12-15-2014

# Understanding Geo-Social Network Patterns: Computation, Visualization, and Usability

Caglar Koylu
*University of South Carolina - Columbia*

# UNDERSTANDING GEO-SOCIAL NETWORK PATTERNS: COMPUTATION, VISUALIZATION, AND USABILITY

by

Caglar Koylu

Bachelor of Science
Middle East Technical University, 2004

Master of Science
Middle East Technical University, 2008

---

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Geography

College of Arts and Sciences

University of South Carolina

2014

Accepted by:

Diansheng Guo, Major Professor

Sarah E. Battersby, Committee Member

Michael E. Hodgson, Committee Member

David B. Hitchcock, Committee Member

Lacy Ford, Vice Provost and Dean of Graduate Studies

## DEDICATION

In memory of my beloved grandfather, İsmail Erdemir

ACKNOWLEDGEMENTS

I am sincerely thankful to all the people who have been supporting, guiding and encouraging me throughout my doctoral program. I am deeply grateful to my advisor Dr. Diansheng Guo who has always been supportive and given me the freedom to pursue various projects. I have learned a great deal from his unique perspective and inspirational advice on research. I would like to thank the committee members, Drs. Sarah Battersby, Michael Hodgson and David Hitchcock for their valuable feedback, insight and time.

I gratefully acknowledge the funding sources that made the development and implementation of my Ph.D. work possible. This dissertation was in large part supported by The National Science Foundation Grant No. 0748813. Additional funding was provided by The Institute of Museum and Library Services Grant No. LG-00-14-0030-14; The U.S. Department of Homeland Security through The Hazards and Vulnerability Research Institute, The Department of Geography, and The Graduate School at The University of South Carolina (USC).

I would like to thank all the faculty, staff and fellow graduate students in the Department of Geography for creating a friendly and motivating work environment. Specifically, I would like to thank Drs. Susan Cutter, Alice Kasakoff, John Kupfer, Amy Mills, and Chris Emrich for all their support, advice and encouragement.

My special thanks go to Mary Windsor for her faithful support during the final stages of this Ph.D. Also, I would like to thank my dear friends, Ian Kramer, Aysegul Yeniaras, Lambert Kramer, John Lauermann, Kevin Ash, Matt Rodgers, Hai Jin, Ke Liao, Xi Zhu, Yuan Huang, Chao Chen, Tianhua Shao, Haiqing Xu, Adam Ereth, Ronnie Schumann, Leanne Sulewski, Balkan Uraz and my cousin Mert Erdemir for their encouragement and support.

Finally, and most importantly, I would like to thank my beloved parents Hürriyet and Süha Köylü for all their love, and support throughout my life. I wouldn't have been able to complete much of what I have done and become who I am without you both in my life.

## ABSTRACT

Geo-social networks are formed by flows of physical entities (e.g., humans, vehicles, sensors, animals), and communication (e.g., information, ideas, innovation) that connect places to places and individuals to individuals. Several major problems remain to be addressed for understanding the complex patterns in geo-social networks. This dissertation makes the following contributions to the theory and methodologies that aim at understanding complex geo-social data by integrating methods of computation, visualization and usability evaluation.

Chapter 2 introduces a novel network-based smoothing approach that addresses the size-difference and small area problem in calculating and mapping locational (graph) measures in spatial interaction networks. The new approach is a generic framework that can be used to smooth various graph measures which help examine multi-space and multi-scale characteristics of geo-social data.

Chapter 3 introduces a space-time visualization approach to discover spatial, temporal and relational patterns in a dynamic geo-social network embedded in space and time. By developing and visualizing a measure of connectedness across space and time, the new approach facilitates the discovery of hot spots (hubs, where connectedness is strong) and the changing patterns of such spots across space and time.

Chapter 4 introduces a series of user evaluations to obtain knowledge on how map readers perceive information presented with flow maps, and how design factors such as

flow line style (curved or straight) and layout characteristics may affect flow map perception and users' performance in addressing different tasks for pattern exploration. The findings of this study have significant implications for iterative design, interaction strategies and further user experiments on flow mapping.

TABLE OF CONTENTS

LIST OF TABLES

## LIST OF FIGURES

## List of Abbreviations

ANOVA ..... Analysis of Variance

CFE ..... Chaffee

GIScience ..... Geographic Information Science

IDW ..... Inverse-Distance Weighting

MAUP ..... The Modifiable Areal Unit Problem

SNA ..... Social Network Analysis

UGC ..... User Generated Content

VGI ..... Volunteered Geographic Information

# CHAPTER 1

## INTRODUCTION

With the increasing availability of affordable long distance travel and advancements in communication technology, in the 90s it was predicted that the effect of distance and geography may cease to play a role in shaping our world (Atkinson, 1998; Cairncross, 2001; Ohmae, 1990). However, the growth of cities in the past two decades has challenged the theory of "death of space" with the evidence of increased centralized investment, and infrastructure development that are mostly urban. Castells (1996) rejected the contention that space will disappear upon the technological advancements in travel and communication, and he described space as "the material support of time-sharing social practices". This new definition of space is built upon the interaction between geography, time and "the network society" through the flows of both physical and intangible phenomenon such as people, commodities, flights, money, information, ideas and innovation. This dissertation uses the concept of geo-social networks to describe and study the complex system created by the flows of physical and intangible phenomenon from an integrated view of geographic information science (GIScience) and network science.

Given the complexity introduced by highly interacting systems of geo-social networks, solving a real-world problem often requires simultaneous consideration of the geographic, temporal and network components that form the system and relationships

among them. For example, disease spread is influenced by many factors such as human mobility, social ties, population dynamics, transportation infrastructure and seasonal changes in weather, and the relationships among these factors. Human migration is driven by not only the availability of jobs or average temperature at a destination place but also the availability of family and friendship ties. During an emergency, knowing how information diffuses among geographic locations and through a network of social actors is crucial for managing situational awareness and assisting emergency management.

Geo-social networks are formed by flows of physical entities (e.g., movements of humans, vehicles, sensors, animals), and communication (e.g., information, ideas, innovation, personal communication) that connect places to places and individuals to individuals. A geo-social network is often large (e.g., county-to-county migration data have 3000 counties and millions of migrants) and involves complex patterns (1) in multiple spaces (e.g., geographic space, network space, multivariate space) with components of spatial information, multivariate factors (e.g., demographics of migrants, types of commodities), and network structures (e.g., connections between individuals, groups and regions) (2) at multiple scales (e.g., national patterns, regional patterns, local patterns); and (3) that are dynamic as their geographic, multivariate, and network characteristics change over time. From here on, the term "geo-social network patterns" will be used to refer to the multi-space, multi-scale and dynamic complexity of patterns in geo-social networks.

Building upon the strong conceptual and methodological overlaps between geography and network science, recent studies (Andris, 2011; Faust & Lovasi, 2012; Luo & MacEachren, 2014) identified the challenges and potential future research for bridging

the gap between network analysis and spatial analysis. Along this path, newly developed computational and visual tools (Gao et al., 2013; Guo, 2009; Kumar, Morstatter, & Liu, 2014; Luo, et al., 2014) have shown great promise for integrating the approaches of geography and network science through the use of human-computer collaboration. Although these approaches provided significant contribution in the area, several major problems remain to be addressed for understanding the complex patterns in geo-social networks. The goal of this dissertation is to address the following challenges. First, there is a need of new approaches that integrate network and spatial analysis to remove or reduce confounding effects such as the problem of size-differences, small area problem and the modifiable areal unit problem (MAUP) (Openshaw, 1983) that conceals true patterns. Second, there is still a lack of research that examines various aspects and interactions within the multi-space, multi-scale and dynamic complexity. Third, although flow maps are commonly used to explore geo-social network patterns, very little is known about how users perceive and use flow maps, and how different flow map designs and tasks influence flow map comprehension.

Addressing the above challenges for the understanding of complex geo-social networks requires the methods of computation, visualization and usability evaluation. This dissertation makes the following three contributions to these efforts. Chapter 2 introduces a new network-based smoothing approach that addresses the size-difference and small area problems in calculating and mapping locational (graph) measures in spatial interaction networks. The new approach is a generic framework that can be used to smooth various graph measures and help examine multi-space and multi-scale characteristics of geo-social data. Chapter 3 introduces a space-time visualization

approach to discover spatial, temporal and relational patterns in a dynamic geo-social network embedded in space and time. By developing and visualizing a measure of connectedness across space and time, the new approach facilitates the discovery of hot spots (hubs, where connectedness is strong) and the changing patterns of such spots across space and time. Chapter 4 introduces a series of user evaluations for understanding: how map readers perceive information presented with flow maps; how major factors for flow map reading such as flow line style (curved or straight) and layout characteristics; and how different tasks for pattern exploration influence flow map perception.

The remainder of this chapter is divided into following sub-sections. Section 1.1 presents a general overview and categorization of geo-social networks. Section 1.2 discusses relevant literature in GIScience and Network Science, and integrated approaches for analyzing geo-social networks. This chapter concludes with a discussion of the dissertation structure.

## 1.1 GEO-SOCIAL NETWORKS

We could categorize geo-social networks into area-based and point-based. In area based networks, the original locations are aggregated into a small set of areas and a cumulative network of flows is created between geographical areas. Area-based networks are often referred to as spatial interaction networks (e.g., county-to-county migration flows, state-to-state commodity flows). Point-based networks are formed by actors at discrete locations. Examples of point-based networks include location-based social networks, networks of social media, and mobility data (e.g., tweets, taxi trips, human mobility and animal movement).

County-to-county migration within the U.S. represents one of the most commonly used area-based geo-social networks. There are over 3000 nodes (counties) and millions of links (migrants). Each link within this network contains an origin county, a destination county, the counts of migrants moved and migrant characteristics (e.g., counts of migrants for each income level that move from that specific origin to that specific destination). Area-based geo-social networks often suffer the modifiable area unit problem and size-difference problem. Original locations are aggregated into a set of areas (e.g., county), and flows represent cumulative movements or connections between those areas. Different aggregation naturally results in a different network and arbitrary aggregation may cause missing major patterns. Also, areal units usually differ in size (e.g., population) which may conceal (instead of reveal) the true underlying spatial and network structures. For example, Los Angeles County has a population over 9 million whereas Loving County in Texas has a population less than 100. As larger counties receive and send more flows, flow volumes between the large and small counties are not directly comparable. Also, area-based networks suffer from the small area problem as flow volume between small areas is unstable as a result of the small populations. To obtain insight into true patterns, confounding effects such as MAUP, size-difference and small area problem need to be addressed.

Social media and networking applications makes it possible to collect large geo-social data from a variety of sources such as Twitter, Foursquare, Flickr, Tumblr and Yelp. Such data naturally produce a point-based geo-social network. For example, 500 million tweets are generated everyday across the world. Approximately 15 million of these tweets have geographic coordinates and about 20 million tweets include a

geographic reference (e.g., place name) in the message content. Such data naturally form massive networks that contain actors (users) and links. Links could be directly formed by each user's connections (e.g., friends, followers, pins) or they could be indirectly derived from interactions between actors such as shared text (e.g., re-tweets, hashtags), videos, images and web pages. Applications such as Twitter and Facebook have hundreds of millions of active users and thus the number of links within such networks easily exceeds billions. Moreover, such networks evolve (change) over space and time as individuals change location, new individuals are added or removed; relations between users develop and change over time; new content is being generated and shared by individuals.

## 1.2 Literature Review

### 1.2.1 Geographic Information Science (GIScience)

Existing methods of spatial interaction analysis in GIScience could be classified as descriptive statistics, modeling approaches and exploratory approaches. Methods of descriptive statistics such as summary tables, histograms, scatter plots, pie-charts, line-charts, and bar-charts (Cadwallader, 1992; Morrill, 1988; Pooley & Turnbull, 1998) have been widely used to provide summaries of spatial interaction data in conjunction with some visual displays. Although these methods are useful in discovering general patterns, they are not effective when the data are large and have multi-space complexity.

#### 1.2.1.1 Spatial Interaction Modeling

Spatial interaction modeling has been widely employed in many research fields. Gravity models are the examples of the earliest spatial interaction models that predict the interaction between two locations using a function of the attributes of those locations and

the distance between them. Gravity models were then replaced by the more general concepts of entropy. Entropy and information theoretical models provided a statistical framework in the prediction of interactions. Models that utilize a probabilistic framework were then widely developed in the fields of spatial econometrics, physics, geography, and transportation modeling. A detailed review of spatial interaction modeling could be found in (Roy & Thill, 2004). Spatial interaction models are theory-driven and they are used to predict the effects and causes of interactions based on some theoretical assumptions. Spatial interaction models are inadequate for analyzing spatial interaction data because of many reasons such as theoretical assumptions, exclusion of critical factors out of the model, and not being able to address the challenge that result from high dimensionality of spatial interaction data. For example, to discover the influence of location characteristics on migration, spatial interaction models (Dorigo & Tobler, 1983; W.H. Frey, Liaw, Xie, & Carlson, 1995) have been developed that predict flows of migrants based on a set of origin-destination characteristics. However, the selection of the origin-destination characteristics within these models are determined by previously known hypotheses and the models incorporate only a small selection of these characteristics while disregarding information that is not included in the model. Other examples of spatial interaction models focus on discovering migration patterns between different sizes of settlements within the urban hierarchy (W. H. Frey, 2005; L. Long, 1988; David A. Plane & Jurjevich, 2009; Pooley, 1979; C. C. Roseman, 1977) and uncovering the influence of location characteristics on attracting or distracting migrants (Dorigo & Tobler, 1983; W.H. Frey et al., 1995).

#### 1.2.1.2 REGIONALIZATION AND GRAPH PARTITIONING

Regionalization is a method to group spatially contiguous regions based on an objective function (e.g., connectivity within regions are maximized) (Wise, Haining, & Ma, 1997). By aggregating the geographic units into regions, a regionalization method could help overcome the limitations of flow mapping such as spurious data variations and discovery of general flow patterns. A range of techniques such as hierarchical clustering, principal components analysis and factor analysis have been used to identify regions in the migration literature (Morrill, 1988; Pandit, 1994; P. Slater, B., 1975; P. B. Slater, 1976, 1984). Although they have been used to identify regional structure in conjunction with domain knowledge, these techniques are not capable of extracting spatially contiguous regional structures and the regionalization strategy does not guarantee consistent patterns (e.g., different groupings result in a lot different flow structure and the decision to regionalize is not objective).

Guo (2009) proposed a spatially constrained hierarchical clustering and partitioning approach to derive spatially contiguous regions in a spatial interaction network in which connections between locations were determined by a set of measures such as modularity. For example, in migration case, modularity measure considers background population to remove the expected flows from the actual flows. This measure is used to define links (connections) of the network and a set of hierarchical clustering methods is used to obtain several hierarchies of regions. Then, combining a graph partitioning with a fine tuning strategy, an objective function which optimizes the within region connectivity is used to derive natural regions (community structures) from the hierarchies of regions. In addition to discovery of natural regions from the flow structure,

and handling the spurious data variations; this approach also offers solutions to visual cluttering problem by generalizing flow patterns to higher abstraction levels using the discovered hierarchy of regions.

#### 1.2.1.3 FLOW MAPPING

Flow maps illustrate the movement of phenomena between pairs of locations (origins and destinations). Slocum et al. (2009) identifies five kinds of flow maps: distributive, network, radial, continuous and telecommunications flow maps. A Distributive flow map can further be categorized into two subcategories based on whether it depicts actual routes of flow or abstract links that connect locations. French cartographer Charles Joseph Minard's flow map depicting the shipping routes of wine exported from France (Robinson, 1967) is one of the earliest examples of a distributive flow map that depict actual routes of flow. On the other hand, Tobler's (1976) flow maps depicting state-to-state migrations are examples of flow maps that include abstract links that connect locations.

The second type of flow maps is a network flow map that depicts flows within networks such as transportation and utility networks. Parks (1987) discusses that a map of general shipping routes is a form of network flow map since they depict flows on the network of shipping route. The main distinction for network flow maps is that the route of flows is more important than the precise flow values exchanged between locations. The third type is a radial flow map that depicts radial pattern of movement to/from each location (as in Color Plate 19.3 in Slocum et al., 2009). The radial pattern is illustrated using a start/snowflake schema in which edges correspond to flows to or from a list of selected locations. The fourth type is a continuous flow map that depicts the movement of

a continuous phenomenon such as wind or ocean currents. Since the magnitude and direction of flows change at any location, continuous flow maps are depicted using unit-vectors.

Finally, the fifth type is telecommunication flow maps which Slocum et al. (2009) describes as flows of telecommunications technology such as the Internet and its associated information spaces. Although they might depict location-to-location interactions, telecommunication flow maps are often created by a graph drawing algorithm or a strategy that places nodes according to an objective such as preserving distinctive patterns (e.g., community structures) in the network. A discussion of graph drawing methods is given in the section 2.1.4. Graph Visualization.

*DESIGN AND ISSUES*

In a flow map, a flow is often depicted as a straight or curved line connecting an origin to a destination. The color and/or width of each line can be used to represent the volume of the flow. The directionality of a flow is commonly displayed using arrows and the right-hand traffic rule that draws a flow line on the right side while the line is pointing to its destination (Guo, 2009). Bezier curves can also be used to draw flow lines where each line is curvy at the origin and straight on the destination end. Also, two divergent colors can be used to distinguish the origin and destination of a flow line (Boyandin et al., 2010).

