes, J. & Slingsby, A. Visualisation of origins, destinations and flows with OD maps *The Cartographic Journal,* **2010**, *47*, 117—129

[16] Noirhomme-Fraiture, M. Visualization of large data sets: The zoom star solution *International Electronic Journal of Symbolic Data Analysis,* **2002**

[17] Kandogan, E. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* **2001**, 107—116

[18] Teoh, S. & Ma, K. StarClass: Interactive visual classification using star coordinates. *Proceedings of the 3rd SIAM International Conference on Data Mining,* **2003**

[19] Klippel, A.; Hardisty, F.; Li, R. & Weaver, C. Colour nhanced star plot glyphs: Can salient shape characteristics be overcome? *Cartographica: The International Journal for Geographic Information and Geovisualization*, **2009**, *44*, 217—231

[20] Mierswa, I.; Wurst, M.; Klinkenberg, R.; Scholz, M., & Euler, T. YALE: Rapid prototyping for complex data mining tasks *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06),* **2006**

[21] Jain, A.; Murty, M. & Flynn, P. Data clustering: A review. *ACM Computing Surveys,* **1999**, *31*, 264—323

[22] Reades, J.; Calabrese, F.; & Ratti, C. Eigenplaces: Analysing cities using the space-time structure of the mobile phone network *Environment and Planning B: Planning and Design,* **2009,** *36*, 824—836

[23] Hulten, C. & Schwab, R. Regional productivity growth in U.S. manufacturing: 1951-78 *The American Economic Review,* **1984**, *74*

[24] Frey, W. Metropolitan America: Beyond the transition *Population Bulletin*, **1990**, *45*, 49

[25] Rodgers, A. Industrial Inertia: A major factor in the location of the steel industry in the United States *Geographical Review,* **1952**, *42*, 56-66

[26] Stouffer, S. Intervening opportunities and competing migrants *Journal of Regional Science,* **1960**, *2*

[27] Galle, O. & Taeuber, K. Metropolitan migration and intervening opportunities *American Sociological Review,* **1966**, 31

[28] Bright, M. & Thomas, D. Interstate migration and intervening opportunities *American Sociological Review,* **1941**, *6*, 773-783

[29] Sturgeon, T. How Silicon Valley came to be. In *Understanding Silicon Valley: Anatomy of an Entrepreneurial Region* Ed. Martin Kenney, *Stanford University Press*, *California,* **2000**

[30] Huff, D. Defining and estimating a trade area *The Journal of Marketing*, **1964**, *28*, 34-38.

CHAPTER 4

Predicting Migration System Dynamics with Conditional and Posterior Probabilities

ABSTRACT

Traditional models of migration assume that migrants move to places of greatest economic incentive, and are more likely to move when current economic conditions 'push' migrants from their origin. Although prospective income at a destination has been a major determining factor for migration in pre-existing migration models, and distance between origin and destination is also a major consideration, we take a new approach with a model that reflects migration 'chaining', where migrants to a city B send information back to their origin city A, and interest other members of A to migrate to B. We isolate the social factors of place-pair synergies through components from Bayes' Law: conditional probability and posterior probability of unique origin/destination migrant volume, and a system-wide probability of unique O/D transfer. These allow us to model social space as well as physical space, rather than physical space alone.

We test these variables' power for predicting future migration against four other predictive models: the traditional gravity model, transit data, airline and trip data, and linear trends. We use a case study of U.S. Migration flows in a system of major cities, given annual data from 1996-2004 to predict city-to-city flows annually for 2005-2008, and find that conditional and posterior probabilities outperform system-wide probabilities, gravity, transit and linear forecast models. These probabilities also exhibit a surprising level of steady-state stationarity, and therefore are a promising avenue for more accurately modeling future migration flows.

Keywords—Migration, Probability, Flows, System Dynamics, Bayes' Law, United States

INTRODUCTION

Accurate inter-city migration models are essential for urban and regional planners and economists because population growth or decline directly affects the infrastructure needed to sustain city life. With proper use of these predictive models, urban planners can better manage for smart growth, and smart shrinkage, or initiatives like Special Economic Zones. Since city populations usually fluctuate in response to migration dynamics, predicting the size of cities in the future requires good estimates of inter-city migration flows.

Traditionally, economists have modeled propensity to migrate from an origin to a destination as the income differential between the two entities. Incorporating Regional Economic theory (Isard 1960), the Euclidean distance between origin and destination is also a major consideration. In this study, we create a model that reflects migration 'chaining', where migrants to a city B send information back to their origin city A, and interest other members of A to migrate to B. Where previous models have aggregated origins and destinations for a large general model, here, we are interested in modeling the unique magnetism between individual place-pairs and consequently, the predictability/reliability of this magnetism, as an effect not of Euclidean space, but of social choice. Echoing Limtanakool et al (2007), we believe that what ties people to particular places today is not travel costs, but social and interpersonal relationships.

We hypothesize that we can best model and predict migration by harnessing elements of Bayes' Law (Bayesian Inference): the conditional and posterior probabilities of moving given specific origins and destination. We apply these probabilities to longitudinal population projections in an effort to predict future interaction between city pairs.

We use a case study of U.S. inter-county migration with year-to-year address changes for more than 200 million people in the United States to represent migrant flows for years 1995 through 2008 and compare the accuracy of a calibrated gravity model, with our Bayesian place-pair model.

The study proceeds as follows: First, we review recent literature in migration, and briefly outline our approach. Section 3 introduces the migration, flight, travel, and geographic datasets. Section 4 describes our methodology and our approach for assessing the predictive power of different models, in two parts. First, we parameterize a gravity model and create our Bayesian components. In Section 5, we compute the predictive power of each of the following models: trip model, flight model, gravity model, linear forecast, and Bayesian technique, their correlation with actual migration data by year, and each method's ability to project the future dynamics of inter-city migration. Finally, we conclude with an overview of our findings.

MODELING MIGRATION

In our review of the literature, we discuss previous models and data used to predict migration volume. This section reviews literature in the field of migration determinants (reasons why we migrate) models, and effects. Finally, we present reasoning for our model.

Previous Models

This section reviews literature in the field of migration determinants (reasons why we migrate) models, and effects. Finally, we present reasoning for our model.

Economic models generally try to give an overarching theory of in and out migration at an aggregate level. Individual cities are not typically considered, but are represented by their characterizing features (like average income). Since individual cities are not usually individualized for these popular models, unique city pairs are certainly less likely to be considered. In other words, migrants are modeled as they move from one type of place to another type of place, but the places are never specific pairs.

1) Chaining: First, 'chaining' behaviour has been acknowledged as a well-known feature of migration, where migrants exhibit a strong tendency to move to places where others at their origin had previously moved, hence a stream of migrants can be found between unique pairs of origins and destinations. Chaining is abundant in migrant behaviour because of its benefits: migrants (1) would like to reunite with friends, and more frequently, a family member (2) receive descriptions about conditions at the destination, and (3) realize that established contacts at the destination can facilitate job and housing processes. (Fawcett 1989)

Regarding communication, the Petersen-Greenwood hypothesis says that communication networks are vital to early and sustaining decisions to migrate. (Tarver 1973) Out-migrants often send job and housing information to friends and family 'back home.' (Hansen 1940, Gottlieb 1997). Historically, letters from 19th century European immigrants informed would-be migrants of U.S. labour-market conditions. (Hansen 1940) In more recent times, Rojas (2010) shows evidence of migration chaining through global telephone calls. Remittances are also expected to have a strong correlation with migrants. (Castells 2005)

Second, regarding established contacts, chaining creates an economy of scale, as moving costs fall with the stock of migrants. (Carrington 1996) Previous migrants frequently provide later migrants with professional opportunities, credit and temporary housing (Scott 1920, Grossman 1989). These pre-established friends reduce the psychological and monetary cost of adapting to a new lifestyle and place. (Massey 1990, Marks 1989)

2) Economic Models: The Neoclassical model of migration argues that labor flows from poor to wealthy areas, and that this volume of migration should lessen as the differential of income between the two cities reaches equilibrium. Based on this notion, economic opportunity is the main parameter for predicting how attractive a place will be for potential migrants. Movers from i to j is most often modeled as a Function of {Distance between i and j, Income at i, Income at j, and Auxiliary factors}. Rogers (1967) tests the hypothesis that 'employment, occupation and salary are the major considerations in any decision to move.' Job-search models indicate that movement occurs when potential migrants can earn higher income for a extended ranges, or a 'lifetime'. (Yezer and Thurston 1976) Migrant models also incorporate data from the US Bureau of Economic Activity, which reports relative employment opportunities (REO) and relative wage rate (RWR), a wage rate adjusted for cost of living (Treyz et al 1993) or per-capita government expenditures (Greenwood and Sweetland 1972). Other models include a wage subsidy to promote a competitive wage, and a migration-restriction police. (Harris and Todaro 1970)

In another light, economic rational has drilled in the idea of cost/benefit. Dorigo and Tobler (1983) illustrate the value of quantifying reasons to 'stay or go' by push and pull factors, while Stouffer's (1940) theory of intervening opportunities is also often incorporated into more robust models.

Aside from income and wage differentials, demographic and living factors or perceived factors (Schachter and Althaus 1982) are taken into account. Also, type of job has been used as input data: Horiba (2000), which could inform social decisions to relocate, as relationships tend to be concentrated within a discipline or within a vocation. (Lee 1970)

Also secondary to economics are considerations like age and gender. (Mangalam and Schwarzweller 1970) The young and single are the most mobile of the population, and therefore the percentage of this demographic at each city is used to predict migration. (Franklin 2003) Weather and climate are also prominent in migration decisions, as well as moving costs. (Schachter and Althaus 1982, Perry 2003) Even historically, most invasions/conquests were related to the people having a nice climate/area. (McNeill 1984) Further, gender is often taken into account in an economic model. Rodgers (1967) finds that the "economic opportunities" hypothesis doesn't explain female migration. Pedraza (1991) and Sassen (1995) illuminate reasons for moving that are inherent to women, and thus, indirectly support a different model for women and men.

Next, another class of non-traditional models consider equilibrium, steady-states and stationarity.

3) Stationarity: Issues of stationarity have been addressed either directly or indirectly through a number of different methods. One method of modeling temporal stationarity has been to assess urban rank-size transition probabilities via Zipf's Law or a Pareto Distribution (Greenwood 1985), temporal Markov chains (e.g. Joseph 1975), geographically interpolated stationarity (Tobler 1978), and eigenvalue decomposition (Reades et al 2009) to determine the steady state stationarity. These studies follow a theory of stationarity that is over 125 years old. (Tobler 1995, Ravenstin 1885) Additional approaches include household choice, (Sassen 2002) network systems (Limtanakool et al 2007) and cellular automata models. (Batty 2007) We follow this trend of disaggregation in our approach, described below.

Arguing that migration literature is biased towards examining the determinants of migration, Greenwood cites migration deterrents like psychological costs. Further arguing against modeling migration 'enablers', Greenwood (1975) writes that job opportunities can be unreliable indicators of impending migration, as high unemployment rates can be a push factor.

Apart from the neoclassical model and additional parameters, Microdata (decisions of individual people) have been used for modeling household choices to migrate over time. (Greenwood 1985). A Markov chain with constraints like a temporally-constant transition probabilities, and that probability for i to j transition is based only on i, has been tested. (Joseph 1975) Furthermore, Network models include Sassen's (1995) 'dominant node concept' of prestige (Limtanakool 2007, Alderson and Beckfield 2004), while node centrality has been used as an indicator of growth, and technological innovation. (Freeman 1977)

We note the important issue of scope and geographic granularity in a study of origin/destination flows. Some migration reports give an overgeneralized picture of U.S. migration, which abstracts confidence intervals for specific cities, and can lead to confusion. For example, in Mississippi, 62.5 thousand out-migrants indicates that the state exhibits a heavy 'push', and so it confused modelers when 51.9 thousand inflow to Mississippi counter-intuitively indicated a strong pull power. (Sjaastad 1962) Others report at a general level: "Migration to Florida was often from the Northeast and Midwest." (Perry 2003) Bearing this in mind, we use the U.S. Metropolitan Area as the geographic origin and destination for our migration flow system.

Our Approach

Our approach tests for stationarity by evaluating the temporal variation of multiple *a priori* migration predictors. We have two types of predictors: transit (airline flights and trips) and gravity model predictors, and transition probability predictors. Types of transition probability predictors include: system-wide city-to-city transition probability, and two Bayesian elements, Conditional and Posterior Probabilities.

We perform two operations with the two sets of predictors. For all tests, we set aside a set of "future" migration years (2005-2008) so we can check our predictions. First we correlate the expected number of migrants, with actual number of migrants to gauge the strength of relationships for 1996-2004. Next, we try to predict "future" migration values for each variable, using linear extrapolation. Finally, since our hypothesis focuses on the Bayesian probabilities, we compute the 10 year correlation between variables and address the stationarity of system variables.

Next, we discuss data sources for our experiments.

Data Sources

We use three different tabular datasets for this experiment, and one spatial dataset containing metropolitan areas. Our analysis is limited to the 48 contiguous states in the United States to prevent the skewing of gravity model distance parameters due to anomalous distances.

A. *Migration Data*

The U.S. Internal Revenue Service (IRS) migration data set is based on the year-to-year changes of address shown on the tax returns contained in the IRS Individual Master File system. (U.S. IRS 2008) Years 1995-1996 through 2007-2008 (Fiscal Year) are available. The data summarize migration patterns by U.S. County, including inflows and outflows. We use the number of Personal Exemptions, where head(s) of household files dependents (usually spouses or children). This values better approximates the residential migrant population, than individual number of tax returns. From these data, we compile a set of asymmetrical, stochastic matrices that will be the basis of our research.

B. *U.S. Metropolitan Airline Flights*

We created our flight data set from the Federal Aviation Administration, which is provided by the U.S. Department of Transportation. (2010) We use data from the earliest available year, 1997, until 2004, the last year of our experiment's "known" datasets. Each value in the flight dataset represents the average number of daily flights for airport pairs. Since year is divided by financial quarter (usually 4 quarters), we average the daily passengers for the 4 quarters, and assign this value to the year. The data's spatial reference is the city name where each major airport is located. Some text mining and changes were needed to allow for these names to match those in the GIS.

C. *Travel Dataset*

The travel dataset was accessed from American Travel Survey (ATS) Survey. (U.S. Dept. of Transportation 1995) Participants were asked to report the number of trips, and destination of trips of 75 miles or more over a two week period. Data is recorded for each instance of individual travel of at least 100 miles away from home and returns. Data is available for 1995.

D. *Spatial Reference*

We define a 'city' as a Consolidated Metropolitan Statistical Areas (CMSA). (U.S. Office of Management and Budget 2000) A CMSA is defined by the Office of Management and Budget as an area "containing a recognized population nucleus and adjacent communities that have a high degree of integration with that nucleus." The flight dataset includes of 97 of these CMSAs, and trip dataset includes 144 CMSAs as shown in Figure 1.

Since the migration data is at the U.S. County Level, grouping movement by metropolitan area naturally omits data (from more rural areas). The resulting number of migrants ranges from a lot of 4.01 million (found in 1996) to a high of 4.97 million, found in 2006. (Table 1)

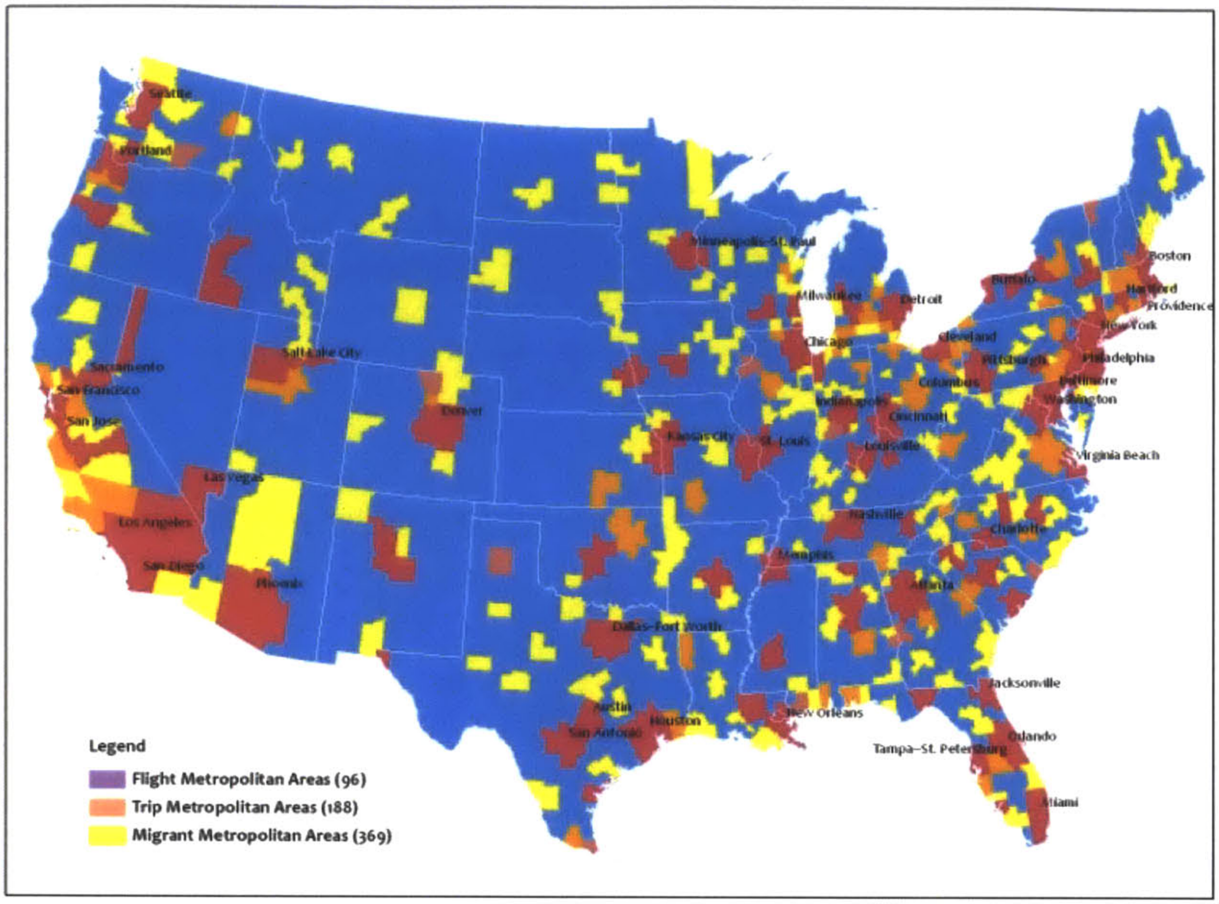


Figure 4.1: Metropolitan Statistical Area Boundaries and Available Data for the Contiguous U.S. **Combined Metropolitan Statistical Area boundaries are coloured to show flight cities in pink, trip cities in orange and cities for migration in yellow.**

TABLE I
METROPOLITAN AREA SYSTEM MIGRANTS

Year	System Migrants	Year	System Migrants
1996	4,012,295	2003	4,296,067
1997	4,105,669	2004	4,389,689
1998	4,151,370	2005	4,612,292
1999	4,117,659	2006	4,970,521
2000	4,312,329	2007	4,712,496
2001	4,387,985	2008	4,681,094
2002	4,398,891		

COMPUTATIONAL METHODS AND RESULTANT DATA

A. *Gravity Model Parameterization*

In this section we outline two methods: the traditional gravity model and the place-pair probability approach. We consider the gravity model to be a standard calculation for geographic flows, and thus our null hypothesis is that the Bayesian approach will not improve on the predictive accuracy of the gravity model. The Bayesian approach for place-pair probabilities is more experimental, and therefore can be considered as the alternative approach we test here. We do two things with the gravity model. First, we ask how has the friction of distance changed over time, as first mentioned in the introduction. To perform this calculation, we calibrate the beta coefficient for the gravity model over the 13 year span.

The gravity model used to estimate migration between points A and B is given as:

$$G_{AB} = \frac{P_A^{\alpha} P_B^{\gamma}}{D_{AB}^{\beta}} \quad (1)$$

where G_{AB} is the expected number of migrants from A to B, P_A and P_B are the populations of A and B, and D_{AB} is the relative distance between the two points of interest. The powers α, β, γ are unit-less parameters solved for using a non-linear least squares algorithm.

To estimate the best-fit values of the powers in the gravity model, a nonlinear least squares fitting routine is employed. The Levenberg-Marquardt method, defined by Levenberg (1944) and Marquardt (1963) is an efficient fitting algorithm for finding local minima in nonlinear χ^2 spaces. We utilize a useful Interactive Data Language implementation of this technique, MPFIT [that requires a user-defined function (of arbitrary form), accepts trial values for the parameters being solved for (α, β, γ), evaluates the model G_{AB} at those values, and returns the corresponding residuals to the data series. MPFIT then uses the information on the model residuals to step through parameter space and locate the minimum in χ^2using a combination of Newton's method and the method of steepest-descent.

B. *Bayesian Model*

To calibrate the Bayesian model, we use data from 1996 until 2004. For this model, we eliminate any city pairs that do not post inter-city migrant flows for each of the 13 years from 2006 to 2008 (inclusive). This process cut out about 3 million migrants and left nearly 49 million migrants for the analysis.

The system-wide probability of transition from location A to location B for any given year is given by,

$$SYS_{AB} = P(M_{AB})_{SYS} = \frac{M_{AB}}{{}^{N_A,N_B}_{i \neq j}(M_{A_iB_j})} \tag{2}$$

where A and B are particular origins and destinations of interest, respectively, and $N_A, N_B$are the total number of origin/destination points.

There are four major parts to the probabilities of a Bayesian model. In this framework we have Prior, Marginal, Conditional and Posterior probabilities. The first two can be calculated simply: The **prior** probability is the chance that any system migrant hails from city B. On the same lines, the **marginal** probability is the chance that any system migrant moves to city A. The **system-wide** probability of this move is the chance that this specific, directed OD migration is chosen out of all system migrations.

The following sets of probabilities concentrate on the individual place-pair probabilities of directed connections. We consider the **conditional** probabilities and the **posterior** probabilities of moving from city A to city B. Our Bayesian model gives the Conditional Probability (COND) as: Given that a migrant's "Origin is City A", what is the probability that her "Destination is city B"? The Posterior Probability (POST) are the chance that: Given that a migrant's "Destination is city B"(D_b), what is the probability that his "Origin is City A" (O_a)? Written, these relationships can be calculated by:

$$COND_{AB} = P(O_a|D_b) = \frac{P\,(D_b|O_a)\,P(O_a)}{P(D_b)} \tag{3}$$

$$POST_{AB} = P\,(D_b|O_a) = \frac{P(O_a|D_b)\,P(D_b)}{P(O_a)} \tag{4}$$

C. *Computational Approach*

1) Transit and Gravity Model Expected and Actual Migration Values Correlation: We first correlate Reported Trips (Tr), Airline Flights (F) and Expected Migration values (G) as calibrated by the Gravity Model with Pearson's R^2 Correlation Coefficient.

The Pearson Correlation Coefficient (R^2) between two data sets a and b is defined as:

$$R^2{}_{ab} = corr\ \ a, b\ \ = \frac{a-a \quad b-b}{\sigma_a\sigma_b} \tag{5}$$

where a and b are the mean values of the corresponding data sets, with standard deviations $\sigma_a$and σ_b.

The correlation between travel and gravity model variables, and migration volume is given for flights $\boldsymbol{F}$, trips $\boldsymbol{Tr}$, and Gravity Model-calibrated Values $\boldsymbol{G}$, with migrants $\boldsymbol{M}$ as $R^2{}_{TrM}$, $R^2{}_{FM}$,and $R^2{}_{GM}$, respectively.

We report these findings individually for our 'known years' (1996-2004) to show the relationship between human visits (flights and trips) and the widely-used gravity model. As mentioned previously, we hypothesize that these variables should show concordance with migration.

2) *Individual Probability Temporal Correlation:* In this section, we assess the correlation of each system-wide and Bayesian component probability value, with the variable value for the same origin/destination pair, at a one year time step. For all $1997 \leq t \leq 2004$, the correlation values for each variable is given as:

$$\begin{gathered} R^2_{SYS\ t\ ,SYS\ t-1} \\ R^2_{COND\ t\ ,COND\ t-1} \\ R^2_{POST\ t\ ,POST\ t-1} \end{gathered} \tag{6}$$

3) *2005- 2008 Time Series Prediction:* Next, we test to see how well our variables can predict future migration volumes. The model is fed data from the next year (distance stays the same, population changes), and the number of migrants it predicts is compared to the number of migrants expected (from the previous year). The difference between these two outcomes is tested with a chi squared test of categorical distribution difference.

Expected migration volume at time *t* for t = 2005-2008 is estimated by extrapolating actual migration values t:

$$E\ M\ _t = a_T\ M\ _t + b_T \tag{7}$$

where a_T and b_T are linear regression parameters from 'known' time values of T for $2004 \geq T \geq 1996$.

Further, values for of expected migration volume between cities A and B ($E\ M_{AB\ t}$) at time *t*, for years 2005-2008, are estimated by:

$$E\ M_{AB\ t} = a_T\ M_{AB\ t} + b_T \tag{8}$$

4) 10 Year Prediction using Individual Probabilities: Our final test using system-wide, conditional and posterior probability values to predict 'hidden' mirrored values in 2006, 2007 and 2008. We perform an R^2 correlation test to assess the stability of these three properties over a ten year lapse. For each value of *t* in {1996, 1997, 1998} and, respectively, *t'* {2006, 2007, 2008}, we compute:

$$\begin{gathered} R^2_{SYS\ t\ ,SYS\ t'} \\ R^2_{COND\ \ t\ ,COND\ \ t'} \\ R^2_{POST\ \ t\ ,POST\ \ t'} \end{gathered} \tag{9}$$

The results of the aforementioned four types of tests are described in the next section.

TABLE II
CALIBRATED GRAVITY MODEL PARAMETERS

Year	Alpha	Gamma	Beta	Linear
1996	*0.4234*	*0.5217*	*1.1982*	*1.1867*
1997	0.4308	0.5185	1.2053	1.1972
1998	0.4355	0.5151	1.2062	1.2077
1999	0.4251	0.5237	1.1926	1.2182
2000	0.4324	0.5310	1.2306	1.2287
2001	0.4322	0.5312	1.2253	1.2392
2002	0.4266	0.5491	1.2563	1.2497
2003	0.4234	0.5571	1.2697	1.2602
2004	0.4208	0.5636	1.2741	1.2707
2005	0.4172	0.5623	1.2538	1.2812
2006	0.4173	0.5548	1.2207	1.2917
2007	0.4190	0.5414	1.1996	1.3022
2008	0.4407	0.5168	1.1960	1.3127

RESULTS

A. Trip and Migration Correlation

The reported trips did not show a clear correlation with the migration dataset. We find the low correlation (a mere 1.6%) curious. This dataset contains information (for example) that shows that the top 5 destinations (trips) for Bostonians in a two week period are: New York, NY (11603), Worcester, MA (5077), New Haven, CT (2587), Bridgeport, CT (2356) and Washington D.C. (1767). Surprisingly, these data do not correlate with migration patterns. However, upon further research, we could find a 24.8% correlation with a power-law. Although our trend is subject to overfitting, we find the best model to be $y = 1.8535x^{0.6529}$.

B. Flight and Gravity Model Correlation with Migration

The predicted migration volume based on airline flights and gravity model (eq. 2) calculations each had, on average, a .561 r^2 correlation, (eq. 5) revealing that each has similar relationship with actual migration data.

TABLE III
AIRLINE FLIGHTS AND GRAVITY MODEL CORRELATION COEFFICIENTS

Year	*Airline Flights*	*Gravity Model*
1996	--	0.561
1997	0.526	0.572
1998	0.545	0.576
1999	0.571	0.567
2000	0.539	0.577
2001	0.523	0.572
2002	0.605	0.556
2003	0.607	0.541
2004	0.571	0.528
Average	0.561	0.561

C. Bayesian Elements

Since we are not comparing two different datasets in the Bayesian Elements model, we use the R^2 tests in (eq. 6) to compute correlation between one-year time-step variable differences for system (eq. 2), conditional (eq. 3) and posterior (eq. 4) probabilities. Table 3 shows that the system variables have a very high correlation with the system probabilities slightly more predictive of future years than conditional or posterior probabilities.

TABLE IV
CORRELATION COEFFICIENTS FOR TRANSITION PROBABILITIES AT A ONE-YEAR TIME STEP

Year	***System***	***Conditional***	***Posterior***
1997	0.992	0.934	0.985
1998	0.994	0.99998	0.984
1999	0.986	0.99995	0.971
2000	0.986	0.99997	0.971
2001	0.993	0.971	0.988
2002	0.981	0.986	0.986
2003	0.991	0.985	0.989
2004	0.993	0.988	0.988
Average	**0.989**	**0.983**	**0.983**

D. Future Predictions

We find that predictive power for extrapolated years (eq. 7, 8) is strongest for the conditional and posterior probabilities, followed by a linear trend and the system probabilities. The strength of linear trend is indicative of the general predictability of the dataset. These results show that there is stability in the relationship of unique city pairs over time.

TABLE V
CORRELATION COEFFICIENTS FOR EXPECTED AND ACTUAL MIGRATION DYNAMICS (LINEAR EXTRAPOLATION)

Year	***Flight***	***Gravity***	***Linear***
2005	0.540	0.524	0.982
2006	0.434	0.500	0.889
2007	0.465	0.525	0.930
2008	0.518	0.575	0.883
Average	***0.489***	***0.531***	***0.921***

Year	***System***	***Conditional***	***Posterior***
2005	*0.977*	*0.979*	*0.986*
2006	*0.878*	*0.957*	*0.979*
2007	*0.903*	*0.956*	*0.973*
2008	*0.824*	*0.954*	*0.967*
Average	***0.895***	***0.962***	***0.976***

E. Ten-Year Predictions and Stability

The ten year predictions (eq. 9) favor the Conditional and Posterior probabilities over that of the entire system. This indicates that the probability of moving from city A to city B out of any moves in the system will change over time, but the probability of moving from A to B, using Bayes' Law, which accounts for prior and marginal probabilities, stays similar over time. (e.g. Table 5) Averages and standard deviations of actual values for all years are plotted in Figure 2.

TABLE VI
CORRELATION COEFFICIENTS FOR EXPECTED AND ACTUAL MIGRATION DYNAMICS (10 YEAR PROJECTION)

Year	***System***	***Conditional***	***Posterior***
2006	0.849	0.962	0.966

2007	0.906	0.954	0.964
2008	0.947	0.956	0.969
Average	***0.900***	***0.958***	***0.967***

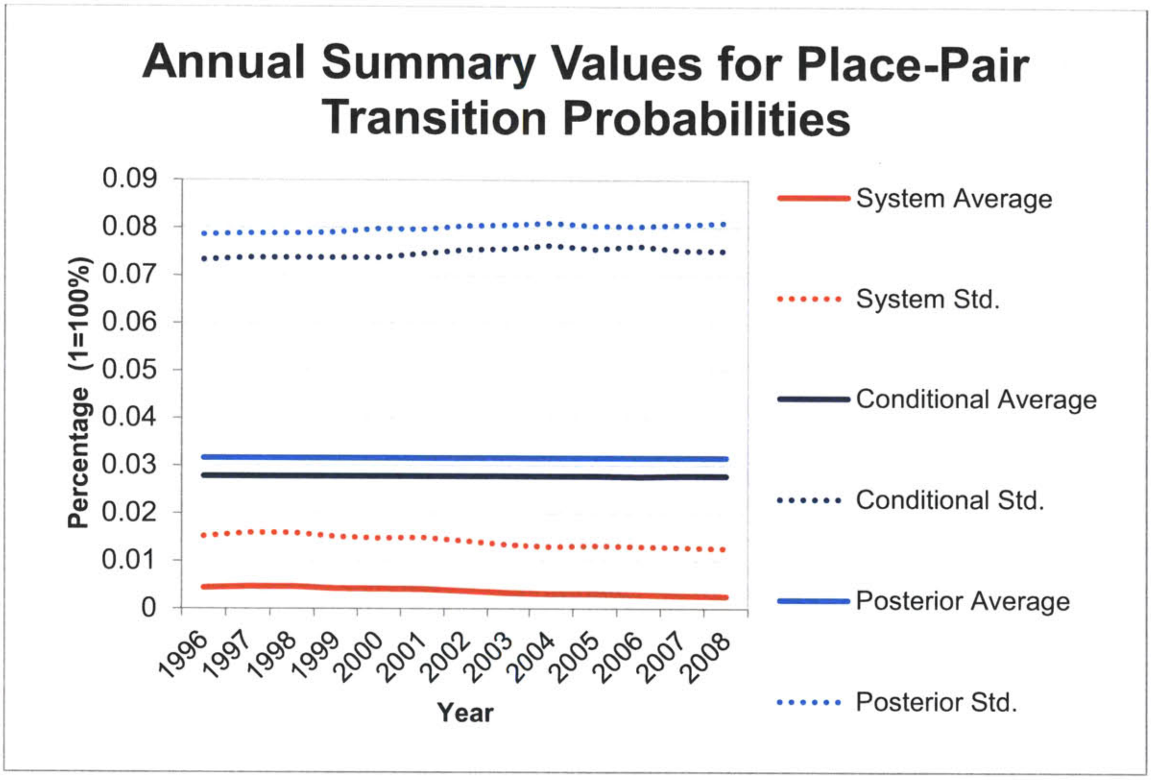


Figure 4.2: Annual Summary Values for Place-Pair Transition Probabilities **Annual summary variables plotted over time show the stability of the conditional and posterior averages.**

TABLE VII

TEMPORAL TRANSITION PROBABILITY AVERAGES AND STANDARD DEVIATIONS

	System		*Conditional*		*Posterior*	
Year	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
1996	0.004365	0.0152	0.027776	0.0733	0.03165	0.0786
1997	0.004633	0.0159	0.027777	0.0738	0.03165	0.0789
1998	0.004637	0.0159	0.027777	0.0738	0.03165	0.0789
1999	0.004256	0.0151	0.027777	0.0738	0.03165	0.0791
2000	0.004163	0.0148	0.027777	0.0738	0.03165	0.0798
2001	0.004127	0.0149	0.027774	0.0746	0.03164	0.0797

2002	0.003713	0.0142	0.027779	0.0754	0.03165	0.0804
2003	0.003297	0.0134	0.027778	0.0756	0.03165	0.0806
2004	0.003081	0.0130	0.027776	0.0763	0.03165	0.0810
2005	0.003103	0.0132	0.027779	0.0755	0.03164	0.0804
2006	0.002918	0.0130	0.027689	0.0761	0.03165	0.0803
2007	0.002733	0.0128	0.027780	0.0752	0.03164	0.0806
2008	0.002547	0.0126	0.027778	0.0751	0.03165	0.0810

To put these findings into context, let us examine a few extreme cases for posterior probabilities.

For example, all migrants to Parkersburg-Marietta WV/OH hail from Columbus OH. Similarly all migrants to Williamsport, PA come from the Philadelphia-Camden-Wilmington, PA-NJ-DE-MD area. Although this could seem unlikely, if one considers low migration numbers, like 10 people—the minimum—it seems more plausible that one small city (like Parkersburg or Williamsport) draws its entire in-migration population from a nearby metropolis (like Columbus or Philadelphia). On the opposite extreme, we can see places that draw very few of its in-migrants from a specific city. For example, migrants from Midland, TX only make up a 10- thousandth of .0001 (.01 per cent) of all migrants to Los Angeles-Long Beach-Santa Ana, CA.

We find that these posterior probabilities are the most reliable over time, followed closely by conditional probabilities.

CONCLUSIONS

Although literature on migration cites interpersonal relationship 'chains' as a major component of predicting migration, models tend to predict migration with income differentials and city pair distance. We test a traditional gravity model, transit predictors, and linear forecasts against a method that isolates a city-pair's unique relationship: system-wide probabilities and Bayesian elements. We use a case study of U.S. metropolitan system migration to predict annual migration for years 2005-2008, based on dynamics from years 1996-2004. We find that instead of using transit flows, population, Euclidean distance, or even migration system probabilities to predict future migration patterns, we can best model and predict migration by harnessing city-pair probabilities, as manifested by unique city-to-city conditional and posterior probabilities of moving. Otherwise stated, our results indicate that the most stable properties of a large migration system are probabilities that overtly give place-pair migrant probabilities for all combinations of cities, normalized by prior and marginal probabilities, as in Bayes' Law. This finding supports modern theory on the importance of social, personal communication as a prominent factor in international and domestic migration, as mentioned in the review of migration theories. With more precise predictive models, we may be able project further into the future, and make better informed decisions for planning.

ACKNOWLEDGEMENT

The author acknowledges Frank Hardisty for supervision, and Sam Halverson for helping calibrate the gravity model. This work was presented by Diansheng Guo at the IEEE 1st International Conference on Spatial Data Mining in July 2011. We thank the National Defense Science and Engineering Graduate Fellowship Program, The MIT Sense*able* City Lab, and The Penn State GeoVISTA Center for their support on this research. A special thank you to David Lee for assisting with data transfer.

REFERENCES

[1] Isard, W. Methods of regional analysis: An introduction to Regional Science *The MIT Press, Cambridge, Massachusetts,* **1960**

[2] Limtanakool N.; Dijst, M. & Schwanen, T. A theoretical framework and methodology for characterising national urban systems on the basis of flows of people: empirical evidence for France and Germany *Urban Studies,* **2007,** *44*

[3] Fawcett, J. Networks, linkages, and migration systems *International Migration Review*, **1989**, *23*, 671—680

[4] Tarver, J. & McLeod, R. A test and modification of Zipf's hypothesis for predicting interstate migration *Demography*, **1973**, *10*, 259—275

[5] Hansen, M. Atlantic migration 1607-1860, *Harvard University Press, Cambridge, Massachusetts* **1940**

[6] Gottlieb, P. Making their own way: Southern blacks' migration to Pittsburgh, 1916-30 *University of Illinois Press, Urbana-Champaign* **1997**

[7] Rojas, F. New York Talk Exchange: Transnational telecommunications in a globalizing *Unpublished Doctoral Dissertation, Massachusetts Institute of Technology, Dept. of Urban Studies and Planning* **2010**

[8] Castells, M. Communication power *Oxford University Press, Oxford,* **2009**

[9] Carrington W.; Detragiache, E. & Vishwanath, T. Migration with endogenous moving costs *The American Economic Review,* **1996,** *86*, 909—930

[10] Scott, E. Negro migration during the war *Oxford University Press, Oxford,* **1920**

[11] Grossman, J. Land of hope: Chicago, Black southerners, and the great migration *University of Chicago Press, Chicago* **1991**

[12] Massey, D.; Alarcón, R.; Durand, J. & Gonzalez, H. Return to Aztlan: The social process of international migration from western Mexico *Univ. of California Press,* **1990**

[13] Marks, C. Farewell--we're good and gone: the great Black migration *Indiana University Press Bloomington, IN,* **1989**

[14] Rogers, A. A regression analysis of interregional migration in California *The Review of Economics and Statistics,* **1967,** *49*, 262—267

[15] Yezer A. & Thurston, L. Migration patterns and income change: implications for the human capital approach to migration *Southern Economic Journal,* **1976,** *42*, 693—702

[16] Treyz, G.; Rickman, D.; Hunt, G. & Greenwood, M. The dynamics of US internal migration *The Review of Economics and Statistics*, **1993**, *75*, 209—214

[17] Greenwood, M. & Sweetland, D. The determinants of migration between standard metropolitan statistical areas *Demography*, **1972,** *9*, 665—681

[18] Harris, J. & Todaro, M. Migration, unemployment and development: a two-sector analysis, *The American Economic Review*, **1970,** *60*, 126—142

[19] Dorigo, G. & Tobler, W. Push-pull migration laws *Annals of the Association of American Geographers*, **1983,** *73*, 1—17

[20] Stouffer, S. Intervening opportunities: a theory relating mobility and distance *American Sociological Review*, **1940,** *5*, 845—867

[21] Schachter, J. & Althaus, P. Neighborhood quality and climate as factors in US net migration patterns, 1974-76 *American Journal of Economics and Sociology* **1982**, 387—400

[22] Horiba, Y. US Interregional Migration and Trade *Murphy Institute Conference on the Political Economy of Migration, Murphy Institute of Political Economy, Tulane University,* **2000**

[23] Lee, E. Migration in relation to education, intellect, and social structure *Population Index, Office of Population Research, Princeton University,* **1970**, 437—444

[24] Mangalam, J. & Schwarzweller, H. Some theoretical guidelines toward a sociology of migration *International Migration Review*, **1970**, *4*, 5—21

[25] Franklin, R. Migration of the young, single, and college educated: 1995 to 2000 *U.S. Census Bureau Report CENSR-12.* **2003**

[26] Perry, M. State-to-State migration flows: 1995 to 2000 *Washington, DC: US Census Bureau, US Department of Commerce. Accessed at: http://www. census. gov/population/www/cen2000/ migration.html* **2003**

[27] McNeill, W. Human migration in historical perspective *Population and Development Review,* **1984,** *10*, 1—18

[28] Pedraza, S. Women and migration: The social consequences of gender *Annual Review of Sociology,* **1991**, 303—325

[29] Sassen, S. Immigration and local labor markets *The Economic Sociology of Immigration, Russell Sage Foundation, New York,* **1995**, 87—127

[30] Greenwood, M. Human migration: theory, models, and empirical studies *Journal of Regional Science*, **1985**, *25*, 521—544

[31] Joseph, G. A Markov analysis of age/sex differences in inter-regional migration in Great Britain *Regional Studies*, **1975**, *9*, 69—78

[32] Tobler, W. Migration fields *Population*, **1978,** 215-232

[33] Reades, J.; Calabrese, F. & Ratti, C. Eigenplaces: analysing cities using the space- time structure of the mobile phone network *Environment and Planning B: Planning and Design,* **2009**, *36*, 824-836

[34] W. Tobler Migration: Ravenstein, Thornthwaite, and beyond *Urban Geography*, **1995**, 16, 327—343

[35] Ravenstein, E. The laws of migration, *Journal of the Stat. Society of London*, **1885,** *48*, 167—235

[36] Batty, M. Cities and Complexity *Cambridge: The MIT Press*, **2007**

[37] Greenwood, M. Research on internal migration in the United States: a survey *Journal of Economic Literature, American Economic Association,* **1975**, *13*, 397—433

[38] Sassen, S. Global networks, linked cities *Brunner-Routledge, London, UK,* **2002**

[39] Limtanakool, N.; Dijst, M. & Schwanen, T. A theoretical framework and methodology for characterising national urban systems on the basis of flows of people: empirical evidence for France and Germany *Urban Studies,* **2007**, *44*, 2123

[40] Alderson, A. & Beckfield, J. Power and position in the world city system *American Journal of Sociology,* **2004**, 811—851

[41] Freeman, L. A set of measures of centrality based on betweenness *Sociometry,* **1977**, *40*, 35—41

[42] Sjaastad, L. The costs and returns of human migration *The Journal of Political Economy,* **1962**, *70*, 80

[43] United States Internal Revenue Service SOI Tax Statistics - County-to-County Migration Data *Accessed Online: http://www.irs.gov/taxstats/indtaxstats/ article/0,,id=96816,00.html,* **2008**

[44] Domestic Airline Fares Consumer Report *U.S. Department of Transportation*, **2010**

[45] "Person OD" American Travel Survey (ATS) *Bureau of Transportation Statistics, U.S. Department of Transportation*, **1995**

[46] Standards for Defining Metropolitan and Micropolitan Statistical Areas. Federal Register 65:82228 – 82238, *United States Office of Management and Budget,* **2000**

[47] Marquardt, D. An algorithm for least-squares estimation of nonlinear parameters *SIAM J. on Applied Mathematics,* **1963**, *11*, 431—441

[48] Levenberg, K. A method for the solution of certain problems in least squares *Quart. Applied Math.,* **1944**, *2*, 164—168

CHAPTER 5

Social and Spatial Patterns in the U.S. House of Representatives

ABSTRACT

We create a set of methods and an implementation of these methods for integration with visual representations of a social network and its associated geographic map. Each agent in the social network has a corresponding location entity on the map that can be highlighted and analyzed. Conversely each location entity can be highlighted and analyzed in terms of the social agents associated with it. The case example we have chosen to illustrate is the friendship and agreement behaviors of elected representatives to the United States House of Representatives for the 111th Congress with one another and the respective political parties. From this character analysis, we can draw ties between certain U.S. cities, regions and demographic groups, and uncover the cities and socio-demographic constituencies that help characterize social groupings. We suggest that many problems in characterizing social behavior would benefit from linked analysis of social agents to spatial entities.

Keywords: Social Network, Spatial Analysis, U.S. House of Representatives, Politics, Congress, Roll Call Votes

INTRODUCTION

There have recently been a number of articles published on the topic of social network analysis, but relatively few account for underlying geographies, interaction spaces, distance and cost barriers that guide social space time coincidences (Diakonis and Mosteller 2006) as even in the digital age, face-to-face contact seems to be what keeps humans close to one another. (Goldenberg and Levy 2009, Scellato et al 2010) Noticing this lack of connection between geographic topology and social relationships, some have called for more attention to fusing spatial and social systems (Loglisci et al 2010, De Montis et al 2010, Xu and Harriss 2008)

At the extreme of this movement, some have argued that there is virtually no social and spatial network research currently being conducted, (Onnela, Arbesman et al 2010) and respond with an early experiment in social and spatial convergence by comparing the betweenness centrality of a node within its social network, to its geographic centrally in physical space. This type of agenda should be perused further, but with deeper consideration of geographic space. One major drawback of these social network-to-geo-network comparisons (as in Onnela, Arbesman et al 2010), is that a network poorly represents geographic space. (Loglisci et al 2010) Geographic is a continuous plane, and plucking a few entities from this plane to represent with nodes can leave out important topological system elements: for example, magnitude of spatial adjacency, area, and boundaries cannot be represented well. Another problem with this type of node-edge configuration is the use of Euclidean distance to represent edge lengths, as this type of distance does not account for intricacies in the physical and built environment that skew straight-line calculations of accessibility from place to place. Traditional social network analyses use Euclidean distance as the metric of choice to show geographic dispersion, but advancements in the social/spatial agenda should account for cost distance as well.

One early experiment models gang fraternization as a social network graph, overlays this graph on areal space, and modularizes their activity space based on pockets of friend and foe affiliations. (Radil et al 2010) This research is a successful early example of efforts to combine social and spatial, although it falls short by disregarding the computational aspects of social network analysis.

Bearing in mind that many data types must be coupled for data mining, we use a multi-relational model (Dzeroski and Lavrac 2001, Malerba 2008) model with a dynamic (Shneiderman 1994) network representation loosely coupled with a cartographic representation of demographic variables and a number of statistical tools, as found in the exploratory spatial data mining (ESDA) and geospatial visual data mining communities.

CASE STUDY

Noting a lack of effective methods for simultaneously analyzing dynamic social networks with orthorectified geographic representations, as packages in the research community like UCINet (Borgatti et al 2002) and PajeK (Batagelj and Mrvar 1998) do not offer this functionality, we couple these two systems with the ArcGIS Geographic Information Systems environment for the interactive exploration of dynamic networks as they are linked with political maps. We illustrate the uses of this computing environment with an example analysis which uncovers hidden patterns in the spatial, social and temporal aspects of the U.S. Legislative system, via a social network of Congressional Roll Call Votes in the U.S. House of Representatives.

Roll call votes have been analyzed for statistical patterns in previous studies. (Clinton et al 2004) In two different studies, researchers employ network science techniques such as hierarchical clustering and modularity to model the system of House committees and subcommittees, in terms of shared members and in terms of party majorities on teach committee (Porter et al 2005, 2007)

We use roll call vote patterns from the House of Representatives during the first session of the 111th United States Congress, which took place from January 6th, 2009 until December 24th, 2009. Speaker Nancy Pelosi (D) was Speaker of the House and President Barack Obama promulgated all enacted statutes except for a statute in early January. Some selected passed legislation includes: extended privileges and compensation to disabled veterans and spouses of military personnel, children's' insurance and adoption incentives, major weapon and tobacco reform, and a number of economic acts that address retirement, investment, foreclosures and credit cards. (U.S. Government Printing Office) We approach the U.S. Congressional Representative social network in a holistic manner, as we do not concentrate on formal groupings, but on the ideology similarities visible through similar voting patterns. Some have noted that this "informality" though seemingly unconstrained, does not show autonomy in representative decision making, but that representatives are driven by affiliated party, and the level of party alignment (as shown by voting records) has varied over time. (Lee 2003, Cox and Poole 2002)

While literature on congressional social ties seems to point to clusters of relationships in the U.S. House of Representatives, few studies have attributed these relationships to similarities in geography or in the demographic make-up of constituents. One such study showed that funding for statewide projects in congress favored populous states (with many representatives), but the authors did not mention whether a potential kinship of intra-state members was a factor in their success. (Snyder and Grose 2000) In terms of constituent demographics—an inherently topological feature—district representatives and their friendships are rarely casually linked to the similar nature of their constituent district demographics. One example links low home ownership to certain representatives. (Macrae 1952) There are certainly complex interactions that drive decision-making and relationships in the House, which represents the largest social network in the three branches of federal government; We can only imaging the social 'balancing act' between pleasing constituents, sponsoring bills, interacting with lobbyists, following party agendas, creating trust networks for communication, collaboration, sharing ideas, garnering support for initiatives, negotiating provisions and maintaining one's own sense of ethics and orthopraxy.

Harkening back to the agenda set by Porter et al (2005, 2007) we build a social network of congressional representatives based on their voting preferences, in order to analyze the holistic network behavior with new methods for complex network analysis. (Watts 2004) Our study is novel in that it brings together a social network, polygonal geographic districts and the demographic constituencies that these network actors represent—to that their decisions and behavior can be analyzed in terms of hidden correlations in friendship, geographic clustering and demographic homophily. To our knowledge, no study has coupled a social network with areal cartographic space while employing founded network science measures.

DATASET AND NETWORK

The United States House of Representatives is a group of 435 members each representing a district in the United States or U.S. Territory. The Congress follows a two-party system. In the House, there are 265 Democrats, 178 Republicans and 1 Independent. (Figure 5.1)

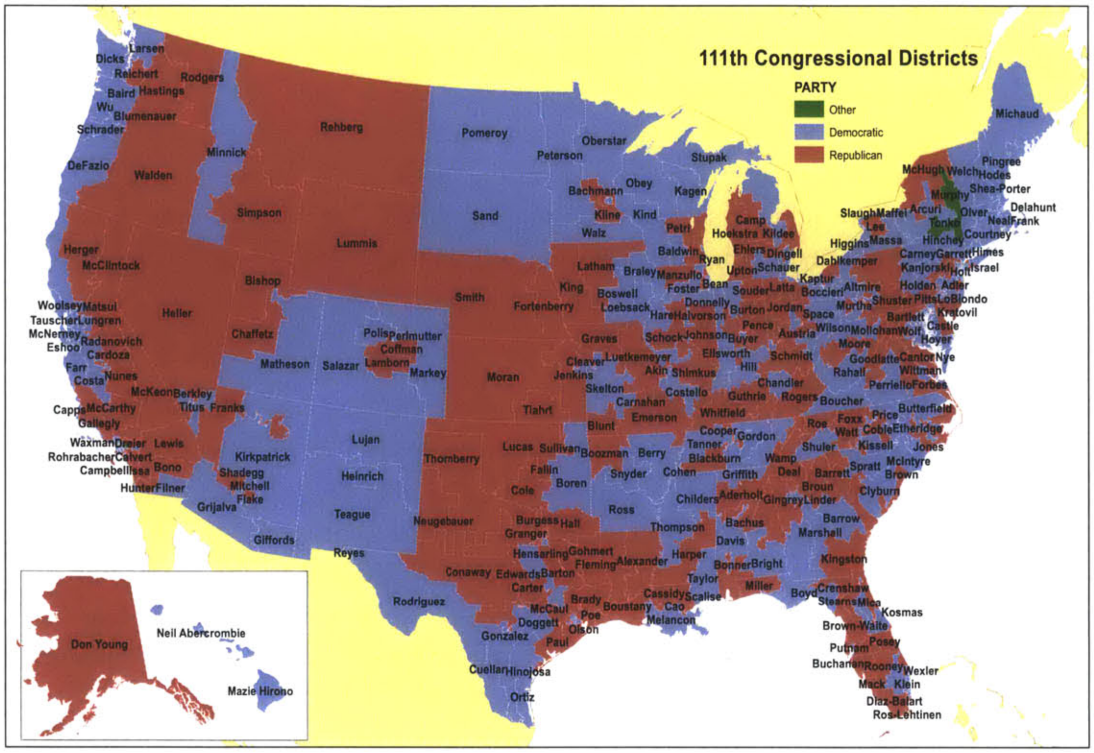


Figure 5.1: Map of U.S. Congressional Districts, Representative Names and Parties **A map of U.S. Congressional Districts, where each district is represented by the corresponding member of congress. Light blue denotes districts with a representative from the Democratic Party (D), and dark red represents districts with a Republican (R) Congressperson. One district in New York State is dark green, to symbolize that its legislator is an Independent.**

A Congressional District based in Cheyenne, Wyoming has the fewest constituencies at nearly 550,000, while the district based in Miramar, Florida has the most constituencies at 1,265,934 (2004 estimate). Between these extremes, each district has a relatively similar number of constituencies with a mean around 826,000. The number of representatives per state is pro-rated by the state's population. Accordingly, California, Texas, New York and Florida have the most representatives at 53, 32, 29 and 25, respectively. States Alaska, Delaware, Montana, North Dakota, South Dakota, Vermont and Wyoming have only one representative per state.

The geographic trends are quantified in two ways. First is the spatial position of competing or cooperating districts, their geographic adjacency, proximity and propensity to form cohesive regions. (Cressie 1992) The tightness of these district clusters are measured with test statistics like Moran's I or Geary's C for statistically significant spatial clusters of behavior groups, Hot & Cold Spot Detection, and LISA (Local Indicators of Spatial Autocorrelation). (Moran 1950, Geary 1954, Anselin 1995) In addition to measuring proximal regions, demographic (feature) clustering shows the correlation between certain constituent social features and U.S. Census information like income, urban areas, racial composition, or areal designations like economic development regions, or statewide alliances. These underlying features of a geographic district or socio-demographic constituency are correlated with relative party-loyal maverick, authority, clique, or neutral behavior of the congressperson in the wide congressional social network.

Many have noted the increasing polarization of the Republican and Democratic parties, and its effects on the congressional system. To quantify the prevalence of same-party geographic neighbors, we use Moran's I test statistic to determine the significance of bipartisan autocorrelation at different locations. Moran's I can be calculated for n observations at location i,j which represents the centroid (geometric center) of the polygonal district.

$$\boldsymbol{I} = \frac{\boldsymbol{N}}{\boldsymbol{S_0}} \frac{i \;\; j w_{ij} \;\; x_i - x \;\;\; x_j - x}{i \;\; x_i - x \;\; ^{2}} \tag{1}$$

Where x is the mean of the variable x, w_{ij} are the elements of the weight matrix and S_0 is the sum of the weights.

REPRESENTATIVE ACTIVITY

We create a network based on the number of agreements between two unique congresspersons during the first session of the 111th U.S. Congress, in 2008, with 445 total members. In the 111th Congress, 10 representatives were replaced, yielding more unique members in our network than a standard House network, which would typically have 435 members. These agents, or nodes, are not considered for individual features like gender, race, age, but are considered for features like political party (Democrat, Republican or Independent).

TABLE I
111TH CONGRESS, FIRST SESSION VOTE FREQUENCY

Number of Votes Casted for the 111th Congress	*Number of Representatives in this Voting Class*
Under 200	7
200 - 500	6
600 - 800	9
800 - 850	5
850 - 900	21
900 - 950	124
950 - 987	272

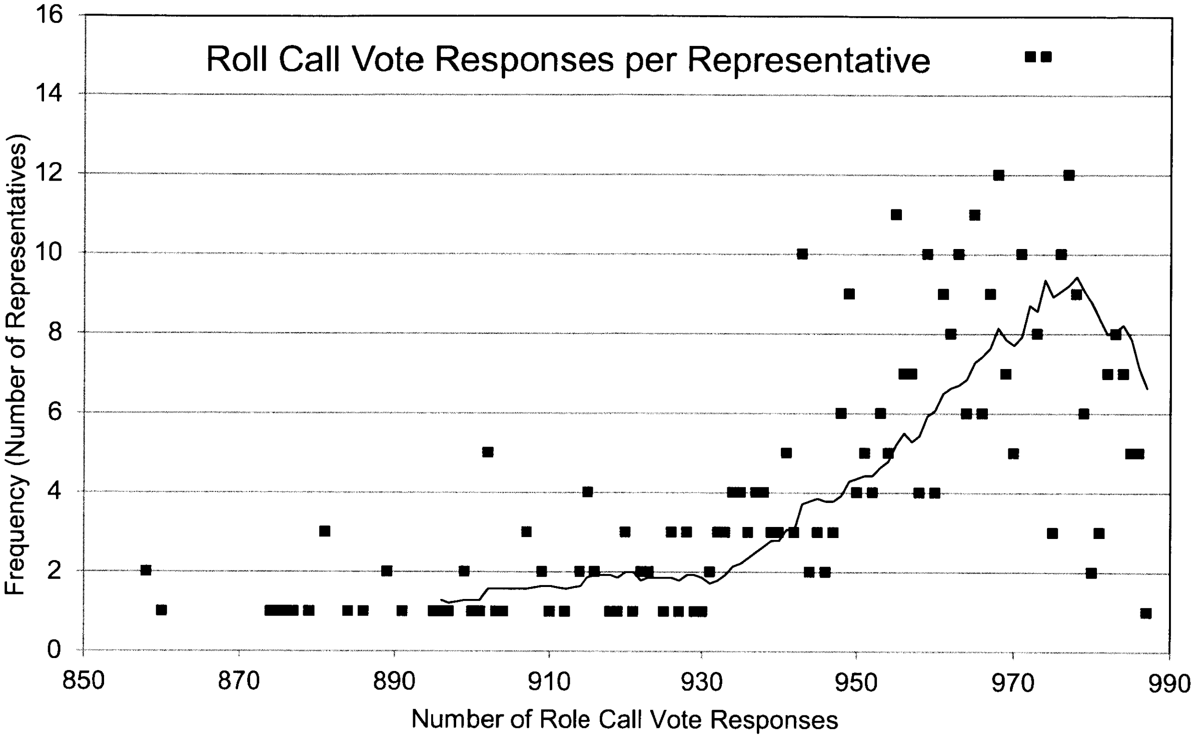


Figure 5.2: Roll Call Vote Responses per Representative **shows the distribution of representatives based on their frequency of voting. We can see that the majority of voters responded to 950 or more votes and about 400 representatives voted for at least 900 out of 985 possible roll call votes. A moving average line is drawn to guide the eye.**

NETWORK FORMATION

The quantified relationship between agents (network edges) are created by tallying 985 Roll Call votes, where an agreement between representative i and representative j is tallied when i and j each vote yay on a roll call vote, or each voting nay on a roll call vote. A disagreement is recorded if an agent votes opposite another agent (e.g. one representative is recorded as having voted Yay and the other Nay), or if one or both agents decline to vote. (The latter make up a relatively small percentage of the dataset). Typically, congressional representatives vote on over 90% of the roll call votes. (Table 1, Figure 5.2)

In terms of building the network, edges are made between node agents under the following conditions: An agreement itself does not result in an edge connection because there are almost 1000 votes (e.g. agent i has almost 1000 opportunities to agree with agent j), we quantify how many agreements a representative pair must have to be considered a friend, and thus, an edge forms between Representatives i and j. Otherwise stated, a pair of representatives would not have a weighted edge of a value x agreements between them, but only a binary 0 or 1 friendship, if enough agreements occur.

To find this threshold of "friendship" we first find the number of agreements between each set of pairs. The distribution of pair-wise agreements shows a bimodal distribution, where most frequently, voters are agreeing on roughly 800-950 bills, or 400-600 bills. (Figure 5.3)

From this distribution, we can infer that there are two distinct types of pairs: those who agree most of the time (friends), and those who disagree more often (not friends). The local minima for this distribution are 300, 660 and 985, indicating that pairs with agreements in the range of 300-659 are not friends and those with 660-985 are friends.

Since it is clear that there are pairs of people whose voting pattern and ideologies can be considered either "clashing" or "cooperating" from the distribution, henceforth, the term friendship (and friend) will be used to describe the relationships between representatives who show aligning interests. We operate on the theoretical principle that stronger ties indicate a relationships where (1) information is transferred faster, (2) ideologies are shared, (3) trust is deeper (4) opinions are heard more readily and, in sum, stronger ties may indicate that there might be a stronger ability to convince each other to vote a certain way—whether implicitly or explicitly. From this distribution we can also infer that since so few pairs fall under about 400 agreements, that around 300 roll call votes address non-controversial issues. Upon further inquiry, we find that these non-controversial issues include congratulating sports figures or motions to postpone.

This division is most probably driven by polarization of the Democratic and Republican Party affiliation in the House. We can assume that same-party pairs are in the agreement range and cross-party pairs (e.g. a Republican and a Democrat) fall into the disagreement range, even though there could certainly be exceptions.

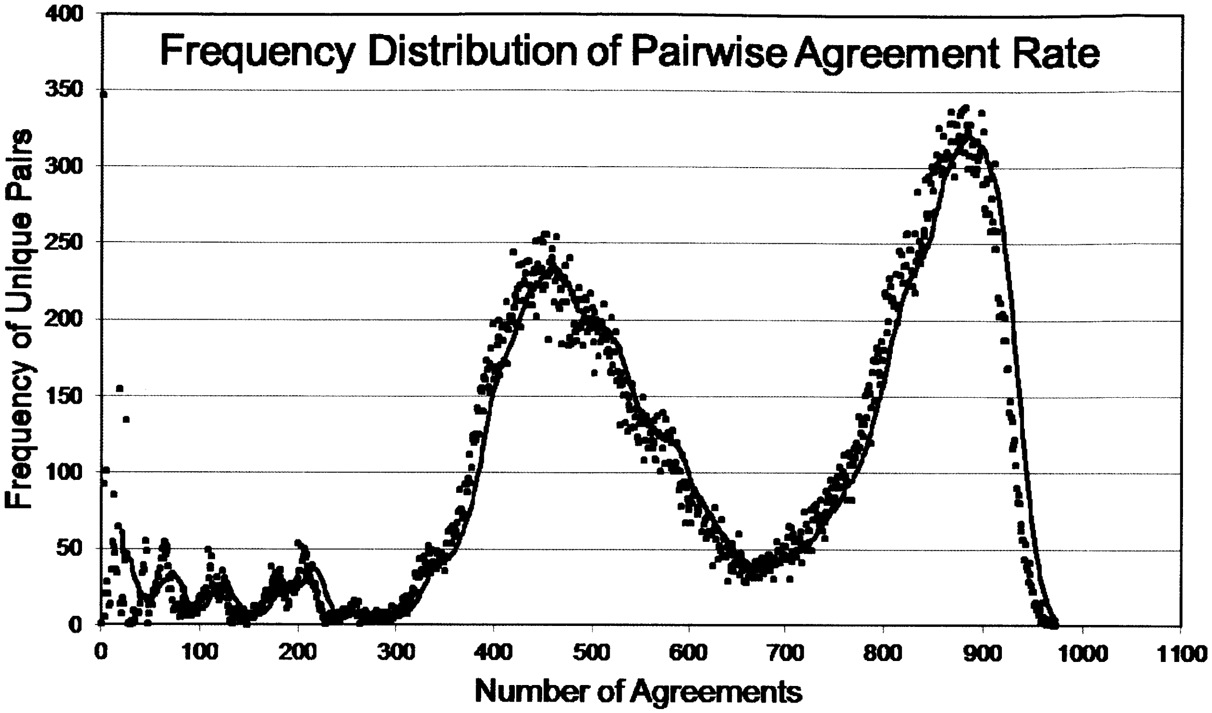


Figure 5.3: Frequency Distribution of Pair-wise Agreement Rate **The allocation of edge weight values among all node pairs depicts a bimodal distribution wherefrom we estimate that 'friends' can be considered as those with at least 660 agreements, and 'not friends' can be considered those with between 300 and 660 votes. We consider those with fewer than 300 agreements 'noisy' pairs. A moving average line is drawn to guide the eye.**

NETWORK CHARACTERISTICS

In this section, we first examine the degree distribution of the Congressional network. (Figure 5.4) The degree distribution shows a two-hump, or bimodal distribution, indicating that two major groups are uncovered. Since a friendship is considered to have more than 660 concurring votes between a pair, we see that this occurs more often for one group of representatives than another. To relate our degree distribution to Figure 5.3, the distribution of pair-wise relationships, it seems as though the representatives with more friends (the upper curve in Figure 5.4) are more likely to occur frequently in the upper curve of the distribution in Figure 5.3. Unearthing these two bimodal distributions supports the hypothesis of major two-party factions within this network.

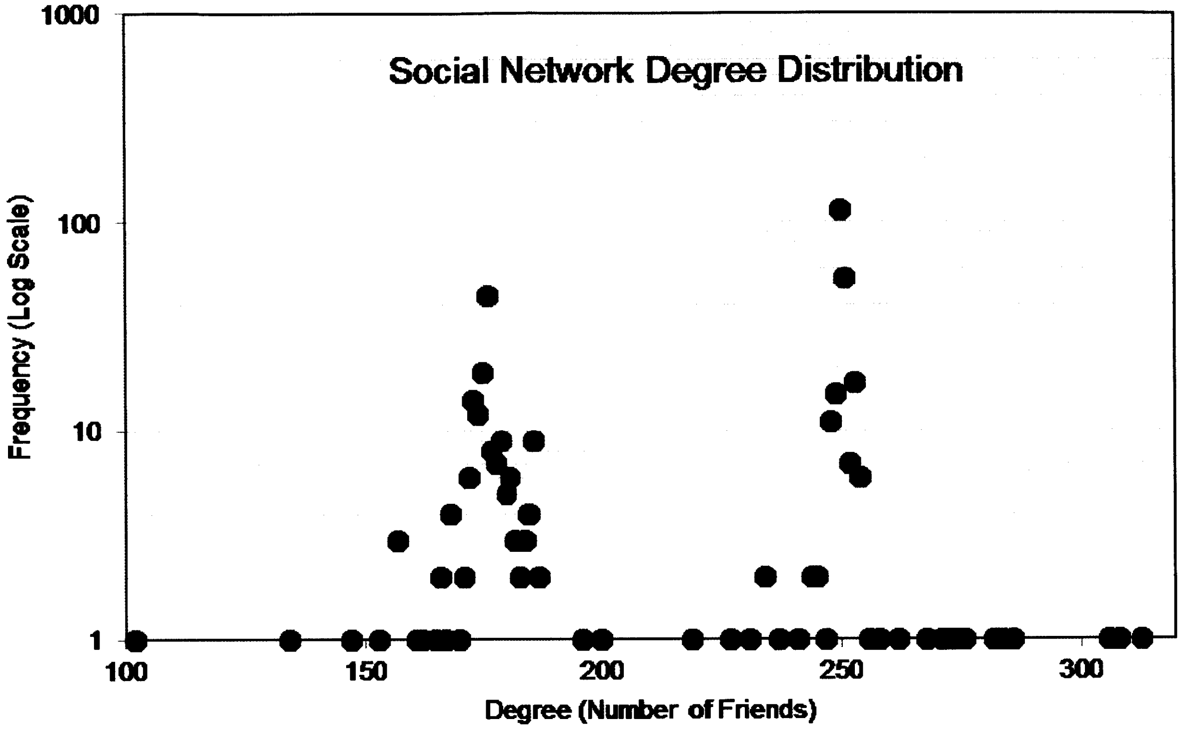


Figure 5.4: Social Network Degree Distribution **shows the degree distribution of agents in our network. We see a bimodal distribution here, indicating that a group of agents have many friends, and a group of agents have relatively few friends.**

Next, we calculate network characteristics of the Congressional network, and analyze them in comparison to an Erdos-Renyi random network, also referred to as a configuration model. (Erdos and Renyi 1959) In the Erdos-Renyi model (henceforth called ER), a predefined number of links are constructed randomly to connect a predefined number of nodes, without regard to underlying relationships or to past linkages in the linking process. Our ER model has the same number of nodes and links as our Congressional model, and can be considered a 'random' model or a 'null hypothesis' of what a network with the same descriptive statistics can be expected to look like. We compare diameter, degree, betweeness and closeness centrality measures between the ER model and the Congress network.

Our network is not directional, meaning that the relationships edge on (i,j) would be the same as the edge of pair (j,i). Since this symmetry also eliminates self-nodes, we calculated the number of pairs possible {max($_{ij}Aij$)} from the number of agents n as:

$$\frac{n\ n-1}{2} \qquad (2)$$

resulting in 98,790 possible edges. This maximum condition would yield a complete graph.

The diameter of a network is an important property for considering spreading and network traversal ease. There are two ways to calculate diameter: the "average" shortest path length, which is the expected number of hops between any two random nodes, and the maximum diameter, which indicates the longest path needed to ensure all are reached. Note that similar average and maximum diameter values can indicate the absence of asymmetric branching.

$$D = \frac{2}{N \ N-1} \qquad {}_{<i,j>} \, l_{ij} \tag{3}$$

$$D = max \ \ l_{ij} \tag{4}$$

Where l_{ij} is the distance between nodes i and j and subscript $< i, \mathrm{j} >$ indicates for all i, j pairs. (Valiente 2002)

The diameter (given by equation 4) of the Congressional network is 4, while the ER random network diameter is 2. We can interpret this finding to mean that the ER model exhibits more "small world" tendencies than the Congressional network, as any pair of nodes in the former network is connected by two hops or fewer. At first, this result seems starling, as a network constructed randomly rarely seems to conduct information more readily than a social network. (Liben-Nowell and Kleinberg 2008, and examples abound) However, due to the many edges in the network, the network is understandably tightly woven, and therefore, easily traversable with two hops. Conversely, the Congressional social network requires 4 hops to travel between two distant node pairs. This also indicates that although the network has the characteristics to be almost clique-like (as the ER model has shown), there is greater than expected separation between some node pairs. We can interpret this to mean that some nodes may branch off of others, and that traversal through these branching bridges may yield slightly higher maximum diameters. However, in perspective, this diameter is small for a social network, as the *average* diameter of a social network is found to be close to 6.6 (Watts and Strogatz 1998) which surely larger than our *maximum* of 4. This comparison of diameters is not necessarily 'fair', as the Congressional network diameter could be larger if the threshold for creating a 'friendship' was higher. Reverting to Figure 5.3, a higher threshold would likely be at the local maximum for the friendship distribution—close to 890 agreements.

Seeing that the congressional social network can be traversed relatively easily, but not as easily as expected for the high edge to node ratio, we next calculated the clustering coefficient for the group. The coefficient represents the number of potential links among a node's direct neighbors. (Johnson 1967) In other words, this statistics can be seen to represent an agent's participation in friend triangles, where two mutual friends are also friends. We find the *average* clustering coefficient CC, by calculating the clustering coefficient C_i for a single node i with degree k_i and K_i actually observed links among i's neighbors, are given as:

$$C_i = \frac{2K_i}{K_i\,(K_i - 1)} for \ (k \geq \ 2) \tag{5}$$

$$C = \ 1 \,/\, N \quad {}_{i=1}^{N} C_i \tag{6}$$

The clustering coefficient of our network is 0.96920, while the clustering coefficient of the EIR model is 0.23828. This high clustering of the network of representatives shows that the group is very tightly connected. In comparison to the ER model, the clustering coefficient of the Congress network shows that the agents in this network have many mutual friends. In fact, there are very few agents who have a friend that he does not share with another friend. The presence of triangles indicates that in most instances, a representative is well-embedded in a system, meaning that she can received news from multiple sources, but that the multiple sources close to her also have access to similar information. This circumstance means that more friends doesn't necessarily mean more robust information, since a friend's information is likely repetitive. The high clustering coefficient also indicates that "branching" occur less frequently in the network. With the triangulated topology, ideas, bills or information can spread through many paths, instead of relying on a string of agents for transmission.

Since the clustering coefficient is much higher than expected via the configuration model, yet, counter intuitively, the diameter is also higher than expected, we are faced with competing conclusions about the network: the social network is very tight, but it is not as traversable as expected. These conclusions lead us to the hypothesis that agents are tightly embedded in the system, but that the system is not a giant component and may have one or more bridges connecting tight clusters. Bearing in mind that the bimodal distributions above (Figures 5.3 and 5.4), and the nature of the U.S. political system, we can hypothesize that clusters may be formed by party affiliation.

Platform Themes

While visual data mining has proven helpful for knowledge discovery and visual pattern recognition, we also incorporate deterministic metrics into the system. The following social network measures are implemented under four major themes: *Community Detection, Popularity, Attachment* and *Spreading.*

Theme A is comprised of a group of *community detection* metrics like cliques (maximally complete subgraphs), hierarchical clusters (groups that form within larger groups) (Bron and Kerbosch 1973, Newman 2004) modularity measures (the relative propensity of a member of a partition to talk to people within the partition). (Jin et al 2001, Newman 2006)

Modularity Q is given as

$$Q = \frac{1}{4m} \quad {}_{ij} \quad A_{\mathrm{ij}} - \frac{k_i kj}{2m} \quad s_i s_j \tag{7}$$

Where m is the number of edges in the network, A_{ij} is an element of the network, $k_i$and k_j are the degrees of vertices i and j. Variables s_i and $s_j$are the groups to which i and j belong. We can determine the number of groups for the algorithm.

Girvan and Newman (2002) give an edge betweenness clustering algorithm for partitioning, which first removes edges with the highest betweenness values, then recalculates betweenness, and repeat until no edges remain. The betweenness centrality C_B of a vertex i is defined as the number of shortest paths between pairs of other vertices that run through i.

$$C_B \; n_i \;\; = \frac{{}_{j<k}\frac{g_{jk} \;\; n_i}{g_{jk}}}{|(g-1)(g-2)/2|} \tag{8}$$

Where g_{jk} is the number of geodesics (shortest paths), between nodes j, k, and $g_{jk}(n_i)$ is the number of shortest paths that travel through node i.

Theme B incorporates *popularity* measures like degree centrality (the number of nodes adjacent to a given node), betweenness centrality (as mentioned, the number of times a node occurs on all system geodesic paths), closeness centrality (the average shortest path between a node and other nodes) and hubs & authorities (agents receiving a relative advantage in information access). (Freeman et al 1991, Kleinberg 1999, Marsden 2002, Ibarra and Andrews 1993) A fundamental theme in network science, degree centrality $C_d$of node i with degree k_i can be calculated by:

$$C_d \;\; i \;\; = \frac{k_i}{n-1} \tag{9}$$

The equation for a node's measure of betweenness centrality can be adapted from equation (8), where the edge variable i is substituted to indicate node i.

Closeness Centrality,C_c, can be measured by:

$$C_c(i) = \frac{1}{{}_{t \in V/_i} d_G \;\; i,t} \tag{10}$$

Where V is total number of vertices in the graph G, t is the number of possible paths, $d_G \;\; i,t$ is the number of geodesic diameters (shortest paths), that pass through i. (Sabidussi 1966) Here, we also include brokerage roles: coordinator, representative, gatekeeper, liaison and itinerant broker. (de Nooy et al 2005)

Theme C represents **attachment behavior**, namely, homophily (how much an ego's attaches to its alters based on a specified attribute, or the correlation between ego attributes and alter attributes) and system clustering coefficients (the density of an agent's open neighborhood). (McPherson and Smith-Lovin 2001) The clustering coefficient is defined in equation (5).

Theme D models **spreading processes**, like time-step propagation, resistance to percolation, cascades and rule-based diffusion—like voter, gossip, SIS, SIR and SISR models. (Watts 2002, Cowan and Jonard 2004) Algorithms for diffusion and traversal select one or more nodes of interest, and increments edge-bound hops on to reachable nodes. These processes include measures of diameter, average diameter, and other metrics of reach.

The following section presents examples of results from each theme.

RESULTS

Overall Structure

The results of this analysis show that the network of congressional representatives is a highly partisan system with separation between the democratic and republican parties. Furthermore, each core in the two-core network contains central party members, intermediaries and bridges between the two networks, and super-loyal or extreme members. We graphically display the network with the force-directed Kamada-Kawai layout found in (Batagelj and Mrvar 1998) and can henceforth visualize the positions of the aforementioned three roles. Bridges or intermediaries are found between the two cores and near the converging area and the more 'super-loyal' members can be found at the far edges of their respective parties.