There are many issues regarding the design of a flow map. First, a flow map can easily become cluttered when it displays a large number of flows. To overcome this problem, interaction operators such as linking (Buja et al., 1996), brushing (Alan M MacEachren, Wachowicz, Edsall, Haug, & Masters, 1999; Shepherd, 1995), filtering and

zooming (Keim, 2002; Shneiderman, 1996) can be used to manipulate the representation in a way that there are less but more "important" flows on the map. Tobler (1987) and Yadav-Pauletti (1996) implemented interactive flow mapping applications that allows selecting a subset of flows, origins and destinations based on user queries. User interactions are helpful in answering task-based questions in interactive applications, however, by only relying on the use of interaction operators does not provide an overview of the data (Andrienko et al., 2007).

Some studies (Holten & van Wijk, 2009; Lambert et al., 2010) employ edge bundling to overcome the cluttering problem in flow maps. Edge bundling is a technique to visually bundle adjacent edges together in such a way that edges are merged into bundles along their joint paths and fanned out at the end. The flow map layout derived by an edge bundling approach is similar to the layout of a distributed flow map depicting actual routes of flows. However, flow lines are determined by an algorithm and do not represent actual routes in the former. Spiral trees (Buchin et al., 2011; Phan, Xiao, Yeh, & Hanrahan, 2005) have been introduced by combining the edge bundling method with spiral layout algorithms. Spiral trees are especially effective when the task is to show connections (e.g., inflows, outflows) of one or a few places. Although edge bundling methods reduce the overall edge crossings by grouping the edges into bundles, it does not provide a solution to recognizing natural flow patterns.

Alternatively, computational and visual tools have been utilized to summarize the flow data and display the most interesting flows. Liu (1995) was one of the earliest scholars to summarize flow data by using projection pursuit methods and then visualize multivariate and spatial aspects of the data in multiple views by utilizing dynamic

brushing technique. Similarly, Yan and Thill (2009) used self-organizing maps (SOM) linked with a flow map to reduce the complexity of the data and visualize multivariate patterns. Guo (2009) incorporated a hierarchical regionalization method into a flow mapping framework to discover a hierarchy of geographical (natural) regions (communities), where there are more flows or connections within regions than across regions. Guo's (2009) regionalization method is effective in summarizing large flow data while preserving major structural patterns.

Second, a flow map is usually dominated by spurious data variations. Spatial interactions form networks which are usually scale-free (Newman, 2003) with a small number of hubs with a larger number of connections. For example, in county-to-county migration network, there are huge variations between the sizes (i.e., population) of counties. Counties with larger populations have larger flows as compared to other regions. However, larger flows do not necessarily indicate interesting patterns. Therefore, a strategy is needed to remove or at least reduce the effect of counties with differing populations.

Iterative proportional fitting procedure (IPFP) (P. Slater, B., 1975) has been widely used to reduce the effect of size on the flow structure. IPFP provides a double standardization and each individual cell in the result matrix shows a relative estimation of the number of people who would migrate from the specific origin to the specific destination (which is identified by the specific value of the cell), if all spatial units (counties) had the same number of in-migrants and out-migrants. Scaling does not change the cross-product ratio of the diagonal elements of the original matrix, and as a result the flow structure is preserved. However, IPFP transformation can distort the relative

significances of nodes in a spatial interaction network in which the variability of node sizes is large (Fischer, Essletzbichler, Gassler, & Trichtl, 1993; Holmes, 1978). Guo and Zhu (2014) recently developed flow smoothing approach to generalize flow maps, remove spurious data variance, normalize flows with control population, and detect high-level patterns that are not discernible with existing approaches.

Third, rather than focusing on individual flows between the spatial units (e.g., county-to-county flows), a flow map should allow visual examination of flows that have similar characteristics (e.g., multivariate components of flows and locations), or natural regions (community structures) that are strongly connected to each other. Therefore, a strategy is needed to discover general flow patterns with multivariate components at higher abstraction levels.

*APPLICATIONS*

Tobler (1987) was the first one to develop a flow mapping application. The application was demonstrated using state-to-state migration flows and the user could use filtering operators to depict one-way migration to/from a particular state by arrows of varying width. While depicting flows between all states, the flows below the mean were removed in order to overcome the cluttering problem. Tobler's original software was later updated to an interactive application that included new features such as colored and scaled arrows, two-way flows and a setting to control the movement volume to be shown (W. Tobler, 2004). While Tobler introduced the major framework for a flow mapping application, other applications were also developed in the meantime. For example, Yadav-Pauletti (1996) developed a migration mapping software that utilized animation with small multiples to depict migration flows over time. Similarly, Thompson and Lavin

(1996) developed an application to automate the generation of animated vector field maps.

Phan et al. (2005) developed a flow mapping application that bundles edges to minimize edge crossings using a hierarchical clustering method. The goal of Phan et al.'s (2005) application was to create a flow map layout thus the user interactions are limited for exploratory visualization. Using node clustering and flow aggregation, Boyandin et al. (2010) introduced an interactive application to analyze temporal changes in migration flows. Boyandin et al.'s (2010) application supports user interactions such as flow and node highlighting, selection and dynamic queries for filtering out flows by their volume and length. Using multiple linked views, Guo (2009) introduced an interactive and integrated flow mapping framework to discover community structures (natural regions), identify multivariate relations of migration flows, and examine the spatial distribution of both flow structures and multivariate patterns. User interactions such as selection-based brushing and linking are provided to allow the user to change the selection and combination of variables to examine different multivariate flow patterns or choose different number of regions to visualize flow patterns at different levels (e.g., local, regional and national).

### 1.2.2 Network Science

In the following subsections, the network science approach which draws on theories and methods of graph theory, statistical mechanics, social network analysis (SNA), data mining and information visualization will be discussed.

#### 1.2.2.1 Structural Characteristics

Graph theoretical measures such as centrality (Bonacich, 1987; Borgatti, 2005; Freeman, 1977; Kolaczyk et al., 2009; Wasserman & Faust, 1994), clustering coefficient (Watts & Strogatz, 1998), assortativity, and disassortativity (Newman, 2003) are often used to understand structural characteristics of social networks. Each measure can inform about a certain type of characteristic and that particular characteristic might not be relevant for all type of networks. To apply these measures to a network requires an initial knowledge of what kind of characteristic to look for.

In order to characterize network structures in terms of the relative importance of nodes, different types of centrality measures have been widely used in social network analysis (Bonacich, 1987; Freeman, 1977; Scott, 2000; Wasserman & Faust, 1994). Major examples of these indices are: point degree centrality in which a node is more central if it has relatively more connections; betweenness centrality in which a node is more central if it lies between the various other points and mediates connections. Moreover, centrality definition is also defined as a function of geodesic distance in the network space. For example, according to closeness centrality a node is more central if it lies at short distances from many nodes. On the other hand, according to competitive distance centrality, if a node is connected to central points, it becomes more central and it transmits this centrality to other points as well (Bonacich, 1987; Scott, 2000).

Centrality measures provide information about the relative importance of the nodes in the network. From another perspective, the focus could also be the neighborhood of nodes. For example, one might want to know if a person (a node) whose friends (neighbors) are also friends (neighbors) to each other. Clustering coefficient

measure could answer this question by characterizing the presence of loops in the network. Another example related to the neighborhood could be to find out if the hubs (people having the greatest number of connections) are well connected to each other (Costa et al., 2007). "Rich-club coefficient" (Zhou & Mondragon, 2004) that shows if nodes of higher degree are more interconnected to each other than nodes with lower degree, could be applied to answer such a question.

Gastner and Newman (2006) showed that there is a strong connection between the topological and geographical features of spatial networks. Several measures of centrality have been used to analyze spatial networks such as migration (Irwin & Hughes, 1992), commuting (Limtanakool, Schwanen, & Dijst, 2009), and habitat connectivity (Estrada & Bodin, 2008; Estrada et al., 2008). Also, similar measures have also been introduced in application-specific domains. For example, in order to discover geographical concentration (i.e., spatial focusing) of interregional migration flows, many indexes have been developed such as Gini index (D. A. Plane & Mulligan, 1997), coefficient variation (Long & National Committee for Research on the 1980 Census., 1988), migration efficiency (D. Plane, A. & Rogerson, 1991).

#### 1.2.2.2 Degree Distributions

Analyzing degree distribution (Barabási & Albert, 1999) is another commonly used method to quantify the network features of interest. The degree distribution, P(k) expresses the fraction of nodes in a network with degree k (Costa et al., 2007). Correlations between different degrees of nodes within a distribution might show important clues about the network structure. For example, if people having a lot of connections (high degree nodes) tend to connect with people that have a few connections

(low degree nodes), then the network could be called disassortative. On the contrary, if people tend to connect to other people with the same number of connections (having same degrees), then the network is called to be an assortative one (Newman, 2002). Several network models have been developed to define topological properties of networks by looking at degree-distributions. Random graphs, small-world, and scale-free graphs are some of the most common network models.

#### 1.2.2.3 Community Detection

As well as using statistical and graph theoretical measures to characterize networks, exploratory approaches have been developed to discover structural patterns (e.g., community structures) within networks. Some examples to these approaches are graph partitioning (Schloegel et al., 2000), hierarchical clustering (Clauset et al., 2004), and mixture models (Newman & Leicht, 2007). These approaches discover certain types of structural patterns such as assortative mixing (e.g., nodes that have many connections tend to be connected to other nodes with many connections) and disassortative mixing (e.g., nodes that have many connections tend to be connected to nodes with less connections) (Newman, 2002). Without depending on any prior knowledge about what to look for, exploratory approaches extract interesting and unknown information hidden in a network.

#### 1.2.2.4 Graph Visualization

Graph visualization is similar to flow mapping in that they are both used to visualize networks. The difference between these two methods is in how they construct the layout of the network (graph). In flow mapping the layout is predetermined by placing nodes at their corresponding geographic coordinates, whereas in graph visualization the layout is

generated by a graph drawing algorithm which places nodes according to an objective such as preserving the structural patterns (e.g., community structures) in the network. Graph drawing methods are widely used to visualize non-spatial networks such as social networks and biological networks.

The same network could be visualized by different layouts for different purposes. For example, when displaying very large networks, force-directed algorithms (Battista, Eades, Tamassia, & Tollis, 1999) are one of the most common methods to obtain an aesthetically pleasing layout by positioning the nodes in a way that all the links are of more or less equal length with fewer intersections. Graph drawing is a combinatorial optimization problem that requires optimization of several parameters such as edge crossings and distance measures between nodes and edges. In addition to force-directed method, there are several other types of algorithms such as multidimensional scaling, stress majorization (E. Gansner, Koren, & North, 2004), cross minimization (Kato, Nagasaki, Doi, & Miyano, 2005), incremental arrangement (Cohen, 1997). In addition to graph layout algorithms, interactive environments that allow visual examination through navigation techniques such as pan-and-zoom, fisheye and topological zooming have been developed (Abello, van Ham, & Krishnan, 2006; E. R. Gansner, Koren, & North, 2005).

### 1.2.3 Integrating Geography with Network Science

In geography and spatial sciences, the complex web of relationships between humans, environment and society are studied through the perspective of space, place and time (Goodchild et al., 2000) with diverse application areas such as global trade (Poon, 1997), human migration (Young, 2002), diffusion of innovation (Maggioni, Nosvelli, & Uberti, 2007), and disease spread (T. Davies, M. & M. Hazelton, L., 2010) . These studies are

similar in that they conceptualize interactions as a function of geographic distance and characteristics between locations, while disregarding the connections between the actors (e.g., countries, migrants, companies, and humans) of the network. On the other hand, theories of network and social sciences emphasize social interactions between the actors of a network while considering the influence of geography in a limited manner. A variety of computational and statistical methods such as graph theoretical measures (Scellato et al., 2011), random graph modeling (Schaefer, 2012), factor analysis (Hipp et al., 2012), simulation (Butts et al., 2012), and regression analysis (Viry, 2012) have considered influence of geography in their analysis of social networks. However, the methodologies introduced by these studies have limited capability in analyzing the spatial, temporal and relational aspects as they treat geography as a background variable.

By examining the similarities between the theories and methodologies of GIScience and Network Science, recent studies (Andris, 2011; Faust & Lovasi, 2012; Luo & MacEachren, 2014) identified the challenges and potential future research for bridging the gap between social network analysis and spatial analysis. Along this path, newly developed methods of computation-visualization framework and visual analytics (Gao et al., 2013; Guo, 2009; Kumar et al., 2014; Luo, et al., 2014) have shown great promise to address the challenge of integrating the approaches of geography and network science through the use of human-computer collaboration.

### 1.2.4 Usability in Geovisualization

Usefulness of a system shows whether the system can be used to achieve the goals of analysis. Utility and usability are sub-components of usefulness. Usability describes ease-of-use and is often measured with five attributes: learnability, error rates, efficiency,

memorability, and satisfaction (Nielsen, 1993; Rubin, 2008). On the other hand, utility describes usefulness and can be evaluated through benchmark tasks or grading of insights.

Traditional testing methods under controlled conditions are not suitable for evaluating the utility of visual tools because of the exploratory nature of visualization (Demsar, 2007). This is mainly because it is hard to define effectiveness or "success" for an exploratory task. Andrienko et al. (2002) suggest users should be given a free hand so that they could ask new questions and generate new hypotheses. However, it is challenging to observe exploratory behaviors of users and interpret the results since they are not replicable.

There are two alternative approaches to evaluating utility of a visualization tool: an objective-based approach and an insight-based approach. An objective-based approach (Roth, 2012) classifies interaction into a set of visual tasks (Demsar, 2007; Etien L. Koua & Kraak, 2004; C. Tobon, 2005; C. m. Tobon, 2002) and compares the effectiveness of users completing those tasks. On the other hand, an insight-based approach (Chang, Ziemkiewicz, Green, & Ribarsky, 2009; North, 2006) captures and grades individual observations about the data or visualization by the participant as an insight, a unit of discovery.

An objective-based approach categorizes interaction into a set of visual tasks the user may wish to complete with the cartographic interface (Roth, 2012). Visual tasks are derivatives of basic visual operators such as identify, compare, associate, etc., and were first introduced by Wehrend and Lewis (1990). Using the objective-based approach, many studies (Aufaure-Portier, 1995; Davies, 1995; Knapp, 1995) decoded the

exploration process into visual tasks (identifying clusters in the data, finding relationships between elements, comparing values at different locations and distinguishing spatial patterns, identifying spatial positions of objects, their spatial distribution and density, etc.) and performed experiments (Koua et al., 2006; C. Tobon, 2005) based on those visual tasks.

An insight-based approach is composed of three phases: defining insight, identifying several measurable characteristics of insight, and establishing methods to recognize insight (Gotz & Zhou, 2008). There is no commonly accepted definition of insight, however, North's (2006) definition of insight as complex, deep, qualitative, unexpected and relevant seems to cover most aspects of insight discussed in the visualization community. Saraiya et al. (2004) categorized measurable characteristics of insight as observation (occurrence of insight), time (when the insight was generated), correctness of the insight and category (types of insight, e.g., insight into data, insight into visualization). A think-aloud protocol can be used to capture insights and quantifiable usability characteristics of each insight is then encoded for analysis.

In addition to the insight-based and objective-based approaches, measures of mental effort and visualization efficiency have been introduced to better understand the perception of graphs using the cognitive load approach. The cognitive load approach evaluates cognitive capacity allocated to accommodate the demand imposed by a visual task by implementing subjective measures such as rating scales (Wierwille & Casali, 1983; Zijlstra, 1993), performance-based measures and physiological measures such as pupillary dilation (De Waard & Studiecentrum, 1996; Paas, Tuovinen, Tabbers, & Van Gerven, 2003). A number of studies that employed empirical testing of mental effort

(Huang, Eades, & Hong, 2009; Paas et al., 2003) suggest that subjective rating scale techniques of perceived mental burden are easy to use, inexpensive, can detect small variations in workload, reliable, and provide decent convergent, construct and discriminate validity. Huang, Eades and Hong (2009) further developed a measure of visualization efficiency which is the difference between cognitive cost (i.e., mental effort and response time) and cognitive gain (response accuracy). According to the definition of visual efficiency, high efficiency is gained with high accuracy and low mental effort and a short response time, whereas low efficiency occurs when low accuracy is associated with high mental effort and a long response time.

## 1.3 Dissertation Structure

This dissertation introduces a theme for understanding complex patterns of geo-social networks using three independent manuscripts. Each of these manuscripts represents a chapter of the dissertation (chapters 2-4). Chapter 2 introduces a generic framework that can be used to smooth various graph measures and is the first attempt that truly considers the flow structure in implementing spatial kernel smoothing in a spatially embedded network. Due to the copyright agreement of the published material, further analyses and discussion of the results are included in Appendix A. Chapter 3 introduces a novel approach to discover spatial and structural patterns among individual locations of a dynamic geo-social network embedded in space and time. By developing and visualizing a measure of connectedness across space and time, the new approach facilitates the discovery of hot spots (hubs, where connectedness is strong) and the changing patterns of such spots across space and time. Chapter 4 introduces a user evaluation to obtain knowledge on how map readers perceive information presented with flow maps, and how

design factors such as flow line style (curved or straight) and layout characteristics may affect flow map perception and users' performance in addressing different tasks for pattern exploration. Chapter 5 concludes with the findings, broad impacts and future research directions.

# CHAPTER 2

## SMOOTHING LOCATIONAL MEASURES IN SPATIAL INTERACTION NETWORKS[1]

[1] Koylu, C., & Guo, D. (2013). Smoothing locational measures in spatial interaction networks. *Computers, Environment and Urban Systems, 41(0), 12-25. doi: http://dx.doi.org/10.1016/j.compenvurbsys.2013.03.001*
Reprinted here with permission of publisher.

## 2.1 ABSTRACT

Spatial interactions such as migration and airline transportation naturally form a location-to-location network (graph) in which a node represents a location (or an area) and a link represents an interaction (flow) between two locations. Locational measures, such as net-flow, centrality, and entropy, are often derived to understand the structural characteristics and the roles of locations in spatial interaction networks. However, due to the small-area problem and the dramatic difference in location sizes (such as population), derived locational measures often exhibit spurious variations, which may conceal the underlying spatial and network structures. This paper introduces a new approach to smoothing locational measures in spatial interaction networks. Different from conventional spatial kernel methods, the new method first smoothes the flows to/from each neighborhood and then calculates its network measure with the smoothed flows. We use county-to-county migration data in the U.S. to demonstrate and evaluate the new smoothing approach. With smoothed net migration rate and entropy measure for each county, we can discover natural regions of attraction (or depletion) and other structural characteristics that the original (unsmoothed) measures fail to reveal. Furthermore, with the new approach, one can also smooth spatial interactions within sub-populations (e.g., different age groups), which are often sparse and impossible to derive meaningful measures if not properly smoothed.

Keywords: smoothing, spatial interaction, spatial network, network measure

## 2.2 INTRODUCTION

Spatial interactions, such as migration and airline travel, naturally form a location-to-location network (graph). In the network a node represents a location (or an area) and a link represents an interaction (flow) between two locations. Locational measures, including both simple ones such as in-flow, out-flow, and net-flow and more complicated ones such as centrality, entropy and assortativity, are often derived to understand the structural characteristics and roles of locations in generating interactions. However, due to the dramatic differences in size (such as population) among locations and the small-area problem, locational measures derived with the original flow data often exhibit spurious variations and may not be able to reveal the true underlying spatial and network structures.

Scaling approaches such as iterative proportional fitting procedure (IPFP) are often employed (Clark, 1982; Pandit, 1994) to remove the confounding effects of origin and destination sizes on flows. However, such transformation procedures may distort the relative significances of nodes in a network (Fischer et al., 1993; Holmes, 1978). Alternatively, several studies have applied existing spatial kernel smoothing methods to remove spurious data variations (Porta et al., 2009; Sohn & Kim, 2010), which treat a locational measure (e.g., centrality) as a regular attribute and apply a traditional spatial kernel smoothing method to directly smooth the derived measure values. However, directly smoothing the measure values may generate unreliable or even misleading results for two main reasons. First, the original measure values may be unstable due to varying unit sizes and small flows between units. Second, traditional smoothing methods do not differentiate flows within and beyond a neighborhood and it is inappropriate to

directly smooth original locational measures. For example, the net flow ratio (i.e., net flow / total flow) for a neighborhood (i.e., a group of contiguous spatial units) cannot be calculated as the average of unit-level net flow ratios within the neighborhood.

We introduce a new approach to smoothing locational measures in spatially embedded networks. For each location, the new method first smoothes the flows to/from that location considering flows to/from its neighborhood and then calculates its locational measure with the smoothed flows. The same procedure is repeated for each location, using the original flows (i.e., the smoothed flows for the previous location are not used). The neighborhood of a location is defined as the minimum set of nearest neighbors that meet a size constraint (such as a minimum population threshold or a distance threshold). To demonstrate the usefulness of the approach, we use the county-to-county migration data in the U.S. and smooth the net migration rate and entropy measure for each county. The smoothed results clearly help discover natural regions of attraction (and depletion) and a variety of structural characteristics that the original measures fail to reveal. Furthermore, we also smooth measures for sub-populations (e.g., different age groups), which can help discover not only distinctive regions of attraction and depletion but also show that attractiveness changes in both geographic space and multivariate space (e.g., migrants of different ages).

## 2.3 Related Work

### 2.3.1 Locational Measures

Locational measures (network/graph measures) have been extensively used in spatial interaction analysis to examine structural characteristics such as centrality (Hughes, 1993; Irwin & Hughes, 1992), entropy (Limtanakool et al., 2009), connectivity (Estrada

& Bodin, 2008), assortativity and disassortativity (Fagiolo, Reyes, & Schiavo, 2009) and weighted clustering coefficient (De Montis, Barthelemy, Chessa, & Vespignani, 2007). Similar measures have also been introduced in application-specific domains such as migration. For example, many index approaches have been developed and used to quantify migration characteristics such as spatial focusing of migration streams (D. A. Plane & Heins, 2003; D. A. Plane & Mulligan, 1997; A. Rogers, 1992; Andrei Rogers & Raymer, 1998; Andrei Rogers & Sweeney, 1998). The index measures are usually derived for each location with the graph data (e.g., migration network). Commonly-used measures include net migration rate (A. Rogers, 1992), Gini index (D. A. Plane & Mulligan, 1997), coefficient variation (L. Long, E., 1988) and migration efficiency (D. Plane, A. & Rogerson, 1991). However, due to the dramatic difference in unit size (e.g., population) and the small-area problem, derived locational measures often exhibit spurious data variations, and may conceal (instead of reveal) the true underlying spatial and network structures.

### 2.3.2 ITERATIVE PROPORTIONAL FITTING PROCEDURE (IPFP)

In order to remove the effects of location sizes on flows and capture patterns that are not necessarily associated with larger volumes, scaling approaches have been employed (Clark, 1982; Pandit, 1994; P. Slater, B., 1975). The most commonly used scaling approach is the iterative proportional fitting procedure (IPFP), which can be used to standardize a migration network by transforming the flows among locations so that all locations have the same inflow and outflow. Scaling does not change the cross-product ratio of the diagonal elements of the original matrix, and as a result the flow structure is preserved. However, IPFP transformation can distort the relative significances of nodes

in a spatial interaction network in which the variability of node sizes is large (Fischer et al., 1993; Holmes, 1978).

### 2.3.3 Kernel Density Estimation and Smoothing

Kernel density estimation or smoothing methods are commonly used for smoothing lattice spatial data, e.g., point- or area-based location attribute data, which are different from connection-based spatial interaction data. A spatial kernel smoothing method recalculates the attribute value of a location using a weighted average of the attribute values of its spatial neighbors (Borruso & Schoier, 2004; Carlos, Shi, Sargent, Tanski, & Berke, 2010), where the weight is calculated considering geographic distance. Alternative to spatial kernel smoothing, locally weighted average smoothing that uses a background value such as population to calculate weights is commonly used in smoothing disease rates (Kafadar, 1994; X. Shi, 2010). Bandwidth and kernel function selection are two important parameters in a spatial kernel smoothing method. The choice of the bandwidth determines the maximum radius (e.g., the extent of the neighborhood) or the number of neighbors that is considered to have an effect on the point of interest. The kernel function determines how each neighboring observation will be weighted and considered in the smoothing process. Previous research on kernel density estimation proved that the performance of the estimation is greatly affected by the choice of the bandwidth while the kernel function usually does not have a significance effect (Bors & Nasios, 2009; Silverman, 1986).

The most commonly used kernel functions include Gaussian kernel, triangular kernel, and Epanechnikov's kernel (Danese, Lazzari, & Murgante, 2008; Wand & Jones, 1995). There are two main types of bandwidth: *fixed* and *adaptive*. In a fixed-bandwidth

approach, the radius that defines the extent of the neighborhood is assumed to be the same throughout the dataset. An adaptive bandwidth allows the radius to vary from one data point to another. Domain knowledge is commonly used to obtain a fixed bandwidth. However, it is widely acknowledged that a fixed bandwidth causes biased estimations for most spatial data sets, where the underlying density often exhibit significant spatial heterogeneity (T. M. Davies & M. L. Hazelton, 2010). Alternatively, various adaptive bandwidth approaches have been developed (Abramson, 1982; Carlos et al., 2010; Sain & Scott, 1996; Yang, Luan, & Li, 2010), which can be categorized into model-based and domain-based approaches.

In model-based bandwidth selection approaches, the goal is to improve a statistical model fit such as in geographically weighted regression. A statistical criterion is often used to provide guidance on selecting an appropriate bandwidth among a large number of possible bandwidth values (D'Amico & Ferrigno, 1990). Cross-validation (CV), Akaike Information Criterion ($AIC_c$) and Bayesian Information Criterion (BIC) are among the most commonly used criteria to select an appropriate bandwidth for local spatial statistics such as geographically weighted regression (Fotheringham, Brunsdon, & Charlton, 2002). In model-based approaches, an appropriate bandwidth is the one that gives the best model fit among a large number of possible bandwidth values. However, model-based approaches are not applicable for spatial smoothing in which there is no statistical model to fit and the goal is to smooth each unit with the neighborhood values. In domain-based bandwidth selection approaches, a relevant attribute (e.g., population) is used to determine the bandwidth. For example, to adapt with the underlying heterogeneous population distribution common in public health research, some studies

(Carlos et al., 2010; Xun Shi, 2009) have utilized a population threshold (i.e., the size for a neighborhood) to determine the adaptive bandwidth. Therefore, the bandwidth stops expanding when the threshold value is reached.

### 2.3.4 Smoothing Network Measures

Traditional smoothing methods introduced above have been adopted and used in transportation analysis research (Porta et al., 2009; Sohn & Kim, 2010) in order to accommodate the neighboring effect in calculating centrality measures. Existing smoothing practices treat the locational network measure (e.g., centrality) as a regular attribute and apply an existing spatial kernel smoothing method to directly smooth each locational measure with neighboring values. However, since a network measure summarizes the structure of the flows incident on a node in a network, it is inappropriate to directly smooth measure values without considering the flow structure within and beyond the neighborhood.

## 2.4 Methodology

The new smoothing approach consists of four steps. First, for a location (node) $s$ in a spatial interaction network, identify its spatial neighborhood $N_s$ based on a geographic distance threshold (fixed-bandwidth) or a size threshold such as a minimum population (adaptive-bandwidth). The neighborhood $N_s$ is represented with a gray circle in Figure 2.1.

Second, *temporarily* remove the flows within the neighborhood, i.e., those with both origin and destination in the same neighborhood. Note that these flows are excluded only for this specific neighborhood and are still eligible for consideration for other neighborhoods. Then we weigh flows from/to the nodes (including $s$) in the

neighborhood based on their distances to location *s*. The result is a smoothed sub-graph, in which lows to/from location *s* are modified considering follows to its neighbors. Figure 2.1(B) illustrates the smoothed sub-graph of a location *s* where flows within $N_s$ are removed and flows to/from $N_s$ (shown by dashed lines) are weighted and partially considered as flows to/from location *s*.

Third, calculate the needed network measure for location *s* with the smoothed sub-graph. In other words, the weighted flows to/from the neighborhood are used in calculating the network measure for the location.

Fourth, repeat the above three-step process for each location (node). After the measure is obtained for a location, the smoothed flows are discarded and their original flows are restored. In other words, the smoothing (Step 2) is only temporary for each neighborhood.

In following subsections, we introduce each of the steps. To demonstrate the approach, we use county-to-county domestic migration data between 1995 and 2000 in the contagious U.S. provided by census surveys, which includes 3075 counties (of the 48 continental states and Washington D.C.) and millions of migrants moving between these counties. Each data record has an origin county, a destination county, the count of migrants, and migrant characteristics, e.g., counts of migrants for each income level or age group that move from the origin to the destination.

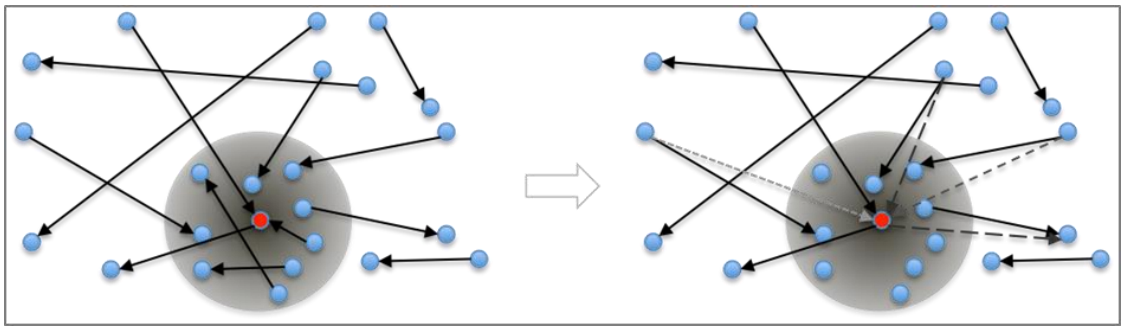


**Figure 2.1:** An illustration of the smoothing approach for spatial interaction data. The left map (A) shows the original data. The map on the right (B) shows smoothed flows related to a location (in red, at the center of the circle) and its neighborhood (gray circle). Dashed lines represent weighted flows to/from the neighborhood that are now partially considered as flows to/from the location in red and used in calculating the network measure for the location.

### 2.4.1 Bandwidth Selection

There are two potential alternatives for choosing the bandwidth. If applicable to the context of the spatial interaction network, a domain-based approach could be employed, which uses an attribute and a threshold value to configure the size of a neighborhood, e.g., the population or total flow of a neighborhood. Alternatively, a data-driven approach could be employed to determine the bandwidth according to the properties of the spatial interaction network. In this research we primarily focus on the first approach (domain-based) to configure neighborhood and discussed the alternative (data-driven) approach in the conclusion section.

In spatial interaction data, locational measures can be sensitive to the volume of flows or population of involved locations. It is more meaningful to make each neighborhood be of a similar and sufficiently large size so that the flows to/from different neighborhoods can be compared. Therefore, we employ a domain-based approach and use a population threshold to determine the adaptive bandwidth (or neighborhood size)

for each unit. Other than population, the total volume of in-flow or out-flow may also be used for defining the size threshold. The choice depends on its applicability to the locational measure. For example, a net migration rate represents the net-flow of a location normalized by its population, in which case it makes sense to make each neighborhood have a similar population.

Let the population threshold be $p$. The neighborhood $N_s$ of a location $s$ is the smallest set of nearest neighbors that has a total population $P(N_s)$ greater than $p$. Specifically, the neighborhood $N_s$ for unit $s$ is constructed with two steps: (1) initially, let $N_s = \{s\}$ and sort all other units based on their distance to $s$; (2) the nearest neighbors are added to $N_s$ until $P(N_s) > p$. The bandwidth for $s$ is then the distance to the farthest unit in its neighborhood $N_s$.

For cases where the population attribute does not exist or is inappropriate for the context of the analysis, alternative variables can be used to define neighborhood (bandwidth). For example, the in-flow entropy measure quantifies the diversity of flows that go into a location. Thus, it is appropriate to use the total in-flow to a neighborhood (excluding flows within the neighborhood) to define the bandwidth in calculating the in-flow entropy measure. Similarly, for the out-flow entropy measure we may use the total volume of out-flows from a neighborhood to define the adaptive bandwidth.

Figure 2.2 illustrates the bandwidth selection process with a simple data set. Let the population threshold $p$ =100. Three nodes $r$, $s$, $t$ are highlighted and their population values are $P(r) = 15$, $P(s) = 40$ and $P(t) = 130$. Since node $t$ is sufficiently large, it forms a neighborhood by itself and thus no smoothing is needed. Nodes $r$ or $s$ need to add

neighbors to meet the threshold $p$. Following the procedure outlined above, we have $N_r$ = $\{r, s, d, a, b, f, g, h\}$ and $N_s = \{s, a, e, r, b, c\}$, with $P(N_r) = 110$ and $P(N_s) = 105$.

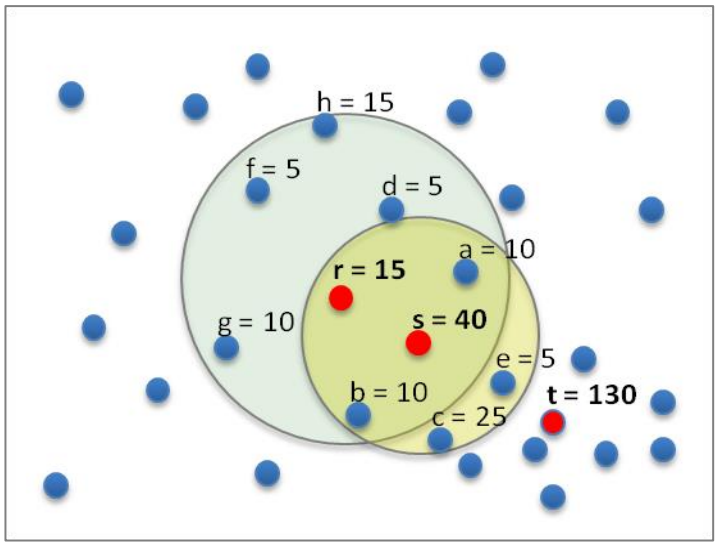



**Figure 2.2:** Illustration of the bandwidth selection process. The neighborhood $N_s$ of a location $s$ is the smallest set of nearest neighbors that has a total population $P(N_s)$ greater than a given population threshold $p$, which is 100 in this example. The map shows the neighborhoods of three locations $r$, $s$, and $t$.

Choosing the population threshold for determining bandwidth involves a tradeoff between over-smoothing and under-smoothing. On one hand, the bandwidth should be sufficiently large to avoid artifacts caused by small neighborhood and under-smoothing. On the other hand, if the neighborhood is too large, interesting local patterns may disappear. Smoothing results change in a predictable way with decreasing/increasing bandwidth, with larger bandwidths generating more smoothed result (we will show the experiments with different bandwidths in Section 2.5.4). For the county-to-county migration data of the U.S., we experimented with different population thresholds to

examine flow patterns at different scales and chose a population threshold of one million, which is about the population of a medium-sized metropolitan area.