In addition to these distinctions, there are also quasi-hierarchical 'bands' of inter and intra-party friendship. The 'frontline' bands are those who interface with the other party at a direct level (e.g. they have a friend in the opposite party). The second band represents those who do not interface with the other party, or have sparse friendships. A rare third band (of which the extreme party members belong) interact mostly with second-band members of their respective parties, and so increasing the network distance between themselves and a member of the opposing party to two or three.

Theme A: Community Detection

The separation between the democratic and republican parties in network and geographic topologies is discussed under Theme A. Also in Theme A, we find that there is little significant evidence of demographic homophily, the propensity to be friends with representatives whose constituencies exhibit similar socio-economic qualities, although party factions can explain some variation.

Party Homophily

There is a definite sense of "inbreeding" with friendships in this network. (Figure 5.5) In fact, out of all connections, 27,496 are republican to republican, 58,412 are democrat to democrat, and merely 4175 connections link members of different parties. Under a model of independence, we could expect these values to be 15303, 31006 and 43773, respectively, meaning that R-R ties are 180% of expected, D-D ties are 188% of expected and D-R/R-D ties are only 10% of what is expected. In other words, representatives are 20.57 times more likely to have a 'friend' in the same party with an odds ratio of 92.142. Only two members of the House, Rep. Minnick of Idaho (D) and Rep. Bright of Alabama (D), are 'mismatched', as they fall in the same modularity group as the Republican cluster.

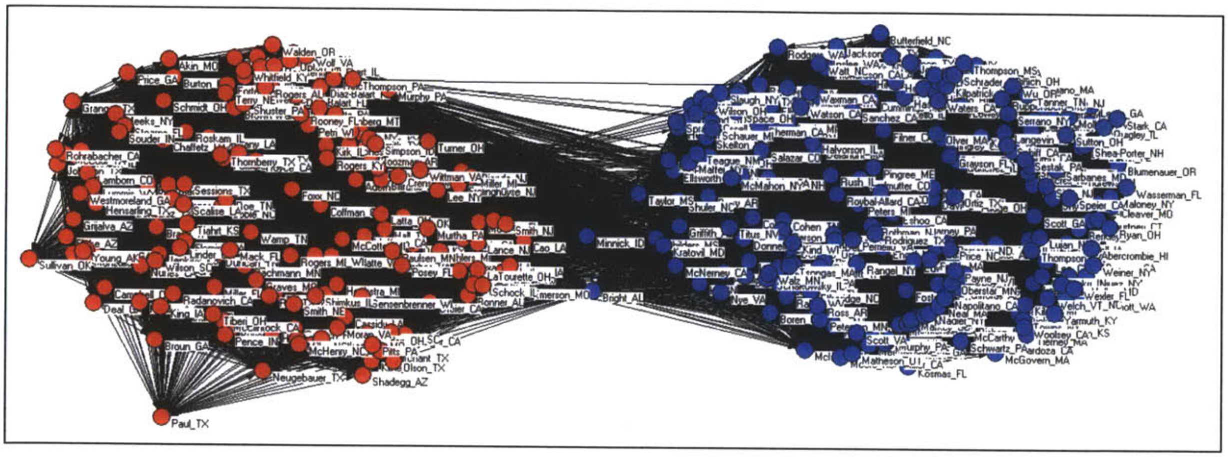


Figure 5.5: Social Network of Congresspersons Force-Directed by Binary Tie **The force-directed display of the congressional network shows two major components of Republicans (red)and Democrats (blue) with notable members between the modules and on the periphery—especially in the Republican party.**

Demographic Homophily

We next look to see if demographic variables play a role in any tie patterns. As mentioned, the most telling characteristic of friendship is the representative's party. Bearing in mind that a member is likely to attach to another member of the same party, perhaps regardless of demographic structure, we first enumerate average party demographics, in order to thwart future bias. (Table 2) The Republican Party representatives' constituencies are, on average, comprised of a higher percent of White citizens, and a lower percent of Black and Hispanic citizens. Republican-led districts are less urban than those led by Democrat representatives. There is little difference between Democrat and Republican-led constituencies in terms of the average of percent of vacant properties, median household income and percentage of constituents receiving social security, indicating that each party represents citizens of elderly status, comparable incomes and economic housing indicators.

TABLE II
DEMOGRAPHICS BY PARTY
(2000 CENSUS, BY CONGRESSIONAL DISTRICT)

	Republican	*Democrat*
Percent White	76.91	64.1
Percent Black	7.83	13.0
Percent Hispanic	7.54	11.1
Percent Urban	73.89	82.8
Percent Vacant Properties	9.51	8.5
Median Household Income (in thousands of dollars)	32.31	30.0

Percent Receiving Social Security	25.93	25.2

In addition, we look at the following variables' role in patterning friendship ties. The test used here is Moran's I statistic for spatial autocorrelation, given an adjacency matrix. Note that in this test, higher values (near 1.0) indicate independence, and lower values (0.0) indicate that a relationship is at play. Results show little evidence of demographic homophily or heterophily in the network. (Table 3)

TABLE III
DESCRIPTIVE DEMOGRAPHIC STATISTICS BY PARTY

Demographic	*Autocorrelation*	*Significance*
Percent White	0.966	0.008
Percent Black	1.018	0.219
Percent Hispanic	1.019	0.177
Percent Urban	0.945	0.037
Percent Vacant Properties	0.987	0.254
Median Household Income	1.005	0.399
Percent Receiving Social Security	0.962	0.046

Geographic patterns do not seem to be determining cause of ties. When Republican and Democrat party-led districts were tested for boundary-adjacent autocorrelation, no significant global values for same-party districts as geographic neighbors or clusters arose. Local Moran's I statistics show values that are not significant. (Table 4)

TABLE IV
LOCAL MORAN'S I SCORES

Parameter	*Republican*	*Democrat*
	Average	*Average*
Local Moran's I Index	1.509	5.812
Local Moran's I Z Score	0.982	2.119
Local Moran's I P Value	0.309	0.326

In sum, modules in the network seem to be formed mostly by party affiliation, and propensity to vote almost completely with the party or mostly within the party of membership. Next, we visit the dynamics of core membership and peripheral affiliation.

Core and Periphery

Results of a core/periphery model showed that 245 member belong to core 1, 168 members belong to core 2, and 15 members can be considered peripheral. Of these 15 members, 6 are democrats, and 9 are republicans. (Figure 5.6) Peripheral members are scattered throughout the U.S., and represent rural and urban areas alike, including large cities, Mesa (Phoenix), Chicago, Atlanta and Lakewood (Los Angeles), as well as non-metropolitan areas like Alaska and rural Georgia. Constituents of these non-core areas may benefit from considering their respective representatives' roles in the social network. Understandably, these peripheral members are also likely to be members of the 4-step transmission chains, which represent the longest hops in the network. More surprising, however, there members also show to have high clustering coefficient values, indicating that they should not be considered disconnected or marginalized in the system. Although the peripheral nodes are notably far from the center of the network's cores, they are not connected weakly but instead show some of the strongest connections and embeddedness in the system.

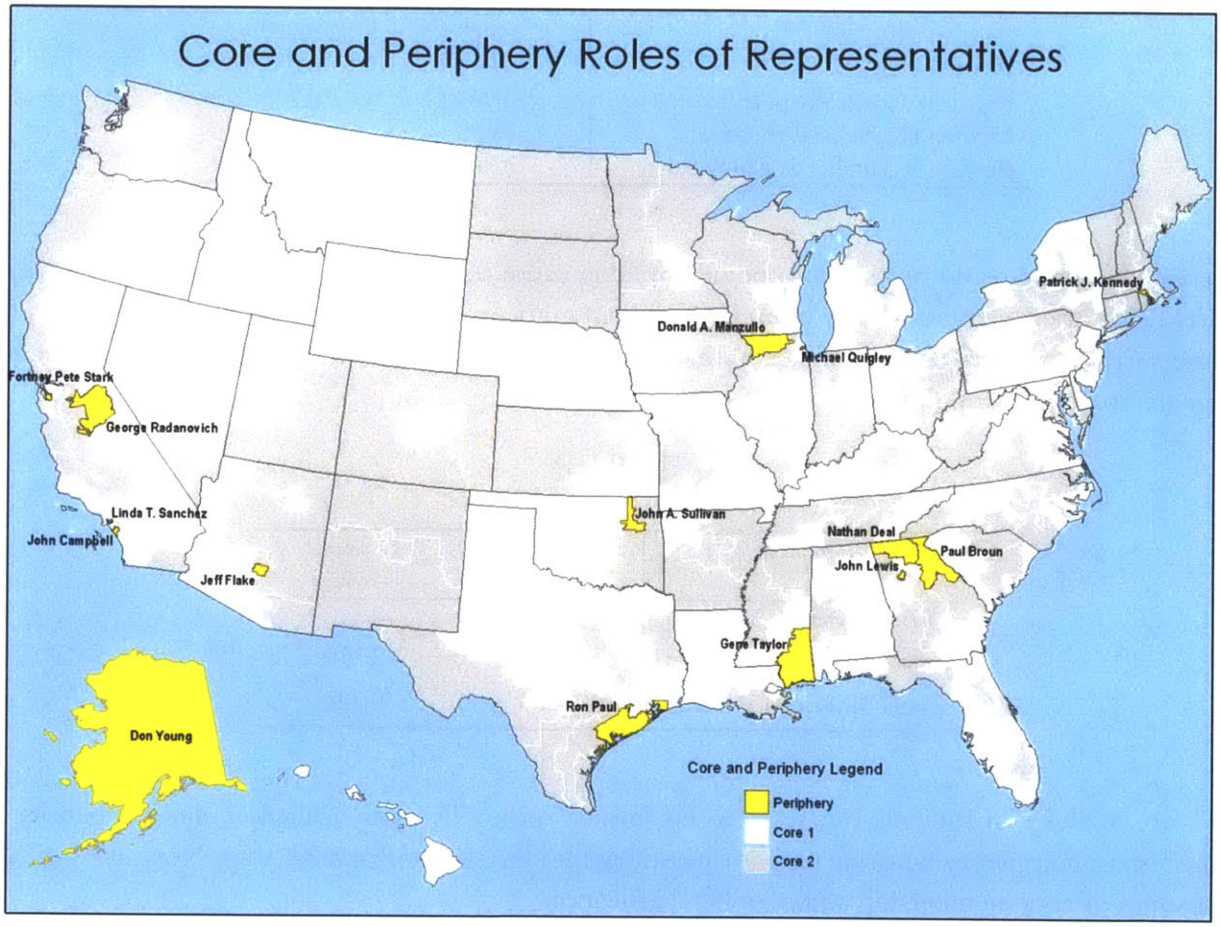


Figure 5.6: Core and Periphery Roles of Representatives **The cartographic representation of peripheral members (in yellow) shows 5 members from the Phoenix-San Francisco Corridor, and a significant cluster in Georgia.**

Theme B: Popularity Measures and Ego Roles

In total, the average number of friends per representative is 220, when not including the 8 people (3 Republicans and 5 Democrats) who do not have friends. When including these representatives, the average number drops to 216. Since there are more Democrats in the system, it follows that Democrat representatives will have a higher average number of friends. We see that on average, Democrats have 264 friends, and Republicans have 171. (Table 5) It naturally follows suit that average degree centrality will be similarly appropriated. When normalized by number of representatives in each party, however, party 'averages' are at .78 and .79, indicating that representatives, on average, have friend values equivalent to 80% of the number of people in their respective parties. In other centrality measures, Democrats have an edge on closeness centrality, again probably due to their numbers, and betweenness centrality, although low, is an effect of the same phenomenon. We discuss each issue further below.

TABLE V
DESCRIPTIVE NETWORK STATISTICS BY PARTY

Parameter	*Republican Average*	*Democrat Average*
Friends (Degree)	171.039	246.523
Friends normalized by number of members in respective party	0.780	0.790
Closeness Centrality	0.526	0.604
Degree Centrality	0.387	0.555
Betweenness Centrality	0.001	0.002

Degree and Friendships

As mentioned, members of the Democratic Party generally have more friends, and higher degrees, because there are more ideologically-similar liberal compatriots for tie formation. The highest degree nodes are located on the forefront of the Democratic Party cluster, as they have a strong network of ties within the Democratic cluster, and augment these relationships with ties in the Republican cluster as well. Since ties are reciprocal, the friends that the high-degree democrats make in the Republican cluster are naturally marked as more moderate members of the GOP. (Figure 5.7)

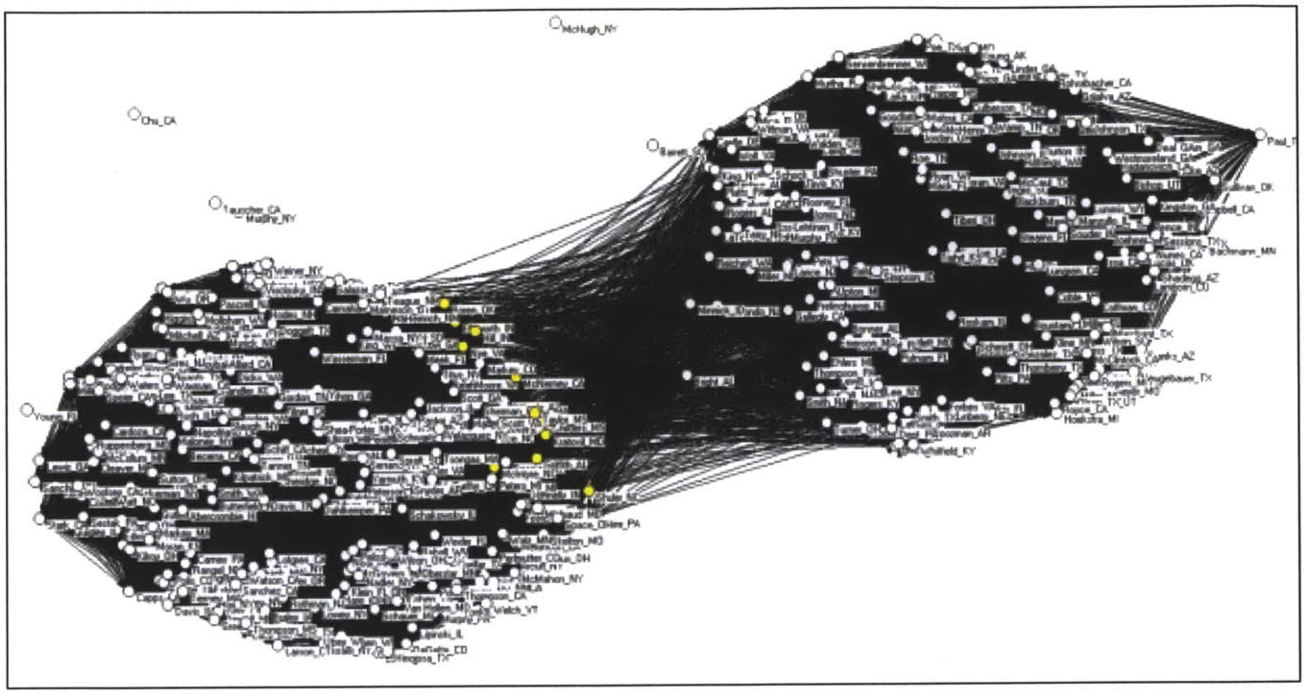


Figure 5.7: Network Location of High Degree Nodes in the Social Network **A force-directed display of the Congressional network shows that high-degree nodes (above one standard deviation from the mean, marked in yellow) are positioned at the 'front line' of the Democratic Party cluster.**

The geographic distribution of the high-degree nodes shows a clustering of representatives in the Midwest and South. No representatives from the New England or urban Boston-New York-D.C. Corridor, Chicago or Los Angeles are present. (Figure 5.8) Additionally, of the four most-represented states, California, Texas, New York and Florida, (totaling 129 representatives) there is only representative marked has having an outstanding number of ties, Rep. McNamey of California.

Representatives of rural areas play a large role in the high-degree group, a surprising finding as each high-degree representative is a member of the Democratic Party, where presence leans toward urban areas. Exurban Areas outside Memphis, Charlotte, Wilmington N.C., Washington, D.C., Pittsburgh, Indianapolis and Chicago are represented, as are two decidedly rural areas, in Eastern Oklahoma (Rep. Boren) and Eastern Arizona (Rep. Kirkpatrick). The overrepresentation of high degree members in exurban Central U.S. may point to interesting findings about the cultural nature of constituents and the friendliness or political moderateness of their elected officials.

When network friends are normalized by number of members in respective parties, new leaders emerge. (Figure 5.9) Rep. Reichert (R-WA), Rep. Miller (R-CA), Rep. Johnson (R-TX), Rep. Wolf (R-VA), LoBiondo (R-NJ), Rep. King (R-NY). When adding the high-friend Republicans, more urban areas arise. Inner city New York, Dallas and Los Angeles are now represented, as are direct suburbs of Philadelphia, D.C. and Seattle.

Combined, these findings are interesting, as rural Democrats, and Urban/Suburban Republicans seem to garner the most friends, indicating that the chemistry surrounding a cross-mixture of rural ideologies and liberal representatives, and urban ideologies and conservative representatives equal the level of moderation needed to attract the most friends.

Next, we discretize network position further, and examine closeness centrality and broker roles in the network.

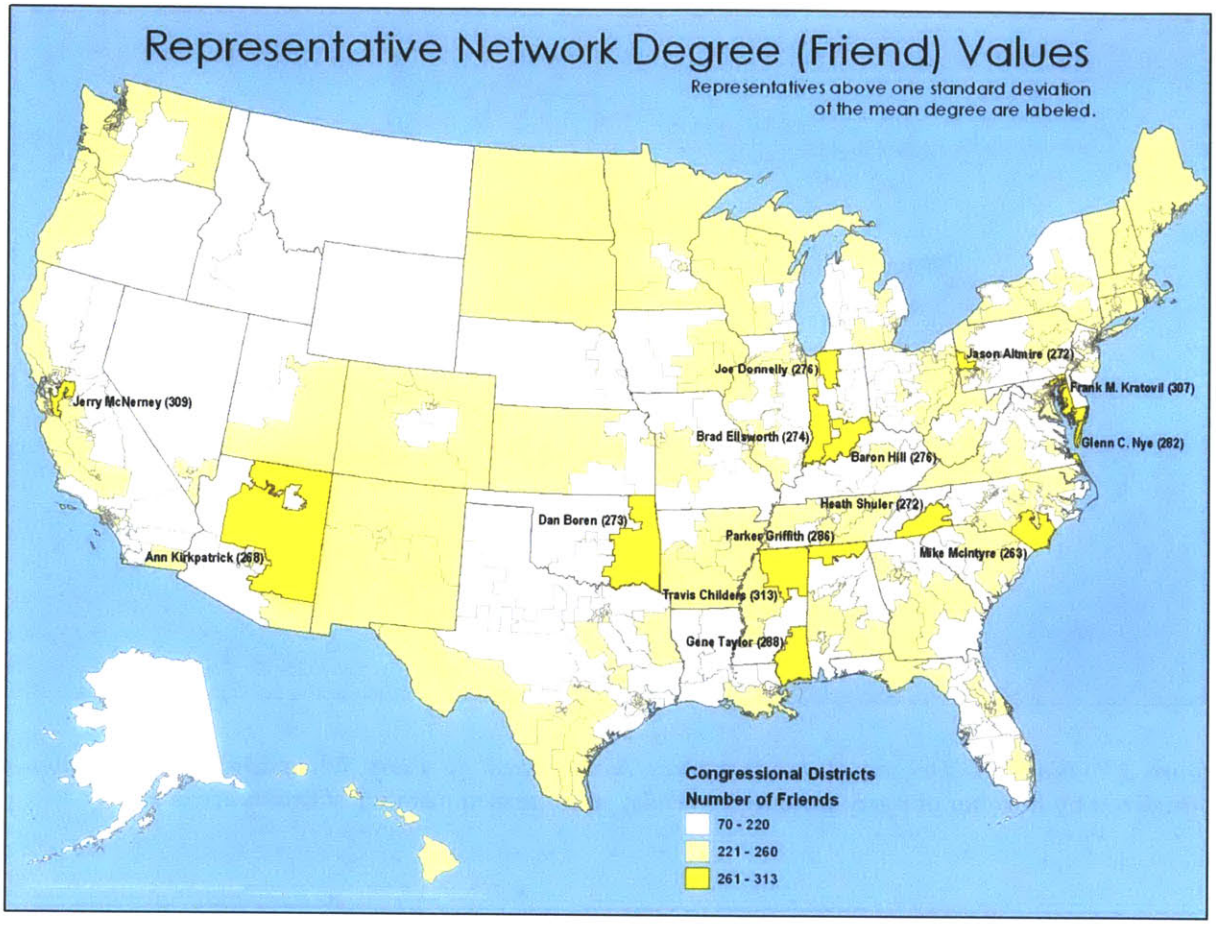


Figure 5.8: Representative Network Degree (Friend) Values **A group of high-degree individuals are centered around the American South and Rust Belt, with a few geographically-peripheral members on the rural Delmarva Peninsula, rural Arizona and Oklahoma, and finally, one geographic outlier in the more urban San Francisco Bay Area.**

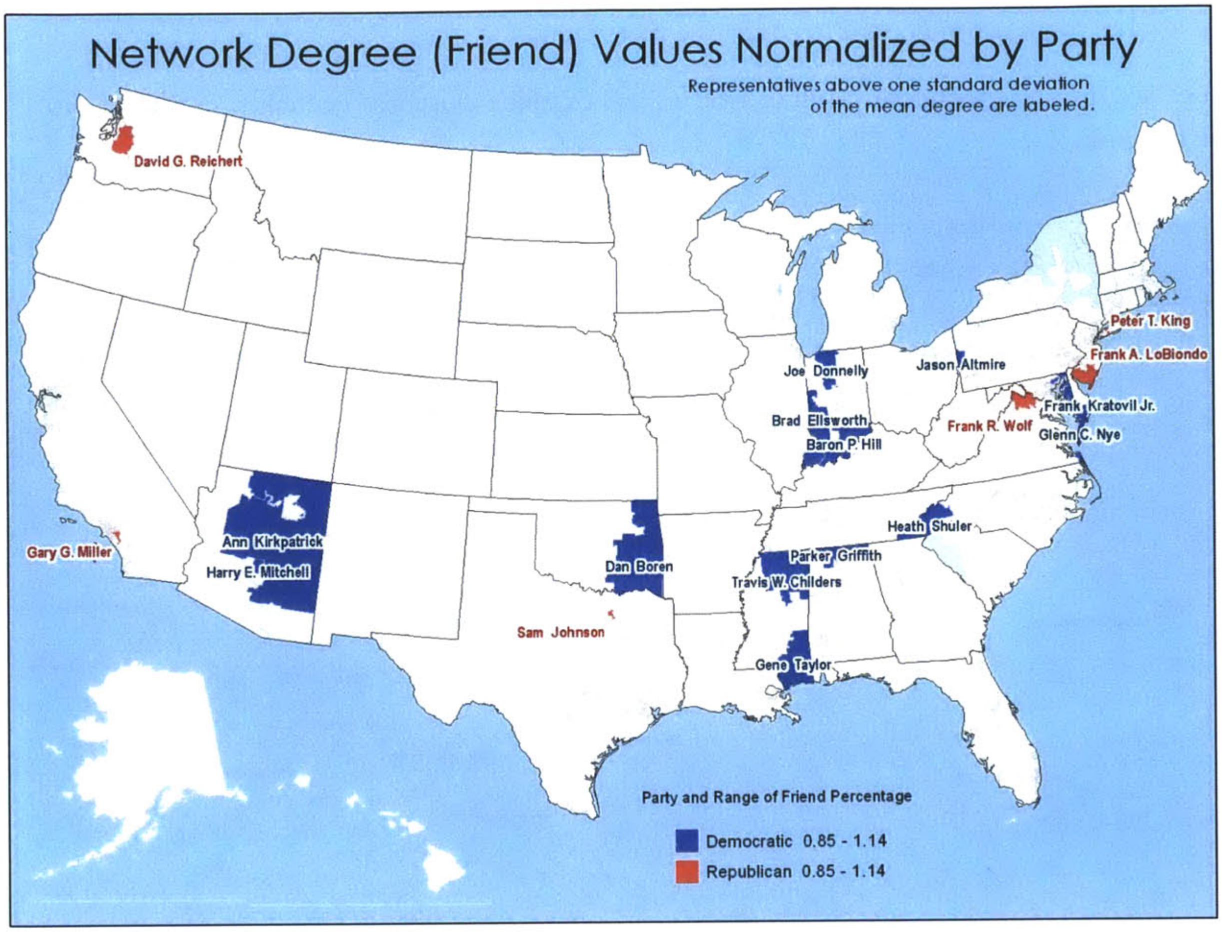


Figure 5.9: Network Degree (Friend) Values Normalized by Party **When high-degree nodes are normalized by number of party-potential friends, an increased number of urban areas arise.**

Closeness and Betweenness Centrality

Closeness centrality metrics are evaluated network and geographic positions two standard deviations above and below the mean values. At two standard deviations above the mean, we see roughly the same characters as in the high-degree friendship analysis. (Figure 5.10) This group of representatives, set at the forefront of the Democratic Party cluster, also exhibits the highest closeness centrality values. This structure indicates that the high-friend representatives are also in a position of powerful reach to other members of the system. At the opposite end, the members with the lowest closeness centrality, those most relatively removed from accessing other members include a number of Republican congresspersons, Rep. Cambpell (CA), Rep. Flake (AZ), Rep. Paul (TX), Rep. Sullivan (OK), and Reps. Deal and Broun (GA). These members may be considered 'too conservative' to gain friends across party lines, or even friends who have friends across party lines. It is also interesting to note that, unlike the juxtaposition between urban Republicans and rural Democrats in the previous section, two Republican representatives, Reps. Campbell and Flake, each represent dense urban areas, but exhibit highly conservative positions in terms of network reach. (Figure 5.11)

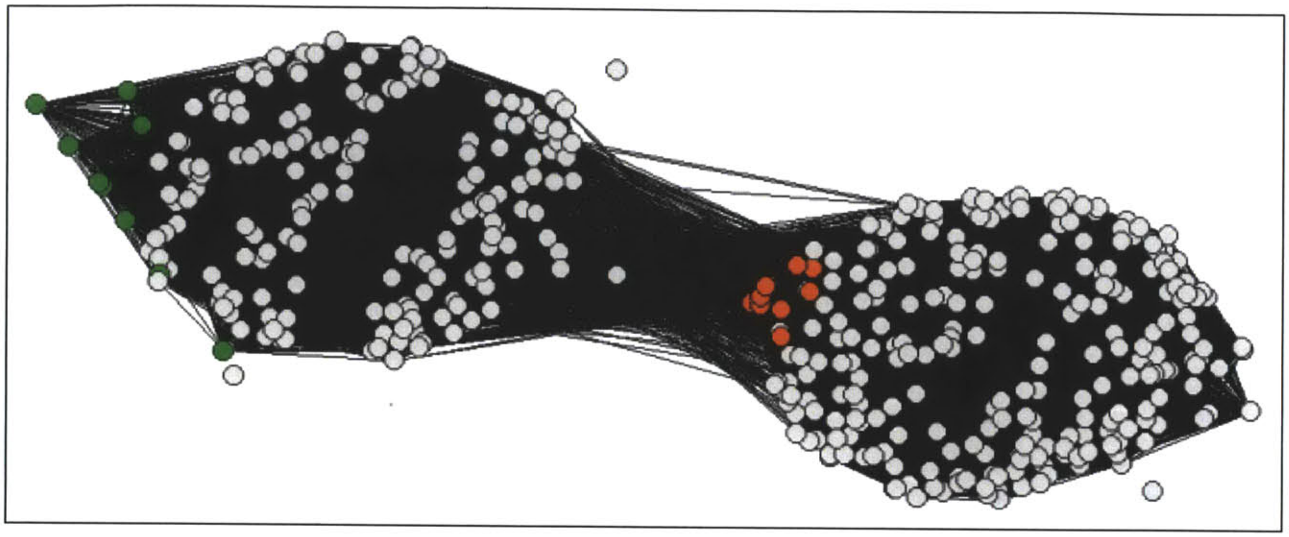


Figure 5.10: Representatives with High and Low Closeness Centrality Measures **Closeness Centrality measures above two standard deviations are shown in orange and below two standard deviations in green. Green and orange nodes are found exclusively in the Republican and Democrat clusters, respectively.**

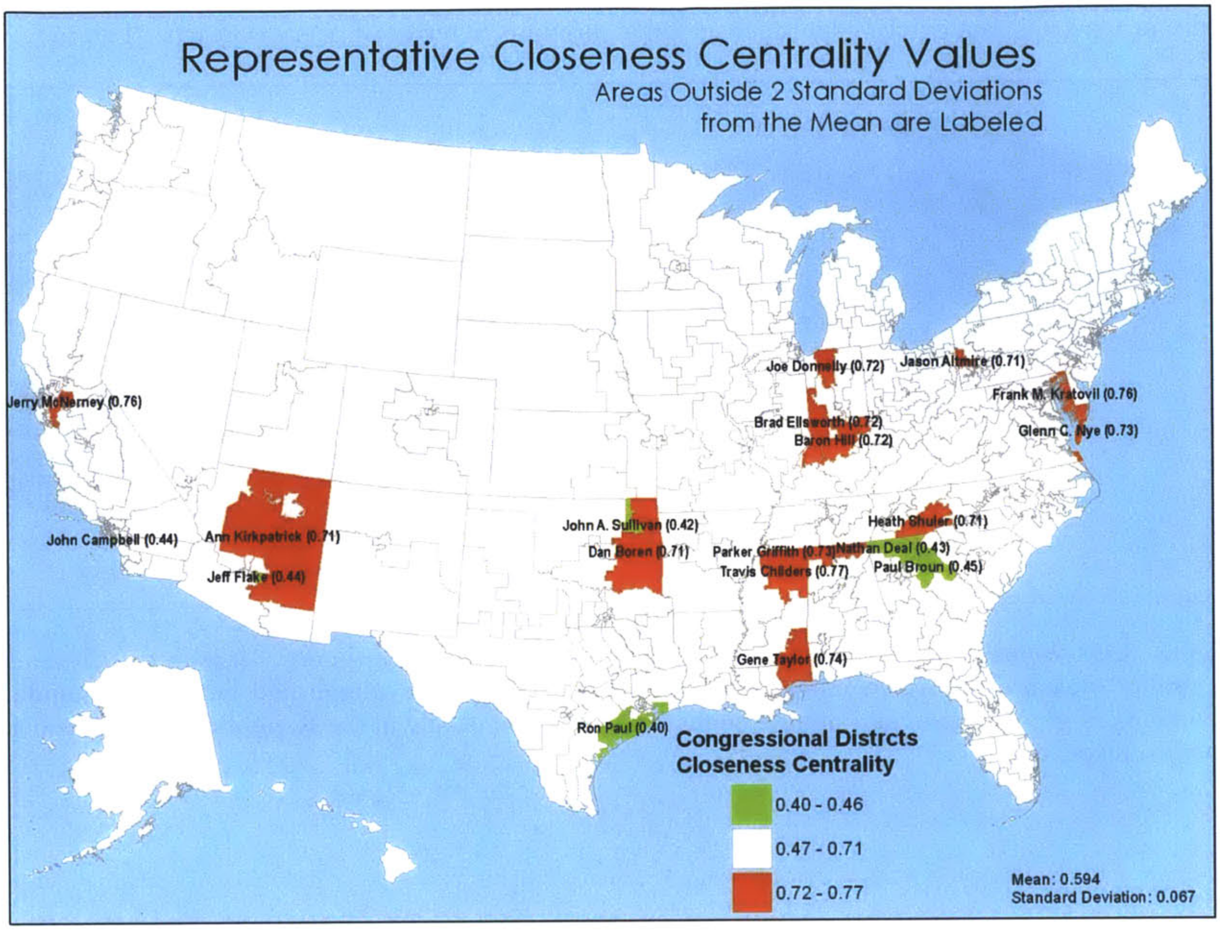


Figure 5.11: Map of Representative Closeness Centrality Values **Closeness Centrality measures above and below two standard deviations complement the network positions in Figure 5.10.**

Leaders in betweenness centrality include Reps. Bright (AL), LoBiondo (NJ), Taylor (MS), Childers (MS), Kratovil (MD), McNerney (CA), with Minnick (ID), Reichert (WA), Schuler (NC), Griffith (AL), Connelly (IN), Hill (IN) and Nye (VA), shown in orange and yellow in Figure 5.12. Representatives in this category also exhibit high-friendships. These representatives also have high brokerage values, especially representative, coordinator and liaison (party independent) brokers.

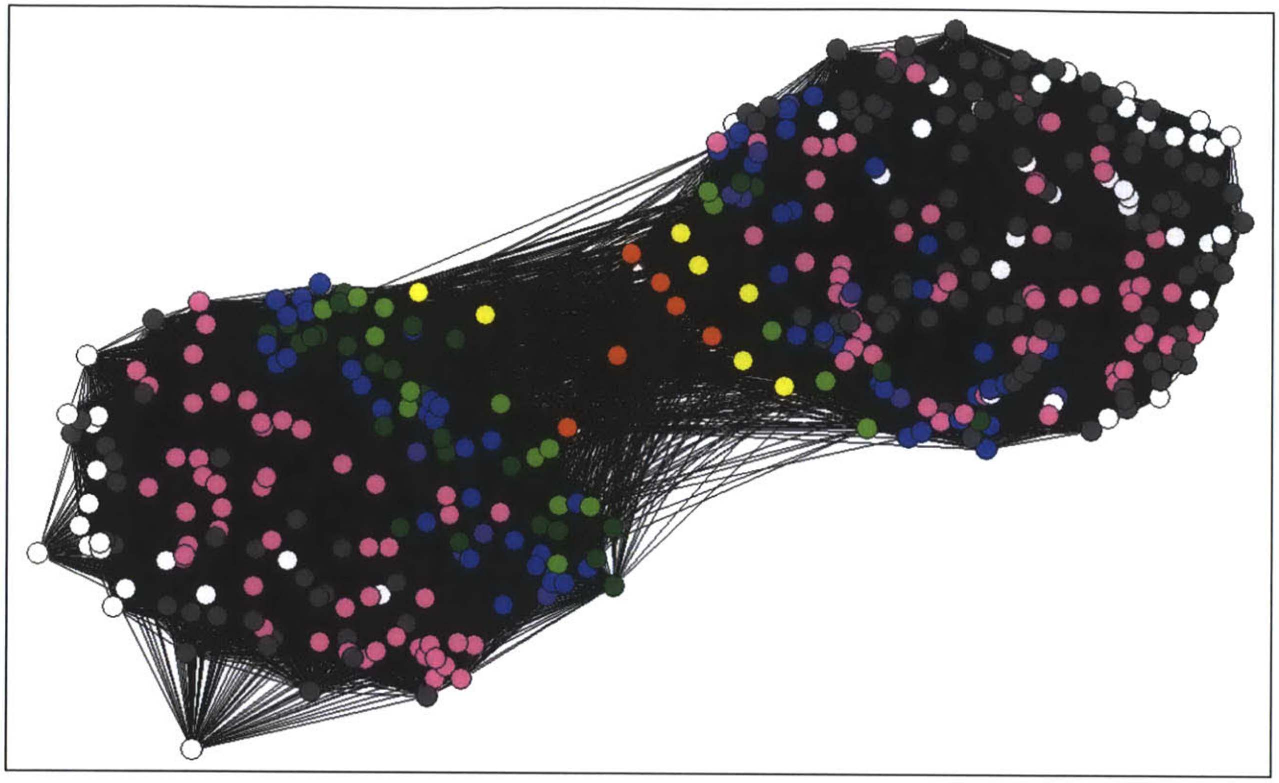


Figure 5.12 Spectrum of High to Low Betweenness Centrality Values **shows the intermediary members with high betweenness centrality values. Values are approximated by the natural log of betweenness centrality multiplied by 10^5, where highest values (7) are represented by red-orange nodes, and continue down the rainbow spectrum to zero (white).**

THEME C: ATTACHMENT BEHAVIOR

Clustering Coefficient

The congressional network is highly clustered. The mean clustering coefficient is .969, with a standard deviation of 0.055, meaning that a 'perfect' clustering coefficient of 1, where all of an ego's friends are also friends with one another, is within one standard deviation of the mean.

Showing that 'popular' or 'well connected' does not necessarily mean 'embedded', members with the highest degree (friendships) also exhibit some of the lowest clustering coefficient values (below two standard deviations below the mean). (Figure 5.13) Those that fit both categories include Reps. Ellsworth, Hill, Altmire, Kimball, Nye, Shuler, Griffith, Taylor, Childers, Boren, and McNamey. (Figure 5.14) These representatives have the highest percentage of relationships with representatives who are not friends with one another. Accordingly, we note two party "rebels" in this category: Rep. Bright (D-AL), Walt Minnick (D-ID). As mentioned above, the centrally located members are comprised of Democratic representatives whose districts are generally rural, and Republican representatives whose districts are suburban. These more moderate members have low clustering coefficients because they have ties on each 'side' of the network, and since their friends hail from two parties, these friends are not likely to be friends with one another. Representatives on the peripheries of the network, and in the 'extreme' sections have clustering coefficients of 1, indicating that their friends belong to a nested clique. Additionally, the presence of these many-member cliques supports not only the fundamental belief that the network is highly bifurcated, but that members of each core are strongly embedded.