### 2.4.2 SMOOTHING FLOWS

For a specific location $s$ and its neighborhood $N_s$, $s \in N_s$, we smooth the flows that go into or out of the neighborhood. Let $B_s$ be the bandwidth. First, flows within $N_s$ are temporarily removed, i.e., a flow is ignored if its origin and destination are both in $N_s$. Removing flows within the neighborhood is necessary because the entire neighborhood is considered as a single unit (i.e., location $s$) in calculating a network measure. Second, a kernel function is incorporated to weigh each flow from/to *Ns* based on the distance between $s$ and the flow origin or destination (whichever is inside *Ns*). In other words, each flow to/from *N*s is partially (according to its weight) considered as a flow to location $s$ even if the flow does not directly involve $s$ in the original data, which essentially reassigns weights to existing edges or adds new edges to location $s$.

The most commonly used kernel functions include the uniform kernel, the Gaussian kernel and triangular kernel (Figure 2.3). Previous research and our experiments show that the smoothing results are not significantly affected by the choice of models (Bors & Nasios, 2009; Silverman, 1986). In this research, we have experimented with the above three models and report the results using the Gaussian kernel.

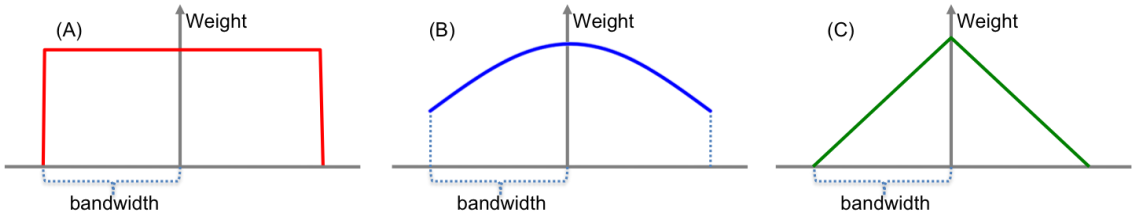



**Figure 2.3:** Three commonly used kernel functions. (A) Uniform: $W_{si} = 1$ if $d_{si} \leq B_s$; else 0. $B_s$ is the bandwidth and $d_{si}$ is the distance between location *s* and its neighbor *i*. (B) Gaussian: $W_{si} = \exp(-(d_{si}/B_s)^2)$ if $d_{si} \leq B_s$; else 0. (C) Triangular: $W_{si} = 1 - |d_{si}/B_s|$, if $d_{si} \leq B_s$; else 0.

In Figure 2.4 we show an example of a smooth graph that includes the connections to/from a location *s* (in red) and its neighborhood $N_s$ (gray circle). In addition to edges *(a, s)* and *(s, k)* that exist in the original data, the smoothed sub-graph for location *s* also has newly added edges *(b, s), (c, s), (j, s)* and *(s, i)*, which will be included in calculating the location measure for *s*. The value for the new "flow" *(b, s)*, for example, is the product of the value of flow *(b, e)* and its weight $W_{se}$ according to a chosen kernel model. Note that flows within $N_s$ are ignored.

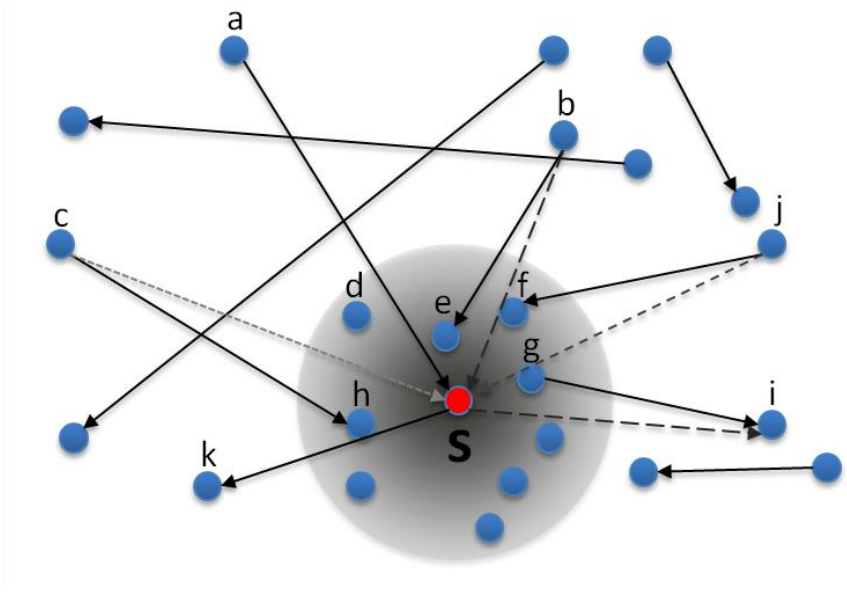



**Figure 2.4:** An illustration of a smoothed sub-graph. Dashed lines are newly added edges.

### 2.4.3 Calculating a Locational Measure

Using the smoothed sub-graphs, it is straightforward to calculate a variety of network measures for the focal location, which are more stable (with less spurious variation) than those calculated without smoothing. Here we use the net migration rate and an entropy measure as two case studies to demonstrate the approach and evaluate its results.

#### 2.4.3.1 Net Migration Rate

Net migration rate is the difference between in-migration (in-flow) and out-migration (out-flow) of an area in a period of time, divided by the population of the area. Net migration rate is usually multiplied by 1000 to represent the number of migrants per 1000 inhabitants. To obtain a smoothed net migration rate for a neighborhood, we smooth the flows for the neighborhood (as introduced earlier), calculate the inflow and outflow of the neighborhood with the smoothed graph, and then divide (inflow – outflow) with the total weighted population of the neighborhood of *s*, denoted by $P(N_s)$. In other words, the same weighting is applied to both the flows and the population.

#### 2.4.3.2 Entropy

The variation of flow volumes across the links to/from a location can provide important insights about the structure of the network and the characteristic of the location. Local entropy measures (Limtanakool et al., 2009) are often used for this purpose. Entropy of a location *s* (i.e., its neighborhood) is calculated using the formula in Equation 2.1:

$$EI_s = -\sum_{j=1}^{J} \frac{x_{sj} ln(x_{sj})}{ln(J-1-n)}$$

**Equation 2.1:** Entropy

where $EI_s$ is the Entropy Index of location $s$, $x_{sj}$ is proportion of flow $sj$ in relation to the total flow connected to $s$, $J$ is the total number of locations in the network, and $n$ is the number of locations inside the neighborhood $N_s$. The maximum number of links that location $s$ may have is J – 1 – n. *EI* measures the variation in the magnitude of interactions across the connections of a node. The index value ranges between 0 and 1. A small inflow entropy value indicates that the flows to the location vary greatly (with large flows from a few locations and small flows from elsewhere), whereas a large inflow entropy value indicates that a location receives similar amount of flows from all locations. Entropy can also be calculated for out-flows or all flows (inflow and outflow together).

With the county-to-county migration data in the U.S., we calculate and map the smoothed net migration rate and the entropy measure for each county, which clearly help discover natural regions of attraction or depletion and a variety of structural characteristics that the original measures fail to reveal. Furthermore, our smoothing method make it possible to calculate measures for a subset of flows (e.g., flows of a specific age group), which are impossible to obtain without smoothing due to the small-area problem.

## 2.5 Results

### 2.5.1 Smoothed Net Migration Rate

In this section, we show the smoothed net migration rates and compare them to the original measures. For the county-to-county migration dataset of the U.S., we chose a population threshold of one million, which approximates the population of a medium-sized metropolitan area. To enable comparison of the two measures, we used a custom

classification for both in which 0 was chosen as the critical midpoint and the Jenks natural breaks classification was applied separately to each side of the midpoint. A diverging color scheme is used to represent different value ranges, with red representing attraction and blue for depletion (i.e., negative net migration rate).

The original net migration rates are shown in Figure 2.5, in which it is difficult to distinguish regions of attraction and depletion because of unstable values caused by the dramatic population differences among counties and the small-area problem in the data. On the contrary, the smoothed net migration rates (with a neighborhood size of one million population) shown in Figure 2.6 can clearly reveal the regions of attraction and depletion with differing magnitudes. For example, major attraction regions (i.e., those of darker red hues) include Florida, Arizona, Greater Las Vegas region, north-east outskirts of the Atlanta metropolitan area, counties surrounding Denver, Dallas, Houston and San Antonio, and the metropolitan counties in North Carolina. On the other hand, large metropolitan counties such as Los Angeles, New York City, Chicago and Miami and rural counties in Montana, North and South Dakota can be easily recognized as regions of depletion. The smoothing results also reveal contrasting patterns locally within metropolitan areas, such as Chicago, Denver, Washington D.C., Dallas and Miami, where the core metropolitan areas have a push effect on migrants while the counties surrounding these core metropolitan areas have a pull effect on migrants as a result of suburbanization and urban sprawl.

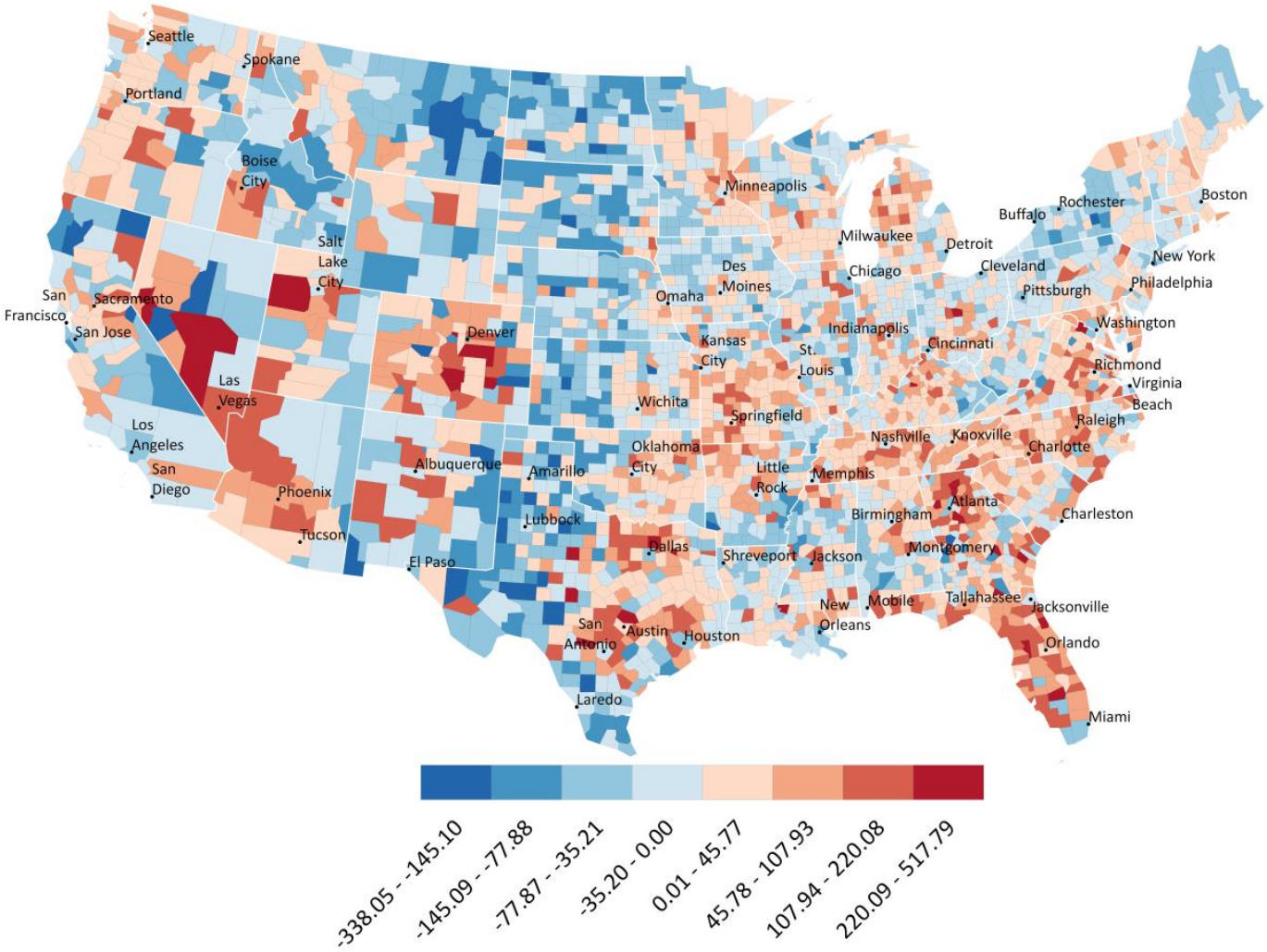



**Figure 2.5:** Original net migration rates

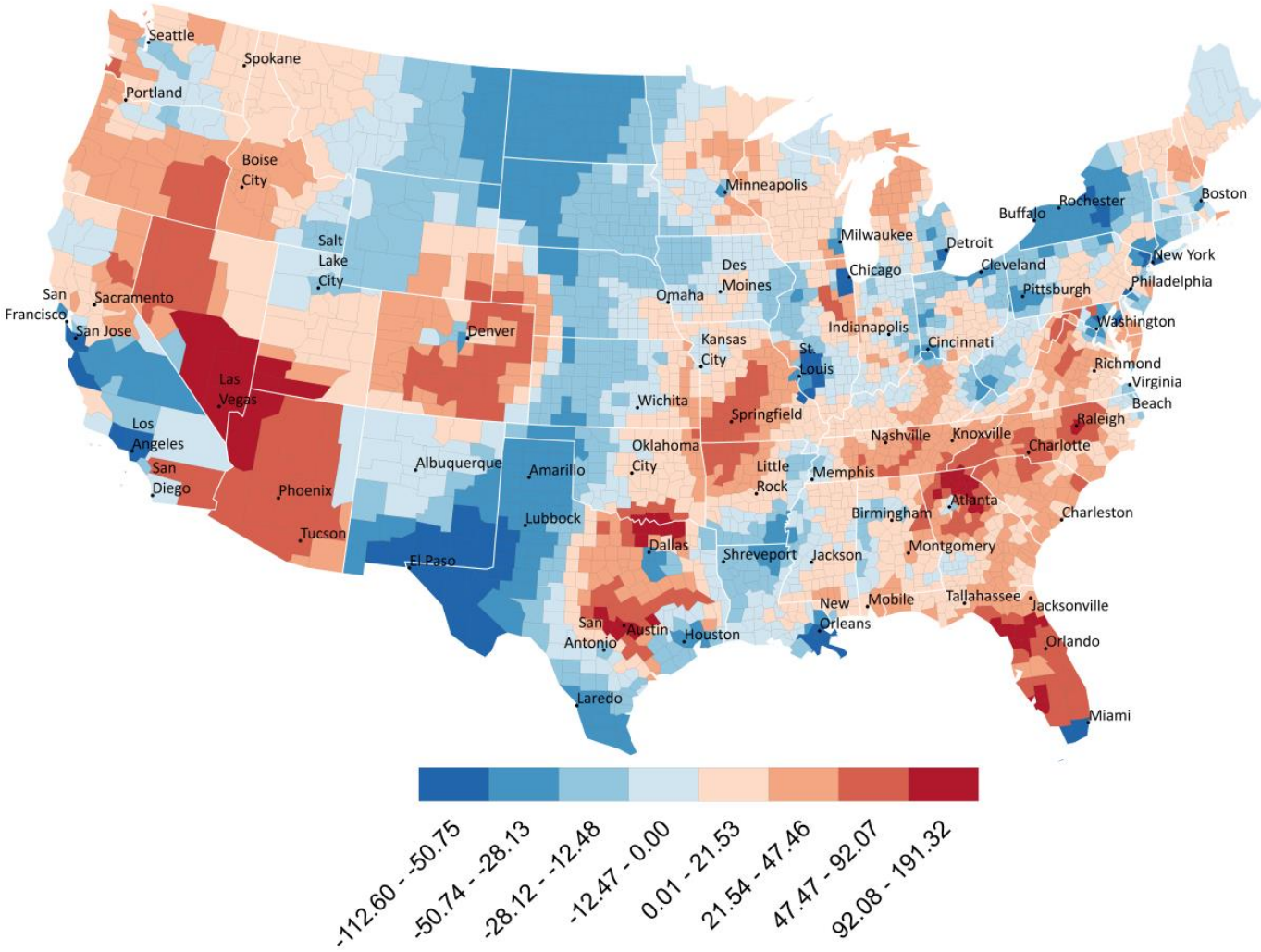



**Figure 2.6:** Smoothed net migration rates

### 2.5.2 Smoothed Net Migration Rate for Sub-Populations

Migration patterns of sub-populations such as different races, ethnicities or age groups are expected to be spatially and structurally different from each other. Locational measures for sub-population flows are even less reliable because of much smaller volumes of flows and small base populations. To illustrate the effectiveness of our approach to overcome this problem, we smooth the flows within each age group, calculate net migration rate with the smoothed flows and compare them to their original net migration rate results. After examining the smoothing results for each age group, we focus on two age groups, namely 20-24 and 25-29; because they have the highest mobility and distinctive migration patterns (we will explain this below in Figure 2.7). The original net migration rates for age groups 20-24 and 25-29 are shown in Figure 2.7 and Figure 2.8. It is difficult to interpret both maps because of unstable measure values that have spurious variation.