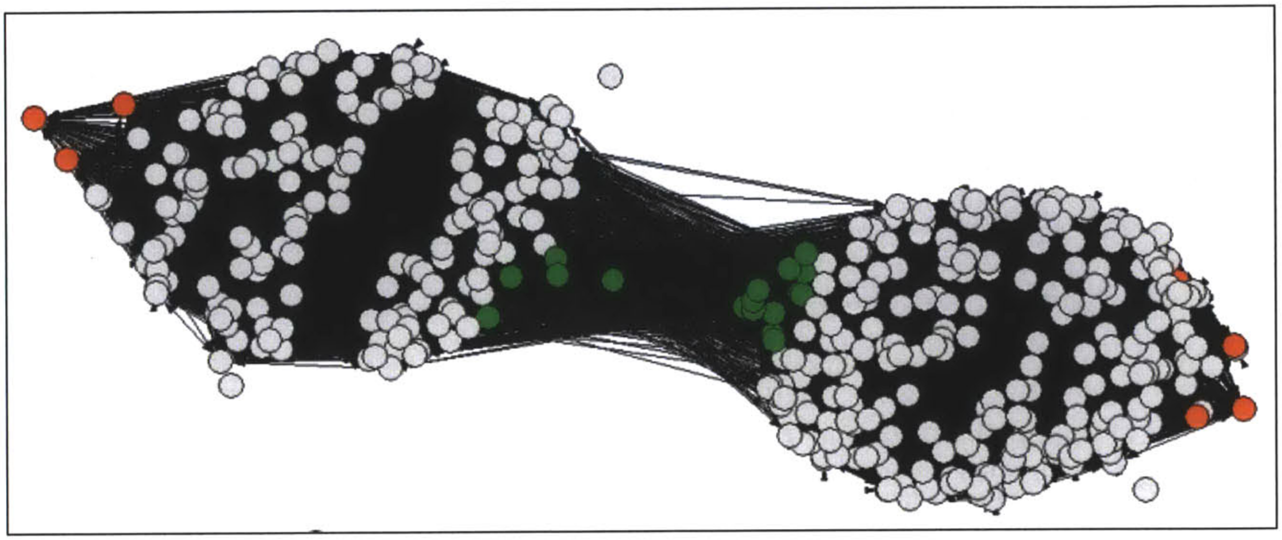


Figure 5.13: Representatives with High and Low Clustering Coefficients **Nodes with a clustering coefficient of 1 are denoted by an orange color and nodes betlow two standard deviations of the mean clustering coefficient are marked in green.**

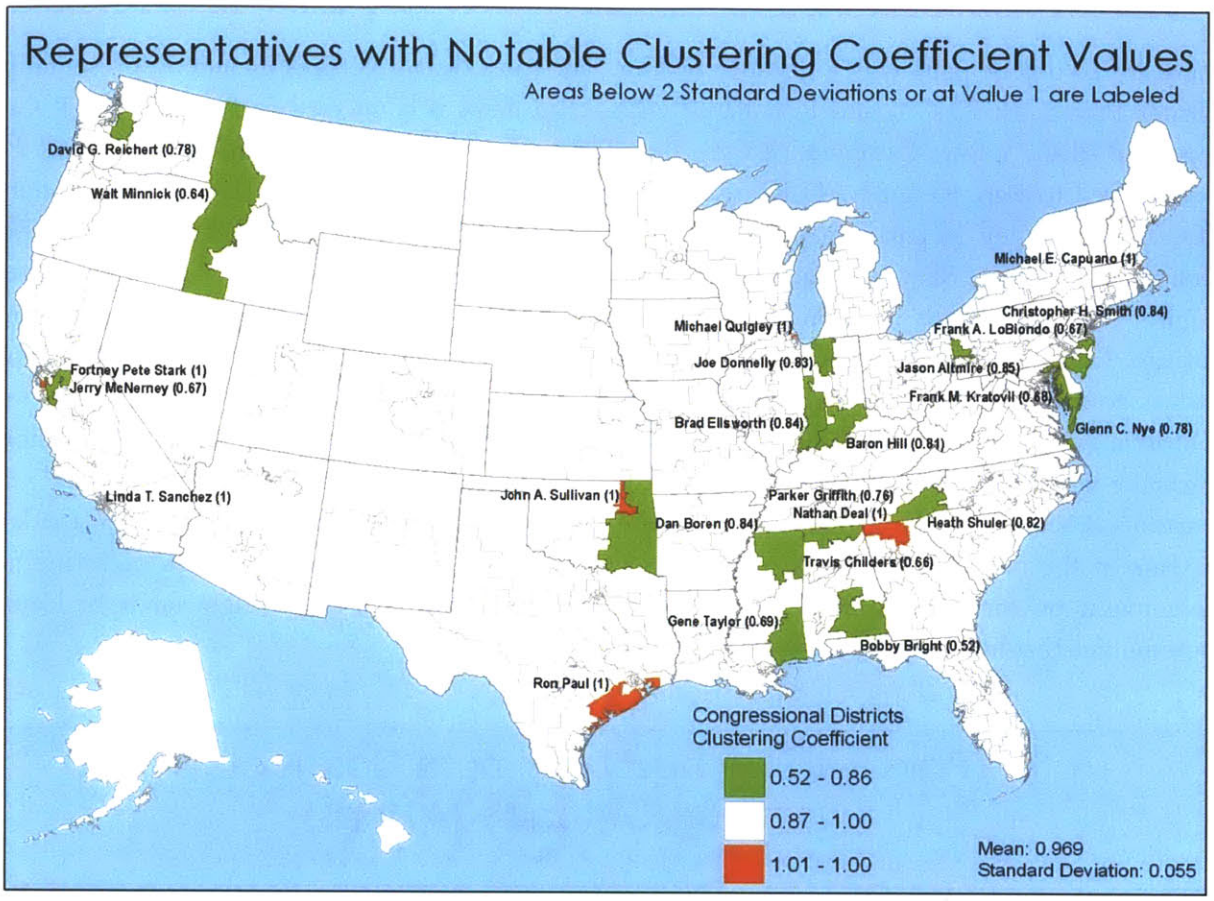


Figure 5.14: Location of Representatives with Notable Clustering Coefficient Values **As a complement to Figure 5.13, areas with a clustering coefficient of 1 are orange and areas betlow two standard deviations of the mean clustering coefficient are green. Here, we see a repeating pattern of high represenation from the Central U.S. accompanied by a number or representatives from the Mid-Atlantic.**

THEME D: DIFFUSION

For the 'diffusion' section, we revert back to findings on network traversal dynamics and structure. As mentioned when comparing our network to the Erdos Reyni configuration model, the maximum diameter of the congressional network is 4. Of each of the pair paths, the frequency of geodesic values (also referred to as distances) is as follows: 94,258 geodesic paths connecting any pairs (i,j or j,i) can be reached in one step. 53,802 geodesic paths connect pairs in 2 steps, 34,676 geodesic paths connect pairs in 3 steps, and 20 geodesic paths connect pairs in 4 steps (the maximum). In addition, 6904 pairs cannot be reached by direct 'hops'. Note here that each of these frequency values can be divided by 2 to represent unique pair frequencies. The average distance between reachable pairs: 1.67.

Most Distant Pairs

There are 10 unique pairs with a distance of four. The most distant vertices include Representatives Fortney Pete Stark (D-CA) and Ron Paul (R-TX). *Rep. Stark* is connected to Rep. Broun (R-GA), Rep. Deal (R-GA), Rep. Campbell (R-CA), Rep. Flake (R-AZ) and Rep. Sullivan (R-OK). *Rep. Paul* is connected to Rep. Kennedy (D-RI), Rep Lewis (D-GA), Rep. Quigley (D-IL) and Rep. Sanchez (D-CA). In addition to pairs with network distance of four that include either Rep. Stark or Rep. Paul, the pair of Rep. Stark and Rep. Paul, at the party extremes, also requires four hops to connect. (Figure 5.15) Rep. Paul's position in deep Texas reaches to very dense urban areas of downtown Chicago, Los Angeles, Atlanta and Providence. Rep. Stark's position in downtown San Francisco, reaches to two rural areas in Georgia (Athens and Ranger) and Tulsa, Oklahoma, and Irvine, California (via Rep. Campbell) and Mesa, Arizona. We note that although a large international university is located in Athens, a factor that usually points to a liberal constituency, this constituency's leader's role in the network is among the most Republican-loyal. Although the later are part of the metropolitan areas of Los Angeles and Phoenix, the constituencies of Irvine and Mesa might be further examined for their anomalistic behavior as surprisingly far-right leaning communities that live near urban centers.

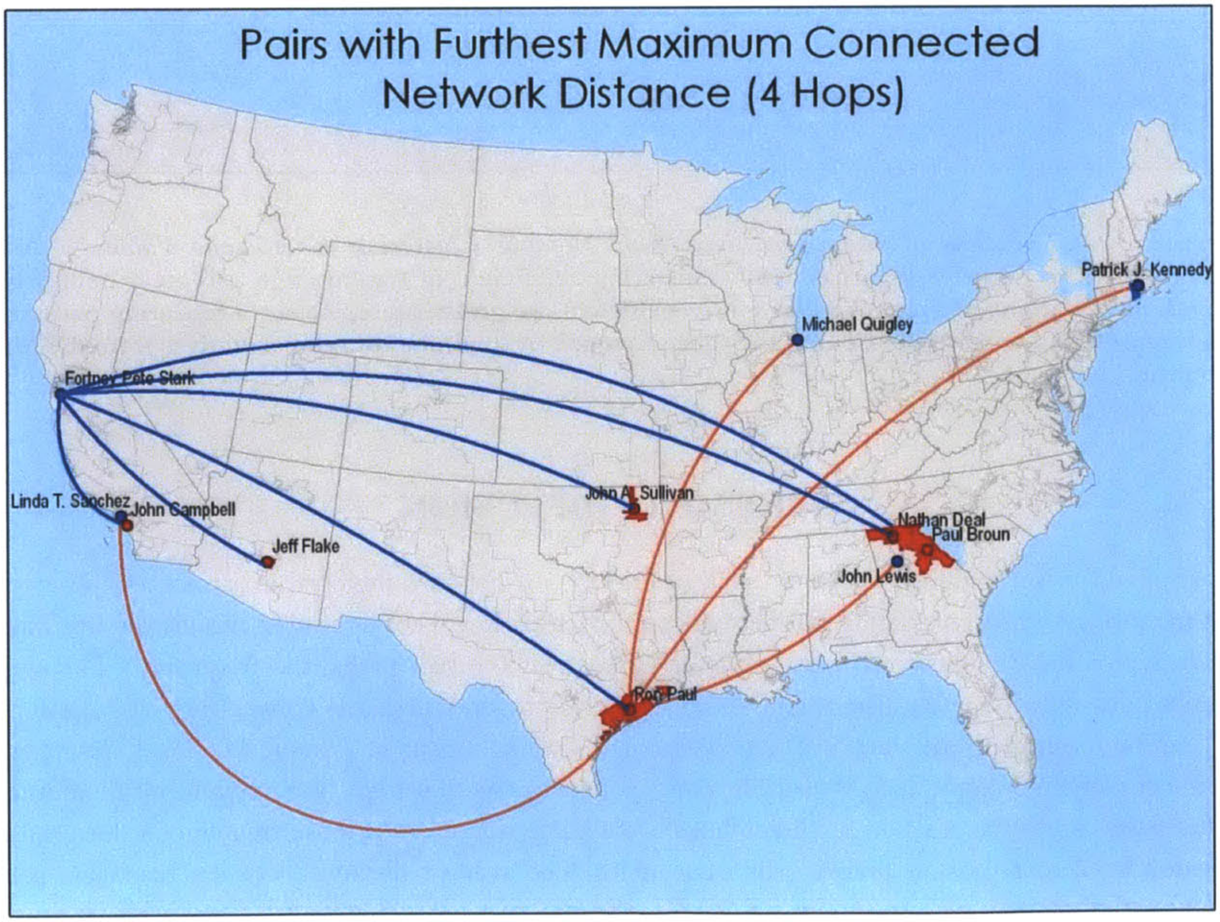


***Figure 5.15: Pairs with Furthest Maximum Connected Network Distance (4 Hops)* shows unique pairs of D-R representative relationships that require the maximum network distance of 4 hops for topological connectivity. Accordingly, these pairs also have the highest cardinality of unique geodesic paths to connection.**

After uncovering perhaps the most saturated or intense representatives in terms of party loyalty, we can view this Republican to Democrat (or vice versa) spectrum from an ego point of view. To show an example of diffusion through the network, we use a case study of a representative from Kansas, the most central state in the contiguous U.S. (chosen for visualization purposes). Representative Lynn Jenkins (R-KS) is based in Topeka, Kansas, the relatively urban capital of the state. Rep. Jackson is located not on the forefront 'band' of the Republican Party cluster nor on the periphery, but generally central in the GOP core. (Figure 5.16)

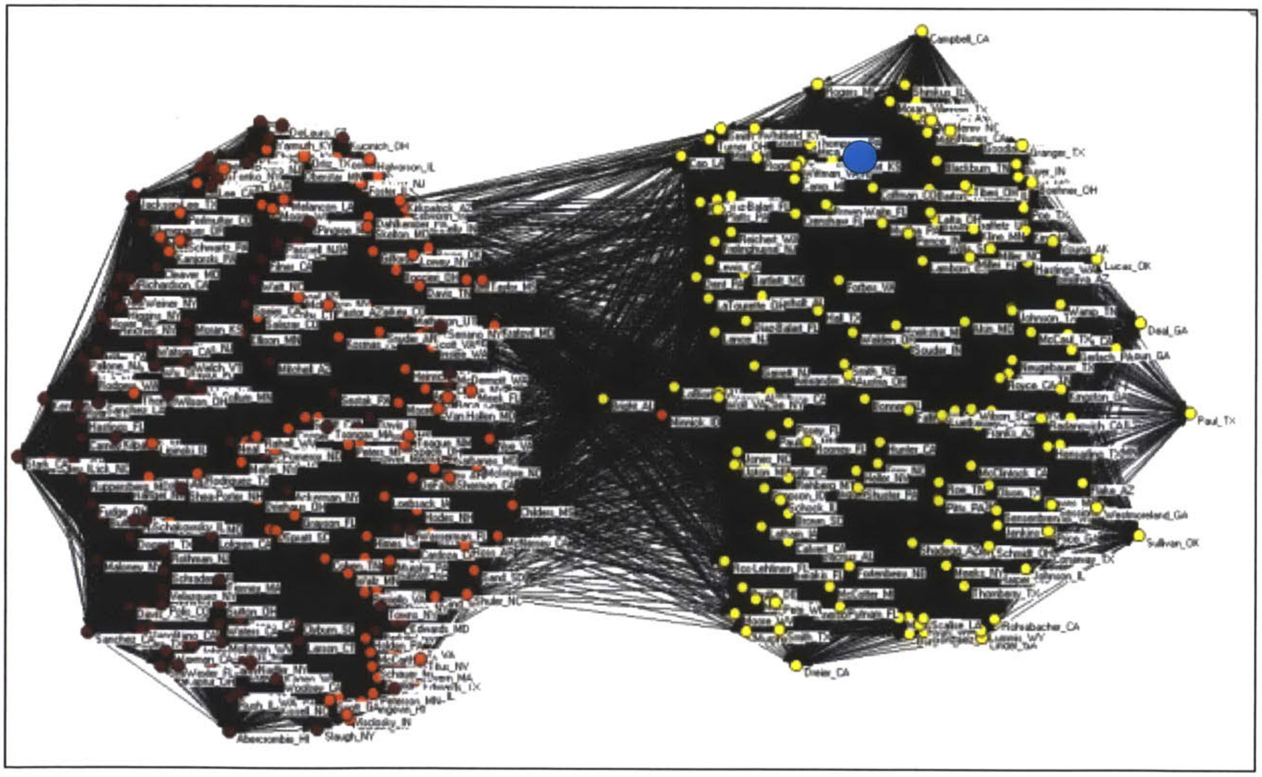


***Figure 5.16: Social Network of Representative Lynn Jenkins* Rep. Lynn Jenkins, our case study Congressperson, is marked with a large urquoise node in the Republican cluster. (Top Right) Her immediate neighbors are yellow, followed by second degree neighbors in orange, and third degree neighbors in maroon.**

Rep. Jenkins reaches each Republican representative in one network step, along with the recurring central member, Rep. Rep. Bright (D-AL). Interestingly, the only member of the Republican modularity group that is not reached is Rep. Minnick (D-ID), where perhaps his Democratic ideologies lean just far enough to the left to obfuscate a potential friendship with Rep. Jenkins. Although Rep. Bright is in a similar position as Rep. Minnick, according to Rep. Jenkins' ego point of view, the divide between conservative and liberal may be made as a cut between Rep. Bright and Rep. Minnick.

Rep. Jenkins' second degree neighbors penetrate about half way into the Democratic core, interestingly overstepping some members close to the frontline border, but then impregnating some deeper core members. We can hypothesize that by the locations of the 2nd degree nodes that their ignition as two-hop neighbors is due to scattered friendships with any Republican (who, naturally, was Rep. Jenkin's first degree neighbor). Here, we are reminded that the force-directed visualization is only a guide for connectivity positioning, and does not promise that graphical closeness means connected. Rep. Jennings reaches the remainder of the Democratic representatives with three hops.

In order to visualize and absorb the geographic topology of Rep. Jennings' power, comfort and reach, we create a choropleth map (Figure 5.17) and a smoothed cartographic representation (Figure 5.18) of the number of "hops" that Rep. Jennings would need to reach the representative in each Congressional district. In this experiment, we first assign the hop number (1, 2 or 3) to each district and then use a Kriging interpolation algorithm to smooth the topology. The result is a map of Rep. Jennings' social reach in the U.S., according to her ideological voting patterns, and subsequent friendships with representatives from around the country. Rep. Jennings has the most reach in Northern Texas and Oklahoma, parts of the Gulf Coast, Appalachia, the Ohio and Indiana Border, Breadbasket, and Nevada/Idaho. If Jennings were interested in running for office in another locale, she may want to consider these places, as constituents in these areas elected representatives with voting patterns similar to her roll call voting record. Rep. Jenkins' social reach is most distant in the New England Area (requiring 3 hops in many places), urban California, and the Pacific Northwest. Additionally, parts of the Midwest and Great Lakes region in Wisconsin, Chicago and Ohio may not be as amenable to Jenkins' ideologies, as they have seemed to elect a representative who voted dissimilarly to Rep. Jenkins, and therefore is 2 to 3 hops from her in the network. Finally, this representation indicates that a good destination for a trip outside the continental U.S. is Alaska rather than Hawaii, as Rep. Jenkins' is socially close to Alaska, but is 2 and 3 hops from Hawaii's representatives.

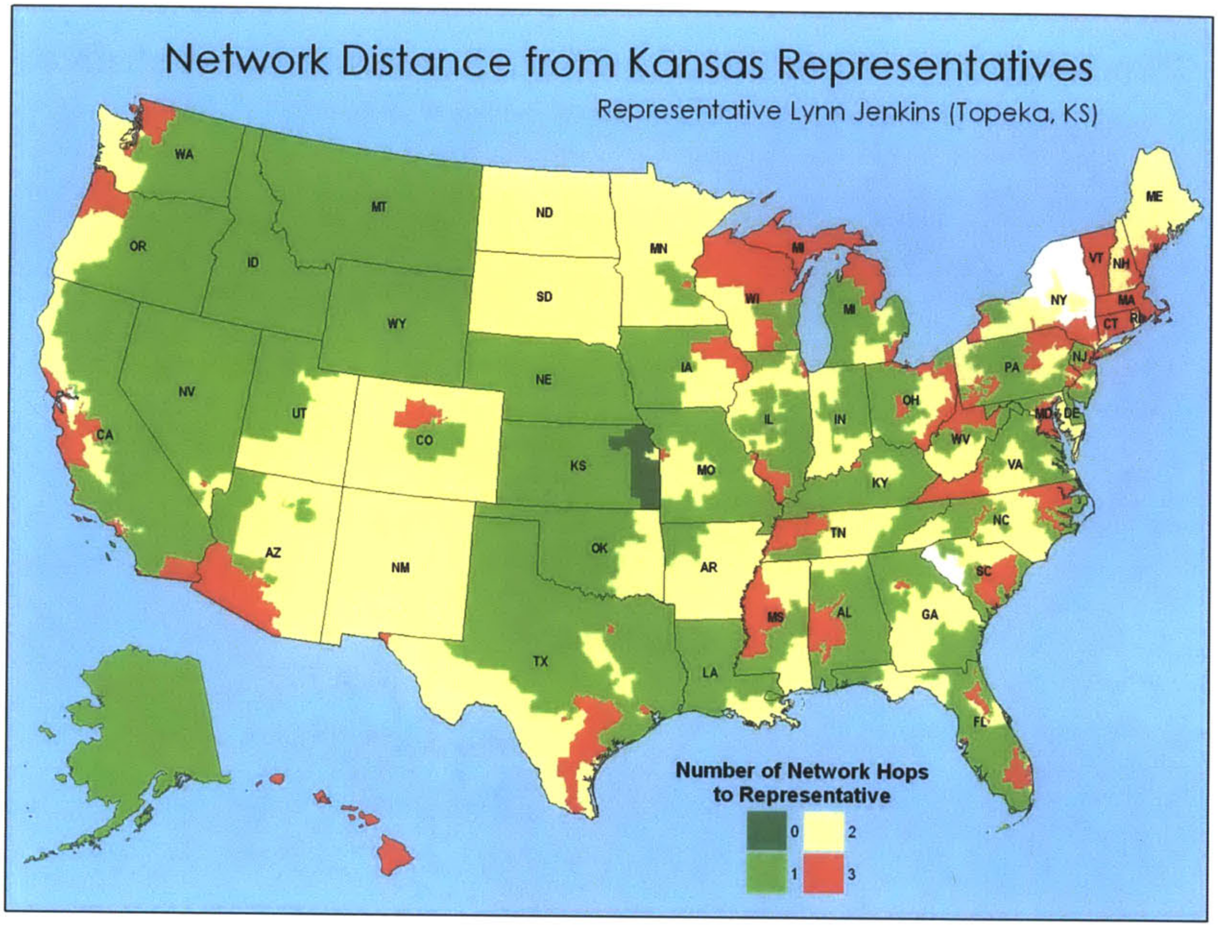


Figure 5.17: Choropleth map of Representatives' Network Distance from Kansas Representative Lynn Jenkins **This map shows the number of network 'hops' Kansas Representatives Lynn Jenkins would have to 'travel' in the network to reach the Representatives in each Congressional District.**

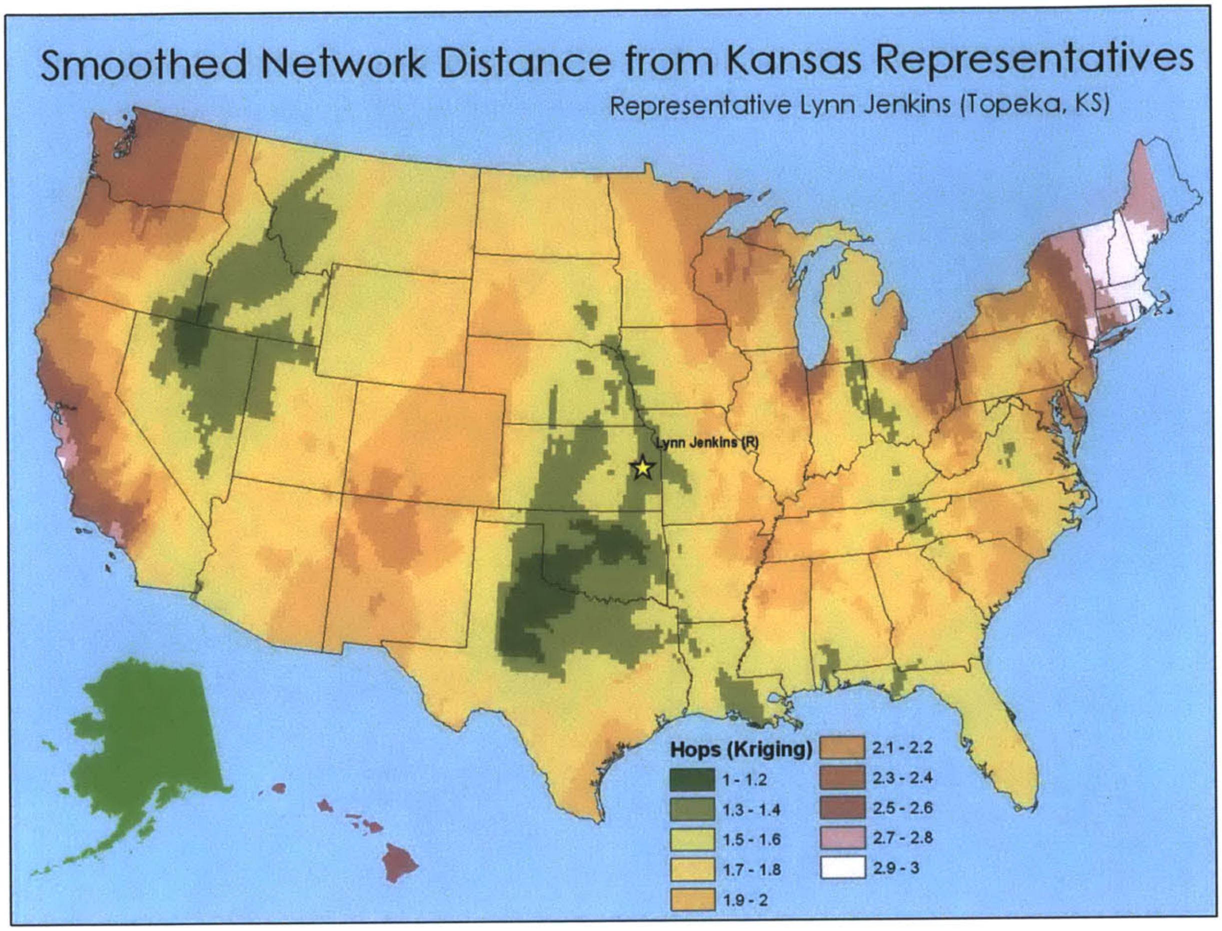


Figure 5.18: Smoothed Network Distance from Kansas Representative Lynn Jenkins **Rep. Jenkins' network hops to each Congressperson is attached to the Congressperson's district, and the values are interpolated for a smooth surface. This surface represents the social 'distance' from Rep. Jenkins to the remainder of the House-represented United States, where green is an ideologically-closer area, and the pink and white are ideologically-further areas. The result is a conceptually-contorted representation of Euclidean distance, based on public policy decisions. The center of Rep. Jenkins' Topeka, Kansas, Congressional district is marked with a star.**

CONCLUSIONS

This article aims to address the lack of social network analysis research that accounts for underlying geographies, interaction spaces, distance and cost barriers that guide social space time coincidences. Further, existing examples of social/spatial analysis underrepresent geographic topologies via disregarding areal space and cost distance.

Here, we use a multi-relational model of a force-directed network representation, loosely coupled with a cartographic representation of demographic variables to illustrate the prospects of tandem social/spatial analysis. We test this system fusion technique with a case study of 445 legislators in the

U.S. House of Representatives for the 111th Congress, and how their friendships correspond to their district demographic and district geographies.

With access to each congressperson's friends, alters, social group, propensity to agree with other certain agents, or to unconventionally float on the periphery of the network, we can analyze the complex dynamics that occur in a network of very different personalities. From these friendship dynamics, we can then learn more about how the 435 unique geographic districts' social groups, with their economic, demographic and locality characteristics, are "secretly" relating to one another through the behavior and ties forged by their elected representative.

We find that tandem social/spatial analysis gives a more robust picture of the relative strength of proximal and fraternal relationships. Geo-located agents have dual identities in a social network because in cartographic space, they are located in a static configuration of geographic places and spaces while simultaneously being 'socially closer' to non-adjacent geo-neighbors in the social network configuration. We hope that continued use of tandem social/spatial network analysis and its resulting metric of 'social distancc' will illuminate otherwise hidden facets of social relationships.

ACKNOWLEDGEMENT

The author gratefully acknowledges Wei Luo for aiding with the social network configuration and equations. The author also acknowledges Frank Hardisty for helpful comments and supervision and Alan MacEachren for providing support. This work is supported in part by the ASEE National Defense Science and Engineering Graduate Fellowship (NDSEG) Program and was conducted while at the Penn State GeoVISTA Center.

REFERENCES

[1] Diakonis, P. & Mosteller, F. Methods for studying coincidences, *Selected Papers of Frederick Mosteller*, **2006**, 605—622

[2] Goldenberg J. & Levy M. Distance is not dead: Social interaction and geographical distance in the Internet Era *ArXiv preprint arXiv:0906.3202*, **2009**

[3] Scellato, S.; Mascolo, C.; Musolesi, M. & Latora, V. Distance matters: Geo-social metrics for online social networks *Proceedings of the 3rd Conference on Online Social Networks,* **2010,** 8

[4] Loglisci, C. Ceci, M. & Malerba, D. Relational learning of disjunctive patterns in spatial networks *Workshop on Dynamic Networks and Knowledge Discovery, (DyNaK), Barcelona, Spain,* **2010**

[5] De Montis, A.; Chessa, A.; Campagna, M.; Caschili, S.; & Deplano, G. Modeling commuting systems through a complex network analysis: A study of the Italian islands of Sardinia and Sicily

Journal of Transport and Land Use, **2010**, *2*, 39—55

[6] Xu, Z. & Harriss, R. Exploring the structure of the US intercity passenger air transportation network: a weighted complex network approach *GeoJournal,* **2008**, *73*, 87—102

[7] Onnela, J.P.; Arbesman, S.; Barabasi, L. & Christakis, N. Geographic constraints on social network groups *ArXiv preprint arXiv:1011.4859,* **2010**

[8] Radil, S.; Flint, C. & Tita, G. Spatializing Social Networks: Using Social Network Analysis to Investigate Geographies of Gang Rivalry, Territoriality, and Violence in Los Angeles *Annals of the Association of American Geographers, Routledge,* **2010**, *100*, 307—326

[9] Dzeroski S. & Lavrac, N. Relational data mining *Springer-Verlag,* **2001**

[10] Malerba, A. Relational perspective on spatial data mining *IJDMMM,* **2008,** *1*, 103—118

[11] Shneiderman, B. Dynamic queries for visual information seeking *IEEE Software,* **1994**, *11*, 70—77

[12] Borgatti, S.; Everett, M. & Freeman, L. UCINET for Windows: Software for social network analysis *Harvard Analytic Technologies,* **2002**

[13] Batagelj, V. & Mrvar, A. Pajek—analysis and visualization of large networks *Graph Drawing,* **2002**, 8—11

[14] Clinton, J.; Jackman, S. & Rivers, D. The statistical analysis of roll call data *American Political Science Review*, **2004,** 98, 355—370

[15] Porter, M.; Muchab, P.; Newman, M. & Friendd, A. Community structure in the United States House of Representatives *Physica A,* **2007,** 386, 414—438

[16] Porter, M.; Muchab, P.; Newman, M. & Warmbrand, C. A network analysis of committees in the U.S. House of Representatives *Proceedings of the National Academy of Sciences*, **2005**, *102*, 7057—7062

[17] Lee, E. Geographic politics in the U.S. House of Representatives: Coalition building and distribution of benefits *American Journal of Political Science,* **2003,** *47*, 714—728

[18] Cox, G. & Poole, K. On measuring partisanship in roll-call voting: The U.S. House of Representatives, 1877-1999 *American Journal of Political Science,* **2002**, *46*, 477—489

[19] Snyder, Jr., J. & Grose, T. Estimating party influence in Congressional roll-call voting *American Journal of Political Science*, **2000**, *44*, 193—211

[20] Macrae, Jr., D., The relation between roll call votes and constituencies in the Massachusetts House of Representatives *The American Political Science Review,* **1952**, *46*, 1046—1055

[21] Watts, D. The New science of networks *Annual Review of Sociology,* **2004**, *30*, 243—270

[22] Cressie, N. Statistics for spatial data *Terra Nova,* **1992,** *4,* 613—617

[23] Moran, P. A test for the serial independence of residuals *Biometrika,* **1950,** *37*, 178—181

[24] Geary, R. The contiguity ratio and statistical mapping *The Incorporated Statistician,* **1954,** *5,* 115—146

[25] Anselin, L. Local indicators of spatial association-LISA *Geographical Analysis,* **1995,** *27*, 93—115

[26] Erdos P. & Renyi, A. On random graphs, *Publ. Math.,* **1959**, 6

[27] Valiente, G. Algorithms on trees and graphs *Springer Verlag,* **2002**

[28] Liben-Nowell, D. & Kleinberg, J. Tracing information flow on a global scale using Internet chain-letter data *Proceedings of the National Academy of Sciences,* **2008,** *105*, 4633

[29] Watts, D. & Strogatz, S. Collective dynamics of 'small-world' networks *Nature, Nature Publishing Group,* **1998,** *393*, 440—442

[30] Johnson, S. Hierarchical clustering schemes *Psychometrika*, **1967**, *32*, 241—253

[31] Bron, C. & Kerbosch, J. Finding all cliques of an undirected graph *Communications of the ACM*, **1973,** *16*, 575—577

[32] Newman, M. Detecting community structure in networks *European Physics Journal B*, **2004**, *38*, 321—330

[33] Jin, E.; Girvan, M. & Newman, M. Structure of growing social networks, *Physical Review E,* **2001**, *64*

[34] Newman, M. Modularity and community structure in networks, *Proceedings of the National Academy of Sciences,* **2006,** *103*, 8577—8582

[35] Girvan, M. & Newman, M. Community structure in social and biological networks *Proceedings of the National Academy of Sciences,* **2002,** *99*, 7821

[36] Freeman, L. Centrality in social networks: Conceptual clarification, *Social Networks,* **1979,** *1*, 215—239

[37] Freeman, L.; Borgatti, S. & White, D. Centrality in valued graphs: A measure of betweenness based on network flow, *Social Networks*, **1991**, *13*, 141—154

[38] Kleinberg, J. Authorative sources in a hyperlinked environment, *Journal of the ACM*, **1999**

[39] Marsden, P. Egocentric and sociocentric measures of network centrality, *Social Networks,* **2002**, *24*, 407—422

[40] Ibarra, H. & Andrews, S. Power, social influence, and sense making: Effects of network centrality and proximity on employee perceptions *Administrative Science Quarterly,* **1993**, 277—303

[41] Sabidussi, G. The centrality index of a graph *Psychometrika,* **1966**, *31*, 581—603

[42] de Nooy, W.; Mrvar, A. & Batagelj, V. Exploratory social network analysis with Pajek, Chapter 7, *Cambridge University Press, Cambridge UK,* **2005**

[43] McPherson, M.; Smith-Lovin, L. & Cook, J. Birds of a feather: Homophily in social networks *Annual review of sociology, Annual Reviews,* **2001**, *27*, 415—444

[44] Watts, D. A simple model of global cascades on random networks *Proceedings of the National Academy of Sciences,* **2002,** *99*, 5766

[45] Cowan, R. & Jonard, N. Network structure and the diffusion of knowledge *Journal of Economic Dynamics and Control,* **2004**, *28*, 1557—1575

CHAPTER 6: DISCUSSION

6.1 REVIEW OF SYSTEMS OF SYSTEMS CONCEPT

To review, we have conducted three experiments in Chapters 3, 4 and 5. In Chapter 1, we introduced the concept of connectivity/ interaction/ flow in GIS, as a data entity that connects two typically non-adjacent places. We further pointed out that a social flow is a flow with the purpose of connecting people to other people, and that this concept is important for both technical and conceptual purposes for understanding how society interacts with one another and with the built environment.

Our main problem is that GIS has grown up outside of Computational Social Science (CSS). Therefore, it is best suited for modeling situation, not connections. Conversely, CSS has grown up outside of GIS, which has resulted in geographic relationships between agents that are often too general, as place is not nominal, ordinal, or a ratio or interval value, but part of a complex underlying system of topological relationships.

In Chapter 2, we note that social flows are challenging to work with in GIS and geocomputation both technologically and conceptually. The chapter begins with some overarching considerations for approaching these challenges, from the standpoint of the history of GIS and its progress, and then outlines eight challenges that are deemed important for the advancement of the 'social distance' field from the seemingly minute details of how to interactively select certain flow lines in a GIS to the overarching concepts of how to create new statistics for this combined field.

In this section, we will review these challenges and suggest ideas for the future in light of having conducted, or metaphorically 'jumped off the diving board" into three experiments in social distance. The purpose of this chapter is to give the reader a "take away" message from which to use in future social distance endeavors.

First, we discuss an analogous scenario, the integration of transportation models into GIS, and compare its lessons with our findings. We harken back to this phase in GIS growth and metamorphosis in order to better understand the path and challenges of integrating new models into GIS. Shortest path and routing algorithms date to the 1970's as commercial software. Early integration with GIS (as exemplified by *ESRI's ArcInfo*) began in the 1980's and early integration of GIS in transportation planning and with CAD (like *Intergraph*) in the 1990's (for example, *TransCAD*). Today, we see transportation software included in packages like ESRI'S ArcGIS as 'toolboxes', like Network Analyst, which provides features like routing algorithms, and tools for locating service facilities.

The layer-based world of GIS has a successful history of integrating network models, when viewed from the vantage point of incorporating transportation models into the GIS environment in computation, simulation and visualization. The finer points of this integration give hope for the integration of social network models into GIS because they provide evidence of effective engineering and retrofitting of graph-based information with planar-based information. We can call the engineering 'effective' because it has enabled fruitful research through the fluid access of both 'worlds' in one GIS view or platform.

Co-location is an obvious reason to combine transit models with the planar geographic offerings of a GIS and the data and theories surrounding spatial studies, as it is not enough to have a road network model in a 'vacuum', since the roads themselves certainly are affected by and effect things like population density, land use, urbanization, demand, and environmental factors like terrain. These considerations are hard to "assign" to an edge in a matrix-type model, and so transportation networks fused with GIS gives a richer picture of the dynamics of urban and rural space, and the dynamics of traversing the space.

The incorporation of transportation models into GIS changed the nature of GIS because the transportation models added a new layer of "network distance" to the Euclidean lattice that spatial models use to situate entities and calculate dispersion. One might even argue that the addition of transit models was a step that helped us conceptualize time as cost into GIS, as metrics for cost/benefit of location could not be calculated by travel time and accessibility measures. Today, one could argue that the explosion of data is a challenge, as boundaries change over time and might not be compatible with all ranges of time. (Table 1)

TABLE I
COST/BENEFIT MEASURES IN TRANSIT AND LAND USE MODELS

Transit Models	***Land Use Models***
Accessibility	Situation
Travel Time	Distance
Traversal Ability	Terrain, Physical Qualities

Meanwhile, from a technical standpoint, it became more difficult to remedy the 'useful' cost distance between two places with their 'actual' distance based on parallels and meridians.