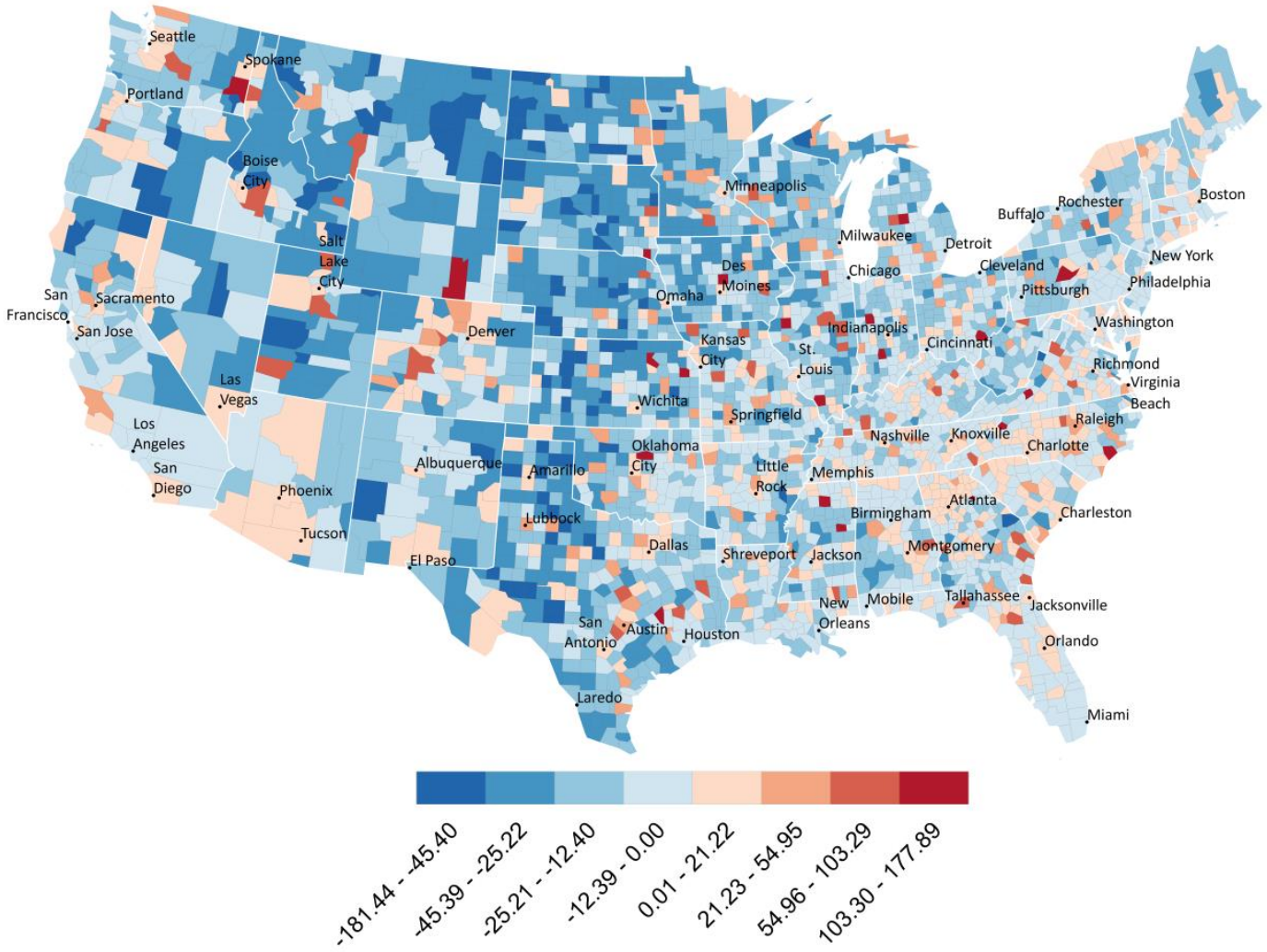



**Figure 2.7:** Original net migration rates for age group 20-24

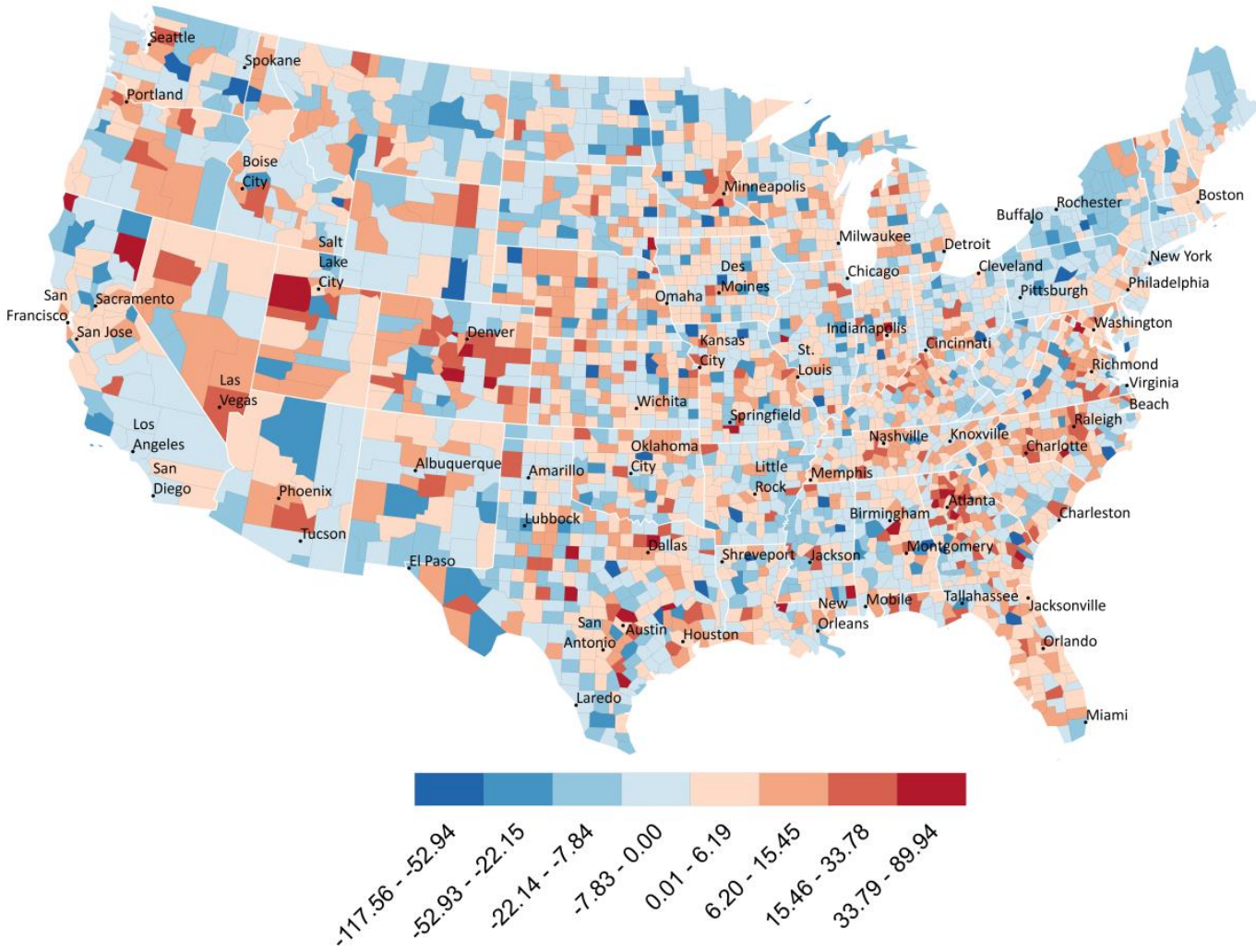



**Figure 2.8:** Original net migration rates for age group 25-29

After we smooth net migration rates within each age group, we use box-plots of the smoothed measure results to give an overall understanding of migration behaviors for different age groups by comparing their distributions (Figure 2.9). One of the most interesting and contrasting patterns that can be observed in Figure 2.9 are those for age 20-24 and age 25-29. On one hand, the age group 20-24 has a large number of outliers with very high positive net migration rates and a larger upper quartile with a median around 0. On the other hand, age group 25-29 has a lower median below 0, a larger lower quartile and some outliers with negative net migration rates. The migration flows within these two age groups are likely related to education and employment specific flows. From Figure 2.9, we may also observe patterns related to elderly migration (David A. Plane & Jurjevich, 2009; Andrei Rogers & Sweeney, 1998). For example, there are outliers that disproportionately attract migrants of age groups 55-75. We can map the net migration rates for each age group to further investigate the observed patterns. Due to limited space, we only show the smoothed results for age groups 20-24 (Figure 2.10) and 25-29 (Figure 2.11).

The smoothing results highlight distinctive patterns that agree with existing migration studies. For example, migration of students and young adults for education and employment purposes (Paul B. Slater, 1976; Whisler, Waldorf, Mulligan, & Plane, 2008) can be seen clearly in Figure 2.10 and Figure 2.11. While metropolitan areas attract age group 25-29 because of employment opportunities, places with a substantial number of universities attract age group 20-24. This divide can be seen in many places across the country. For example, in Texas, though the region surrounding Austin attracts age group 20-24, there is an opposite tendency among age group 25-29 to move away from this

region and possibly targeting the Dallas Metropolitan area. A similar pattern is also observed in Florida. Because of the presence of many university campuses, the region that includes counties around Tallahassee, Gainesville and Jacksonville in Florida attracts age group 20-24, whereas age group 25-29 migrate from this region, targeting the Orlando and Miami Metropolitan areas for jobs. Also, metropolitan areas including Las Vegas, Atlanta, Raleigh, Charlotte, Denver and Minneapolis also attract both of the age groups 20-24 and 25-29

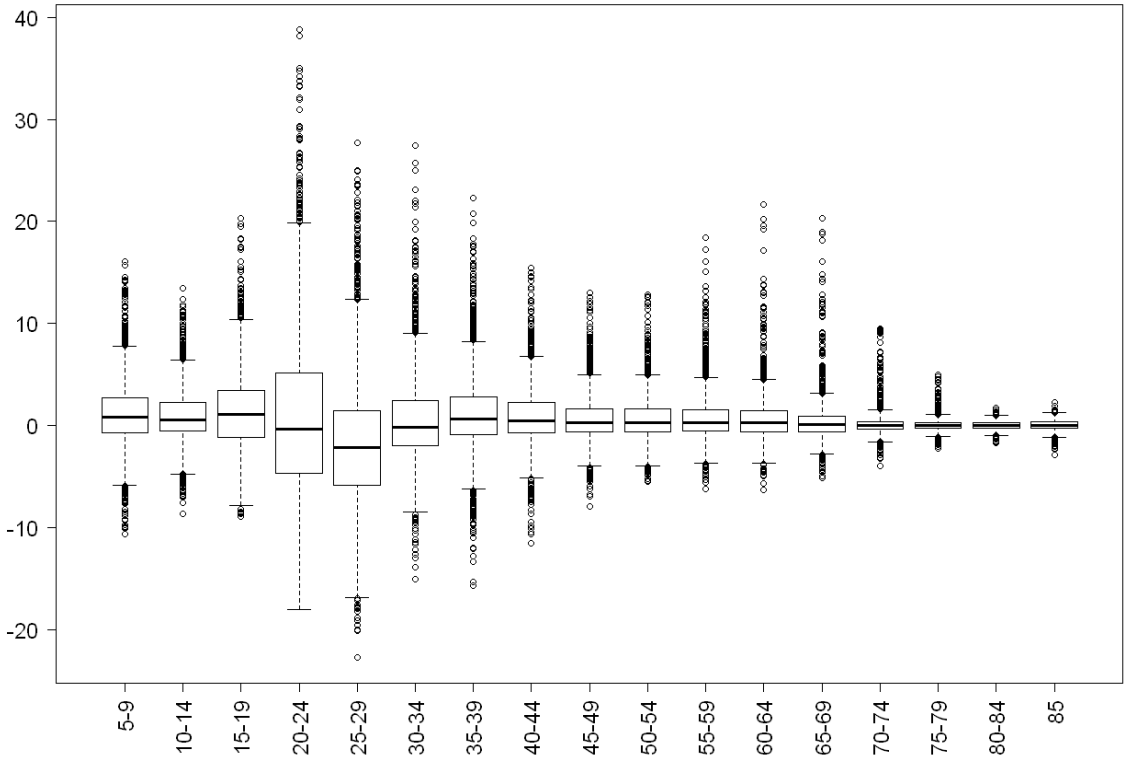



**Figure 2.9:** Box-plots of smoothed net migration rate results for age groups

.