The transit models were mostly integrated through vector layers, where roads were 'tacked' to their longitude/latitude paths, and transit lines were similarly orthorectified to match with underlying geography. Depending on the precision needed for the roads, more or fewer reference points are stored in a transit line file. These transportation networks also have points that are used to associate one vector with another vector: some are intersections, others are stations or points of interest. A 'null hypothesis' for roads may be that once the road lines are laid, vehicles could travel unconstrained through the network, but with time, the 'rules' to road (and other) networks needed to be more stringent in order to accurately describe the phenomenon of travel in the built environment. Integrators developed rules, like "not all road line crossings indicate an intersection", and "not all lines allow for traffic in both directions", "not all lines are traversable at all times." To operationalize these constraints on the network, special infrastructure had to be in place, and so database engineers had to become sensitive to accommodating one-to-many or many-to-many relationships, and event, route or multiple path features. These considerations resulted in tools like *ESRI's* Geodatabase, Network Feature Layer, Network Analyst (as mentioned) and *Caliper Trans CAD's* interoperability with shapefiles.

This integration, however worthwhile, was not an overnight process. Instead it seems to have necessitated cross-dialogues between fields of Transportation/Civil Engineering and GIS/ Geography. Although today we may take these multi-disciplinary relationships for granted, when transportation models were being integrated into GIS environments, the avenues of discussion between transportation engineers and GIS software engineers was not well-paved, although a council between, within and among them seems to be a good way to learn more about the built environment.

Following in the footsteps of Transportation and GIS cooperation, a major thesis of this dissertation is that we find that there are both endogenous and exogenous factors that are central to the future of GIS and CSS integration namely (1) increased exposure of GIS to other fields and (2) improved technological infrastructure. The first factor was concluded after an extensive review of literature found in this dissertation; the second factor was concluded after attempting to conduct our three experiment/case study chapters.

In the next section, we first explore the dialog between GIS and CSS Analysts, and in the following section, some suggestions for improving technological infrastructure through software.

6.2 Suggestions for GIS and CSS Communities

The main argument in this work is that there is much to be gained for urban planners if GIS models and social science models can be employed in a larger body of a new computational system and a new field of research. This new field is markedly not an effort to plug in social science concepts into a post-developed GIS world, nor an effort to mention GIS to the developed world of network science, but the beginning of a culture of research that includes both social and spatial systems.

We believe that there are many interesting futures for this type of social/spatial research, but so far, we have not provided enough evidence to suggest that progress in this new field is obtainable, and thus the conclusion outlines some potential guidelines and recommendations that may serve as beginning blueprints for the conceptualization of a research group.

Benefits of Drawing from both Social and Geographic Systems

It is helpful to draw from both sides of CSS and GIS for particular urban planning problems because CSS researchers can explain factors of human behavior individually and in groups, and GIS analysts can explain how humans traverse and use space. Together, new correlations can be found that uncover how the human condition is related to the built environment, and we can use these patterns, correlations and cause-effect findings to inform planning decisions. This new fusion can birth a batch of statistics that show how to calculate whether a group of linked friends in a social network are also significantly clustered in a geographically or show whether an agent's decision to patronize a nightclub (for example) is driven by least-cost time or by a critical mass of friends in proximity. The fusion can also inform us of biases about aggregating groups of people, different answers based on different geographic scopes or social subsets, problems like causation/correlation, data types, ecological fallacies, ethics and privacy.

Suggestions for Developing a Field

It is not our wish to dilute the value of social or spatial experts, as a life dedicated to following either will most certainly yield insights that an interdisciplinary researcher may not dig deep enough to reach. But we believe that a new arena for asking new questions may yield insights that cannot be reached from devotion to one realm.

Furthermore, this dissertation does not chide a disinterest of 'spatial' for social researchers or the 'social' for spatial researchers. Indeed, a GIS researcher or user could spend his entire career without seeing any of the connections, flows, or human relationships that occur in the space he models. Without this activity, the built environment is very incomplete, but this may be necessary for modeling some kinds of processes, like sedimentary flows. Conversely, a social scientist can conduct her research for a long tenure without any knowledge of the topological landscape that may actually drive the relationships she studies, as physical models may not be helpful for studying family conversations or human dimensions of love. Still, when research questions and phenomena could be tied to both people and place, a rigorous literature should be available for support: Else a social system's spreading behavior may be carelessly 'guessed' by a GIS analyst or a social scientist may 'guess' about the complex nature of transit in city growth when she assesses social capital. When laws that govern social or spatial systems can be applied to one another's models, communication between scientists is helpful, but a haven for this unique dynamic symbiosis may help more, since the new social/spatial field may be equipped to handle esoteric problems. This leads us to our next question: How do we equip this field?

We are not trying to change the GIS world, but to promote a new world that approaches the two fields together. It may be the job of geographers and urban planners, to educate the social science community about the biases they may have towards the facility of connectivity, and their perception of the potential unimportance of the built environment. Furthermore, the social science community should be taught about how physical obstacles can hinder activity or opportunities, and the assumptions they are making by operating their models on a smooth spatial plane.

A new field of social/spatial systems, social flow analysis, and the conjuring of social distance metrics includes partnerships, collaboration and insight with experts in other fields. Pragmatically, dialog and method-sharing should occur among geographers, urban planners, operations researchers, and those in the social network and network science community. In order to mediate differences across cross-cutting subjects, collaborators may be wise to approach those with tacit knowledge of the field, in addition to learning more textbook fundamentals. In fact, respect for tacit knowledge may be crucial for fostering interdisciplinary research, since experts have deep roots in their own experiences and abilities in their respective fields. Leveraging these resources can lead to a deeper understanding of concepts, models, research designs, software, data and ethics in each field, as they intertwine, and build something that allows us to enable insights that come from each side.

THE ROLE OF THE GIS ANALYST

Informing CSS from the GIS Side

What we may be missing from Social forces is physical proximity, and its effect of hindering or enabling interpersonal relations. With social side, there is only so much we can understand without the physical constraints, like where can someone live and still commute to their profession, or how far apart can friends live to visit one another in a day's time. To paraphrase human geographer, Harm De Blij, 'geography finds unexpected linkages that are unmatched in other fields' (2005) meaning that spatial analysis has the power to uncover correlations between a wide variety of variables (like pollution and disease, weather and age, or student achievement and housing prices) that other fields may not be able harness. An important part of spatial analysis is the technology, data, geographic information systems and science (GIS and GIScience) that enables the pairings of variables as they coincide geographically. In this section we use GIS Analyst and Geographer slightly interchangeably, but use the latter for more general field issues, and the former for technical concerns.

Capabilities

To this field, Geographers have much to contribute. Geographic Information Systems analysts have many capabilities, and have learned to (1) represent certain phenomena (as raster or vector), (2) combine layers for meaningful answers to place questions, (3) model flows, transit, and cost-systems over geographic space. What factors enable movement, and what are the lag times of variables like hills or bridges. Geographers can contribute descriptive and analytical concepts in the form of tools and knowledge. Geographers have a good sense of spatial statistics like mapping clusters and hotspots, finding patterns in geographic space, understanding autocorrelation, performing geographically weighted regression, raster algebra, solving logistics and network problems, and finding descriptive data statistics like central features, mean centers and directional distributions. Also, remotely sensed data can be managed and extracted by GIS analysts in many different forms, in one example, analysts can capture real-time temporal changes, say of a crowd formation.

There are many geographic theories and models from which computational social scientist can derive hypotheses instead of relying on an exploratory knowledge discovery process. Von Thunen's theories hailed the ring configuration of a city and its ability to survive independently of other influences, good or services. Christaller's (1933) Central Place Theory, outlined the configuration between hubs, hinterlands and their patterned relationships and dependencies. Losch (1944) that says a political border produces an effect identical to that of increasing the distance. Alfred Weber's classic [Least Cost] "Theory of the Location of Industries" (1909) states that the process of locating an industry should optimize low material cost, low travel time of skilled and unskilled labor and agglomeration. Further, as mentioned in Chapter 2, William Reilly's Law of Retail Gravitation (1953) that supposes that places have magnetic qualities and 'pulls' was later expanded into the Huff Model (1964), which defines parameters and variables for service area calculation. Gras (1922) suggests a variety of ways for delineating social homogeneity. Theories of Waldo Tobler (abound) on the Rules of Geography are also catamount to this field.

Geographers are especially good at enumerating influencing factors when given a study area because they are trained to think about physical, human and political issues, as well as fuzziness, uncertainty and fleeting phenomena or probabilistic phenomena like synoptic meteorological activity. Geographers can give insight into which boundaries are 'important' for social phenomena, like a state border, a highway border, or a mountain chain, and the isolating effects of both. Geographic borders do not seem to be possible to represent in a social network. A social scientist might be able to model the propagation of drug usage or obesity, or just simply the friendship ties, in the U.S. Appalachian Region, but a geographer could explain how or why or when this structure formed. The geographer could discuss the industry, isolating factors of the mountain range, economic stagnation, German heritage, music/entertainment, cuisine/nutrition, value system, access to universities and hospitals, and other issues that the field teaches its students to simultaneously hold in mind while analyzing a place. These questions can tell us in what context are people associate with one another and what are the variables that drive these associations.

Weaknesses

Candidly, online mapping capabilities have diluted a bit of the clout of Geographic Information Systems, which seems to be an idea that very few in the field are out-rightly willing to discuss, or address, even in order to maintain a set of specialties. Instead the GIS community seems to continue to defend their tacit knowledge, access to perform powerful operations and create models in the GIS realm, and the capabilities of the field that that are more difficult (like the work of Noel Cressie in advanced spatial statistics)—although the later are still only digested by GIS analysts with high-level mathematical literacy. For instance, shortest path, geocoding and calculating travel times are already becoming trivial due to online resources. If the availability of reliable GIS layers increases, and tools for managing these layers (like buffer and intersect), GIS may have to re-polish its description of specialties.

In our experience, the perception of Geography as a discipline and geographers seem to be going through challenging times recently, and since the field encompasses so much, it is still having trouble defining itself. As one professor put it, "having a Department of Geography is like having a Department of Temporality." From this analogy we can see how the concept of Time must take a lot to understand—there are so many dimensions, scaling laws, et cetera, but its applications are virtually endless. Outside of the research community, to most non-geographers, Geography seems to be a 'thing of the past', and many mistake the field for Cartography, or evoke an archaic image of Magellan or Columbus charting new territory. Geographers continue to make significant contributions in peer-reviewed literature, conferences are well-attended, and the Advanced

Placement Geography program is sustained in many U.S. high schools. This perception surely has an effect on the level of solicitation that geographers experience from other, especially new, disciplines, even though successful "geo-interventions" have offered easy solutions to the problems that researchers in other fields sometimes see as frustrating or impossible.

With this in mind, we now turn a concerned eye to one prominent, if not sole piece of tandem quantitative analysis of a social network and a geographic plane, and then turn to the efforts of Computational Social Scientists and their potential contributions to the field.

Examining Current Efforts to Include Geography in Social Network Analysis

One of the first works that combines computational social network analysis with the agents' location in geographic space is a recent article that states that "The structure of dyadic social interactions is known to depend on geography, for example, as shown by the decay of friendship probability with distance." (Onnela, Arbesman et al 2011) Are there drawbacks from equating 'geography' with straight-line distance? A distance metric can only be applied to pairs at a time (ex. Jon is 5 km from Ken), where these researchers are looking to describe the relationship between multiple people with only straight line distance. In our example, Jon can be 5km from Ken and from Lenny, but we do not know if Ken and Lenny are neighbors for Jon to easily visit, or if Ken and Lenny are on opposite sides of Jon, indicating he must choose whom to visit, or that he is the 'hot spot' for their meeting point. Certainly these positions will yield different dynamics on the friendships. For those who specialize in the scaling laws that teach us that more of x gives us more of y but that more of x+1 yields much more of y, their 'geographies' do not incorporate any issues showing the increased strength or gravitational pull of a conglomeration of people of interest. Even though a visual of a two-dimensional geographic plane was used in this exercise, we wonder if the authors make the best decision about what constitutes a spatial cluster, as there would be different results under different geographic scopes. We also wonder if the two-dimensional plane, since it represents geographic space, has infrastructure or natural features; both of which, when included could change the results significantly.

Some promising work in this seems to be that of Sevtsuk and Mekonnen (2011) in computing network measures within urban form, who produced a new toolbar for the ArcGIS environment. However suitable for new calibrations of network distance, it may not be as useful for the integration of loosely-spatial social networks.

Suggestions for Operationalization

The research involved three dissertation case studies and opening chapters shows that sociological literature takes virtually no advantage of GIS capabilities and, location and place do not seem to be primary considerations in Computational Social Science. Of course this is not a "mistake", as the maturing field has its own rationales and, moreover, good research to show for its decisions, but we believe inattention may be a bit naïve, considering the copious evidence that place and environment directly affects quality of lives relationships, families, and health.

For GIS to be relevant to the field of Computational Social Science, and vice versa, it may be beneficial for GIS analysts to explain their specific capabilities, proposed role and value as collaborators in scientists' varying agendas. As opposed to the case with civil engineers, our conclusion is that it may not be as easy to convince social scientists to invest in GIS, because the synergy is less clear. To a civil or transportation engineer at a highway site, it becomes immediately evident that there are natural and built features surrounding the road, as well as traffic flows and mode choices. To a social scientist, these features are not as visible.

Therefore, we suggest that to foster dialog and integrated research between the two fields, GIS analysts may present themselves as a special type of statisticians to the sociology community for the following reasons. First, sociologists are often familiar with some statistics and many times understand what capabilities a statistical analysis can provide, but invariably seek the help of a trained statistician for more involved problems. Secondly, Statistics is like GIS in that some capabilities are accessible to anyone with basic math/technical skills: for example, (in statistics) an online bimodal distribution applet or a t value table, or (in GIS) free earth-view software that displays geographic data or a web service that geocodes addresses on demand. Although some geographers naturally fear the 'misuse' of these tools as used by the general public, there seems to be little apprehension by the statistics community that statistics applets and freeware is compromising the field of statistics, or the public's knowledge generated from these tools.

Third, GIS analysts can offer a similar set of services when collaborating with the social researcher, including critical statistics (ex. weighted mean of a point distribution) tests of statistical significance (ex. hot or cold spots, autocorrelation), probabilities and predictions (ex. temporal growth, buffer/intersect analysis), and tasks like entity distributions/ dispersion, and directional means. Like the statistician may find that with 95% confidence coffee intake increases the number of friendly conversations at a mixer by a factor of 2.0-2.5, the GIS analyst can make a non-trivial contribution to social data analysis, by finding that, for example, a community member that lives over 10 miles from the nearest high school has an activity space of 150% larger activity space. Each of these "findings", can aid in proving or disproving hypotheses, hunches or theories of a sociologist that may otherwise remain just educated guesses, observations or anecdotes. Finding a concrete answer that also responds to the question "by how much," gives the social scientist credibility, better proof, and a better idea of the actual effects of their perceived phenomenon on resulting human behavior.

Geographers have a good sense of spatial statistics like mapping clusters and hotspots, finding patterns in geographic space, understanding autocorrelation, performing geographically weighted regression, raster algebra, solving logistics and network problems, and finding descriptive data statistics like central features, mean centers and directional distributions. Also, remotely sensed data can be managed and extracted by GIS analysts in many different forms, in one example, analysts can capture real-time temporal changes, say of a crowd formation.

After the integration of transit models in GIS, the two fields could have a more natural dialog within the standardized world of rigorous research. These advancements could include formalizing terminology in journal articles, grant proposal guidelines, textbooks, classes, workshops, conferences, software, and user manuals. A book like Meyer and Miller's Urban Transportation Planning (2001) is a good example of how to integrate two fields with added value.

As raised in the challenges section, GIS and CSS have a problem of lack of lingua franca. The above considerations for standardizing names in the most communicated and trusted arenas (the academy and large software corporations). In addition, when exposing CSS researchers to GIS, we suggest that proponents of the integration when developing GIS concepts that are intrinsic to other fields (like 'betweenness' or 'modularity'), perhaps 'geo' or 'spatial' onto terms like clustering in order to differentiate from the meaning of these words in other fields.

To be thorough, we can offer insight into the opposite relationship, as to the willingness of GIS analysts to embrace sociology. Seemingly, social science's relevance in GIS may be a much easier question, as the literature in Geography and some Urban Planning shows that GIS analysts seem to invite a host of sociological players and actors on their layer 'stages,' including epidemiology, race/gender issues, class issues and accessibility via travel-time. Led by researchers like Mei-Po Kwan, whose work in social geography issues has ignited new thinking in the field, (2007) it is not a far-fetched hypothesis that GIS analysts in Urban Planning and other fields are amenable to, if not hungry for more robust data and knowledge about social phenomena.

THE ROLE OF THE COMPUTATIONAL SOCIAL SCIENTIST

What may be missing from Geography are topics like Social Capital, Social Relationships, Human Decisions, Sense of Community, Assimilation and other Social Forces. Some of these findings stem from the work of Arrow's Impossibility Theorem (1952) and Nash (for example, 1950), which were novel in modeling collective choice. The work of Arrow and Nash was insightful in understanding how humans are driven to optimal individual choices based on the stalemate of (or lack thereof) different alternatives. Indeed, much of theory building is calibrating micro behavior based on abstraction and extrapolating (via the model) to "explain" macro results. This reasoning is especially applicable to social science research because of the state of technology, as there is much data on microbehaviors that need careful analysis in order to scale to macrobehavior.

Some social science examples from which we learn more about collective and individual behaviors are, (as mentioned in the introductory chapter): Segregation or "not in my back yard" (Schelling 1978), The Strength of Weak Ties (Granovetter 1973), Power Relationships (Blau 1986), Small World properties, (Watts and Strogatz 1998), Social capital (Burt 1998, Fernandez et al 2000), Preferential attachment (Barabasi, Albert 2000), Homophily (McPherson et al 2001), Community structure (Girvan and Newman 2002), Social network resilience (Albert 2000, Ebel et al 2002, Watts 2002), Information Spreading (Granovetter 2003), Teamwork and Incentives (Kearns et al 2006), Influences of smoking and obesity (Christakis and Fowler 2007, 2008), Trust propagation (Watts and Dodds 2007), Innovation (Bettencourt et al 2007) Predicting future friendships (Clauset et al 2008) Pluralistic Ignorance (Centola and Wilier 2009), Strength of Weak Ties (Arbesman et al 2008) and Prisoner's Dilemma and trust (Salganik and Watts 2008). Other models include those that can explain the social variables associated with suicide, domestic violence, fads, eating disorders, bilingualism, and voting. Since each of these phenomena affect and occur in places, and society is connected to these places, GIS models that can incorporate these variables can be of high value.

In particular, social networks and models are special because they have autonomous agents making decisions, so the micro behavior itself is sociologically and psychologically complex. Moreover, the scientists involved in computational social network analysis are prepared to deal with high-level complexities, mathematics and modeling techniques. There is much tacit knowledge to characterizing and drawing information from social network data that can be used to enrich geographic information, like explaining why people cluster in certain areas, or why migration patterns follow certain trajectories. When conducting the 'trip chaining' research in Chapter 4 of the dissertation, it became apparent that in previous research, geographers were concerned with variables of population size, distance and income to predict to where migrants would migrate. However, while conducting this research under the notion that sociological theory, trip chaining in particular, could be driving the dynamics of the system of inter-city migration, the results correlated strongly with the previous findings of the computational social scientists. Without the integration of CSS explanations of human behavior, the geographic patterns in Chapter 5 may still rely on the traditional economic geography models to explain variation within the system. Therefore, we need both the spatial and social systems in order to predict or explain which migration patterns peak or die off at any given time.

Next, GIS models do not have a particularly strong mechanism for modeling how human agents relate to one another on a person-to-person level. Since Urban Planning is in some part a service to help quality of life, and to cater to the social, health and professional needs of humans, it would make sense that planners should have access to social network analysis (SNA) data models and findings.

The field of SNA is moving quickly and has become an increasingly widespread field since roughly the beginning of the millennium. Social networks have come of age through a rich set of properties that have been well matched to the data of the decade, as captured mostly from cell phones and the Internet. Our first two chapters show that current researchers most involved in the progress and propagation of this field have applied physics, biology or health backgrounds, although researchers from these subjects certainly do not comprise the whole SNA research community. From conducting research in this dissertation, we find that SNA researchers seem to be distinguished by strong computational and mathematical skills. As a result, the field has seen rigorous, prolific research, much of which has seen its way into highly regarded publications almost in a constant stream. Perhaps due to the novelty of the datasets, strong sense of empiricism and problem framing, or the conclusive results, this field has been admired and deservedly blessed with a great deal of exposure. Given this electric atmosphere, what in this field could be best applied to and operationalized in the new social/spatial realm?

The field of Social Network Analysis (SNA) is the type of Computational Social Science methods/ models/metrics that we are most concerned with at this juncture, as the field seems to be currently dominated by relationships models as conceptualized as networks or graphs.

Dexterity in the Graph Domain

The expertise of graph theory (or network) concepts can be very helpful, as most geographers (even those dealing with transit networks) do not have exposure to the kinds of descriptive metrics employed on graphs. For example, structural properties of social networks are marked with: markedly more edges than nodes, branching, clusters and bridges (less likely are giant components, trees, chains, and cycles) allowing for social flows and connectivity measures. SNA researchers know how to interpret graphs, characterize graphs, and have a trained eye for noting special properties of some visualized configurations. This is a time-saving tactic for a beginner, because the more senior or expert analyst can suggest the type of degree distribution and diffusion properties that are important for understanding the system dynamics. Trained experts can suggest what parts of the data could be considered 'noise', how to create and interpret meaningful partitions and sub graphs, and how to describe the system holistically. Experts also exhibit tacit knowledge in the smaller scope of the network: a researcher can elicit the nodes that play key roles, through centrality, popularity or important roles, or which links are integral to the structure, susceptible to attack, and highly traversed. Specific metrics can be found in (Jackson 2008).

Concepts

Just as geographers have concepts like "Distance Decay", which can be marked with calculated parameters, SNA concepts include Scale Free, Small World, and Random Networks. In a Small world network, the presence of even just a few long edges makes for much easier, fluid, shorter, traversal through the system. (Watts and Strogatz 1998) The Scale Free network has been tied to the mentality of "preferential attachment", where high-degree nodes are most attractive. Here, many nodes have a small degree, and few nodes have high degree values. (Barabasi and Albert 1999, Albert et al 2000) The Random Graph, or Configuration Model, has also been used as a null hypothesis or variable manipulation technique, as these graphs are construction to configure without specific node-edge blueprints. (Erdos and Renyi 1959) These concepts are well-known to the SNA community as they drive much of the structure of human relationships. GIS Analysts and Geographers interested in social/spatial research and Social Distance could benefit from learning the large-scale organization, or macro-behavior of micro-decisions.

Prediction

Like geographers have prediction models, probabilities for forthcoming behaviors are found in network science, for example "future friendships" can be discovered under the conditions that a pair of nodes has 1) many common neighbors, 2) many short paths between them, or if 3) the product of their degrees is large. (Clauset et al 2008) Unlike expansion in geography, network models in some domains, like the Internet, often lend themselves to "redundant wiring" (Albert and Barabasi 2000), where current relationships and popularity may be further augmented with future joining agents.

Weaknesses

Describing the advent of Computational Social Science, Lazar et al (2008) write: "A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors." These behaviors have certainly been uncovered in that we now have insight to what large-scale behavior 'looks like' and some interpretation of this behavior. These findings have been shared outside the research community to reach massive audiences. The recent work of few popular authors and their book(s) includes those of Nicholas Christakis and James Fowler (Connected), Albert-László Barabasi (Bursts, Linked), Duncan Watts (Six Degrees) and Len Fisher (The Perfect Swarm), and even the more pop culture writings like that of Malcolm Gladwell (The Tipping Point, Blink) discusses how we are all "connected," how and cite beautiful scaling laws for the complex systems that involve humans.

The notion of social science as translated to the masses through literature in the 2000's seems to focus on uncovering powerful universal laws that are purported to be magnificently surprising and applicable to 'everything':

Connected: The surprising power of our social networks and how they shape our lives (2009)
Linked: How everything is connected to everything else and what it means (2003)
Bursts: The hidden pattern behind everything we do (2010)
Six degrees: The science of a connected age (2004)
The perfect swarm: The science of complexity in everyday life (2009)
Blink: The power of thinking without thinking (2005)
The tipping point: How little things can make a big difference (2000)

Many journal articles come to the same conclusions, often resulting in a log-log representation of a scaling law between network agents' variables. In their discussion sections, the novelty of the methodology is rehashed, and sometimes discussion for future research or applications to other enormous datasets is mentioned.

Again, studies of this design are almost fluidly permeating if not saturating computational literature. It is still unclear, however, if these studies are helping people, and how they are applied to bring about positive change. The articles do not appear to be as concerned with the outcome of their research in terms of altruistic concerns, and if so, there is little insight as to how this may be realized. Perhaps those familiar with the process of model, to finding, to policy implementation in GIS research may help some future network concepts in the Social Distance field reach fruition in terms of changing behavior for the better.

Each field can also benefit from one another in terms of how to manage large datasets and make helpful visualizations. When considering social and spatial as standalone systems, there is much evidence for the successful modeling of social processes with behavioral, agent-based, game-driven or discrete choice models, and geographic processes with a proven toolbox of spatial statistics and agent-based processes such as the work of Batty (2007) in cellular automata and other modeling capabilities. But to model social and spatial as a system of systems, where social/cognitive choices are represented spatially, current approaches seem to fall short.

When studying migration in our second case-study in Chapter 4, we find that geographers tend to create gravity models instead of turning to sociological literature. Thus, we recommend that geographers modeling phenomena like migration, telephone calls, and socially-focused activity improve their efforts to consult the findings of sociologists in order to model reasons for geospatial activity. Since there are few socially predicted flow methods, we suggest that predictors use current evidence. In the short term: Follow individual trends and cohorts, predict patronization this way. (data from GPS, cell phones, sociology, 'common' knowledge) In the long term: Treat place-to-place pairs as unique, as current and past relationships seem to be able to predict new conditional and posterior probabilities. For example, the concept of Trip Chaining has been documented by researchers like Greenwood, Castells, Rojas, Sassen, and Carrington (as cited in previous chapters). In the future, international monetary remittances might be a good way to understand geographic patterns of labor flow and migration, as following the sociological theory that workers send funds home to provide for their families in more deprived economic circumstances.

Next, we visit some technical and computational improvements with social flow data, based on a priming discussion of the author's response to the three case study chapters.

6.3 Experiences with Case Study Chapters

We now reflect on our experience with three examples of the type of research that could feed into the field of social/spatial systems and the geographic analysis of social flows, focusing on the convergence of the social and spatial systems and recognizing the benefits of the complex systems that are available, and not easily translated from one to the other.

At an initial glance it may seem that the three case studies may have little in common. After all, we first discuss the unsupervised classification of varying graph structure 'star' forms, and how this process can help us better typecast or categorize places based on their geometric activity. The second chapter tries to predict future migration patterns in five different ways, and relies mostly on linear extrapolation, a non-linear gravity model and Bayesian probability, and finds that cities themselves have seemingly steady relationships over time. These inter-city relationship often depend on the seeding of connections between the two places (which can spike or taper off over time) more than the sheer distance between cities, although colinearity is not surprising, as there is an implicit relationship between the seeding and distance. The third chapter analyzes U.S. Congressmen and their ideological friendships with one another. We find that each agent has his or her own unique 'layer' of the United States, and that these topologies would not have been evident without the geo-implementation a social network and its dynamics. We also find that tangential relationships in a social sphere are not clearly realized as closeness or clusters in the geographic topology.

What could possibly be a unifying theme that connects spatial 'stars', migration and Congressmen? The reasoning that draws the three case study chapters together in this thesis is that social behavior is underutilized by GIS analysts and Urban Planners. Since a model is an abstraction of reality, and we want to simulate a purposeful simulation that captures real phenomena, we confront this underutilization by suggesting a new model of systems of systems in order to take advantage of both spheres through three research case study chapters. Next, we revisit the three chapters, and how they were able to tackle the major challenges listed in Chapter 2. (Table 2)

TABLE II
CHALLENGES ADDRESSED BY CHAPTER CASE STUDIES

	Chapter Name	***Key Challenge Addressed***	**Secondary Challenges Addressed**	
3	Visualizing Migration Dynamics Using Weighted Radial Variation	Muddled **visualization**	Difficulty characterizing system **nodes**	Inability to employ **spatial analytic methods**
4	Predicting Migration System Dynamics with Conditional and Prior Probabilities	Few socially-driven flow **prediction** methods	Inability to employ **spatial analytic methods**	
5	Social and Spatial Patterns in the U.S. House of Representatives	Lack of Visual-Analytic Systems for Exploratory Spatial Data Analysis	Muddled **visualization**	Inability to employ **spatial analytic methods**

Now, we make more formal suggestions for making this field available and useful to other researchers. We want to first leverage knowledge by asking critically, what can geographers really contribute to this field, and what can computational social scientists / network scientists contribute? What is the atmosphere between these two researchers? Based on our experiences, we now share our suggestions on priorities that would facilitate and make new research easier to conduct. We then address how to disseminate and formalize the information through pedagogy, and also how it should be approached through theoretical avenues of questioning (epistemology) and also modeling (abstraction).

Our first chapter, *Visualizing Migration Dynamics Using Weighted Radial Variation* most broadly addresses the issue of muddled visualization in a geographic flow system. Seeing that the mapping of human county-to-county migration flows in 2008 in the Continental U.S. overlapped to the point of undetectable patterns, a classification method was applied so that each county had one single variable, instead of up to over 1000 vectors each. This chapter also helped defined system nodes, because each value of incoming migration flows (e.g. each vector's direction, distance and magnitude) was included in the classification process. This process avoided the alternative of classifying nodes by their summary features, such as total number of neighbors or average incoming migrants per origin.

This processes also broached the issue of inability to employ spatial analytic methods, as although places in the flow system are not adjacent to one another, the k-means unsupervised classification method was successful in treating the counties as nominal entities and categorizing them aspatially. Future work in this area is to compare classifiers (like SOM) and compare results with a 'null hypothesis' of categories—as derived from place typology datasets from *ESRI's* Community Tapestry, *Claritas* or index via Census data. Another important next step in this type of analysis is to find a way to incorporate the actual pair IDs of flow system 'friends' (e.g. two counties that connect) so that second degree neighbors of a node in focus can be traced, and perhaps these findings may show new patterns and geographic directions of connectivity. Additionally, researchers may be interested in a better way to address how to mitigate the number and type of possible geometric star configurations with the underlying geometry of the land. For example, star patterns centered around Miami, Florida have a specific North and Northwest fan, and thus fall into a category of places with a 'limited attraction' power, e.g. a lower variety of places with incoming migrants. In reality, Miami attracts from many diverse places, and thus, its limited radial compass flow geometry should be tempered with the magnitude and cardinality of flow origins.

Difficulties surrounding this chapter included choosing a classifier, choosing an array schema, and interpreting the results. The final difficulty should be taken with special care, as a researcher can often "explain" wrong data or meaningless results. Also, attaching meaning to each of the graph patterns was challenging, as the difference between cities that draw a two, three or five-spoke migration graph may be great, slight, or non-existent, but it is hard to know. Reflecting upon this research, the processes governing the migrant system in *Visualizing Migration Dynamics Using Weighted Radial Variation* were challenging to address, as the method conceptualized migrant flows as a geometric objects. One successful result was the newfound ability to find similar and dissimilar places by their unique origin-drawing power signals. We may now ask, why do we want this information? We find that some cities (like Baltimore and Pittsburgh) are similar and others are (like Baltimore and Washington D.C.) different. Pittsburgh's recent economic decline, in juxtaposition to Washington D.C.'s economic growth, would certainly affect its ability to draw migrants, and a similar graph in Baltimore may be an early indication of financial instability. We also see that suburbs play different roles in contributing migrants or drawing migrants from their interior urban cores. This finding may help metropolitan planning organizations better understand the social dynamics and connectivity between separate components of the city, where traditional methods usually rely on average income, race and other demographic factors to typecast suburbs.

Proof for future economic prediction can be tenuous, but nonetheless, this geographic, and geometric, evidence could provide supporting evidence to bolster purely numeric predictions. To obtain this evidence, a new data mining classification process was applied to geographic space, and although it certainly has drawbacks (most notably the current lack of cross-validation data from which to compare our findings), its unique approach to managing flow data may serve as a platform for flow mining advances and spark new ideas in the Spatial Data Mining and Urban Planning fields. It is important that these methods and subsequent results are tested for their reliability and usefulness in the research and professional spheres, as unsupervised methods can easily divide similar phenomena, and group dissimilar phenomena without leaving indicators of mismatch.

Next, *Predicting Migration System Dynamics with Conditional and Prior Probabilities* pushes against the problem of few socially-driven flow prediction methods in Urban Planning by modeling unique city-to-city migration probabilities instead of predicting flow volume as a function of characteristics of origin and/or destination cities. To predict the unique city-to-city transfer probabilities, this chapter employs the methods of Bayesian probability modeling and linear extrapolation. The motivation for modeling migration in this way draws from research in migrant chaining, where potential migrants most often move to a destination location due to the social contacts they have in that location. Given that, most likely, the social contacts in the destination were once from the potential migrant's origin, their travels are accounted for in data from years past. Data that shows that two places have a steady conditional and posterior probability of hosting unique O-D flows between them would support the trip-chaining model over other economic models, which use income and job availability at origins and destinations, or other push and pull factors to predict the number of migrants between two places.

As mentioned in the challenges section, many flow models pose a problem of geographic adjacency, wherein spatial statistics rely on geographic adjacency, this model was able to use a non-traditional spatial probability method, where the cities in the model were no more than nominal entities. More importantly, this model is "social" because it takes a human decision to move to locales of specific character into account, and the synergy of city-to-city pairs, instead of the sheer distance and population—not social choices—used in the traditional gravity model.

Reflecting on *Predicting Migration System Dynamics with Conditional and Prior Probabilities,* there is not enough empirical evidence to conclude that movers change location due to social chaining, but that if social chaining was the main reason, there is evidence to support this hypothesis. Proponents of the economic model may argue that their models are laden with variables so that the model with the best fit is automatically equipped to explain *why* the migration occurs (e.g. because of more jobs at destination or better weather). However, it does not seem that these variables are as important as the promise of a stable friend, family member or social network at a potential destination. This may tell us more about human nature, reasons why we take risks, our value system, trust, willingness to help, and our perception about the relative richness of place. We can guess that the potential dangers or fears of uprooting may be doused by trusted friends, and that the tabulated prosperity of a city is not as much of a pull factor as the proof of even a single prosperous contact, or the potential of life together. The fact, additionally, that the posterior probabilities (given that a mover has moved to D, what is the probability that he is from O) were more stable than the conditional probabilities, was an interesting finding for planning purposes, as a city mayor can be confident that his incoming migrant cohort will be comprised of citizens from virtually the same percentage of each city each year. This reliability may help a metropolitan planning organization better host incoming citizens.

Moving forward, we may want to compare this migration phenomena to that of commuting. Without the permanence moving homes, the commuting phenomena may not need the blessing of friends to be realized in geographic space. It would also be interesting to see whether phone calls, e-mails or web searches could predict migration. As for scope, looking at migration on an international level, by perhaps tracing remittances, the migration dynamics of the global economy may be just as predictable as the inter-city dynamics seen here.

Finally, the chapter *Social and Spatial Patterns in the U.S. House of Representatives* breaks new ground by fusing the traditionally polarized methods of social network analysis and spatial analysis via a loosely coupled visual-analytic system in which to explore spatial, social and demographic patterns within the 'friendship' system of the U.S. Congress. This experiment shows that the characteristics of network agents, like their attachment behavior, embeddedness within the community, and relative 'importance' or popularity in the whole system, can be linked with other factors like geographic situation, and demographic features of a congressperson's constituency. Here, we look at economic, racial, density and age factors, and geographic clustering measures. This experiment's results, namely its statistics, produce clearer visualizations than would be possible by overlaying the social network directly on the map (as the social network is a tight nest of nodes and edges when visualized). Also, although the social network shows the connection of people who represent adjacent space, their connections (via the social network) tie non-adjacent entities together, which poses the same problem that most flow data imposes on spatial statistics. This chapter also theoretically touches upon the issues of underdeveloped theories of edge assignment, as a "representative network", as shown here, is not typical for the field of GIS, and when edges between the social network are imposed on a map, this creates new meaning for geographic connectivity.

Just as *Social and Spatial Patterns in the U.S. House of Representatives* only touches upon undeveloped theories of edge assignment, two additional challenges from chapter two are not overtly addressed in the three chapters. These include: a lack of GIS infrastructure for node-edge-node storage as a singular entity, and difficulty developing a lingua franca and taxonomies for the field. For the former, a technical approach may be needed to set up a database infrastructure to hold and symbolize multi-structured objects. The definition and instantiation of a new object "the flow" may be more trivial than its implementation and compatibility with current GIS storage and retrieval infrastructure, software and sharing. It would be prudent to conduct a case study with "real" data in order to demonstrate the benefits of using the "flow" object in terms of analysis, interactive spatial selection, joining, table, statistical and SQL operations and discover benefits of these new operations in an urban planning or spatial decision support scenario. In order to address the difficulty developing a lingua franca for the field, it may be best to survey experts in both domains, arrive at semantic roadblocks and agreements, and keep these findings at hand when writing for the convergent communities. Through the standardization of lingua through textbooks and glossaries, scientists and practitioners will have a reference for assistance. Also, on a faster temporal scale, conference proceedings and journal articles may look to use the same encyclopedia of terms, in order to propagate the standardization of ontologies and semantics in the field.

In conclusion, we attempted to characterize system nodes not by one feature, but by their unique signatures. This was generally successful, but not without room for improvement. The three chapters did not make great strides with respect to differentiating between when to represent a flow as origin destination or a physical trace. In fact, Chapter 5 may have "wrongly" tied friendships together over a map, indicating the that states in between were involved in the friendship, but this did not inhibit the map's message. This issue may be more pressing in other studies. The inability to employ spatial analytic methods also continues to remain a problem, but with solutions abound. Since spatial analysis depends mostly on adjacency or proximity, a table of linked network elements can be considered 'proximal' or 'adjacent', and these statistical methods can be employed as such. The lack of prediction methods of socially-driven flows was addressed in Chapter 4 with a promising outcome. Although theories of migration chaining were not proven, we find strong evidence that would fit the hypothesis. New methods in this category may grow organically, where the straying from traditional economic models to place-based models could be realized through a data driven approach. Just as our tandem model approach seems to be best conceptualized as a "system of systems". Issues of Standardization were not a major issue in these chapters. When describing cross-cutting words like 'topology' or 'clustering', which apply to both worlds, the definition was simply clarified for disambiguation. However, when teaching these methods to others, it would be important to distinguish between terms that exist in both worlds.