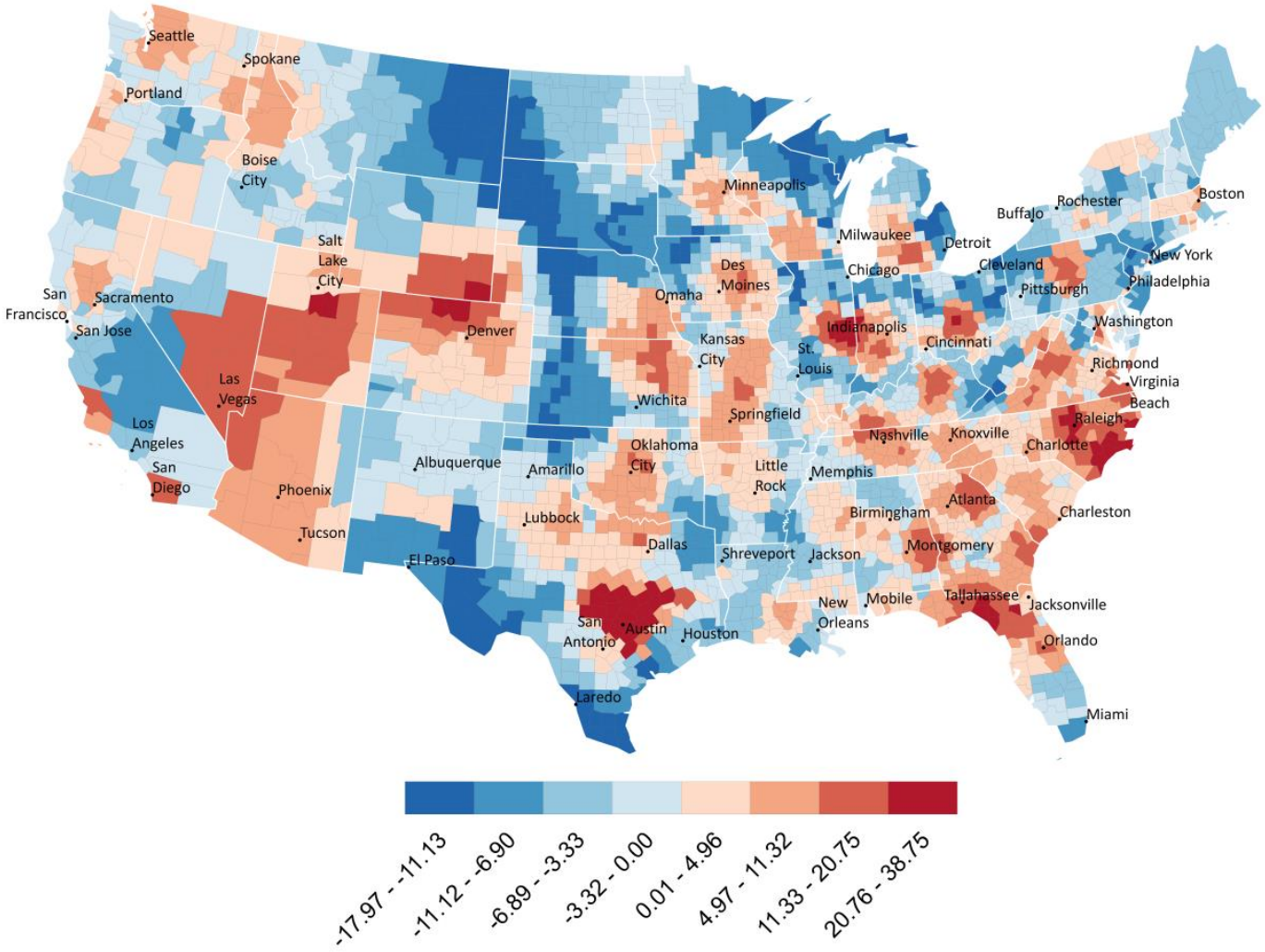



**Figure 2.10:** Smoothed net migration rate for age group 20-24

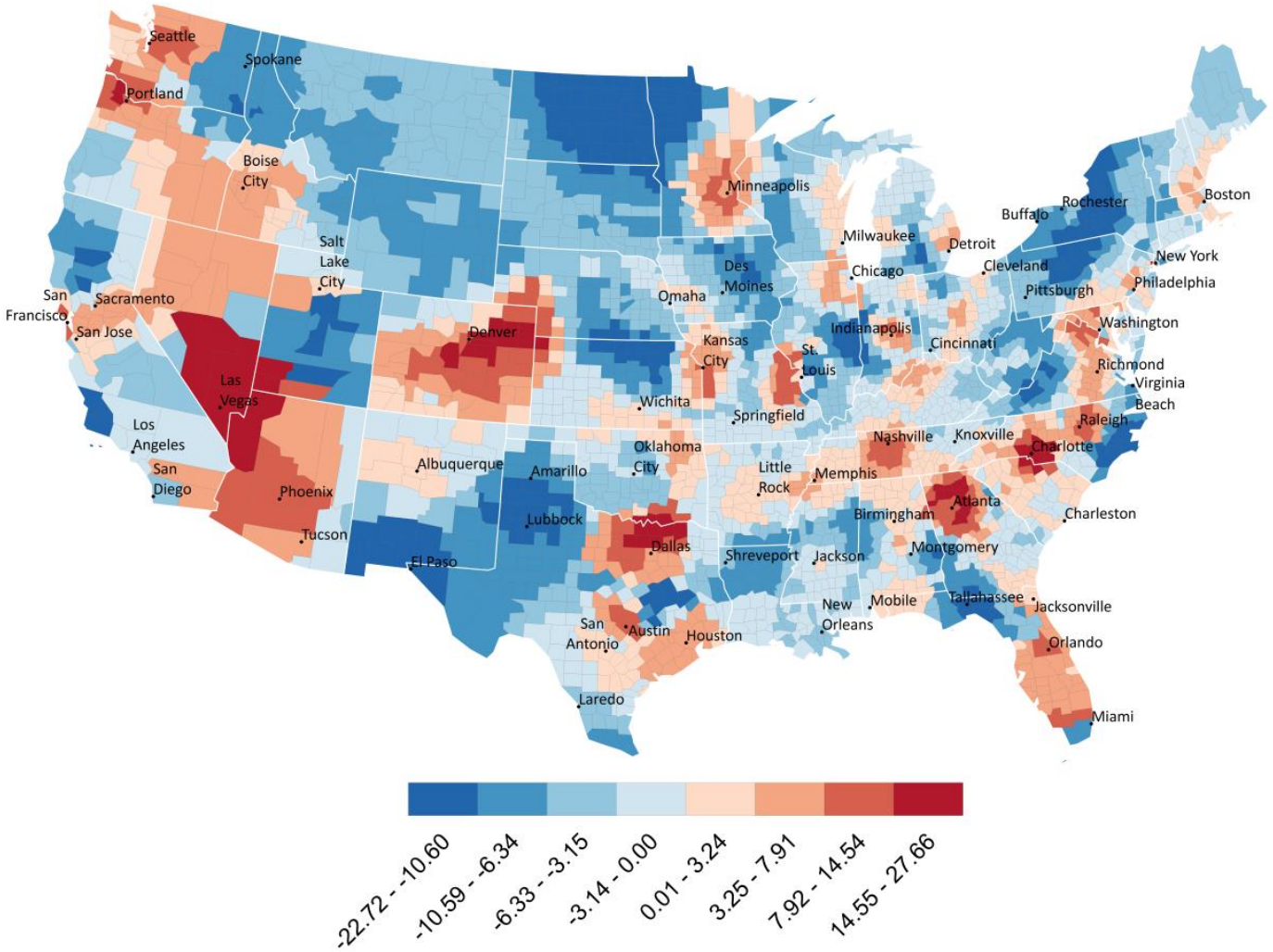



**Figure 2.11:** Smoothed net migration rate for age group 25-29

### 2.5.3 SMOOTHED ENTROPY

In addition to discovering regions of attraction and depletion, it is also important to gain insight into the structure of flows. A variety of measures such as entropy (Limtanakool et al., 2009), Gini index (D. A. Plane & Mulligan, 1997) and coefficient variation (L. Long, E., 1988) could be used to measure the diversity of flow volumes among the links to/from a location. In this section, we use the inflow and outflow entropy measures to capture the differentiation of the magnitude between the links to/from each location. We also compare the smoothed entropy measures to their original measure results. We use the total volume of in-flow to determine the neighborhood in calculating in-flow entropy whereas we use the total volume of out-flow to determine the neighborhood in calculating out-flow entropy.

To balance between over-smoothing and under-smoothing we heuristically chose 100,000 as a threshold volume both for in-flow and out-flow values to determine the neighborhood for calculating in-flow and out-flow entropy measures. To enable comparison, we again use the Jenks natural breaks classification with a sequential color scheme, with darker colors of red representing low entropy values (which highlight spatially focused (targeted) flows) and darker colors of blue for high entropy values (which show more evenly spread flows to/from the other locations).

The original inflow entropy and outflow entropy are shown in Figure 2.12 and Figure 2.13, respectively. Both maps are difficult to interpret because of large and spurious variation in measure values. Moreover, the entropy values correlate with size and, as a result, smaller counties have always relatively low entropy since they normally have much less links than larger counties (see Equation 2.1). Similarly, larger places tend

to have connections with many other places, and hence the entropy value is larger (i.e., more evenly spread).

The smoothed inflow entropy values in Figure 2.14 represent distinctive and different patterns from their original unsmoothed measures. It is interesting to see that the low inflow entropy clusters (red color) center around a number of major cities (such as Atlanta, Houston, San Diego, Chicago, etc.) but exclude the urban counties at each cluster center. The low entropy values indicate that these places draw focused flows, i.e., major flows are from certain places. On the other hand, clusters of high entropy values (blue color) represent places that receive migrants in similar volumes from many locations (i.e., more evenly spread).

From Figure 2.14, we also observe contrasting patterns within some regions. For example, while the core counties of the Chicago, Houston, San Antonio and Dallas metropolitan areas have high entropy values, the counties surrounding these metropolitan cores have low entropy values. This could potentially be explained by the tendency of metropolitan cores to attract migrants (especially young adults) in similar volumes from many places in the country as opposed to the tendency of the outskirts attracting migrants (e.g., families and retirees who prefer suburban lifestyle) from metropolitan cores disproportionately more than they attract migrants from other places.

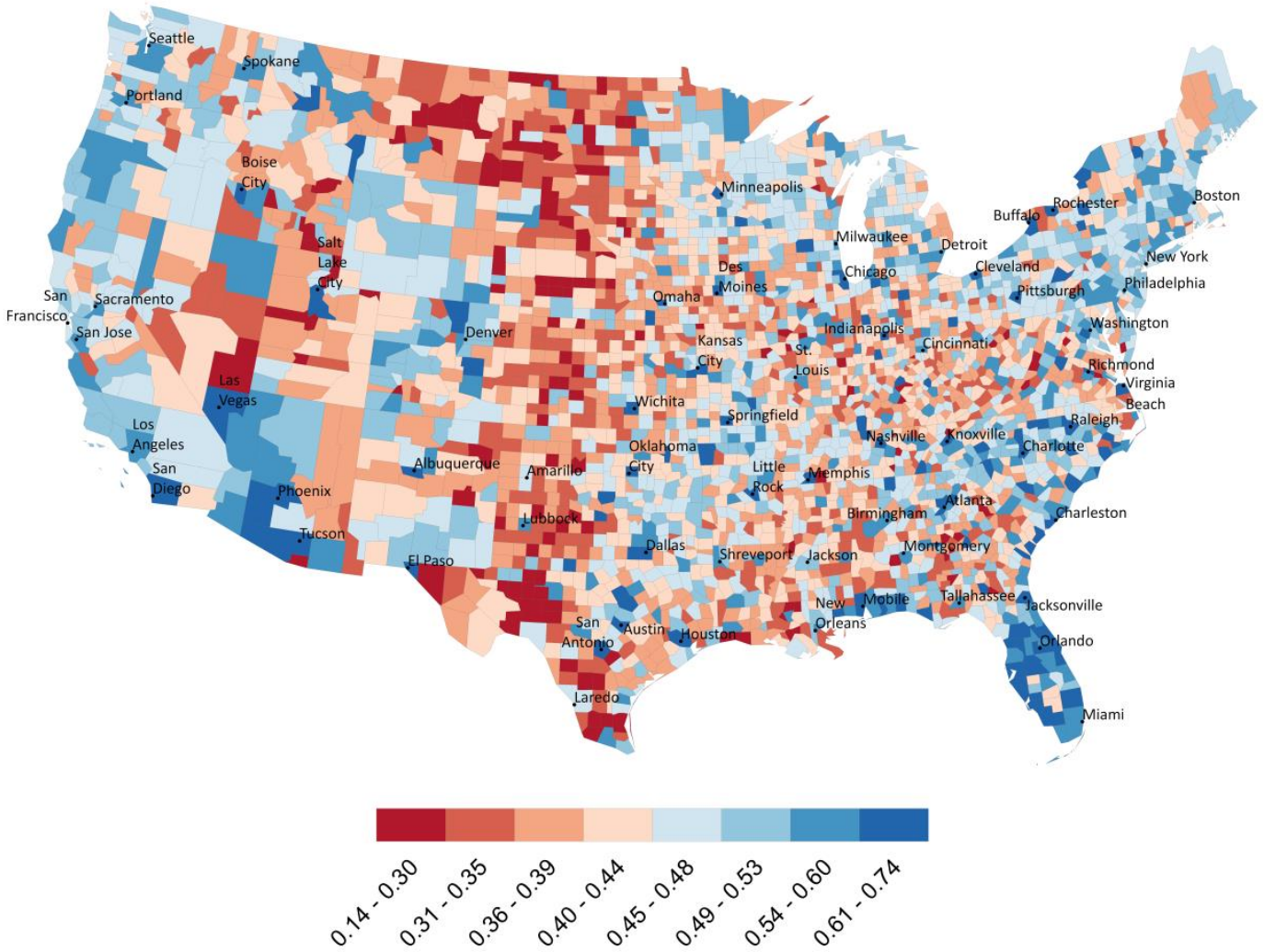



**Figure 2.12:** Original in-flow entropy values.

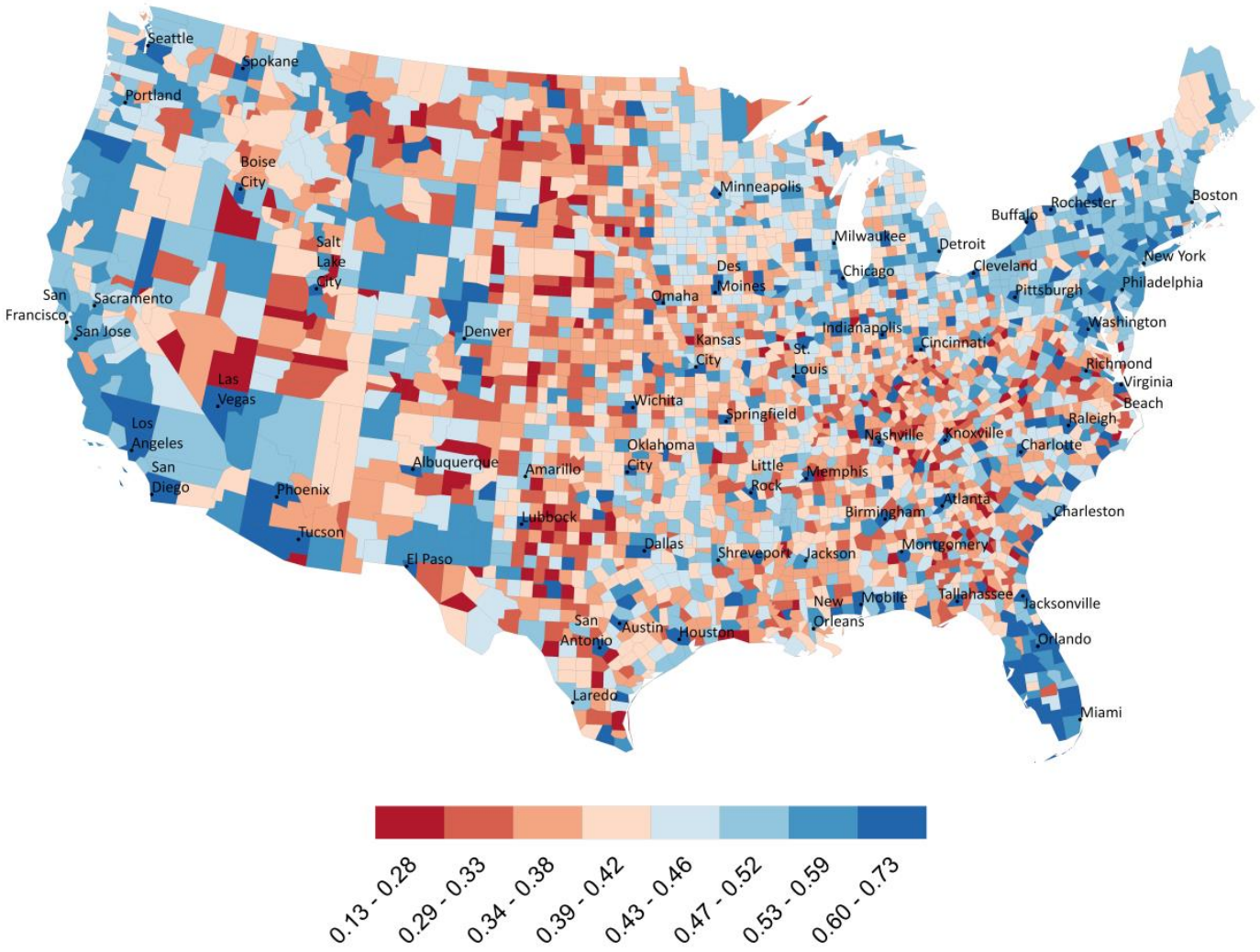



**Figure 2.13:** Original out-flow entropy values.

The overall extents of the clusters in the smoothed out-flow entropy map (Figure 2.15) are similar to the ones in the smoothed in-flow entropy map. However, there are local differences between the clusters of in-flow and out-flow entropy values. For example, in the Dallas, Atlanta and Chicago metropolitan areas, we observe lower in-flow entropy values for the periphery of some metropolitan areas and higher in-flow entropy values for the metropolitan cores. However, we observe the opposite of this pattern in the outflow entropy map where the cores of Dallas, Atlanta and Chicago metropolitan areas have lower out-flow entropy as opposed to the counties surrounding them. Thus, this pattern indicates that migrants leaving these cores are more targeted (focused) towards a fewer number of places in much higher volumes. In addition to these contrasting patterns, we observe that high outflow entropy clusters match the high inflow entropy clusters, indicating that migrants leaving these places do not target certain areas and migrants moving into these places come from many locations in similar volumes.

### 2.5.4 Sensitivity Analysis

In this section we evaluate the sensitivity to population thresholds and compare the results of our approach (smoothing local network and then calculating the measure) and the conventional approach (calculating measures and then smoothing measures). Due to limited space we only present the sensitivity analysis results for smoothing net migration rate.

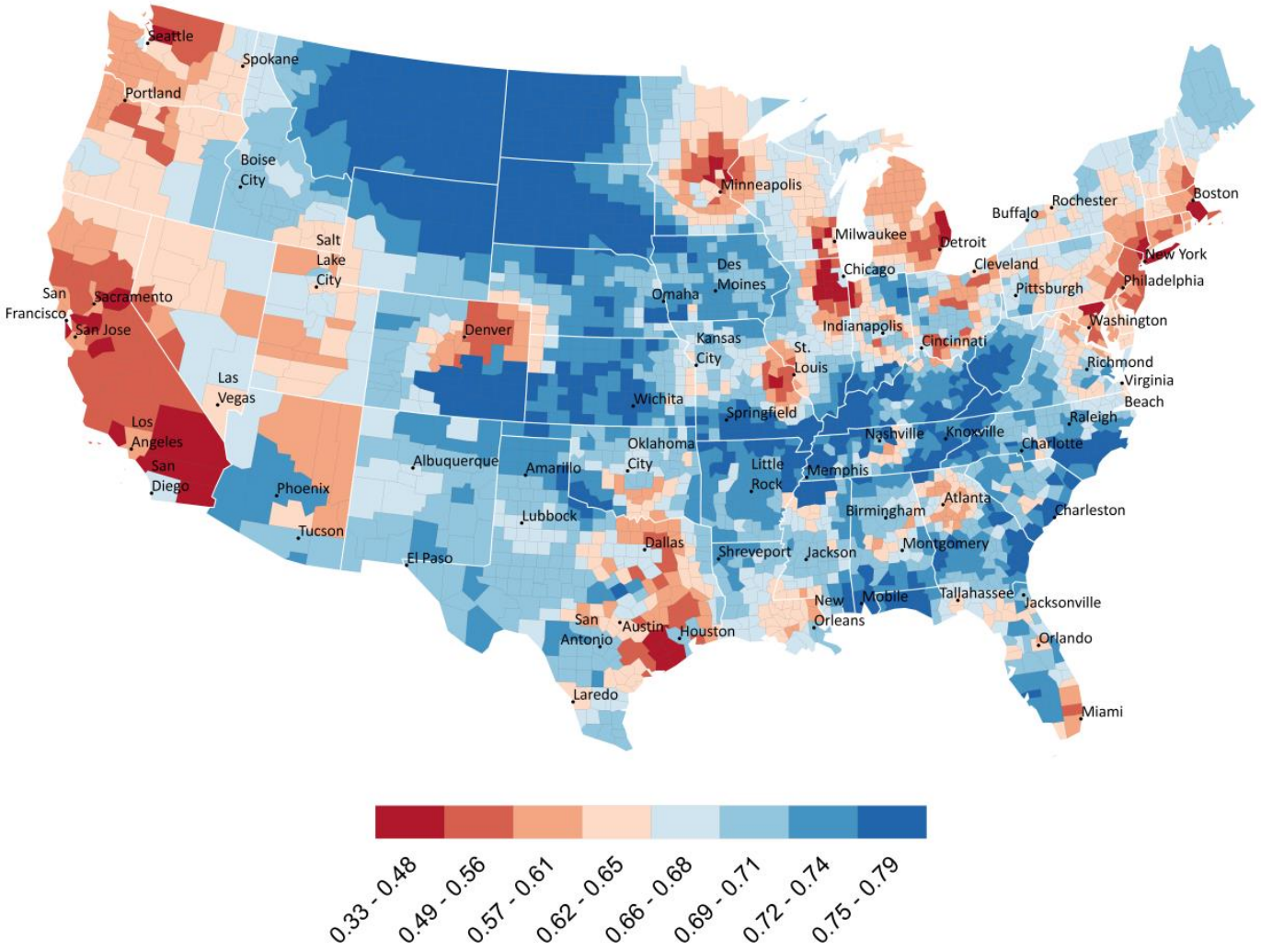



**Figure 2.14:** Smoothed in-flow entropy

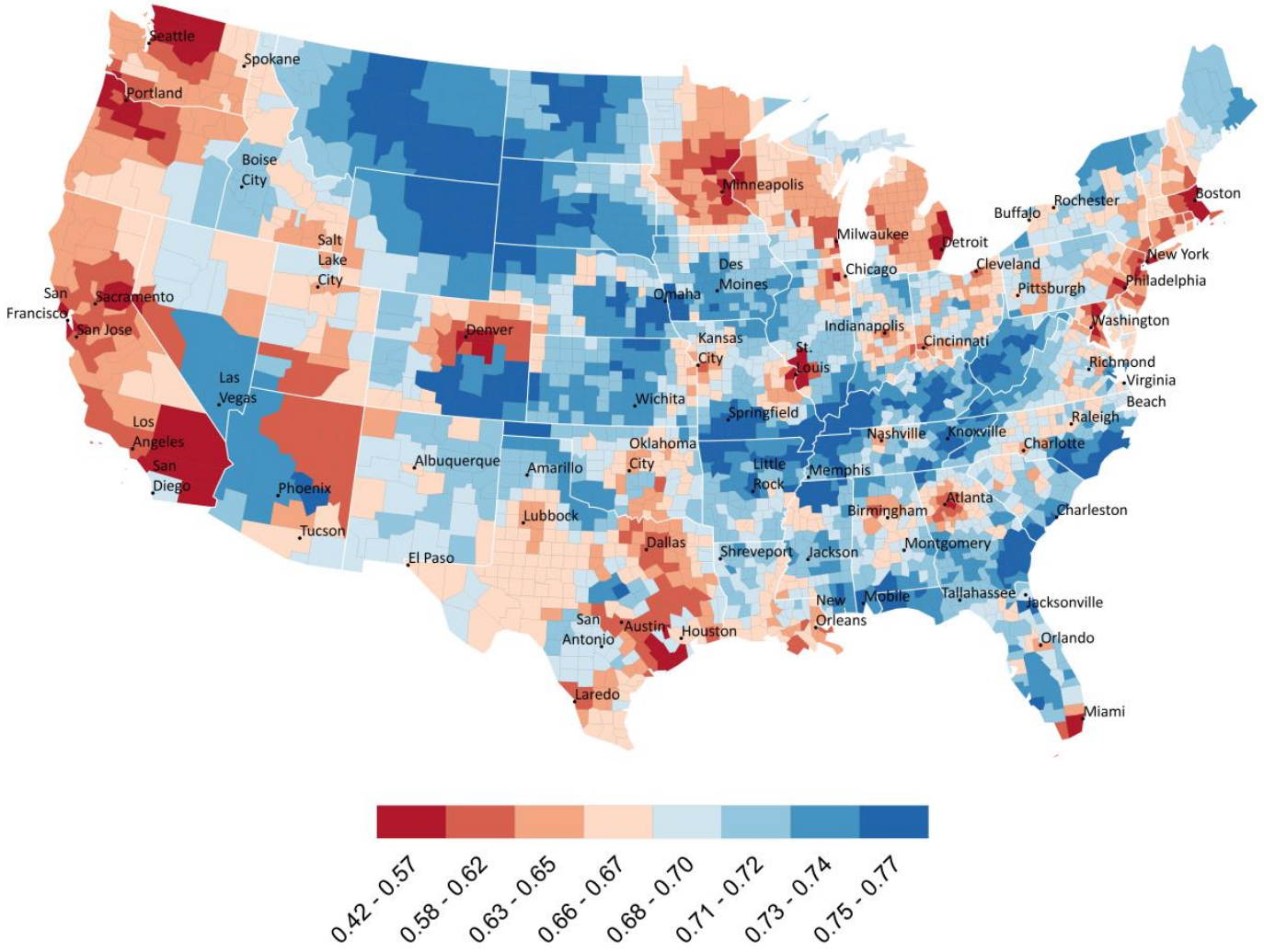



**Figure 2.15:** Smoothed out-flow entropy

In order to select an optimal population threshold (bandwidth) that reveals general patterns in the data and reduces the instability caused by small populations (X. Shi et al., 2007), we experimented with a series of population thresholds using both the conventional smoothing approach and our approach. Both approaches use the same spatial kernel and the same bandwidth. We plot the variance of smoothed rates from both approaches using a series of population thresholds.

Plot B in Figure 2.16 shows the total variance of the resulting rates, whereas Plot A shows the difference in variances between two consecutive thresholds. As expected, Plot B shows that variance decreases as population bandwidth increases. Our approach produces rates with less variance than the conventional result since the latter still uses the original instable rates caused by small base population. Although the general trend is that variance decreases with larger thresholds, Plot A reveals several thresholds where the variance reaches a local minimum, including 300k, 800k, 1 million and 1.5 million. Smoothing results at these different thresholds show patterns at different scale levels, from finer to coarser resolutions. In this paper, we only present the results with a threshold of 1 million, which approximates the population of a medium-sized metropolitan area.

### 2.5.5 Comparison with Conventional Methods

In addition to analyzing the sensitivity to population thresholds, we also compare our approach to a conventional smoothing approach using net migration rate and inflow entropy measures. Figure 2.17 shows the two results (conventional approach vs. our approach) for smoothed net migration rates of for all population. In order to allow comparison, both methods use the same bandwidth (i.e., one million) and the same spatial

kernel function (i.e., Gaussian). The overall patterns are similar in both maps. However, for the conventional approach the effect of small base populations can still be observed in many places such as the surrounding counties of Salt Lake City, UT, Albuquerque, NM and Houston, TX (Figure 2.17, Left), where smoothed rates are affected by the original unstable rates (see Figure 2.5) and the flows within the neighborhood. Our approach eliminates the effect of small base populations by treating the neighborhood as whole, removing internal flows, and calculating the measure based on smoothed network (Figure 2.6 and Figure 2.17 (right)).

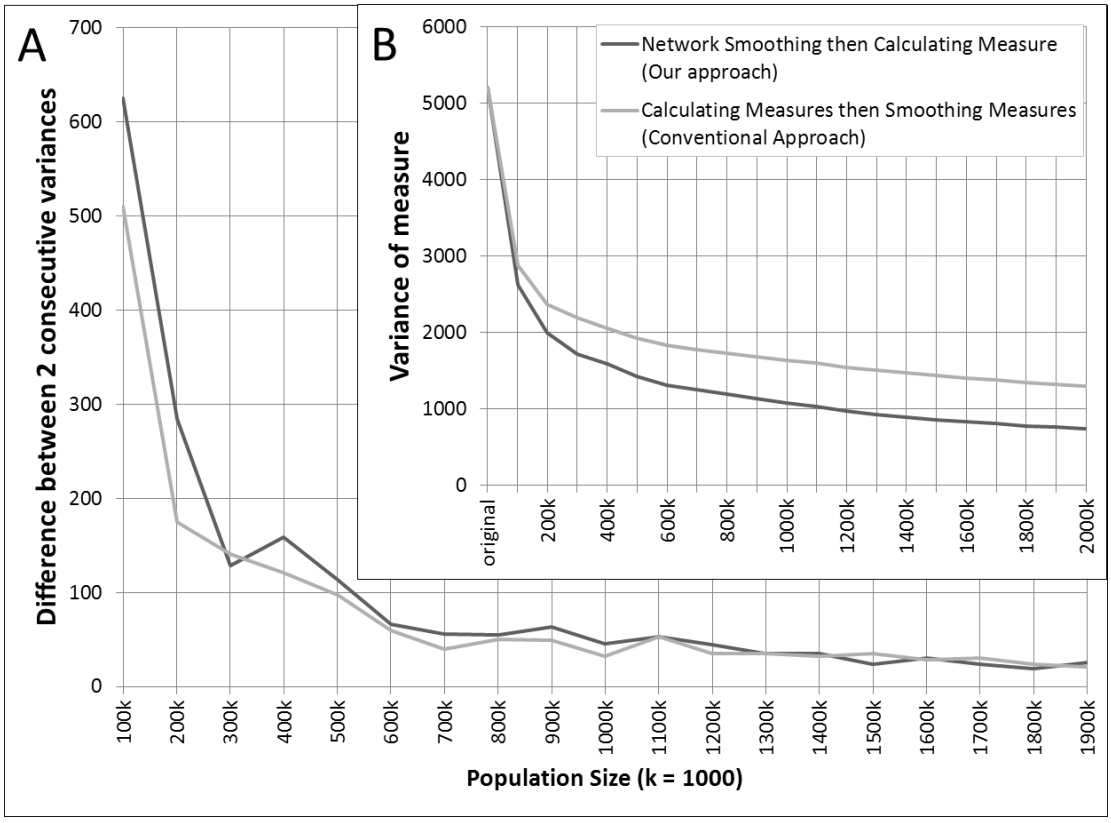



**Figure 2.16:** The variance of smoothed net migration rates for a series of population sizes. (A) The difference in variance between two consecutive thresholds. (B) The total variance for each population size (threshold).

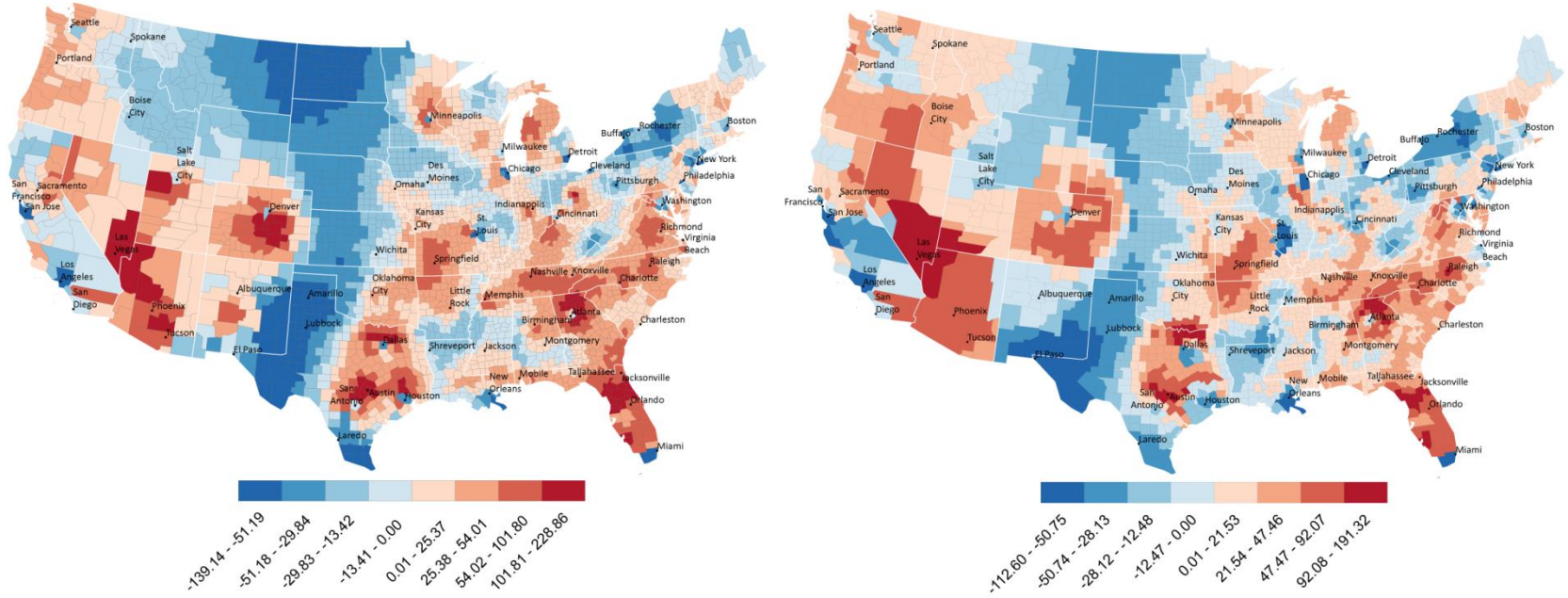



**Figure 2.17:** Comparison of conventional smoothing result (left) and our result (right) for net migration rates. The overall patterns are similar but there are significant local differences between the two results.

The effect of small base populations is more dramatic for the entropy measure, causing small areas to have very small entropy values due to the sparse flows to/from those areas. This can easily be seen in the original measure result as well as the smoothing result of the conventional approach (Figure 2.18, Left), which produces large clusters of low entropy values which are highly correlated with the presence of small counties and their unstable rates (see Figure 2.12). Our approach, on the other hand, first smoothes the network related to a neighborhood and then calculates its entropy measure. As such, our approach reduces the effect of the small-area problem and reveals spatial clusters of low inflow entropy values, indicating places that draw focused in-migration flows (Figure 2.18, Right), which is dramatically different from the result of the conventional approach. Such differences also exist for smoothing the net migration rates of different age groups as shown in Section 2.5.2.

## 2.6 Discussion and Conclusion

Spatial interaction datasets with a relatively small number of observations for most origin-destination pairs suffer greatly from spurious data variations and as a result locational measures calculated for such datasets become unreliable. To demonstrate the usefulness of the approach, we smoothed the net migration rates for all migrants and for migrants of different age groups. We also smoothed in-flow and out-flow entropy measures (1) to show the applicability of our method to smooth network measures; and (2) to capture the variation in the magnitude of flows that each location has. The method can be used to smooth a variety of locational measures such as centrality, chi-square and flow efficiency.

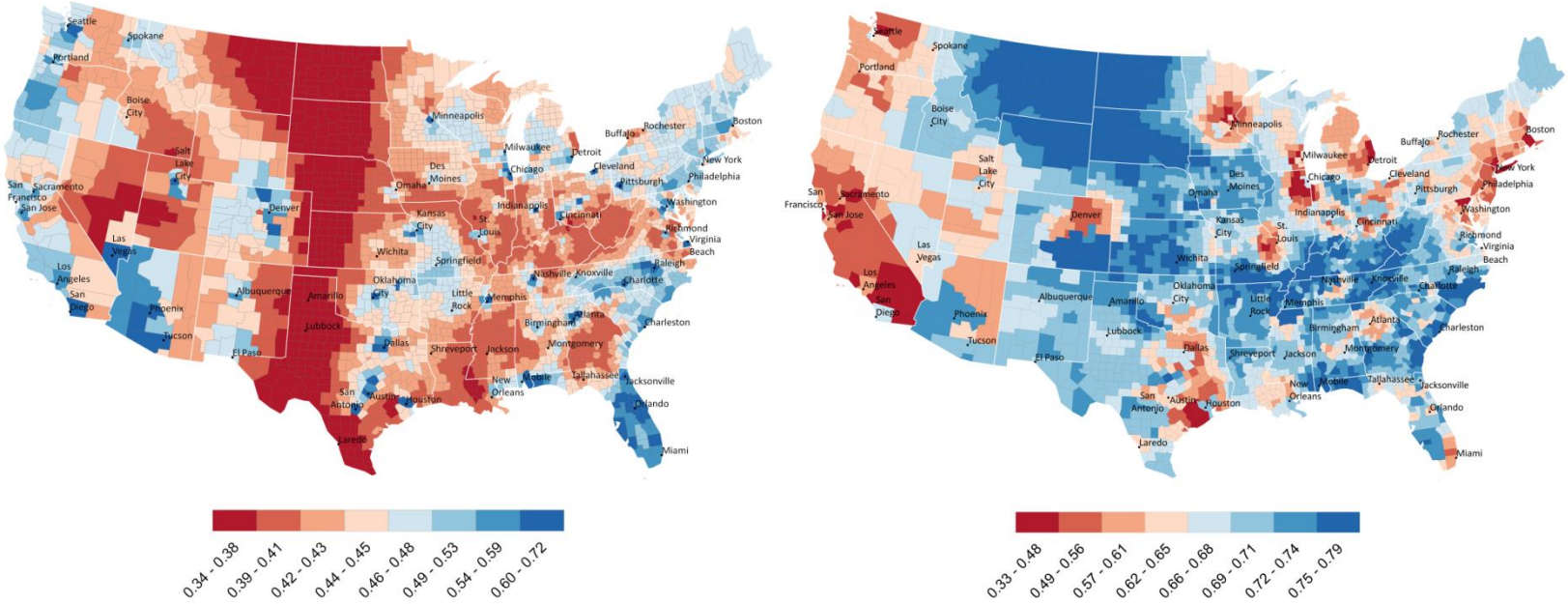



**Figure 2.18:** Comparison of conventional smoothing result (left map) and our result (right map) for inflow entropy. The patterns are dramatically different.

It is important to note that a locational measure can only represent one aspect of the spatial and/or structural characteristics of a location in the network. More insight can be gained through analyzing the relationships between different measure results. For example, if we compare the results of smoothed net migration rate (Figure 2.6) and smoothed in-flow entropy (Figure 2.14), we could discover overlapping clusters such as the coincidence of high entropy clusters with clusters of high net migration rate in the east coast from Virginia to Florida in addition to the coastal areas in the east of New Orleans; and throughout most of the counties in the states of Arizona and Florida. The overlapping of these clusters could help reveal regions that attract migrants from diverse places (as oppose to places that only receive migrants from certain places). We limit our analysis scope since the goal is to present the new smoothing approach rather than carry out a comprehensive analysis of the migration dataset.

We conducted a variance-driven sensitivity analysis to evaluate a series of population thresholds to examine their effect on smoothing result. The analysis showed that smoothing results are consistent in overall patterns. A large population threshold highlights global patterns such as at the national scale while a smaller threshold can better reveal local patterns. For example, a threshold of two million population shows the Southeast as a homogeneous region of attraction, whereas a threshold of 300,000 population shows downtown Atlanta as a place of depletion and its surroundings as places of attraction.