From these case study experiences, we find new meaning to what it really means to 'combine' social and spatial, and how important it is to find the right interaction between components of both worlds. In other words, combining the social system with the geographic system in a meaningful way (e.g. the system of systems), so that parts of the system dynamics can surface to explain certain phenomena requires not just understanding each system on its own and applying its respective dictums to the same problem, but perhaps 'convolving' the system variables in order to consider the right combination of intermixed variables to explain behavior. Indeed, the social and spatial components and systems need to talk to each other in order to track desire, action and then movement, and furthermore, to do all this with sensory input from the environment and cost-benefit factors. Although we find here that this dialog can resemble the 'Chicken or the Egg' paradox, we believe this confusion further vindicates that the systems may be best viewed as systems that are in intertwined in constant feedback loop that communicates spatio-temporal factors that enable and constrain social relationships in geographic space.

In an analogy, neuroscientists study the brain system, and physiologists versed in kinesiology study body movement. And so to understand why or how the body moves, one might not be successful by seeking out only the separate theories of neuroscience with the theories of physiology, but the intricate ways in which the brain decides to signal the body, and how the body responds within its own limits of movement, and also senses its environment to signal information back to the brain. This feedback loop is an example of a system of systems that may draw an even deeper connection to the social/spatial convergence of systems: We can consider space as a somewhat constrained corpus. Like the body, which cannot grow an extra limb or become giant, the politically and physical geography earth imposes similar limitations where we must 'work within boundaries'. Conversely, like the brain's ethereal directives whose power is most often realized only with the manipulation of the corpus, we can consider communication, relationships and decisions as the action points of social interaction which are mostly realized in tandem with geographic space.

With this in mind, we now give some recommendations for improvements with social flow data types, structures, software and computation.

6.4 Technical and Computational Improvements with Social Flow Data

The section on technical and computational improvements is split into a few subsections, based on more specific guidelines. We realize that the building block of the social distance world is the "social flow", and so we focus first on the treatment of flow data.

Data Distributions and Data Types

Data distributions (for example, the distribution of miles in which people travel to attend a concert) attached to spatial entities are a challenge in GIS (and non-spatial data) that has been approached before. The U.S. Census manages distributions, for example, the age distribution of citizens in a tabulation area, by counting the number of people at 5-year increments within the area, and storing the values as separate scalar characteristics.

In Chapter 3, the distributions of interest are: Distribution of edge distances (straight line, or travel time), distribution of edge angles, and distribution of edge magnitudes, the latter of which can have many magnitude values of interest (e.g. incoming and outgoing migration as two separate magnitudes, or a subset of the number of migrants of a certain age, etc.) When we treated these distributions as vector arrays, we were able to classify node places, and partially solve issues of muddled visualization and characterizing system nodes. However, some issues arise, for example: how many classes to choose, which classification method is best, and how do we know if the classes represent the place groups in a way that is meaningful to planners.

We suggest following the work of data mining and pattern recognition specialists, who have developed bootstrapping, class splitting and joining methods to permute the sets and re-test for goodness of fit within the class. Regarding if results are meaningful to planners, it might be interesting to cross-reference and compare with other datasets from companies like *Claritas* or ESRI's Community Tapestry.

In addition to visualization and place characterization, it should also be taught that flow data needs to be handled with a vigilant eye, as a flow itself is a series of endpoints until meaning is ascribed to the flow. As mentioned as one of the eight problems in Chapter 2, some flows are traces where the traveller interacts directly with his environment, like a pedestrian, whereas a communication flow could travel from the same beginning and end points with virtually no interaction with the ground beneath the flow line. (Figure 6.1)

To remedy this problem we should add geodatabase options like “flow type”, where a user could input the flow ‘type’, and the GIS could give the flows certain auto characteristics, like it does now for “topology” considerations.

For communications, the lines can automatically draw to the background, and perhaps be given a certain transparency level, and not interact with the physical trajectories. Further, friendships and social relationships that span geographic space can be visualized or stored as a physical 'tie' between the landing spots of the 'friends.' In Chapter 5, we see this phenomenon when we find the biggest 'enemies' in Congress. (Figure 6.2)

Similarly, for flights, lines can travel through the surface, but not intersect with the roads or underlying infrastructure. This might be helpful for spatial selection and computation, as the length of the call or flight will be primarily used for statistics instead of physically-bound values like tortuousity or buffer statistics. For large datasets, subsetting the lines for a sample that can be drawn quickly and more easily visualized might be a popular automatic option, just like 'Building Pyramids' is a suggestion for large raster images in some software platforms.

Train or subway data might be treated differently, as there are usually many stops between origin and destination. When indicating that the flow is part of a longer system of nodes and links, the system will recognize that the links (train tracks) themselves between the stations are not as important as the stations themselves for social flow, unless used in a terrain or CAD type planning project. The trajectories of buses might fall into this class of flows.

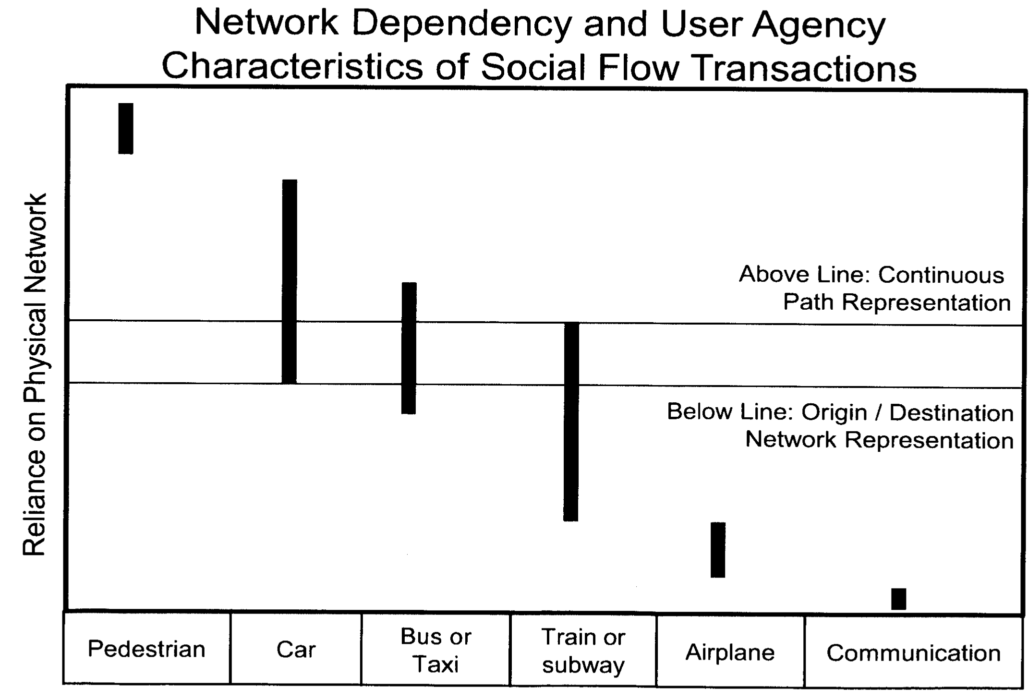


Figure 6.1: Network Dependency and User Agency Characteristics of Social Flow Transactions **This figure shows the scale of trajectory types over spectrum of embededdness in surrounding built environment.**

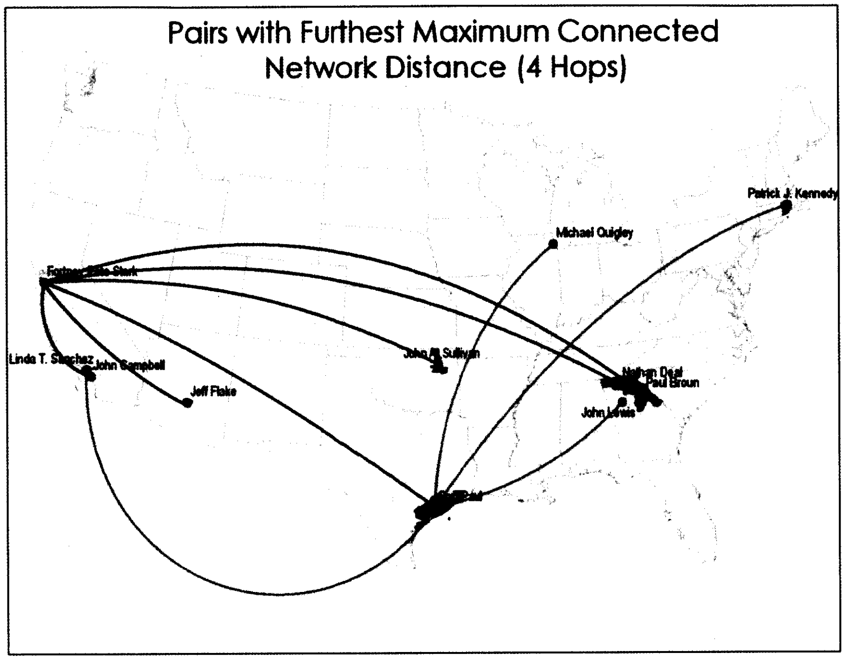


Figure 6.2: Pairs with Furthest Maximum Connected Network Distance **Congressional pairs are shown by red and blue lines, as these lines are overlaid over geographic space. However, the colocation of each trajectory with the space beneath it is not necessarily relevant, as the network connector is completely decoupled from geographic space.**

For cars, pedestrians, bicyclists, some taxis and buses, the trajectory of the flow should be integrated more deeply into the layers of the GIS. These flows should be automatically layered on top of other flow data to emphasize integration with the surrounding environment. Origin and destinations are important parts of these flow entities, but the actual trace is of note as well, and it should be clear that a user is susceptible to the flow space in which he or she travels more so than other modes, meaning that everything from toxic release to minute navigation decisions should be considered more primarily for this class than for communications, flight, train or subway infrastructure.

These decisions and provisions of each class of flow are not necessarily intuitive. In order to best know how to combine layers for meaningful answers to place questions, how to model flows, transit, and cost-systems over geographic space, what factors enable movement, and what are the lag times of variables like hills or bridges, we should look to geographers well versed in these concepts to contribute descriptive and analytical concepts in the form of tools and knowledge.

Data Structures

After evaluating node features, link meaning and characteristics, we discussed the need for better data manipulation and recording, but not particularly the data structures involved. The GIS approach relies on point-patterns, clustering and layering of different geographic phenomena and proximity to these phenomena, but not necessarily flows. Manipulating these flows, in spatial groups or individually, is not an easy task at all times. Spatial joins and spatial selections rely on proximity, but not all flows that are collocated with the layer beneath the flow should be associated with that layer. Selection could also be based on strength of connection, connection density, or the interaction of the human with the environment, which is usually tied closely with model of travel.

Some spatial selections involve more than 3 tables joined: For example, if the user wants to select the cities in Nebraska that flow to Seattle as a second degree neighbor. Also, if the user wants to find the trajectories that connect any place within a buffer of a road, with a place in a California county, she should be able to select the points of interest interactively on a map, and automatically be able to find the links that connect them. If a number of trajectories are highlighted on the map, and the origins and destinations that host the trajectory need to be selected, an automatic 'select node' would be helpful, as an intersect function to pick up any intersecting nodes may select underlying nodes that are collocated with the line, but not attached to it.

Selection commands that might be helpful: "Is a node of", "Is within a distance of a node of", "Is an endpoint of", "Is a 2nd (3rd...) degree neighbor of", "Is an edge of", "Are connected by", "Connects". Some commands that may not have as much use in a network system are "Crosses the path of", "contains" or simply "intersects" as the intersection may not be meaningful in some domains. As we can see from Figure 6.3, some of the trajectories intersect with underlying states, but have no relationship with the state itself, except for that it lies between the two pairs of Congresspersons, who interact mostly in an altogether different location: Washington D.C.

Tables should be able to join one another spatially based on the aforementioned selection commands. For example, if one wanted to select the trajectories that pass through New Jersey but do not have a node endpoint in New Jersey, he does not have this option with the current spatial join unless subsetting the trajectories, re-selecting, table joining with corresponding edge nodes, re-subsetting based on these nodes, and re-joining based on the table. Another factor that would help is if nodes and edges could be stored in the same table for easier manipulation.

In this light, based on data engineering in Chapters 3 and 5, we make four suggestions: (1) More diverse spatial selection terms, (2) Smarter Spatial Joins (join only IF….), (3) More intuitive multi-table selections: selection will highlight both tables and (4) Ability to View different data types in one table for sorting and selection.

Suggestions for Software Development

The development team for GIS/CSS software should be primarily comprised of GIS analysts and computer programmers, who have a sense of visualization, human-user interaction, file standards, and software design. Specifically, geographers and sociologists may be the researchers who test the software for clarity (ease of use) and accuracy (reliable results, based on previous findings). A tradeoff from developing the software this way

The results from the challenges section and three dissertation chapters (focusing primarily on Chapter 5) is that there are insufficient tools for good GIS/CSS analysis. (Figures 6.3 and 6.4) When we tried to select parts of our network that aligned with a geographic selection, tables of numbers had to be exported as a text file and then re-selected between programs. This lack of interoperability between the GIS and CSS (social network analysis) programs significantly reduced speed and functionality and even more important, the interactive capabilities—like dynamic subsetting—that give power to the exploratory analysis.

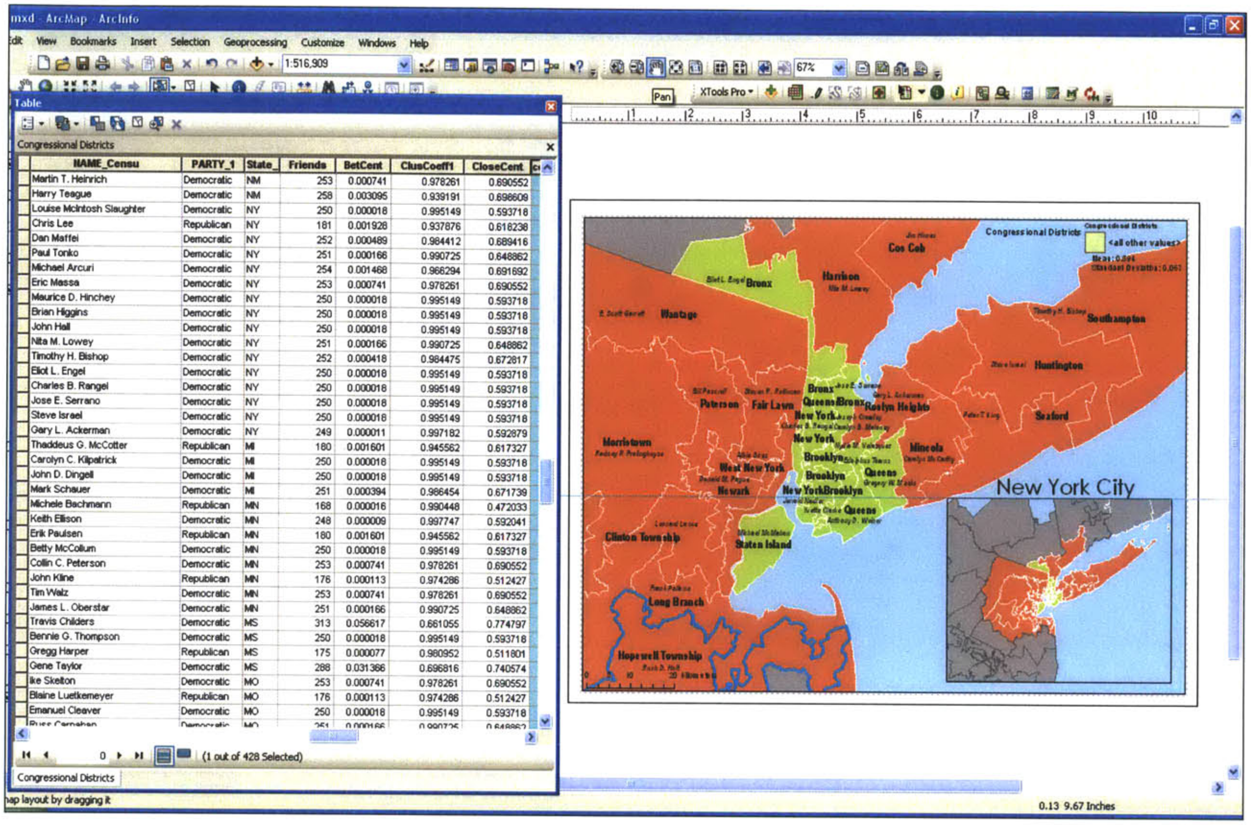
Adding capabilities to familiar software packages is an important component for operationalizing social flows. It is understandable for software to lag behind research in terms of providing comprehensive and consistent packages, but the availability of individual tools should not be compromised. (Unwin and Unwin 1998) Currently, the cost of using a social/spatial system is quite high, mostly due to lack of computational techniques and lack of computational environments and software. Nedovic-Budic (2000) points out two problems in integrating GIS into planning organizations: establishing systems and realizing expected benefits. Additionally, GIS users in planning don't seem to take advantage of sophisticated analyses, but use GIS mostly for data processing and mapping. Researchers need better tools to feel comfortable putting social flow analysis and network / graph theoretical analysis, and modeling human decisions and relationships under this heading.

To address this issue, we look more closely at Chapter 5: "Social and Spatial Patterns in the U.S. House of Representatives", as it begs most fervently for an interactive exploratory system, since the datasets are large and many patterns can be explored. According to Davis et al (1993), a knowledge representation fills two roles that can be addressed under the heading of a lack of visual-analytic systems In Chapter 5, we find that a visual-analytical systems that includes direct manipulation of discrete variables in the social and spatial datasets would have aided the exploration needed to find patterns in the network and map datasets. We believe that a stand-alone tool, with a java-based component map and social network component would solve this problem. Other, less favored options for creating this system could be using *Google's Google Maps* with *IBM's Many Eyes* visualization tools for matrices. These programs could be joined with a tandem infrastructure that allows for drop down menus and interactive map selection. A program like R could be useful because it imports some map controls and GIS capabilities, while allowing for networks to sit in the same platform and view. Another benefit of the Open-Source R is that it allows for many data types to communicate with one another. A software program like *MATLAB* is strong in linear algebra and could handle large datasets, matrix transformations and operations, but is not as strong in geography or dynamic networks. For these, we might enlist, as mentioned, *Pajek, Gephi, UCINet* or a similar program.

Following the capabilities of packages in the research community like *UCINet* (Borgatti et al 2002) and Pajek (Batagelj and Mrvar 2003), we suggest that the tools, operations and computational abilities, should be paired with spatial statistics for use in one environment. Statistical components should be readily available in these systems, if not calculated beforehand. From the network science community, computational methods of *community detection* such as cliques, hierarchical clusters, modularity measures; *popularity* measures such as degree centrality, betweenness centrality, flow betweenness, and hubs & authorities; *spreading processes* such as time-step propagation, resistance to percolation, cascades and rule-based diffusion; and *attachment behavior* such as homophily and clustering coefficients. (Jackson 2008) To compute the statistical significance of spatial patterns, the spatial position of competing or cooperating districts, their geographic adjacency, proximity and propensity to form cohesive regions, cluster detection measured by *Moran's I* (Moran 1950) or *Geary's C* (Geary 1954) Hot & Cold Spot Detection (Cressie 1992), and LISA (Local Indicators of Spatial Autocorrelation) (Anselin 1995). In addition to measuring proximal regions, demographic (feature) clustering shows the correlation between certain social features and U.S. Census information like of Income, Urban Areas, Racial Percentages, so social network agents can sometimes be associated with demographic features, with precaution for assuming ecological fallacy and inaccurate variable correlations.

For the Congressional System, we recommend a system of nodes in a force-directed social network that also corresponds (either one to one, many to one, one to many, or many to many) to projected district map. Currently in a GIS platform like *ArcGIS,* it is challenging to see how the Congresspersons interact with one another. As we view social flows as a "system of systems", interactive selection of visual patterns can relate one system to another. For example, selecting a certain group of agents in the feature space of a force-directed network can highlight their corresponding locations on a map. Conversely, selections of regions on a map could yield a pattern of clustering or dispersion within the force-directed network.

In Chapter 5, the lack of interactive selection seriously hindered the ability of the user to find a relationship between the social system and the geographic system. We suggest attention to the literature in the geovisualization community (as outlined in Chatper 2), especially the work of MacEachren, Anselin, Tobler and Fotheringham, and their successes in ESDA through statistical plots and calculations, maps, symbology, zooming, panning, scrolling, focusing, interactive selecting and "linking and brushing." We also support efforts of software like the *GeoVISTA Studio, GWR, FlowMapper* and *GeoDA*, and suggest that in this tradition, interactive dynamic environments be created for users to explore social and spatial configurations, and statistical properties in a single view. With their many windows of plots, graphs and exploratory techniques (like Parallel Coordinate Plots and Star Glyphs) along with rich map capabilities, these platforms seem to be the best examples of the systems that will bring social flows successfully into the hands of sociologists and GIS analysts alike.

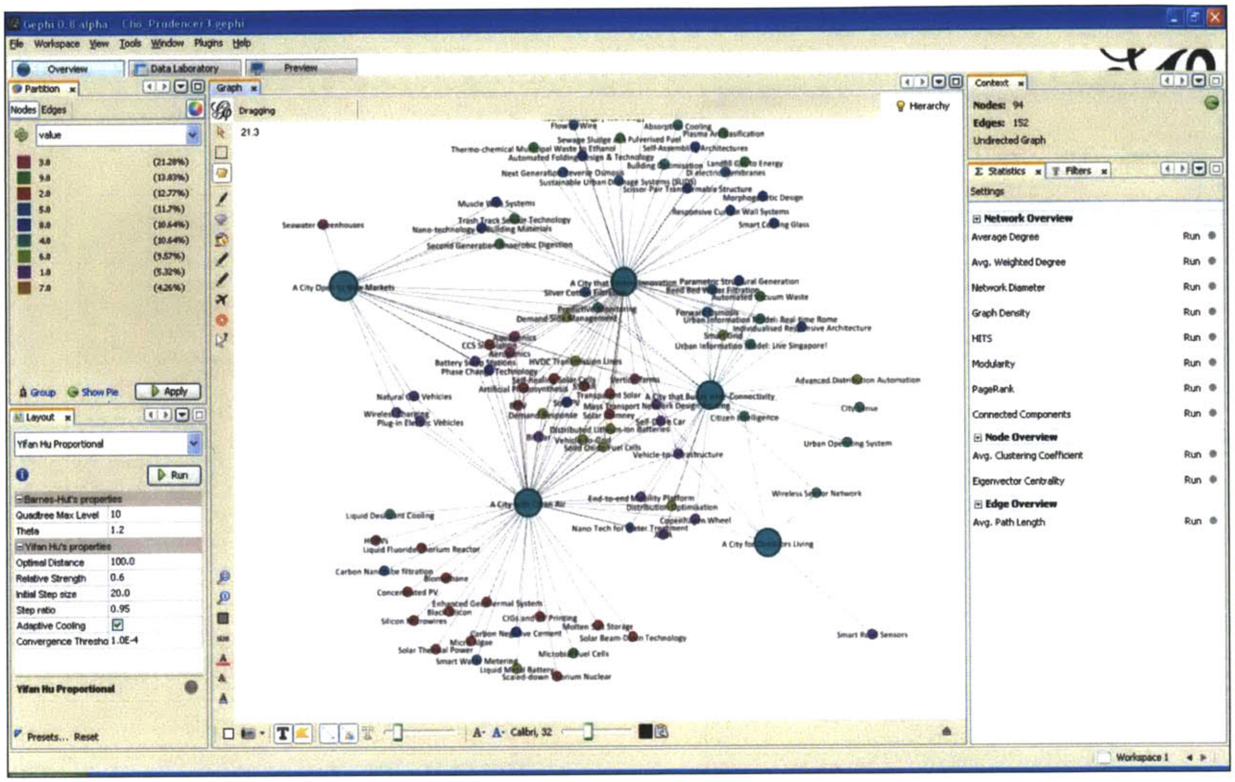


Congressional Districts

NAME_Censu	PARTY_1	State_	Friends	BetCent	ClusCoeff1	CloseCent
Martin T. Heinrich	Democratic	NM	253	0.000741	0.978261	0.690552
Harry Teague	Democratic	NM	258	0.003095	0.939191	0.698609
Louise McIntosh Slaughter	Democratic	NY	250	0.000018	0.995149	0.593718
Chris Lee	Republican	NY	181	0.001928	0.937876	0.616238
Dan Maffei	Democratic	NY	252	0.000489	0.984412	0.689416
Paul Tonko	Democratic	NY	251	0.000166	0.990725	0.648862
Michael Arcuri	Democratic	NY	254	0.001468	0.966294	0.691692
Eric Massa	Democratic	NY	253	0.000741	0.978261	0.690552
Maurice D. Hinchey	Democratic	NY	250	0.000018	0.995149	0.593718
Brian Higgins	Democratic	NY	250	0.000018	0.995149	0.593718
John Hall	Democratic	NY	250	0.000018	0.995149	0.593718
Nita M. Lowey	Democratic	NY	251	0.000166	0.990725	0.648862
Timothy H. Bishop	Democratic	NY	252	0.000418	0.984475	0.672817
Eliot L. Engel	Democratic	NY	250	0.000018	0.995149	0.593718
Charles B. Rangel	Democratic	NY	250	0.000018	0.995149	0.593718
Jose E. Serrano	Democratic	NY	250	0.000018	0.995149	0.593718
Steve Israel	Democratic	NY	250	0.000018	0.995149	0.593718
Gary L. Ackerman	Democratic	NY	249	0.000011	0.997182	0.592879
Thaddeus G. McCotter	Republican	MI	180	0.001601	0.945562	0.617327
Carolyn C. Kilpatrick	Democratic	MI	250	0.000018	0.995149	0.593718
John D. Dingell	Democratic	MI	250	0.000018	0.995149	0.593718
Mark Schauer	Democratic	MI	251	0.000394	0.986454	0.671739
Michele Bachmann	Republican	MN	168	0.000016	0.990448	0.472033
Keith Ellison	Democratic	MN	248	0.000009	0.997747	0.592041
Erik Paulsen	Republican	MN	180	0.001601	0.945562	0.617327
Betty McCollum	Democratic	MN	250	0.000018	0.995149	0.593718
Collin C. Peterson	Democratic	MN	253	0.000741	0.978261	0.690552
John Kline	Republican	MN	176	0.000113	0.974286	0.512427
Tim Walz	Democratic	MN	253	0.000741	0.978261	0.690552
James L. Oberstar	Democratic	MN	251	0.000166	0.990725	0.648862
Travis Childers	Democratic	MS	313	0.056617	0.661055	0.774797
Bennie G. Thompson	Democratic	MS	250	0.000018	0.995149	0.593718
Gregg Harper	Republican	MS	175	0.000077	0.980952	0.511801
Gene Taylor	Democratic	MS	288	0.031366	0.696816	0.740574
Ike Skelton	Democratic	MO	253	0.000741	0.978261	0.690552
Blaine Luetkemeyer	Republican	MO	176	0.000113	0.974286	0.512427
Emanuel Cleaver	Democratic	MO	250	0.000018	0.995149	0.593718
Russ Carnahan	Democratic	MO	251	0.000166	0.990725	0.648862

(1 out of 428 Selected)

Figure 6.3: Table and New York City Region Map View of Congresspersons and Districts **A table of Congresspersons can be displayed in a GIS environment, but the ties between the Congresspersons cannot be displayed as easily, if at all.**



value

3.0	(21.28%)
9.0	(13.83%)
2.0	(12.77%)
5.0	(11.7%)
8.0	(10.64%)
4.0	(10.64%)
6.0	(9.57%)
1.0	(5.32%)
7.0	(4.26%)

Yifan Hu Proportional

Barnes-Hut's properties	
Quadtree Max Level	10
Theta	1.2
Yifan Hu's properties	
Optimal Distance	100.0
Relative Strength	0.6
Initial Step size	20.0
Step ratio	0.95
Adaptive Cooling	☑
Convergence Thresho	1.0E-4

Yifan Hu Proportional

Nodes: 94
Edges: 152
Undirected Graph

Network Overview
Average Degree Run
Avg. Weighted Degree Run
Network Diameter Run
Graph Density Run
HITS Run
Modularity Run
PageRank Run
Connected Components Run
Node Overview
Avg. Clustering Coefficient Run
Eigenvector Centrality Run
Edge Overview
Avg. Path Length Run

Figure 6.4: View of a Network Graph Diagram in the Gephi Software Environment **In Network and Graph Analysis software package *Gephi*, we can see interaction of entities, but without advanced spatial reference and the benefit of overlaying other geographic layers.**

When we attempted to model an individual social network with a cartographic representation, the dialog between the two elements is not transitive or smoothly connected. This chapter required many iterations of sending data from *Pajek* or *UCINet* to *Microsoft Access* or *Microsoft Excel*, and then into *ArcGIS* for analysis. Using visual tools from two loosely coupled environments made exploration nearly impossible, and so knowledge was discovered only by deterministic statistics like 'clustering coefficient', which were calculated in one environment, and then transferred to the GIS for map representation. Another problem with this configuration is that the network environment favored only categorical data for choropleth visualization, making it difficult to use traditional color schemes to represent congruence between agents and geographic entities. Other ideas include a cartogram to alter geographic space to show feature "closeness."

The Congressional system could have been significantly easier to work with via dynamic query-based visualization. Since districts and people have so many different characteristics, the current methods required copious subsetting in both *Pajek* and *ArcGIS*. Once a subset had been declared in one domain, then the list would be carried onto the other domain to subset the corresponding entities. Understandably, this took a significant amount of time and was prone to accidents in transferring. Therefore, query options, via singular, group, union and intersect functions are essential for these systems. A dynamic query-based system gives the opportunity to select certain traces with certain features—like all traces that start or end in a Central Business District between 8AM and 9AM.

Recent work by Guo (2010) on a flow mapping environment that combines unsupervised classification and automatic selection with flow data may solve a few challenges above. This system makes a significant contribution, as it undeniably facilitates innovative interactive ESDA for flow data, but does not consider space for the integration of a social network. In summary, the iterative subsetting was arduous and very inefficient. A potential solution would be to put the network and map visualizations in the same view, and allow the users to detect patterns via linking and brushing, like the aforementioned work of the *GeoVISTA* Center and its Java-based software environment, *GeoViz Toolkit.* Given the experiences from the case studies, we suggest an inter-communicable coupled system—where files can be shared easily. Perhaps this can be realized with an add-on for a GIS environment in order to accelerate evaluation by taking advantage of new technology, and to realize the system of systems approach that supports cross-domain 'selection'.

Computation

How do we compute spatial relationships with connectivity data like the predicament shown in Figure 6.5? Each entity has a spatial reference, but a seemingly competition 'interaction' reference with non-adjacent geographies. It seems as though Origin/Destination matrices (where columns and rows are nominal place entities (ex. Michigan)) are the best way to process interaction data. This can be done in an environment like *R, Matlab, ArcGIS* or another helpful Linear Algebra or network processing software. The O/D matrices should contain connectivity metrics, like migration, call or travel flow. After statistics are computed, rows and columns must be re-matched with their matching nominal entity in the GIS, in order to regain the spatial characteristics of the nominal entity. Some statistics, like average betweenness for each entity can be stored as a single-value entity that can sit in one column. However, pair-wise 'ij' values must be stored as n(j) features in columns of a row i, for all i values in a matrix, a system where matrix transposes might be involved in order to manipulate both i and j as geographic features (as columns are only features of the rows, and cannot be tied to geographic space in a normal geodatabase). The matrix-based analysis should include the following tools, many of which are cited in the three case study chapters: Min Cut / Max Flow, Degreeness, Modularity, Community Detection, Centrality: Closeness and Betweenness, Network Clustering, Cliques and K-cores, Clustering Coefficients, Network Distance, Broker Relationships, Bottlenecks, Diffusion, Percolation and Homophily Measures based on Chosen Characteristics. This discussion has many facets and exciting prospects, which are discussed very limitedly here.

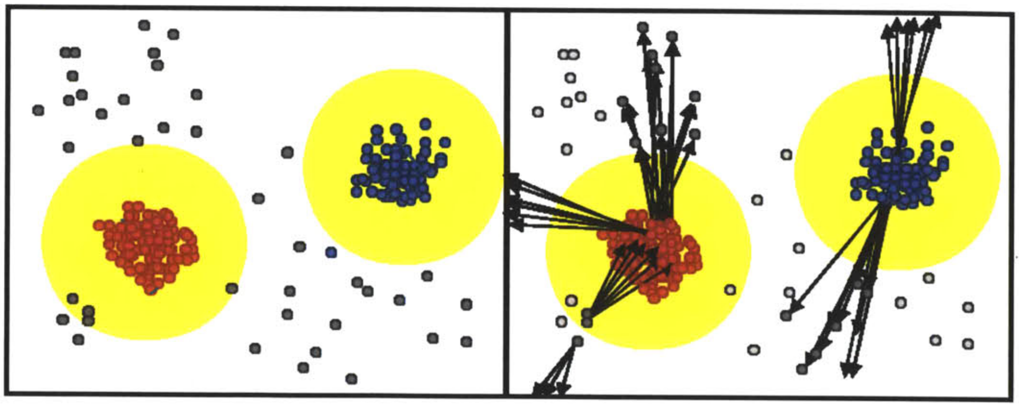


Figure 6.5: City Clusters and Potential Connections **Euclidean proximity may not fully represent the social 'closeness' of places, as represented here by lines that represent the strongest bonds of inter-community connectivity.**

In order to further computational social flow research, we recommend following the work of Waldo Tobler, Peter Nijkamp, Stewart Fotheringham and Michael Batty for modern work in flow networks. To reiterate some of the findings in Chapters 1 and 2, we find interesting work in road networks (Limtanakool et al 2009, Xu and Sui 2007), and airline networks treated as non-Euclidean networks (Xu and Sui 2007) with properties like: Small-world (Guimera et al 2005), rich-club, (Xu and Harriss 2008) Scale-free properties (DeMontis et al 2010), and relative 'diameter' (Barrat et al. 2004). We also find the same kinds of parameters applied to railway systems (Sen et al. 2003), taxis (Liu et al 2010), subway systems (Latora and Marchiori 2002) and general hub and spoke configurations. (O'Kelley 1998)

Furthermore, in communications we recommend the work of (Giannotti et al 2007, Nanni al 2006, Gonzalez et al 2008, Liu et al 2010) in trajectory mining, as well as those working on affixing social network topology onto polygonal geographic space (Radil, et al 2010) social network centrality vs. geographic centrality (Onnela, Arbesman et al. 2011) telephone network partitioning: (Green 1955, Davies 1979, Ratti et al 2010), the theoretical of Torrens (2010) and technical work of Glennon (2010) and Minimum Spanning Trees (Assuncao et al 2005, Guo 2006).

In the final chapter, we offer a possible framework for putting the spatial and social together, readdress urban planning goals and privacy, and offer final remarks.

CHAPTER 7: CONCLUSIONS

7.1 THEORIES FOR EMBEDDED RELATIONSHIPS

Integrating CSS and GIS models means that we can better model people within in the built environment, and will, at the very least, allow us to better understand the recursive patterns formed in built form and urban environment that affect and are effected by human behavior, condition, relationships, personality, choices, goals, ideology and socialization. Once these patterns surface, we can make better sense of the fragments of geographic phenomena, (like subway ridership or real estate value), that we are piecing together in the digital age.

After analyzing our three chapters, we find that it takes (1) transit/migration and (2) communication/ exchange of ideas to connect humans over space. We also find that the human *inclination* and the *energy invested* to be connected over space has intricacies that should not be overlooked, (like shared ideology or a blood relationship) and cannot be described by GIS models alone.

The examples in the three case study chapters in this dissertation are attempts to marry appropriate elements from the two fields in order to innovate in GIS and Urban Planning. In analyzing the work done in this dissertation, it seems like future work may benefit from a better idea of how 'social' and 'spatial' pieces actually fit together. More specifically, how does transportation, location/distance, communication and social relationships fit together? Is there a unifying theory, or a piecemeal theory we can turn to in order to put these phenomena into context?

There are simple laws that govern physics; Physicists are lauded for and constantly in pursuit of these laws. We may be doing a disservice to social and geographic studies to search for clear, 'beautiful' laws, because we risk oversimplification, and that, at least on a micro level, human behavior seems to be better described qualitatively than with numbers. Nevertheless, one answer to the above questions of a "unifying theory" may be to introduce a simple system **not** in order to explain "how everything social/spatial works", but perhaps to provide a framework from which to contextualize and approach complex problems that call for both social and spatial variables.

The results of the three chapters, characterizing places by their geographic power to draw from different magnitudes of distances, uncovering evidence for special city-to-city relationships driven by migrants, and finding geographic patterns in the social relationships of congresspersons, and transitively, the ideologies and constituencies they represent, give us a better understanding of how spatial variables fit into relationships. For this theory, we also draw from other literatures listed in this dissertation, especially those involving phone call, online chat and email analysis, like that of literature previously discussed in the Computational Social Science and Communications research communities.

Relationships in Geographic Space

One of the difficult problems with fusing/improving the dialog between GIS and Social Sciences is that it is not necessarily clear how important the built environment or geography is in creating, propagating or sustaining social relationships.

It seems that people usually meet for the first time in person, and thus are collocated—even for a short amount of time. Relatively few meet for the first time on the Internet, evidenced partly by (as mentioned) a scant .4% of Facebook relationships are those that have never met. (Mayer and Puller 2008) Other ways of "meeting" are through business ties, where phone and email may help establish a 'socio-professional' tie. Once a tie is created, it needs to be somehow sustained.

It seems as though a tie between two people is sustained in some elastic linear combination of three variables: (A) face-to-face meeting aka co-location, (B) telecommunications, and (C) a natural desire to keep the tie. (Figure 7.1) We use the word 'elastic', because sometimes, a strong desire to keep the tie or very frequent telecommunications can make up for less face-to-face meeting, or more face-to-face meeting can make up for less natural desire or telecommunications, et cetera. Sociologists can explain much of the "natural desire" or inclination to sustain a tie, and geographers can explain how easy or difficult it is for a tie to meet in person. Geographers can also show the distance and magnitude of telecommunications flows that tie people together in the aggregate. Telecommunications seems to act as a sustainer of the relationship, it is a necessary element for (1) logistical meet-up plans, and (2) an invaluable source of information flow that helps each person in the relationship evaluate their natural inclination to stay tied to the person. Although the theories behind "what is a tie" can be disputed ad nauseum, for our purposes, *a tie is a social relationship that would want support from good urban planning.*

This triumvirate of action-states may help us link together how a relationship flourishes or prospers in regard to geographic space. Naturally, closer people stay in better touch, and as distance between people grows, conversations get longer (Lambiotte et al 2008 and Leskovec and Horvitz 2008). A natural inclination to be friends can grow from telecommunications (emails, letters...) and face-to-face contact alike. Also, people who telecommunicate are very likely to want to meet up; people who have met up are likely to want to telecommunicate in the future.

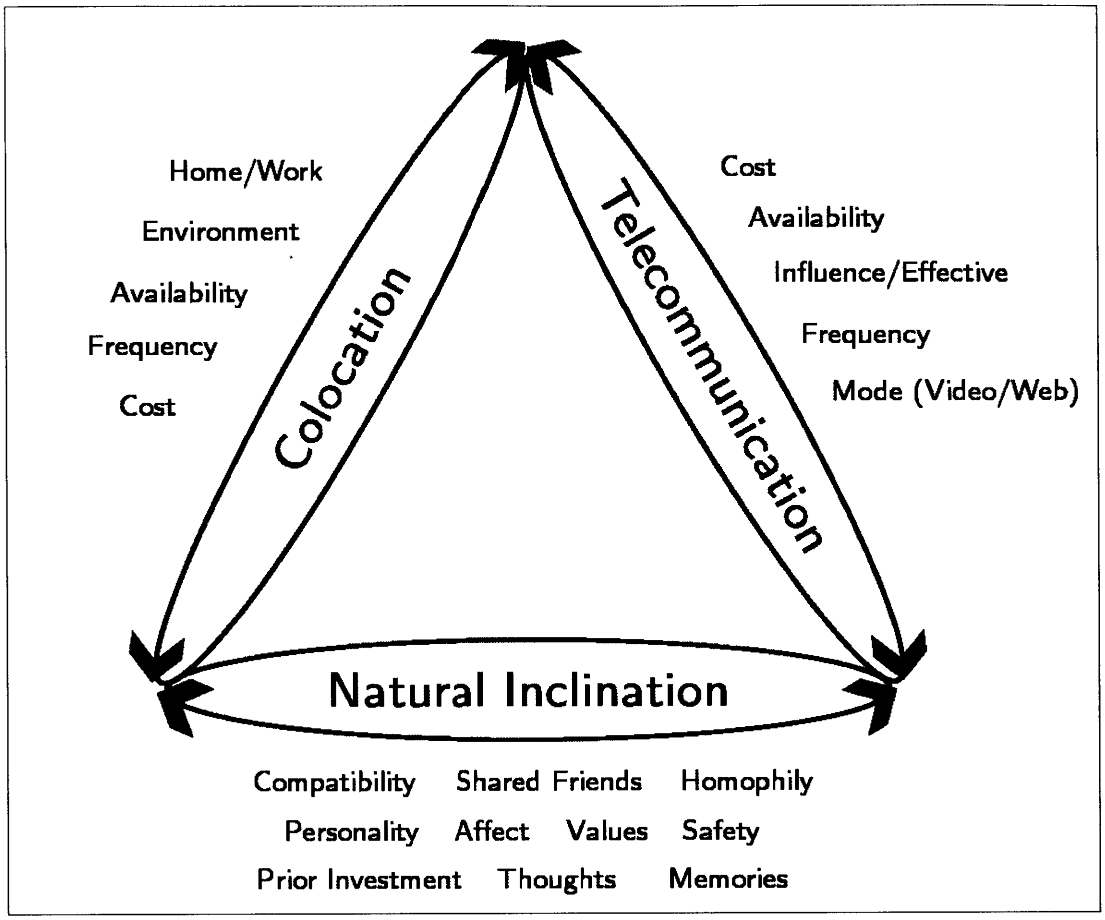


Figure 7.1: A Tri-cyclic Model of Variables that Sustain Social Relationships **We can model three major variables that sustain a social relationship: co-location, telecommunication and natural inclination, as well as the factors that affect each variable (listed outside triangle).**

The variables that might hinder or enable these relationship-sustaining processes are for co-location: location, environment type, availability of transportation service, the frequency of travel as permitted by a job or family, and the monetary, time or psychological cost of travel. (Figure 7.1) Planners with powerful GIS data and tools have access to, and can derive a significant amount of this information of travel variables between a pair of agents, given only a set of longitude and latitude locations.

Variables that might hinder or enable telecommunications are a bit similar: telecommunications cost (time and monetary cost), availability of factors like WiFi or telephone signals, the influence/effectiveness of the conversation (for example, are the participants paying enough attention), the frequency of conversation, and mode (for example, paper letters are not as fast as e-mail; video chat tends to yield a better experience for its users than telephone because of its enhanced sensory involvement). (Figure 7.1) Both geographers and sociologists are helpful in explaining telecommunications data because of, respectively, the location-based nature of the flows, and the content and substantive analysis of transmitted information.

Variables that can hinder or enable natural inclinations are usually aspatial and thus lend themselves to the expertise of the social scientists, although some demographic or ideological data can be derived from Census or business information, wherein a geographer is useful. These variables are different in nature than colocation and telecommunications, and are often more difficult to tease out. The compatibility between two people often relies on how they are related, the prior investment made in their relationship, the number of shared friends, issues of age-race homophily, personality type (for example, extroverted people), emotions, feelings and affect, values (education, music, religion), safety and trust, and the "reflection" of these relationships like thoughts, memories, or even dreams.

What is the role of distance or place in this system? What is the role of telecommunications in this system? How much does natural inclination matter in sustaining a relationship? Can point-to-point travel predict telecommunications? Can telecommunications predict point-to-point travel? If so, what are the specific variables that affect these predictive powers? One example hypothesis is that distance is an important variable in this relationship. One piece of evidence of the A-B relationship is remittance flows; where a money transfer is an indication of former or future travel. In order to learn more about the inner-workings of this system displayed in Figure 7.1, we might suggest some changes in current technology for integrating complexities in Spatial and Social Sciences.

This structure may seem interesting because it can scale from macro to micro: On the macro level, it is friendly to "big data" like geocoded cell phone travel datasets, where records indicate location transfers, and telecommunications connections. On the most micro level, it can be used to map the temporal "state" changes of a pair in a social relationship, and it can also be used as ingredients for describing a relationship. For example, in a temporal sphere, a professor and her student can be traced at each moment as being collocated, telecommunicating, or just having a natural inclination of a professional mentor-protégé relationship with similar research interests. In the components relationship, a person can classify his or her relationship with others in terms of percentage of colocation, telecommunications and natural inclination activity that the pair incurs.

We only present bookends of the capabilities of this model to organize behavior, and its strength may lie in its ability to help uncover what exactly fosters relationships. Of course, the scalability of this model transfers to medium-sized groups like that of a constrained geographic area or group of interest. We will hopefully further pursue this type of framework, and its potential to organize a hairball of rigorous social and spatial research conducted in the late 20th and early 21st centuries.

With these ideas in mind, we discuss the goals of urban planning, and how the ideas in this dissertation might help foster these goals.

7.2 CONNECTIONS AND URBAN PLANNING GOALS

First, let us recount our motivation for pursuing a research agenda: There is great richness in the science of understanding the urban landscape, and an equally great richness the science of understanding how humans interact with one another. Scientists and engineers put much effort into modeling scenarios, landscapes, people, movement, and relationships. If the goal of these efforts is to learn information that is helpful for professionals like policy makers, planners and businesspeople, it may be more constructive and a better use of resources to look at these scenarios in a new perspective, by fusing elements of social and spatial research. Currently, we scientists and engineers take on a lot of problems to solve but may not be equipped with the right tools to answer their questions. Furthermore, they might not even know that they can ask the kinds of questions they may silently wonder, because these technologies are not readily available.

The bigger picture of this message lies in the assumption that the goal of Urban Planning, Sociological and Human Geographic research is to help us learn more about the human experience, and how to support this experience, whether by sustaining good practices, providing better opportunities, or improving conditions that foster a safe, healthy, high quality of life. We model the built environment so urban planners can answer questions like these about space and place: What kind of income groups and ethnic groups live here? How much access to resources like groceries, jobs and park space do people have? Does this population need a car for mobility? Are youth at risk of joining organized crime units or gangs? Do the sightlines have an aesthetically-pleasing effect? Are pollution or environmental factors harming people in this space? Will community members be affected by blackouts or contaminated water?

How have scientists and researchers tried to reach the goals of improving society? We visit different types of research and methods whose goal, in some form, has been to benefit society.

For the past centuries, since these issues matriculated into consideration (Levy 2003), we have attempted to reach these goals through a number of different methods, and we revisit some of the quantitative, qualitative, spatial and social methods of today, and the weaknesses of these methods that could be addressed by social/spatial research.

In history, protecting people and giving voice to citizens has often been subject to the often dictator-like wills of monarchal, ecclesiastical or parliamentary rulers. Outside of this trend, altruistic measures can certainly be spotted, for example the advanced aqueduct system in Ancient Rome or the concerns of the U.S. Constitution on protecting citizens, are each evidence of thoughtful planning for society, and Jacob Riis' 1901 exposé on urban poverty in *How the Other Half Lives* is an early example of reflections that could help planners. More modern research has brought this trend to a new level, as evidenced by the nature of research and publications with the agenda to uncover society's injustices and avenues for improvement. One high-impact example is Rachel Carson's 1961 *Silent Spring,* which highlights human impact on the environment, and an example in current research is Brabyn and Skelly's 2002 review of access to hospitals using GIS techniques.

Cognitive Models

One way for understanding the human experience is through cognitive models. Cognitive models try to find, for example, the causes that drive us to attend a friend's party knowing certain people may also attend, telephone a friend or relative that has not be present for some time, send an email to an esteemed professor for a job application, befriend a student from another neighborhood, attend an emotional new wedding of a former girlfriend, move home with one's parents, or visit someone in prison. These models also try to classify and understand feelings, affect, reactions and stimuli, or predict how humans gain information or adopt fads.

What is missing from cognitive models? Largely, planners have not taken advantage of these models. Psychological and behavioral complexities like self-awareness, ego, ethos, desires, emotions, reactions or self-consciousness have not been undertaken in even the most discrete agent-based geography models. This lack of integration may be due to the relative inappropriateness of characterizing physical objects with characteristics designed to describe human behavior. For example, a road does not avoid another road because it has low self-confidence; a train is not late because it is concerned about arriving with a certain appearance; cell towers are not placed nearby because they are friends; and bridges do not connect places due their desire to be included. But geographers modeling fluvial dynamics may not be able to apply their hydrological equations to the spreading of ideas or the adoption of fads which often require 'complex contagion' dynamics. (Centola and Macy 2005)

Understanding the convergence, communication, transitivity and placement of people, however, may need to take inhibitions, personality traits and human behavior into account. Racial discrimination has been noted as a factor for neighborhood segregation, but only static census tracks have been used as evidence of this separation, never migrant traces. (Schelling 1978) Furthermore, geographic neighborhood segregation has rarely been seen from the point of view of social networks, but generally perceived as homophilic or xenophobic, or the effect of racial stereotypes.

If we can tie together macro geographic flow systems with micro-decisions, research on artificial intelligence, computational social science, affective computing, personality, mental illness, and social norms may be integrated into social flows and social distance. Issues of personality types, (Mischel 1999), memories (Tulving 1983), affect (Picard 2002) or levels of mental processes that scale from "innate, instinctive urges and drives" to higher level thinking of "values, censors, ideals and taboos" (Minksy 2005) are important factors for making choices and decisions, planning, searching, weighing options, and envisioning future scenarios like a decision to bike to work or take a trip to the store for milk, each of which are geographic in nature).

Citizens are likely to emulate one another's behavior, and this emotional convergence usually relies on the mimicking of expressions, vocalizations, posture and movements of another person, where colocation is necessary for mimicking most of these things, and vocalization requires communication or telecommunication. (Hatfield et al 1993, Hatfield and Rapson 2000) Additionally, the public can learn about a new fashion or innovation, but it is only when friends or a critical mass of other people display the novelty or fashion that the adoption is realized. (Centola and Macy 2005) From what geographies do pairs send affectively-charged emails or phone calls? Would text analysis or voice recognition spot this affect between Manhattan firms, or between two small rural towns? In one slightly satirical example, sensing joy after a political election may be an indication of less-than-average migrant outflow. The geographic pockets of most elevated fear levels after the September 11th attacks may have correlated with the phone calls made to loved ones in the cities of New York and Washington D.C.

Quantitative Social Science

Another way we have learned about the human experience and how to support the experience, is through quantitative analysis that explains social phenomena. On the social side, models that 'calculate' the probabilities of human behavior give society clues as to how to mitigate their relationships for better quality of life. As we have mentioned, Christakis and Fowler (2007) find that smoking cessation can be related to the cessation of a certain relative, co-worker or spouse, but less so to that of a neighbor.

What are the limitations of this research framework? While using a social network to calculate probabilities and predict future impact of relationships can be considered a revolutionary field in itself, the discoveries may not be complete without incorporating the environment in which smokers or non-smokers must navigate, and how this affects their choice to smoke and their social influences' choice to smoke. Are cigarettes accessible? Are they expensive? An expert geographer may automatically report that a study conducted in Framingham, Massachusetts (like those of Christakis and Fowlers research) could hail a variety of lifestyles: people from struggling industrial towns, wealthy Boston suburbs, the quiet Berkshires or crime-filled inner-city neighborhoods. When incorporating geographies and demographic factors into these studies, it may show that a slightly, or altogether different set of variables account for smoking cessation, and that future policy should be directed towards these targets.

Geographic Information Systems Analysis

Additionally, discoveries about the human experience have been made by researchers conducting GIS operations that combine demographic data and physical phenomena. These methods have uncovered the places where an inhabitant would be at risk of isolation and vulnerability, and that these factors are very much tied to geographic space. The work of Susan Cutter in describing the landscape of environmental injustice and vulnerability to hazards is a prime example of the computational marriage of physical and demographic factors in order to inform policymakers, emergency services and urban planners about protecting citizens against disasters and risk.

Are there shortcomings of this approach as well? A leader in the field, Cutter and her work has drifted towards better understanding of social characteristics, indeed her 1996 article, *Vulnerability to Environmental Hazards* was followed by *Social Vulnerability to Environmental Hazards* (Cutter et al 2003), but the 'social' semantic is still limited to demographic Census variables: averaged, pro-rated , and re-agglomerated over space. Indices are often concocted, manipulated and turned into 'scores.' In the later article, 42 variables are reduced to 11 factors in order to explain about three quarters of variance in a model, but without consideration of the behaviors that spreadsheet Census data do not show, leaving the picture of social vulnerability incomplete. Hazards research often includes behavior that does not account for geography. In studying Hurricane Katrina, Elliott and Pais (2006) show that variables of religious faith was a major factor in safety via the decision to evacuate, and that the aftermath of increased drinking, swearing, and gun purchases now affect the social landscape of the New Orleans region. Similarly, Eisenman et al (2007) show through interviews that ties to family, friends, and community affected evacuation, in addition to transportation, access to shelter.

We can simulate the tradeoff between cost, time and distance, but it is hard to simulate the tradeoff between cost and social richness, friends or family. We can create an urban economic model that can show which variables may best explain or cause urban phenomena, but since relationships are hard to represent as a scalar, social dynamics do not fit easily into these models, meaning that certain variables that may have a stronghold on choices cannot be accounted for with current economic models. Also, secondary activities like urban clusters can still emerge from these models, but are hard to predict and hard to explain without including social variables.

Politics

Budget cuts for support industries like *Planned Parenthood* can also affect people and places, (as the relocation of certain services inhibits accessibility), family planning may go unassisted in communities that rely on certain education and amenities. An "AIDS village" in China's Henan Province, is home to nearly 3800 villagers, with an HIV rate of over 30%. This is just one example of the many villages where poor Chinese farmers were exposed to the virus when giving blood to enhance their meager incomes. Like many other epidemiological studies, social and spatial components are involved, and can hopefully help doctors and policy analysts control disease and focus on infected community members.

Illegal immigration has social components, as well as spatial components, and may be understood better when variables surrounding different opportunities and relationships via contacts or geographic lushness are understood. Understanding the 2005 Race Riots in Paris, France are important for geographers because the discontentment of certain ethnic groups was tied not only to their social ties and proximity (which propagated the explosive riots), but to the nature of suburban life outside of Paris and the lack of opportunities for those in exurban communities. With social models, planners may be able to increase their sensitivity to human rights issues.

Also on the subject of race, the number of Black men in prison (Boothe 2010) is growing to comprise surprising proportions of inmates. The *Million Dollar Block Project* from *Columbia University's Spatial Information Design Lab* shows a number of city blocks that produced inmates, whose government expenditures when summing by block exceeds one million dollars to incarcerate yearly.

This project is evidence that geographic clustering is present in crime patterns. This problem may be best served by understanding the friendships, social mechanisms and influences that result in incarceration; also, the nature of neighborhoods that host these social mechanisms, and often poor education, drug abuse, crime and unemployment significantly contribute to this issue. One may argue that the social ties are so grounded in geographic space, that if neighborhoods were more integrated, then this percentage might differ.

Traditional community jobs are being jeopardized by industries moving overseas and the closure of face-to-face stores, like the recent dissolution of video rental outlets, and the patronization of big box stores instead of locally-owned commerce and products. While these economic variables make their mark on cities and neighborhoods, they make the landscape less accessible to some, as a car is almost necessary for a trip to large –Mart stores. On the social end, they impact the serendipitous and sensory aspects of human relationships, and leave only the frustration of unemployment.

We often approach some issues in planning through a land-use model. Two issues that could be improved with a better understanding of social and spatial combined are shelter locations and charter schools. The best locations for family shelters would be outside of drug and alcohol networks, but accessible to jobs. Some shelter patrons, or workers from low income neighborhoods must endure long bus rides, or walks along busy suburban streets—where it seems that everyone else has a car—in order to take advantage of high paying jobs or the safety of the shelter. It leads us to wonder if this 'trade-off' of safety values, economic survival and access can be remedied to provide people with better opportunities. In another trade-off, charter schools draw promising students from failing school districts, to the advantage of the student, but to the poor-get-poorer disadvantage of the school district. This is a decidedly geographic decision system, where students are defined by their social capabilities, talent, achievement and progress, but are subject to the address of their home for admission into a charter school. Two equally talented students may be accepted or turned down based on their residence's location in public school boundaries.

Urban Sociology and Architecture

Researchers have tried to address the issues of quality of life and improving quality of life through learning about the configuration of the built environment. Researchers, city designers and developers ask: What kinds of spaces are allowing relationships to foster? What kinds of spaces are hindering relationships? What kinds of activities are being practiced in certain places? In one example, Sevtsuk and Ratti (2005) describe the use of different spaces on the MIT campus given WiFi usage data. From this experiment, the authors were able to guess whether buildings were residential, classroom or business-type areas, and also to gage how to make the campus more amenable to activities, like late night studying, that are changing with the ubiquity of laptops and wireless devices. In an unpublished work, Rojas (2005) finds that wireless users in New York City are drawn to venues with electrical outlets but try not to compromise other benefits in this search for a quiet workspace.

Some of this current research may be missing elements the integration of the human experience. City development and design researchers have taken strides to reassess traditional and neoclassical model of development planners in the digital revolution, and with new design theories and objectives. Albeit productive, designers seem to have taken sufficient interest in the socio-emotive relationships that drive decisions in the built environment. With an increased understanding, designers could take things like obesity into account when asking how to develop a city. They may consider how many residents have traditional 9-5 jobs verses the non-traditional hours of entrepreneurs, night-owl lawyers and emergency doctors, and ask whether residents actually want to be close to work for convenience, or to keep work and home in separate spheres for domestic peace of mind. Do residents of a certain place desire to be elsewhere—like a 'suitcase college' where students travel home for the weekend or a military base that grants R&R (rest and relaxation) weeks off to active duty members? If so, planners may support desires to visit loved ones, or preferable locations with infrastructure that facilitates transit out of the city to popular destinations.

What this research may be missing is the consideration of microbehavior, goals (like productivity), peer pressure, socialization and group dynamics in order to really dig deeper into the mechanisms driving decisions to utilize space. Are students studying because they like the space or because they feel lonely or need support with difficult math problems? Is a park not a favorable place for an independent writer or lawyer to study because of the elements, or because of a park's reputation for being a ground for dodgy acts involving drugs or couples because it is hard to see park activity from the road? If illegal drugs are being used, for example, they are passed hand-to-hand and traverse social systems, but also circulate in a specific spatial area. In order to eradicate drug propagation, should we target the people or the spaces?

Urban sociology research may benefit from approaching transactions and activity as phenomena that need people and place. In the illegal drug example, picture a drug-laden gang eradication team. The team may be best to approach the situation realizing that dangerous relationships accumulate based on both people and place: as proximity and mentality. The geography and the social weights provide the cost/benefit landscape for a potential new gang member and his decision to join. Geographically: the gang is present in his neighborhood, the neighborhood's condition has provided the playground for the gang, the neighborhood's safety has decreased because of the gang, and the cycle of social/spatial/social/spatial continues.

Qualitative Sociology

We have learned more about the human experience and how to support this experience through interviews, surveys and observation. There are countless examples of these types of research projects, but two prominent works on society and its effect on certain individuals due to class, race or economic status are Laureu's *Unequal Childhoods* (2003) or MacLeod's *Ain't No Makin' It* (1987), which uncover the significant differences in child rearing as a function of economic status and race, and the Sisyphean plight of two ethnic groups in a housing project, respectively.

But what is missing from this qualitative research? In Lareau's and MacLeod's work, neighborhoods play a large part of child development, and family opportunities, but are mentioned only in narrative. Decidedly sociological in nature, these two research agendas could be made even more complete, and causes of inequality, inopportunity and injustice could be better verified if geographic factors were included as the culprits of society's unfairness. Conversely, the planning mechanisms that look to benefit people may not be paying close enough attention to these sociological accounts. According to Lareau, low-income families are more likely to live close to relatives, or have multiple relatives living in one home. Conversely, high-income families are generally more detached from their non-nuclear relatives. A planner might benefit from understanding this sociology dynamic, because he or she could then build developments and homes that accommodate multi-family communities.

We see that human opportunity is tied both to relationships and environment. Since this goal of social research is to foster opportunity and quality of life, this dissertation also asserts that these relationships and the environment are also intertwined. With limited resources and lofty but important goals, this is an important decision, but with Social Distance metrics the distinction between action on people or place need not be made—as new models can show the dynamics of both systems.

If our goal is to help people, can we accelerate this process through research? We suggest that researchers interested in participating or leading research in social/spatial endeavors to learn more about society should educate themselves in both realms: the social and the spatial. We now report on important considerations on privacy and ethics in planning, with these new technologies in the atmosphere.

7.3 Ethics and Privacy

On the lines of epistemology and abstraction come a significant amount of critical thinking, metacognition, and careful scrutinizing of the impact of these developments on the research community and its potential beneficiaries. The following section addresses some concerns, though we believe that more concerns will arise after more research is conducted and more researchers get involved. As an MIT professor stated about privacy and the Internet in a recent article, "When cars were first introduced, they created tens of thousands of highway deaths per year until we learned more about best practices like stop lights, and we created technology, like disk brakes, that made it safer. But we didn't stop using cars in the meantime." (Hickins 2011) With this in mind, we try to foresee concerns for the future.

Normative and Positivist Ethics

Data & Technology

Governing bodies and firms with power or influence can impose rules to regulate amenities, goods, services and facilities that become part of everyday life. Some seemingly shared values are things like minimizing harm on the environment, creating jobs for families to have an income, and reducing crime. These kinds of Positivist claims are usually statistically proven, or hard to refute. Some transportation engineers can have goals that fall in the positivist realm: for example, their goal may be to have trains run faster and more efficiently, as fluid transportation is a shared priority that is certainly hard to refute. Here, complex ethical considerations are conveniently avoidable under this Technocratic approach.

The use and availability of new data leaves society at increased risk of being controlled by those who could exploit sensitive individual and aggregate data, and so, we should raise new questions of ethics and governance. Especially with the new combination of social and spatial, not only is it clear where people are at specific times, but we can infer their activities, attitudes, behaviors and future patterns. As Urban Scientists become better at learning how people use the city, human behavior might be scrutinized and understood at a more candid level and then possibly manipulated through changing the built environment, economic or social systems. This ability to manipulate, and then furthermore to intimately track the change in behavior through urban sensing and data records, gives a new level of power to governing bodies. Therefore in the Digital Age, we should be careful to reflect critically on whether the purposes driving the decisions of institutions are in a society's common interest, or if these purposes promote a value system that is not held by the recipient society, in part or in whole.

Individual vs. Society

Some Theocratic governments act on religious values, disallowing certain foods, media, drinks, or illegalizing education and travel for women, although these normative decisions may not be agreed upon by individual citizens but in the interest of upholding the religious values of those in power. Whether Theocrats are acting out of benevolence or not (e.g. banning alcohol can be good for one's health), their normative rules are applied for the masses, with little reception to the values of the individual. In a different kind of example, some have argued that China's One-Child Policy is in the best interest of the future of the People's Republic, as it diverts the path to an overpopulated community that risks poverty, lack of food, and tight spaces and unemployment. However, for individual families, this sanction was not always favorable. When a 2008 earthquake destroyed—among many devastated buildings—a schoolhouse in Sichuan, many parents were left in the most poignant despair, not only because of the utmost tragedy of losing a child—but this was in many cases, the only child in the family, a sanction that was the only choice for many couples who could not afford high taxes of having more children. Of course this tragedy could not be predicted, but it does exemplify how normative values of a country might not always uphold the values of the individual.

We also see here that sometimes these offerings can be set up to directly or indirectly give benefits to those who comply, leaving those who do not comply with little access to these benefits. Even if incentives are provided, these incentives (or similarly, the cost of complying) may not be in the "best interest" of the people, where, by *best interest*, we mean helping citizens live with their value system intact. Some advertizing bodies use location-based services to target and convince special groups of people to prescribe to a certain product or service that may not be in their best interest (like unhealthy food).

Normative Issues in City Form

In terms of changing the built environment, there is a "chicken or the egg" problem with many planning decisions where it is unclear whether the built environment is configured a certain way (perhaps motivated by normative judgement) and inhabitants adjust and become accustomed to this way, or that the built environment grows to accommodate the relatively aspatial social and economic needs of the people, without need to be explicit about the normative values that determine the environment.

Pertaining especially to this dissertation is a relatively simple problem, with a complex solution: Since our dissertation concentrates on social flow data, should planners augment transit and infrastructure in places that have high flows? Or should planners invest in bridging neighborhoods where there is very little connectivity in order to create a more cohesive, and perhaps less segregated, city? Similarly, should developers building mixed income housing intermix parcels of pricy and subsidized houses to fill seemingly-harmless goals of combating economic class stratification with more face-to-face contact, or should the developer create more homogenous clusters of housing ties, in order to follow the natural preference of homophilic social relations that we have seen many times in sociology literature.

In one example of flow tradeoff, the METRO Subway system in Washington, D.C. is famously 'missing' a stop in the densely-commercial neighborhood of Georgetown. A popular topic of argument, some uphold that the physical terrain in Georgetown was not fit for underground tunneling, while others argue that residents of the relatively high end area of Georgetown did not want the D.C. community in 'other' neighborhoods to access their neighborhood so easily through the facile public transit system (a myth later debunked by Schrag 2006). Although Schrag shows it to be largely untrue, the Georgetown community's *hypothetical* desire to keep their area a bit more exclusive is a case where inaccessibility is a value of the citizens in that area. Complying with the wishes of the residents seems easy until it is realize that cost of keeping Georgetown off the Metro System makes travel time longer and harder for many of the low-income retail and home-service workers who rely on public transportation to work in the Georgetown neighborhood—among other issues.

Technology and Privacy

Privacy issues are also a focal point of the challenges in implementing social flows. Since many endeavors involve anomalies and potentially traceable human paths and communications behaviors, responsible measures should be made to ensure the anonymity of users and statistical measures to eliminate singularly probable events. The U.S. Census is a good model to follow—some data from some Census tracts or areas is not disclosed as there are not enough citizens within that area to protect a single household or few households from being having person information (like income) disclosed. Other measures like International Review Board (IRB) approval for use of human subjects should be used when studying human movement, cell phone analysis or behavior, in order to rigorously meet scientific standards, and get the input of law, privacy and research specialists for the safeguarding of research.

The main tools for the capture of the aforementioned phenomena in urban sensing are in the family of GPS, RFID, cellular and sensor networks. Sensing the elements of the built environment in an urban setting is equally important, as it provides the backdrop for motion and activity. Satellites, airplanes, and even hot air balloons are used to digitally capture images of terrain, buildings, traffic, green space, various land cover elements and weather systems. These capabilities prompted the discovery of ruined survey towers on Russia's Trans-Siberian railroad (Clarke and Cloud 2000), and are widely recognized for spotting crops of illegal drug with infrared sensors. The following are some concerns related to data sensing, collection and use.

Commercial Capture

Credit, Debit and even subway transit cards are now often RFID-enabled, meaning that an individual's trace can be tracked with sensors that store the tag's unique ID. Monmonier (2002) warned about retail privacy factors: "Location is a powerful key for relating disparate databanks and unearthing information about possessions, spending habits and an assortment of behaviors and preferences, real or imagined."

In Transit

In fact, governments use remote sensing technologies to monitor speeding (and subsequently alert a police car to issue tickets) along with wire loops and traffic cameras. (Monmonier 2002) Similarly, digital toll passes deducts a fee upon usage (thus connected to credit card and bank account information) issues tickets, while red light cameras can also issue tickets. These policies may provide a tax base to enhance budgets and keep the roads (hypothetically) safer, but perhaps at the cost of psychological unrest and potential resentment of citizens.

Data Fusion

Not only is the privacy of physical latitude/longitude location at risk, but location can be paired with geographically aggregated demographic information (like Census data) and henceforth, data miners can statistically pick out an agent as an anomaly, or can guess consumer habits from association analysis. Mash-up type data fusion can result in extra information tied to one's identity because of their place of residence, work, vacation, or other frequented spot. (Ferreira 2008) Location-based service (LBS) like batchgeocoder.com can find someone's geographic location from a shipping addresses or (more roughly) an IP-address. For a two-mode commuter, the fusion of a train schedule with one's daily drive-home route could give perpetrators precise time windows in which to locate a target or commit a theft at an empty home.

Remotely Sensed Data

Processing remotely sensed data involves orthorectification—or the process of scaling certain elements so that a 2D image of a 3D curved surface is useful for measurement. Infrared cameras can be used to measure the urban heat island and natural features. While these two applications may not raise many privacy issues, these technologies can also capture license plates—from orbit—as well as faces, and clothing. (Longley et al 2005)

Trajectories

In their trajectory mining endeavors others concentrate mostly on the algorithmic side of trajectory analysis, citing these trajectories as "moving objects." (Giannotti et al 2007) These semantics connote a computer science problem (and contribution), rather than a geographic or sociological problem, laced with privacy issues.

Following the work of Bruno Latour, (see for example 1987, 1993) we consider representations of ourselves and surrounding environment as objects of scientific investigation, and that the representation of subjects is politicized by those given **power** to speak for us. Since social data is now more readily available, the 'power' of speaking about the masses and their connections over space may be leaning towards those who can successful turn data into knowledge and disseminate this knowledge through an upright media. When conducting research that may be shared or deemed rigorous, privacy and ethical issues should be increasingly guarded as the realization of emotive, demographic agent-based social flows become available for research.

From a fundamental GIS point of view, one strategy is to stress aggregation. A benefit to aggregation is that users are grouped together without unique identifiers. Some drawbacks: Population disaggregation is an issue in GIS, because once something is aggregated it is hard to disaggregate and prorate variable statistics to smaller geographic units. Analysis can lead to ecological fallacies if comparing two aggregate groups. Additionally, mean and median (and other group statistics) are less robust and informative as raw individual data.

Restricting the use of "mashable" data is very difficult when datasets have a spatial nature. One "good" but potentially dangerous application of these layering techniques is Langford and Higgs' dasymmetric model. The authors couple a topological USGS ordinance survey with demographic neighborhood polygons. From there, the population is allocated to the buildings. Using this method showed that accessibility to health care was lower than expected in rural areas, which has implications for policy makers. (Langford and Higgs 2006) This approach could have exposed pin-pointing building footprint location. In conclusion, new data and tools over new learning, but also require more complicated rules and data protection and permission issues, and so we may see a rise in the research and new laws of regulative bodies, society and data communities.

7.4 CONCLUSIONS

As mentioned, the main goal of advancing the Social Distance field is to support the altruistic ethics in urban planning, like mobility and accessibility contributions since ability to maximize professional, health, consumer, social, cultural, welfare, religious or recreational opportunities is tied with one's accessible environment. The purpose of using digital technology to model these scenarios is to enhance the possibility of improving confidence in cause for action; cases for development funds and research grants; empirical evidence for qualitative observations, surveys and interviews; and information about the extent of a problem‘s threat on the community.

It is not entirely clear what will happen with each use of a new system of systems, but the benefits draw from the combined knowledge and new knowledge available from the cross-pollinated information. Our world today has many voices that go unheard while decisions about their environment are made around them. Integrating social behavior into models of how the built environment is used can help empower those who do not often choose how their landscape will look, and what opportunities it will provide, especially when the underpowered are hampered by economic or cultural limitations.

New models may also be important at this juncture because the concept of distance and its technical uses have not been widely revisited since the before the Digital Era. In popular reading, Thomas Friedman has proclaimed that the world is flat, and Richard Florida, that the world is spiky, but progress in geographic distance metrics may not be well suited for the digitally-connected world, much less for measuring the reverberations and modern emergent theories of digital connectivity and telecommunications. Manhattan and Cost Distance concepts are well over 100 years old; the antiquity of Euclidean Distance is evidenced in part by its namesake, Euclid. The age of these measures is not a mark of weakness, but the opposite, evidence of their essential role in science, yet the rapid changes in technology beg for advancements in standards of measurement that account for societal links.

We also realize that Geography has issues of immutability. Once something is put into geographic space (ex. a school) it is difficult to move it, it is harder to optimize things, and to make accessibility more of an issue. The built environment is adaptable, but its current configurations should be taken seriously, as mutating infrastructure is non-trivial and often costly. Conversely, agents in a social system do things that are not always understandable. Much time and effort may be put into understanding a chaotic system, or wrongly guessing as to causation of behavior.

The cost of incorporating new models into urban planning may be high, but the benefit may be higher. Social distance can tell us more about the constraining and enabling factors that hinder, facilitate or sustain relationships. Without geography, there is a limited amount we can understand about who someone could comfortably make a trip to visit in a day's time. Without geography, someone's lifestyle cannot be well-understood, because it is unclear what opportunities or obstacles his or her landscape could provide

The neglect of the formalization of social flow and social distance methodology may stunt the geospatial models from which important decisions are made. Without this measure of distance the current global platforms on which models are built will remain distorted. A central concept in measuring topological inconsistencies, not only does distance decay at an uneven rate when considering cost, but this distance decay is exacerbated, if not convolved, to produce an even more skewed and spiky, and uncertain topology that follows the invisible patterns of human relationships and their waxing and waning in Cartesian space. The temporal and probabilistic regularity of these patterns is not known, which results in a shaky stage from which to build models—an understandable detractor. But this should not hamper using social space as a stage for urban activity, as these live, breathing actions are what electrify the human tapestry. The way in which we approach and use these new metrics must be scrutinized in order to ensure quality and rigor in the scientific community.

In summary, this research quantitatively models and describes the combined landscape of geographic space and human decisions by calculating movement, communication patterns, and involvement in social media. We call these patterns 'social flows,' and try to show that in some cases, measuring these flows gives us a sense of 'social distance' and can paint a more colorful picture of the synergistic behaviors of humans in the built environment than studies that use only measurements of Euclidean or cost distance. Furthermore, this dissertation attempts to explain why GIS analysts should invest in 'systems of systems' and explore the convergence of spatial systems and social systems, despite the challenges of inter-system dialog required for engineering new models.

We find that GIS models are not currently well-equipped to derive evidence of the human condition and goals within the playground of the built environment, and current reasoning for the human condition and social relationships does not seem to herald the influences, enabling or limiting powers of geographic space. Past models have modeled socio-geographic connectivity through expansion laws of physics, convergent econometric models, each usually under the assumption that interaction probability is simply determined by proximal closeness.

This new convergence of social and spatial systems may help us better answer these questions, by providing a more accurate representation of how citizens interact with one another, and how these interactions are synergistically dependent on and impress upon the built environment. If this picture is realized, we can find better ways to build infrastructure that listens to the user's needs instead of providing a fixed path that the user cannot adapt. Understandably, there are major challenges to put this new thinking into action and solve the problems ahead, but with the advancements of automated data collection, online resources, computational social science, and complex systems analysis, we are approaching these questions at a very exciting time.

Imagine a world where we can more easily allocate resources not to evenly cover a map, but to better connect humans to one another and to what they value—as human connections are the most important connections of all.

ACKNOWLEDGEMENT

We acknowledge the helpful comments of Joseph Ferreira Jr., Marta Gonzalez, Frank Hardisty and Bo Morgan in this section, and thank Ben T. Hall for assistance in editing. Thank you to Nick Andris, Robyn Andris and Prudence Robinson for logistics.

BIBLIOGRAPHY

REFERENCES FOR
CHAPTER 1: INTRODUCTION

[1] Blau, P. Exchange and power in social life *Transaction Publishers,* **1986**

[2] Schelling, T. Micromotives and macrobehavior *Norton, New York,* **1978**

[3] Arrow, K. An extension of the basic theorems of classical welfare economics *Cowles Commission for Research in Economics: University of Chicago,* **1952**

[4] Nash, J. Equilibrium points in n-person games *Proceedings of the National Academy of Sciences of the United States of America,* **1950**, *36*, 48-49

[5] Kearns, M.; Suri, S. & Montfort, N. An experimental study of the coloring problem on human subject networks *Science,* **2006**, *313*, 824

[6] Salganik, M. & Watts, D. Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market *Social Psychology Quarterly, American Sociological Association,* **2008**, *71*, 338-355

[7] McPherson, M.; Smith-Lovin, L. & Cook, J. Birds of a feather: Homophily in social networks *Annual review of sociology, Annual Reviews,* **2001**, *27*, 415-444

[8] Centola, D. & Wilier, R. The Emperor's Dilemma *Theories of Social Order: A Reader, Stanford Social Sciences,* **2009**, 276

[9] Christakis, N. & Fowler, J. The spread of obesity in a large social network over 32 years *New England Journal of Medicine,* **2007**, *357*, 370

[10] Christakis, N. & Fowler, J. The collective dynamics of smoking in a large social network *New England Journal of Medicine,* **2008**, *358*, 2249

[11] Lyons, R. The spread of evidence-poor medicine via flawed social-network analysis *Statistics, Politics and Policy,* **2011**, *2*

[12] Watts, D. & Dodds, P. Influentials, networks, and public opinion formation *Journal of Consumer Research,* **2007**, *34*, 441-458

[13] Girvan, M. & Newman, M. Community structure in social and biological networks *Proceedings of the National Academy of Sciences,* **2002**, *99*, 7821

[14] Watts, D. & Strogatz, S. Collective dynamics of 'small-world' networks *Nature,* **1998,** *393*, 440-442

[15] Clauset, A., Moore, C. & Newman, M. Hierarchical structure and the prediction of missing links in networks *Nature,* **2008,** *453*, 98-101

[16] Watts, D. A simple model of global cascades on random networks *Proceedings of the National Academy of Sciences of the United States of America,* **2002,** *99*, 5766

[17] Albert, R.; Jeong, H. & Barabasi, A. Error and attack tolerance of complex networks *Nature,* **2000,** *406*, 378-382

[18] Ebel, H.; Mielsch, L. & Bornholdt, S. Scale-free topology of e-mail networks *Physical Review E,* **2002,** *66*, 35103

[19] Granovetter, M. The strength of weak ties *American Journal of Sociology,* **1973,** *78*, 1360

[20] Arbesman, S.; Kleinberg, J. & Strogatz, S. An Explanation of Superlinear Scaling for Innovation in Cities *Imprint,* **2008,** *30*, 5

[21] Burt, R. The gender of social capital *Rationality and Society, Sage Periodicals Press,* **1998,** *10*, 5-46

[22] Fernandez, R.; Castilla, E. & Moore, P. Social capital at work: Networks and employment at a phone center *American journal of Sociology,* **2000,** *105*, 1288-1356

[23] Granovetter, M. Social Science: Ignorance, Knowledge, and Outcomes in a Small World *Science,* **2003,** *301*, 773

[24] Ben-Akiva, M. & Lerman, S. Discrete choice analysis: theory and application to travel demand *The MIT Press, Cambridge, Massachusetts,* **1985**

[25] Ben-Akiva, M. & Boccara, B. Discrete choice models with latent choice sets *International Journal of Research in Marketing,* **1995,** *12*, 9-24

[26] Limtanakool, N.; Dijst, M. & Schwanen, T. A theoretical framework and methodology for characterising national urban systems on the basis of flows of people: empirical evidence for France and Germany *Urban Studies,* **2007,** *44*, 2123

[27] Rietveld, P. & Janssen, L. Telephone calls and communication barriers *The Annals of Regional Science,* **1990,** *24*, 307-318

[28] Eagle, N.; Pentland, A. & Lazer, D. Inferring friendship network structure by using mobile phone data *Proceedings of the National Academy of Sciences,* **2009,** *106*, 15274

[29] Larson, R. & Odoni, A. Urban operations research. *Prentice-Hall, Englewood Cliffs, New* Jersey **1981**

[30] De Blij, H.; Murphy, A. & Fouberg, E. Human geography: people, place, and culture *Wiley, Hoboken, New Jersey,* **2007**

[31] Milgram, S. The small world problem *Psychology Today,* **1967,** *2,* 60-67

[32] Travers, J. & Milgram, S. An experimental study of the small world problem *Sociometry,* **1969,** *32,* 425-443

[33] Limtanakool, N.; Schwanen, T. & Dijst, M. Developments in the Dutch urban system on the basis of flows *Regional Studies: The Journal of the Regional Studies Association,* **2009,** *43,* 179-196

[34] Xu, Z. & Sui, D. Small-world characteristics on transportation networks: a perspective from network autocorrelation *Journal of Geographical Systems,* **2007,** *9,* 189-205

[35] Guimera, R.; Mossa, S.; Turtschi, A. & Amaral, L. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles *Proceedings of the National Academy of Sciences,* **2005,** *102,* 7794

[36] Xu, Z. & Harriss, R. Exploring the structure of the US intercity passenger air transportation network: a weighted complex network approach *GeoJournal,* **2008,** *73,* 87-102

[37] De Montis, A.; Chessa, A.; Campagna, M.; Caschili, S. & Deplano, G. Modeling commuting systems through a complex network analysis *Journal of Transport and Land Use,* **2010,** *2,* 39-55

[38] O'Kelly, M. A geographer's analysis of hub-and-spoke networks *Journal of Transport Geography,* **1998,** *6,* 171-186

[39] Barrat, A.; Barthelemy, M.; Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks *Proceedings of the National Academy of Sciences*, **2004**, 101, 3747

[40] Sen, P.; Dasgupta, S.; Chatterjee, A.; Sreeram, P.; Mukherjee, G. & Manna, S. Small-world properties of the Indian railway network *Physical Review E,* **2003,** *67,* 036106

[41] Latora, V. & Marchiori, M. Is the Boston subway a small-world network? *Physica A: Statistical Mechanics and its Applications*, **2002**, 314, 109-113

[42] Tobler, W. Automation and cartography *Geographical Review,* **1959**, 526-534

[43] Tobler, W. Migration fields *Population,* **1978,** 215-232

[44] Tobler, W. Experiments in migration mapping by computer *Cartography and Geographic Information Science,* **1987,** *14,* 155-163

[45] Haggett, P. & Chorley, R. Network analysis in geography, *Edward Arnold,* **1969**

[46] Giannotti, F.; Nanni, M.; Pinelli, F. & Pedreschi, D. Trajectory pattern mining *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* **2007**, 330-339

[47] Nanni, M. & Pedreschi, D. Time-focused clustering of trajectories of moving objects *Journal of Intelligent Information Systems,* **2006**, *27*, 267-289

[48] González, M.; Hidalgo, C. & Barabási, A. Understanding individual human mobility patterns *Nature,* **2008**, *453*, 779-782

[49] Liu, L.; Andris, C. & Ratti, C. Uncovering cabdrivers' behavior patterns from their digital traces *Computers, Environment and Urban Systems*, **2010**, *34*, 541-548

[50] Radil, S.; Flint, C. & Tita, G. Spatializing Social Networks: Using social network analysis to investigate geographies of gang rivalry, territoriality, and violence in Los Angeles *Annals of the Association of American Geographers,* **2010**, *100*, 307-326

[51] Onnela, J.; Arbesman, S.; Barabási, A. & Christakis, N. Geographic constraints on social network groups *Arxiv preprint arXiv:1011.4859,* **2010**

[52] Ratti, C.; Sobolevsky, S.; Calabrese, F.; Andris, C.; Reades, J.; Martino, M.; Claxton, R. and Strogatz, S. Redrawing the map of Great Britain from a network of human interactions *PloS one,* **2010**, *5*, e14248

[53] Davies, W. Urban connectivity in Montana *The Annals of Regional Science, Springer,* **1979**, *13*, 29-46

[54] Green, H. L., Hinterland boundaries of New York City and Boston in Southern New England, *Economic Geography*, **1955**, *31*, 283--300

[55] Scellato, S.; Mascolo, C.; Musolesi, M. & Latora, V. Distance matters: Geo-social metrics for online social networks *3rd Workshop on Online Social Networks- WOSN,* **2010**

[56] Leskovec, J. & Horvitz, E. Planetary-scale views on a large instant-messaging network *Proceeding of the 17th international conference on World Wide Web,* **2008**, 915-924

[57] Liben-Nowell, D.; Novak, J.; Kumar, R.; Raghavan, P. & Tomkins, A. Geographic routing in social networks *Proceedings of the National Academy of Sciences,* **2005**, *102*, 11623

[58] Lambiotte, R.; Blondel, V.; de Kerchove, C.; Huens, E.; Prieur, C.; Smoreda, Z. & Van Dooren, P. Geographical dispersal of mobile communication networks *Physica A: Statistical Mechanics and its Applications,* **2008**, *387*, 5317-5325

[59] Mayer, A. & Puller, S. The old boy (and girl) network: Social network formation on university campuses *Journal of Public Economics,* **2008**, *92*, 329-347

[60] Goldenberg, J. & Levy, M. Distance Is Not Dead: Social interaction and geographical distance in the Internet Era *Arxiv preprint arXiv:0906.3202,* **2009**

[61] Bettencourt, L.; Lobo, J. & Strumsky, D. Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size *Research Policy,* **2007***, 36*, 107-120

[62] Torrens P., Geography and Computational Social Science *GeoJournal*, **2010** *75*, 133-148

[63] Zook, M. The constraints and benefits of space and time in digital social networks *2010 Specialist Meeting—Spatio-Temporal Constraints on Social Networks, Santa Barbara, California, USA,* **2010**

[64] Lefebvre, H. & Enders, M. Reflections on the Politics of Space *Antipode, Wiley Online Library,* **1976***, 8*, 30-37

[65] Gastner, M. & Newman, M. The spatial structure of networks *The European Physical Journal B-Condensed Matter and Complex Systems,* **2006***, 49*, 247-252

[66] Batty, M. Cities and complexity *MIT Press, Cambridge, Massachusetts,* **2005**

[67] Kwan, M. Mobile communications, social networks, and urban travel: hypertext as a new metaphor for conceptualizing spatial interaction *The Professional Geographer,* **2007***, 59*, 434-446

[68] Cummins, S.; Curtis, S.; Diez-Roux, A. & Macintyre, S. Understanding and representing 'place' in health research: A relational approach *Social Science & Medicine,* **2007***, 65*, 1825-1838

[69] Friedman, T. The World is Flat: A brief history of the twenty-first century *Picador USA*, **2007**

[70] Castells, M. The Rise of the Network Society: The Information Age: Economy, Society, and Culture *Wiley-Blackwell, Cambridge, Massachusetts, USA,* **2009**

[71] Mitchell, W. City of bits: space, place, and the infobahn *The MIT Press, Cambridge, Massachusetts* **1996**

[72] Sassen, S. Global networks, linked cities *Routledge, New York, New York*, **2001**

[73] Sassen, S. A sociology of globalization *Norton, New York, New York* **2007**

REFERENCES FOR
CHAPTER 2: CHALLENGES

[1] Longley, P.; Goodchild, M.; Maguire, D. & Rhind, D. Geographical information systems and science *John Wiley & Sons Inc*, **2005**

[2] Strahler, A. Physical geography *Wiley-India,* **2007**

[3] Jensen, J. Introductory Digital Image Processing: A Remote Sensing Perspective *Prentice Hall, Upper Saddle River, NJ, USA,* **1995**

[4] Cowen, D. J. GIS versus CAD versus DBMS: What Are the Differences? *Photogrammetric Engineering and Remote Sensing*, **1988**, 54, 1551-1555

[5] Schon, D. The reflective practitioner *Basic Books New York*, **1983**

[6] Tomlinson, R. Thinking about GIS: geographic information system planning for managers *ESRI Press, Redlands, California,* **2007**

[7] Peet, R. Social theory, postmodernism, and the critique of development *Space and Social Theory: Interpreting Modernity and Postmodernity,* **1997**, 72-87

[8] Soja, E. The socio-spatial dialectic *Annals of the Association of American Geographers,* **1980***, 70*, 207-225

[9] Lefebvre, H. & Enders, M. Reflections on the Politics of Space *Antipode, Wiley Online Library,* **1976***, 8*, 30-37

[10] Shneiderman, B. Dynamic queries for visual information seeking *IEEE Software,* **1994***, 11*, 70-77

[11] Couclelis, H. From cellular automata to urban models: new principles for model development and implementation *Environment and Planning B,* **1997***, 24*, 165-174

[12] Woodcock, C. & Gopal, S. Fuzzy set theory and thematic maps: accuracy assessment and area estimation *International Journal of Geographical Information Science,* **2000***, 14*, 153-172

[13] Miller, H. & Han, J. Geographic data mining and knowledge discovery *CRC,* **2001**

[14] Batty, M. Cities and Complexity *MIT Press, Cambridge, Massachusetts,* **2007**

[15] Goodchild, M. & Janelle, D. Toward critical spatial thinking in the social sciences and humanities *GeoJournal,* **2010**, 75, 3-13

[16] Openshaw, S. & Openshaw, C. Artificial intelligence in geography *New York, John Wiley & Sons,* **1997**

[17] Assunção, R.; Neves, M.; Câmara, G. & Freitas, C. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees *International Journal of Geographical Information Science,* **2006,** *20*, 797-811

[18] Guo, D. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP) *International Journal of Geographical Information Science,* **2008,** *22*, 801-823

[19] Pickles, J. Tool or science? GIS, technoscience, and the theoretical turn *Annals of the Association of American Geographers*, **1997**, 87, 363-372

[20] French, S. & Wiggins, L. California planning agency experiences with automated mapping and geographic information systems *Environment and Planning B,* **1990,** *17*, 441-450

[21] Nedovic-Budic, Z. Evaluating the effects of GIS technology: Review of methods *Journal of Planning Literature*, **1999**, 13, 284

[22] Budic, Z. Effectiveness of Geographic Information Systems in local planning *Journal of the American Planning Association,* **1994**, 60, 244-263

[23] Wright, D.; Goodchild, M. & Proctor, J. GIS: Tool or science *Annals of the Association of American Geographers,* **1997**, 87, 346—362

[24] Unwin, A. & Unwin, D. Spatial data analysis with local statistics *Journal of the Royal Statistical Society: Series D (The Statistician),* **1998,** *47*, 415-421

[25] Nedovic-Budic, Z. Geographic Information Science implications for urban and regional planning *URISA-Washington DC-, The University of Wisconsin Press*, **2000**, 12, 81-93

[26] Batty, M. A decade of GIS: What next? *Environment and Planning B*, **2002**, 29, 2, 157-158

[27] The English Indices of Deprivation 2007. *Social Disadvantage Research Centre (SDRC) and Department of Social Policy and Social work at the University of Oxford. Department for Communities and Local Government, England, United Kingdom*, **2008**

[28] Boots, B. & A. Getis. Point pattern analysis *Sage University Paper Series on Quantitative Applications in the Social Sciences, Series no. 07-001.* **1988**

[29] Getis, A. Interactive modeling using second-order analysis *Environment and Planning A*, **1984** 16, 173-183

[30] Getis, A. & Ord, J. The analysis of spatial association by use of distance statistics *Geographical Analysis* **1992**, 24, 3

[31] Openshaw, S. & Taylor, P. The Modifiable Areal Unit Problem *In: Quantitative Geography: A British View.* Ed. N. Wrigley and R. Bennett: *Routledge and Kegan Paul, London,* **1981**, 60-69

[32] Bailey, T. & Gatrell, A. Interactive spatial data analysis *Longman Scientific & Technical Harlow, UK,* **1995**

[33] Guo, D.; Chen, J.; MacEachren, A. & Liao, K. A visualization system for space-time and multivariate patterns (vis-stamp) *IEEE Transactions on Visualization and Computer Graphics*, **2006**, *12*, 1461-1474

[34] Calabrese, F.; Reades, J.; & Ratti, C. Eigenplaces: Segmenting space through digital signatures *IEEE Pervasive Computing*, **2009**, *9*, 78-84

[35] Andris, C. Weighted Radial Variation for node feature classification. *Arxiv preprint arXiv:1102.4873,* **2011**

[36] Skupin, A. Where do you want to go today [In attribute space]? In: *Societies and Cities in the Age of Instant Access* Ed: Miller, H. *Springer,* **2007**

[37] Fruchterman, T. & Reingold, E. Graph drawing by force-directed placement *Software: Practice and Experience* **1991**, *21*, 1129-1164

[38] Adams, P. A taxonomy for communication geography *Progress in Human Geography*, **2011**, *35*, 37-57

[39] Tobler. W. Automation and cartography *Geographical Review*, **1959**, *49*, 526–534

[40] Tobler, W. Experiments in migration mapping by computer *Cartography and Geographic Information Science*, **1987**, *14*, 155–163

[41] FlowMapper Software *Center for Spatially Integrated Social Science, University of California—Santa Barbara,* **2005**

[42] Hirtle, S. & Jonides, J. Evidence of hierarchies in cognitive maps *Memory & Cognition,* **1985,** *13*, 208-217

[43] Golledge, R. The nature of geographic knowledge *Annals of the Association of American Geographers*, **2002**, *92*, 1-14

[44] Montello, D. The perception and cognition of environmental distance: Direct sources of information In: *Spatial Information Theory: A Theoretical Basis for GIS*, **1997**, 297-311

[45] Battersby, S. & Montello, D. Area estimation of world regions and the projection of the global-scale cognitive map *Annals of the Association of American Geographers,* **2009**, *99*, 273-291

[46] Slocum, T. Thematic cartography and visualization *Prentice Hall Upper Saddle River, NJ,* **1999**

[47] Harrower, M. & Brewer, C. Colorbrewer.org: an online tool for selecting colour schemes for maps *Cartographic Journal,* **2003**, *40*, 27-37

[48] Plewe, B. GIS online: Information retrieval, mapping, and the Internet *OnWord Press*, **1997**

[49] Shneiderman, B. & Plaisant, C. Designing the user interface *Addison-Wesley Reading, Massachusetts,* **1998**

[50] Davis, R.; Shrobe, H. & Szolovits, P. What is a knowledge representation? *AI Magazine*, **1993**, *14*, 17

[51] Takatsuka, M. & Gahegan, M. GeoVISTA Studio: A codeless visual programming environment for geoscientific data analysis and visualization *Computers & Geosciences*, **2002**, *28*, 1131-1144

[52] Anselin, L.; Syabri, I. & Kho, Y. GeoDa: An introduction to spatial data analysis *Geographical Analysis*, **2006,** 38, 5-22

[53] Tukey, J. We need both exploratory and confirmatory *The American Statistician,* **1980**, *34*, 23-25

[54] Keim, D. Information visualization and visual data mining *IEEE Transactions on Visualization and Computer Graphics,* **2002**, *8*, 1-8

[55] MacEachren, A. & Kraak, M. Exploratory cartographic visualization: advancing the agenda *Computers and Geosciences,* **1997**, *23*, 335-343

[56] MacEachren, A.; Gahegan, M.; Pike, W.; Brewer, C.; Cai, G.; Lengerich, E. & Hardisty, F. Geovisualization for knowledge construction and decision support IEEE *Computer Graphics and Applications*, **2004**, *24*, 13—17

[57] MacEachren, A. & Taylor, D. Visualization in modern cartography *Pergamon,* **1994**

[58] Plaisant, C.; Shneiderman, B.; Doan, K. & Bruns, T. Interface and data architecture for query preview in networked information systems *ACM Transactions on Information Systems (TOIS),* **1999**, *17*, 341

[59] Guo, D. Flow mapping and multivariate visualization of large spatial interaction data *IEEE Transactions on Visualization and Computer Graphics,* **2010**, *15*, 1041-1048

[60] Borgatti, S.; Everett, M. & Freeman, L. UCINET for Windows: Software for social network analysis *Harvard Analytic Technologies,* **2002**

[61] Batagelj, V. & Mrvar, A. Pajek—analysis and visualization of large networks *Graph Drawing,* **2002**, 8-11

[62] Jackson, M. Social and economic networks *Princeton University Press,* **2008**

[63] Moran, P. A test for the serial independence of residuals *Biometrika,* **1950**, *37*, 178-181

[64] Geary, R. The contiguity ratio and statistical mapping *The Incorporated Statistician,* **1954**, *5*, 115-146

[65] Ord, K. Estimation methods for models of spatial interaction *Journal of the American Statistical Association*, **1975**, 70, 120-126

[66] Cressie, N. Statistics for spatial data *Terra Nova,* **1992**, *4*, 613-617

[67] Anselin, L. Local indicators of spatial association-LISA *Geographical Analysis,* **1995**, *27*, 93—115

[68] Glennon, A. Creating and validating object-oriented geographic data models: Modeling flow within GIS *Transactions in GIS*, **2010**, *14*, 23-42

[69] Doytsher, Y.; Galon, B. & Kanza, Y. Querying geo-social data by bridging spatial networks and social networks *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, **2010**, 39-46

[70] Tobler, W. On the first law of geography: A reply *Annals of the Association of American Geographers,* **2004**, *94*, 304-310

[71] Goodchild, M. & Glennon, A. Representation and Computation of Geographic Dynamics *Understanding dynamics of geographic domains, CRC,* **2008**, 13

[72] Cova, T. & Goodchild, M. Extending geographical representation to include fields of spatial objects *International Journal of Geographical Information Science,* **2002**, *16*, 509-532

[73] Ferreira, J. Database management tools for planning *Journal of the American Planning Association*, **1990**, *56*, 78-84

[74] Unwin, A. & Unwin, D. Exploratory spatial data analysis with local statistics *Journal of the Royal Statistical Society. Series D (The Statistician),* **1998**, *47*, 415-421

[75] Lösch, A. Die räumliche ordnung der wirtschaft *G. Fischer,* **1944**

[76] Weber, A. Uber den Standort der Industrien. Erster Teil. Reine Theorie der Standorte *Mohr, Tübingen,* **1909**

[77] Reilly, W. The Law of Retail Gravitation *New York: Pilsbury Publishers, Inc,* **1953**

[78] Gras, N. The development of metropolitan economy in Europe and America *The American Historical Review,* **1922**, *27*, 695-708

[79] Andrews, R. Mechanics of the urban economic base: Special problems of base identification *Land Economics,* **1954**, *30*, 260-269

[80] Ben-Akiva, M. & Lerman, S. Discrete choice analysis: theory and application to travel demand *The MIT Press, Cambridge, Massachusetts,* **1985**

[81] Karlsson, C. & Olsson, M. The identification of functional regions: theory, methods, and applications *The Annals of Regional Science,* **2006**, *40*, 1-18

[82] Limtanakool, N.; Dijst, M. & Schwanen, T. A theoretical framework and methodology for characterising national urban systems on the basis of flows of people: empirical evidence for France and Germany *Urban Studies,* **2007**, *44*, 2123

[83] Davies, W. Urban connectivity in Montana *The Annals of Regional Science, Springer,* **1979**, *13*, 29-46

[84] Holmes, J. Telephone traffic dispersion and nodal regionalisation in the Australian states *Australian Geographical Studies,* **1983**, *21, 231-250*

[85] Straffin Jr, P. Linear algebra in geography: eigenvectors of networks *Mathematics Magazine,* **1980**, 269-276

[86] Castells, M. The Rise of the Network Society. *Wiley-Blackwell, Cambridge, Massachusetts,* **2000**

[87] Worboys, M. & Duckham, M. GIS: A computing perspective *CRC,* **2004**

[88] Fonseca, F., Egenhofer, M., Davis C., & Borges, F. Ontologies and knowledge sharing in urban GIS, *Computers, Environment and Urban Systems*, **2000,** *24*, 251-271

REFERENCES FOR
CHAPTER 6: DISCUSSION

[1] De Blij, H. Why geography matters: three challenges facing America: climate change, the rise of China, and global terrorism *Oxford University Press,* **2005**

[2] Von Thünen, J., *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationaloekonomie Jena,* **1910**

[3] Christaller, W. Die zentralen Orte in Süddeutschland. *Gustav Fischer, Jena,* **1933**

[4] Lösch, A. Die räumliche ordnung der wirtschaft *G. Fischer,* **1944**

[5] Weber, A. Theory of the Location of Industries [translated by CJ Friedrich from Weber's 1909 book] *Chicago: The University of Chicago Press,* **1929**

[6] Reilly, W. The Law of Retail Gravitation *New York: Pilsbury Publishers, Inc,* **1953**

[7] Huff, D. Defining and estimating a trading area *The Journal of Marketing,* **1964**, *28*, 34-38

[8] Gras, N. An introduction to economic history *Harper & Brothers,* **1922**

[9] Onnela, J.; Arbesman, S.; González, M.; Barabási, A. & Christakis, N. Geographic constraints on social network groups *PloS one,* **2011**, *6*, e16939

[10] Sevtsuk, A. and Mekonnen, M. Urban network analysis: A toolbox for ArcGIS 10. *Massachusetts Institute of Technology City Form Research Group, Accessed at: http://cityform.mit.edu /projects/urban network - analysis.html*, **2011**

[11] Meyer, M. & Miller, E. Urban transportation planning: a decision-oriented approach *McGraw-Hill New York,* **2001**

[12] Kwan, M. Mobile communications, social networks, and urban travel: hypertext as a new metaphor for conceptualizing spatial interaction *The Professional Geographer,* **2007**, *59*, 434-446

[13] Arrow, K. An extension of the basic theorems of classical welfare economics *Cowles Commission for Research in Economics: University of Chicago*, **1952**

[14] Nash, J. Equilibrium points in n-person games *Proceedings of the National Academy of Sciences,* **1950**, *36*, 48-49

[15] Schelling, T. Micromotives and macrobehavior *Norton New York,* **1978**

[16] Granovetter, M. The strength of weak ties *American Journal of Sociology,* **1973**, *78*, 1360

[17] Blau, P. Exchange and power in social life *Transaction Publishers,* **1986**

[18] Burt, R. The gender of social capital *Rationality and Society,* **1998,** *10*, 5-46

[19] Fernandez, R.; Castilla, E. & Moore, P. Social capital at work: Networks and employment at a phone center *American Journal of Sociology, JSTOR,* **2000,** *105*, 1288-1356

[20] Barabasi, A. & Albert, R. Emergence of scaling in random networks *Science,* **1999,** *286*, 509

[21] McPherson, M.; Smith-Lovin, L. & Cook, J. Birds of a feather: Homophily in social networks *Annual review of sociology,* **2001,** *27*, 415-444

[22] Girvan, M. & Newman, M. Community structure in social and biological networks *Proceedings of the National Academy of Sciences,* **2002,** *99*, 7821

[23] Ebel, H.; Mielsch, L. & Bornholdt, S. Scale-free topology of e-mail networks *Physical Review E, APS,* **2002,** *66*, 35103

[24] Albert, R.; Jeong, H. & Barabasi, A. Error and attack tolerance of complex networks *Nature,* **2000,** *406*, 378-382

[25] Granovetter, M. Social Science: Ignorance, Knowledge, and Outcomes in a Small World *Science,* **2003,** *301*, 773

[26] Kearns, M.; Suri, S. & Montfort, N. An experimental study of the coloring problem on human subject networks *Science,* **2006,** *313*, 824

[27] Christakis, N. & Fowler, J. The spread of obesity in a large social network over 32 years *New England Journal of Medicine,* **2007,** *357*, 370

[28] Christakis, N. & Fowler, J. The collective dynamics of smoking in a large social network *New England Journal of Medicine,* **2008,** *358*, 2249

[29] Watts, D. & Dodds, P. Influentials, networks, and public opinion formation *Journal of Consumer Research,* **2007,** *34*, 441-458

[30] Clauset, A., Moore, C. & Newman, M. Hierarchical structure and the prediction of missing links in networks *Nature,* **2008,** *453*, 98-101

[31] Centola, D. & Wilier, R. The Emperor's Dilemma *Theories of Social Order: A Reader, Stanford Social Sciences,* **2009**, 276

[32] Bettencourt, L.; Lobo, J. & Strumsky, D. Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size *Research Policy,* **2007,** *36*, 107-120

[33] Arbesman, S.; Kleinberg, J. & Strogatz, S. An Explanation of Superlinear Scaling for Innovation in Cities *Imprint,* **2008,** *30*, 5

[34] Salganik, M. & Watts, D. Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market *Social Psychology Quarterly,* **2008,** *71*, 338-355

[35] Jackson, M. Social and Economic Networks *Princeton University Press,* **2008**

[36] Watts, D. & Strogatz, S. Collective dynamics of 'small-world' networks *Nature,* **1998,** *393*, 440-442

[37] Erdos, P. & Renyi, A. On random graphs I *Publ. Math. Debrecen,* **1959,** *6*, 156

[38] Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabasi, A.; Brewer, D., C. N.; Contractor, N.; Fowler, J.; Gutmann, M. et al. Life in the network: the coming age of computational social science *Science,* **2009,** *323*, 721

[39] Christakis, N. & Fowler, J. Connected: The surprising power of our social networks and how they shape our lives *Little, Brown and Company,* **2009**

[40] Barabási, A. Linked: How everything is connected to everything else and what it means *Plume,* **2003**

[41] Barabási, A. Bursts: The hidden pattern behind everything we do *EP Dutton,* **2010**

[42] Watts, D. Six degrees: The science of a connected age *WW Norton & Company,* **2004**

[43] Fisher, L. The perfect swarm: The science of complexity in everyday life *Basic Books,* **2009**

[44] Gladwell, M. Blink: The power of thinking without thinking *Little, Brown and Company,* **2005**

[45] Gladwell, M. The tipping point: How little things can make a big difference *Little, Brown and Company,* **2000**

[46] Batty, M. Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals *The MIT Press, Cambridge, Massachusetts,* **2007**

[47] Unwin, A. & Unwin, D. Spatial Data Analysis with Local Statistics *Journal of the Royal Statistical Society: Series D (The Statistician),* **1998**, 47, 415-421

[48] Nedovic-Budic, Z. Geographic information science implications for urban and regional planning *URISA-Washington Dc-, The University of Wisconsin Press,* **2000**, 12, 81-93

[49] Davis, R.; Shrobe, H. & Szolovits, P. What is a knowledge representation? *AI Magazine,* **1993,** *14*, 17

[50] Borgatti, S.; Everett, M. & Freeman, L. UCINET for Windows: Software for social network analysis *Harvard Analytic Technologies,* **2002**

[51] Batagelj, V. & Mrvar, A. Pajek—analysis and visualization of large networks *Graph Drawing,* **2002**, 8-11

[52] Moran, P. A test for the serial independence of residuals *Biometrika,* **1950**, *37*, 178-181

[53] Geary, R. The contiguity ratio and statistical mapping *The Incorporated Statistician,* **1954**, *5*, 115-146

[54] Cressie, N. Statistics for spatial data *Terra Nova, Wiley Online Library,* **1992**, *4*, 613-617

[55] Anselin, L. Local indicators of spatial association-LISA *Geographical Analysis,* **1995**, *27*, 93—115

[56] Limtanakool, N.; Schwanen, T. & Dijst, M. Developments in the Dutch urban system on the basis of flows *Regional Studies: The Journal of the Regional Studies Association,* **2009**, *43*, 179-196

[57] Xu, Z. & Sui, D. Small-world characteristics on transportation networks: a perspective from network autocorrelation *Journal of Geographical Systems,* **2007**, *9*, 189-205

[58] Liu, L.; Andris, C. & Ratti, C. Uncovering cabdrivers' behavior patterns from their digital traces *Computers, Environment and Urban Systems*, **2010**, *34*, 541-548

[59] Xu, Z. & Harriss, R. Exploring the structure of the US intercity passenger air transportation network: a weighted complex network approach *GeoJournal,* **2008**, *73*, 87-102

[60] De Montis, A.; Chessa, A.; Campagna, M.; Caschili, S. & Deplano, G. Modeling commuting systems through a complex network analysis *Journal of Transport and Land Use,* **2010**, *2*, 39-55

[61] Barrat, A.; Barthelemy, M.; Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks *Proceedings of the National Academy of Sciences*, **2004**, 101, 3747

[62] Sen, P.; Dasgupta, S.; Chatterjee, A.; Sreeram, P.; Mukherjee, G. & Manna, S. Small-world properties of the Indian railway network *Physical Review E,* **2003**, *67*, 036106

[63] Latora, V. & Marchiori, M. Is the Boston subway a small-world network? *Physica A: Statistical Mechanics and its Applications*, **2002**, 314, 109-113

[64] O'Kelly, M. A geographer's analysis of hub-and-spoke networks *Journal of Transport Geography,* **1998**, *6*, 171-186

[65] Radil, S.; Flint, C. & Tita, G. Spatializing Social Networks: Using social network analysis to investigate geographies of gang rivalry, territoriality, and violence in Los Angeles *Annals of the Association of American Geographers,* **2010**, *100*, 307-326

[66] Onnela, J.; Arbesman, S.; Barabási, A. & Christakis, N. Geographic constraints on social network groups *Arxiv preprint arXiv:1011.4859,* **2010**

[67] Ratti, C.; Sobolevsky, S.; Calabrese, F.; Andris, C.; Reades, J.; Martino, M.; Claxton, R. and Strogatz, S. Redrawing the map of Great Britain from a network of human interactions *PloS One,* **2010**, *5*, e14248

[68] Davies, W. Urban connectivity in Montana *The Annals of Regional Science, Springer,* **1979**, *13*, 29-46

[69] Green, H. L., Hinterland boundaries of New York City and Boston in Southern New England, *Economic Geography*, **1955**, *31*, 283—300

[70] Torrens P., Geography and Computational Social Science *GeoJournal*, **2010**, *75*, 133-148

[71] Assunção, R.; Neves, M.; Câmara, G. & Freitas, C. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees *International Journal of Geographical Information Science,* **2006**, *20*, 797-811

[72] Guo, D. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP) *International Journal of Geographical Information Science,* **2008**, *22*, 801-823

References for
Chapter 7: Conclusion

[1] Mayer, A. & Puller, S. The old boy (and girl) network: Social network formation on university campuses *Journal of Public Economics,* **2008**, *92*, 329-347

[2] Lambiotte, R.; Blondel, V.; de Kerchove, C.; Huens, E.; Prieur, C.; Smoreda, Z. & Van Dooren, P. Geographical dispersal of mobile communication networks *Physica A: Statistical Mechanics and its Applications,* **2008**, *387*, 5317-5325

[3] Leskovec, J. & Horvitz, E. Planetary-scale views on a large instant-messaging network *Proceeding of the 17th international conference on World Wide Web,* **2008**, 915-924

[4] Levy, J. Contemporary Urban Planning *Prentice Hall, New Jersey,* **2003**

[5] Riis, J. How the other half lives: Studies among the tenements of New York *Dover Publications,* **1901**

[6] Carson, R. Silent spring *New York: Houghton Mifflin,* **1962**

[7] Brabyn, L. and Skelly, C. Modeling population access to New Zealand public hospitals *International Journal of Health Geographics,* **2002**, 1, 3

[8] Centola, D. & Macy, M. Complex Contagions and the Weakness of Long Ties 1 *American Journal of Sociology,* **2007**, *113*, 702-734

[9] Schelling, T. Micromotives and macrobehavior *Norton, New York,* **1978**

[10] Mischel, W. Introduction to Personality, Sixth Edition *Harcourt Brace, Fort Worth, Texas* **1999**

[11] Minksy, M. Society of Mind, *MIT Press, Cambridge, Massachusetts* **2005**

[12] Tulving, E. Elements of Episodic Memory *Oxford: Clarendon Press,* **1983**

[13] Picard, R. Affective Computing *Cambridge: MIT Press, Cambridge, Massachusetts* **2002**

[14] Hatfield, E., Cacioppo, J. T., & Rapson, R. L. Emotional contagion. *Cambridge University Press, New York,* **1993**

[15] Hatfield, E. & Rapson, R. L. Emotional contagion. In W. E. Craighead & C. B. Nemeroff (Eds.). The Corsini encyclopedia of psychology and behavioral science. *John Wiley & Sons, New York,* **2000** 493-495.

[16] Christakis, N. & Fowler, J. The spread of obesity in a large social network over 32 years *New England Journal of Medicine,* **2007**, 357- 370

[17] Cutter, S. Vulnerability to environmental hazards *Progress in Human Geography,* **1996**, *20*, 529

[18] Cutter, S.; Boruff, B. & Shirley, W. Social Vulnerability to Environmental Hazards* *Social Science Quarterly,* **2003**, *84*, 242-261

[19] Elliott, J. & Pais, J. Race, class, and Hurricane Katrina: Social differences in human responses to disaster *Social Science Research, Elsevier,* **2006**, 35, 295-321

[20] Eisenman, D.; Cordasco, K.; Asch, S.; Golden, J. & Glik, D. Disaster planning and risk communication with vulnerable communities: lessons from Hurricane Katrina *American Journal of Public Health, Am Public Health Assoc,* **2007**, *97*-109

[21] Boothe, D. Why are so Many Black Men in Prison? *USA: Full Surface Publishing,* **2007**

[22] Sevtsuk, A. & Ratti, C. iSPOTS. How Wireless Technology is Changing Life on the MIT Campus *9th International Conference on Computers in Urban Management and Urban Planning, University College London, London,* **2005**

[23] Rojas, F. Unpublished work, First Year Paper *Massachusetts Institute of Technology, Dept. of Urban Studies and Planning, City Development and Design,* **2005**

[24] Lareau, A. Unequal childhoods: Class, race, and family life *University of California Press,* **2003**

[25] Macleod, J. Ain't no makin' it: Leveled aspirations in a low-income community *Boulder, CO: Westview Press,* **1987**

[26] Hickins, M. MIT Prof: Data Privacy Is Your Problem (or Asset) *May 19th Technology News and Insights,* **2011**

[27] Schrag, Z. The great society subway: A history of the Washington Metro *The John Hopkins University Press, Baltimore, Maryland,* **2006**

[28] Monmonier, M. Spying with maps *University of Chicago Press,* **2002**

[29] Clarke, K. & Cloud, J. On the origins of analytical cartography *Cartography and Geographic Information Science, Cartography and Geographic Information Society,* **2000**, *27*, 195-204

[30] Ferreira Jr, J. Comment on Drummond and French: GIS Evolution: Are We Messed Up by Mashups? *Journal of the American Planning Association,* **2008**, *74*, 177-179

[31] Longley, P.; Goodchild, M.; Maguire, D. & Rhind, D. Geographical information systems and science *John Wiley & Sons Inc*, **2005**

[32] Giannotti, F.; Nanni, M.; Pinelli, F. & Pedreschi, D. Trajectory pattern mining *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining,* **2007**, 330-339

[33] Latour, B. Science in action *Harvard University Press,* **1987**

[34] Latour, B. We have never been modern *Harvard University Press,* **1993**

[35] Langford, M. & Higgs, G. Measuring potential access to primary healthcare services: the influence of alternative spatial representations of population *The Professional Geographer*, **2006**, *58*, 294–306