In this paper, we employed a domain-based approach and used population or inflow/outflow to select an adaptive bandwidth. For other types of spatial interaction data where an attribute such as population doesn't exist or using a population-based threshold

is inappropriate for the context of the analysis, one can employ a data-driven approach to select a bandwidth based on the properties of the network. One potential approach is to employ a graph partitioning method (Guo, 2009) to discover community structures and natural regions (groups of spatially contiguous and strongly connected units). The size of the discovered regions can provide important information for determining the size of the neighborhood (bandwidth). Our experiments with different bandwidth values showed that the smoothing results are not sensitive to small changes in bandwidth and that results with different bandwidths usually reveal patterns at different scales.

# CHAPTER 3

## MAPPING FAMILY CONNECTEDNESS ACROSS SPACE AND TIME[2]

[2] Koylu, C., Guo, D., Kasakoff, A., & Adams, J. W. (2014). Mapping family connectedness across space and time. Cartography and Geographic Information Science, 41(1), 14-26. Reprinted here with permission of publisher.

## 3.1 ABSTRACT

Understanding the structure and evolution of family networks embedded in space and time is crucial for various fields such as disaster evacuation planning and provision of care to the elderly. Computation and visualization can potentially play a key role in analyzing and understanding such networks. Graph visualization methods are effective in discovering network patterns; however, they have inadequate capability in discovering spatial and temporal patterns of connections in a network especially when the network exists and changes across space and time. We introduce a measure of family connectedness that summarizes the dynamic relationships in a family network by taking into account the distance (how far individuals live apart), time (the duration of individuals' coexistence within a neighborhood), and the relationship (kinship or kin proximity) between each pair of individuals. By mapping the family connectedness over a series of time intervals, the method facilitates the discovery of hot spots (hubs) where family connectedness is strong and the changing patterns of such spots across space and time. We demonstrate our approach using a data set of nine families from the US North. Our results highlight that family connectedness reflects changing demographic processes such as migration and population growth.

Keywords: space-time visualization, family connectedness, network measure, social network, family tree

## 3.2 INTRODUCTION

The interaction between geography and social relationships has long been studied by researchers (Festinger, Schachter, & Back, 1963; Hägerstrand, 1976; Michelson, 1970). Due to the wide use of social networking applications (e.g., Facebook and LinkedIn) and genealogy applications (e.g., Family Search and Ancestry), large social networks with geographic information have become increasingly available. Using such data, recent studies have proposed new ways of quantifying relationships, some of which make use of geography to infer social interactions (Backstrom, Sun, & Marlow, 2010; Crandall et al., 2010), while others examine how geography and migration (or movement) influence relationships between individuals (Onnela et al., 2011; Phithakkitnukoon et al., 2011). Understanding of how relationships (e.g., kinship, friendship) evolve across space and time is crucial for decision-making in various fields such as disaster evacuation planning and provision of care to the elderly.

In a social network each individual is represented by a node and each edge represents the relationship between two individuals. The weight of an edge can be quantified in a variety of ways such as the degree of kinship in a family tree; co-authorship in a scientific collaboration network; and the frequency of phone calls, text messages or emails exchanged in a communication network. A social network is dynamic because it evolves (changes) over space and time as individuals move (migrate), new individuals are added or removed, and relationships develop and change over time. In this paper, we use the term “dynamic geo-social network” to refer to a dynamic social network embedded in space and time. Understanding the changing aspects of a dynamic

geo-social network requires methods that can simultaneously account for the spatial, temporal and relational (network) dimensions of the network.

In order to understand the dynamics of social networks embedded in geographic space, a variety of computational and statistical methods such as graph theoretical measures (Scellato et al., 2011), random graph modeling (Schaefer, 2012), factor analysis (Hipp et al., 2012), simulation (Butts et al., 2012), and regression analysis (Viry, 2012) have been introduced by studies in social networks. A similarity between these studies is that they consider geography as a background variable to interpret the results of network analysis. However, the methodologies introduced by these studies have limited capability in analyzing the spatial, temporal and relational aspects of dynamic geo-social networks.

With the advancement of graph drawing algorithms, current methods of graph visualization ( Lewis, Gonzalez, & Kaufman, 2012; Patil, 2011) are effective in discovering network (connection) patterns, e.g., clusters of connected members, or commonalities between friends who share interests and groups in a social networking application. However, existing graph visualization methods are inadequate for discovering the spatial and temporal patterns in social networks. On the other hand, spatiotemporal visualization methods (G. Andrienko et al., 2010; Fyfe, Holdsworth, & Weaver, 2009) have been successfully applied to identify temporal variation of spatial patterns, which often do not adequately consider the network dimension (connections between individuals). Therefore, there is still a lack of methodology that can incorporate the relational aspect (sophisticated relations between individuals) of geographically embedded and time-varying social networks.

We introduce a measure and mapping approach to analyze connectedness in a dynamic family network and its changing patterns across space and time. Our approach differs from the current methods in that it takes into account the time that each pair of individuals spend together, the distance that they live apart, and the strength of their relationship (e.g., the degree of kinship). To demonstrate the approach, we use a dataset of family trees derived from the published genealogies of nine families in the US North over a span of 300 years. The data also include information on migration of individuals. The remainder of the paper is organized as follows. First, we review the related work in the next section. We then introduce our data and describe our methodology in detail. Finally, we present the results and conclude with a summary and a discussion for the future research.

## 3.3 Related Work

This article introduces a methodology to understand the spatial, temporal and relational (network) aspects of a dynamic geo-social network. A dynamic geo-social network evolves (changes) over space and time as the actors of the network move (migrate), new actors are added or removed, and relationships between the actors develop and change over time. We demonstrate our approach using a dynamic family network embedded in space and time. Previous approaches to analyzing geo-social networks span a variety of themes and methodologies. In this section, we review the studies that aim at bridging social network analysis and spatial analysis in certain aspects.

### 3.3.1 Computational and Statistical Methods

There are various studies on geo-social networks within the social network domain. For example, a number of studies (Daraganova et al., 2012; Lomi & Pallotti, 2012; Sailer &

McCulloh, 2012; Schaefer, 2012) used exponential random graph models to account for geographic embeddedness of individuals in modeling social networks and investigate the effects of social and spatial distance on the network structure. Doreian and Conti (2012) analyzed a set of empirical networks to understand how networks are shaped by social and spatial contexts using a variety of modeling strategies. Butts et al. (2012) conducted an exploratory simulation study to examine the influence of spatial variability of background population on the network structure and the social ties.

Viry (2012) examined the relationship between spatial dispersion of personal networks, residential mobility and network composition by conducting regression analyses. Cho, Myers and Leskovec (2011) and Scellato et al. (2011) focused on online geo-social networks to describe the relationship between geography and social interaction using graph theoretical methods. Similarly, Radil et al. (2010) introduced a spatialized positional analysis to reveal spatial patterns of social relations. Hipp et al. (2012), and Mennis and Mason (2012) performed factor analyses to delineate neighborhood boundaries by taking into account the density of social ties and the physical distances between the members of a social network. A similarity between the studies that focus on geo-social networks in the social network domain is that they consider geography as a background variable to interpret the results of network analysis. However, the methodologies introduced by these studies have limited capability in analyzing the spatial, temporal and relational aspects of dynamic social networks.

### 3.3.2 VISUALIZATION

Alternative to modeling, graph theoretical and statistical approaches, network visualization methods have been developed to examine the dynamic nature of social

networks. Dynamic network visualization methods allow the discovery of complex patterns in a network over time using animation (network movies) (Moody et al., 2005) and "small multiple displays" (Robertson et al., 2008). The layout of a graph is constructed by a graph drawing algorithm which often places nodes (individuals) that have strong relationships closer to each other. To enhance the perception of changes in a sequence of graph layouts, a collection of methods are developed by considering additional criteria such as minimizing edge crossings and ensuring repeatability and stability (Bender-deMoll & McFarland, 2006). However, such graph layouts represent only the topological structure of the network while disregarding its geographic dimension.

To incorporate a geographic dimension into the network space, a number of studies (Faust et al., 2000; Nag, 2009; Todo et al., 2011) mapped actors (people) based on their geographic location and drew edges between those actors using different width and color intensity to reflect relative strength of each relationship. However, a graph layout that positions nodes based on their geographic coordinates suffers from the visual cluttering problem. Moreover, with a relatively large network, it is difficult to perceive network structures that involve multiple dimensions (i.e., space, time, and social connections). Because social network data are highly dynamic, it is challenging to reveal how social relationships change across geographies and time by simply displaying a sequence of graphs.

Alternatively, some studies (Luo et al., 2011; Onnela et al., 2011) introduced integrated approaches that use dynamically linked views of network space and geographic space and allow user interactions to demonstrate the interplay of topological

structure and geography. Discovering the interaction between geography and the network is useful in extracting micro-scale (individual level) patterns. However; there is also a need to summarize spatial, temporal as well as relational aspects of such networks in order to provide a general overview of the data.

Hägerstrand (1976) introduced a space-time framework to conceptualize and represent human interactions over space and time. Adopting this framework, many spatiotemporal visualization approaches (e.g., space-time path, density surface, computational and interactive approaches) have been introduced (Aigner et al., 2011). The space-time path approach (Chen et al., 2011; J. Y. Lee & Kwan, 2011) identifies human activity patterns in a social network by visualizing individuals' paths in a three dimensional surface. Alternatively, the density surface approach summarizes the activity patterns by a density surface which is represented with either an animated sequence of continuous surfaces (Rana & Dykes, 2003) or a three-dimensional surface of the space-time continuum (Demšar & Virrantaus, 2010; Nakaya & Yano, 2010). Additionally, some computational and interactive approaches such as self-organizing maps (Agarwal & Skupin, 2008) have been used to identify temporal variation of spatial patterns.

The space-time approach by Shaw et al. (2008), Fyfe et al. (2009) and Andrienko et al. (2010) examine geo-social interaction patterns across space and time, but it does not adequately consider the network dimension (connections between individuals). Therefore, there is still a lack of methodology that can incorporate the relational aspect (sophisticated relations between individuals) of geographically embedded and time-varying social networks. Another challenge in analyzing geographically embedded and time-varying social networks is the small area problem, where a single node or

connection is often too small (with insufficient data) for deriving stable statistical measures. Koylu and Guo (2013) introduced a smoothing approach to mapping graph measures in geographic space. In this research, we introduce a different space-time smoothing or interpolation method for visualizing both network measures and social relations in space and time.

## 3.4 DATA

To demonstrate our approach, we use family tree data derived from published genealogies of nine families from the US North over a span of three hundred years. These books were compiled by family members with the help of professional genealogists. More information on migration has been added by linking the genealogies to the U.S. censuses using data from Ancestry.com. A series of demographic events (e.g., births, deaths, migrations) were coded from the genealogy, including the places where events had occurred. From these event locations and dates we can infer the migration paths of each individual in the families.

For the simplicity of methodology presentation and result explanation, in this paper we only report the analysis results with the Chaffee family (Chaffee, 1909), which was selected over eight other genealogies on the basis of better temporal resolution and information on migration. The Chaffee family includes 1225 males descended in the male line from the founder who came to Hull, Massachusetts from England in 1635. All men born into the family up to 1860 were included along with all siblings of men born through 1840. There were 2387 geo-coded moves and 856 distinct locations where the family members lived in 296 years.

The family data involve only males because women changed their names at marriage; they were more difficult to follow. Although life expectancy changed over time, it was largely due to changes in infant and child mortality (Kasakoff & Adams, 2000). This study included only men who survived to at least age 20. If we included those who had died young, we might have biased the study towards families with high infant and child mortality. Life expectancy at age 20 was remarkably moved westward, albeit at greater and greater distances (Egerbladh et al., 2007).

Information on moves comes from records of vital events. If an event occurred in a place where a person had not previously lived, the move was assigned at a date close to the vital event. Most moves occurred before the vital event and thus the dates are approximate. The most accurate move dates come from the child bearing years because this population had children approximately every two years. Also only about 65% of the men had death dates recorded in the genealogy. For the rest, the last date on record was considered a death date. The animation of the migration of nine families including the Chaffee (CFE) family in the U.S. can be viewed at the link: http://129.252.37.169:8400/flowvis/trajectories/index.html(Koylu, 2013a).

## 3.5 METHODOLOGY

We introduce a measure and mapping approach to analyze the relationships in a family network embedded in space and time. Given a space-time window, the measure quantifies the family connectedness of each individual, considering his/her kinship to other family members coexisted in the window, their geographic distances, and the time duration of their coexistences. We then interpolate the measure values for all locations, map a series of space-time windows to examine the changing dynamics of the family

relationships across space and time. Specifically, the approach consists of three steps. First, the time dimension is partitioned into a sequence of time intervals. Second, within each time interval, we calculate the measure of family connectedness for each individual at each location where he/she was present, considering the closeness (the degree of kinship) of his/her connections he/she has within a geographic distance threshold and the temporal duration of each connection. Third, given the family connectedness value for each individual at each unique location within a time window, a surface of family connectedness is produced using a smoothing and interpolation method based on inverse-distance weighting. In the following subsections, we introduce each of the steps.

### 3.5.1 Time Interval

To allow for a temporal analysis of connectedness in a family network, one can employ a data-driven approach such as sliding windows, top-down or bottom-up segmentation algorithms (Keogh et al., 2001; Warren Liao, 2005) to obtain time intervals. For our case study, we employed a domain-specific approach to partition time series data into equal intervals and reflect meaningful stages of the family tree data. Because some patterns may fall between time windows and not appear, we use a sliding window approach.

In the family dataset, the minimum period needed for a connection (co-existence) to occur is one year. On average a man is 35 years old when a son is born and 20 years is nearly the smallest generation, i.e., the youngest a man might be when he has a son. Also, a period of 20 years divides the life course into meaningful stages: age 1-20 would be before marriage, child bearing should stop by age 60 (Adams & Kasakoff, 1984). So people in different 20 year windows should be in different life stages. Therefore, we partition the data into a time window (interval) length of 20 years. Theoretically, we can

move this 20-year window one year at a time to obtain a smooth time series. To reduce the size of time series (and data redundancy), we move the window 10 years each step. In other words, there is a 10-year (i.e., 50%) overlap between neighboring time windows.

Figure 3.1 shows a sample subset of a family tree data to illustrate the measure calculation. The horizontal axis represents the time periods of individuals (i.e., grandfather A, father AA, uncle AB, uncle AC and son AAA) whereas the vertical axis represents the locations (i.e., Loc 1, Loc 2, Loc 3, Loc 4) of those individuals in those time periods. For example, AC lived in Location 3 between 1674 and 1700 whereas AB lived in Location 4 between 1672 and 1685, moved to Location 3 and lived there between 1685 and 1700. Additionally, the solid vertical lines represent the beginning and the end of time intervals: 1675-1695 and 1685-1705.

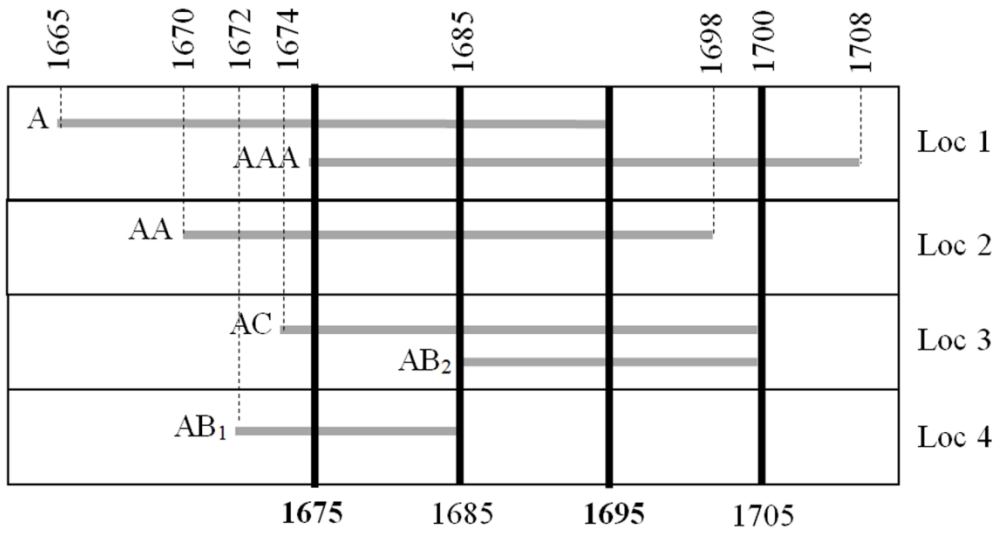



**Figure 3.1:** A sample subset of a family network. The horizontal axis illustrates time and the vertical axis represents unique locations (i.e., Loc 1, Loc 2, Loc 3, and Loc 4). An individual at a location is represented with a horizontal line with a beginning and an ending year. For example, AB1 refers to the period that AB lived at location 4 between 1672 and 1685, whereas AB2 refers to the period that AB lived at location 3 between 1685 and 1700.

### 3.5.2 FAMILY CONNECTEDNESS

We argue that a potential spatial interaction between two individuals in a time period is often dependent on how close those individuals are to each other both in terms of their geographic and kin proximity. While we use geographic proximity to form a territory of potential spatial interaction for each individual, we conceptualize kin proximity by the closeness of the relationship (e.g., degree of kinship) between two individuals. Naturally, the potential for spatial interaction between individuals change across time as individuals move, new individuals are added or removed, and relationships develop and change over time. By taking into account the time-varying relationship between geographic and kin proximity between individuals, and the time duration of their co-existence, we introduce a measure of family connectedness as a proxy for potential spatial interaction.

For each time window, we derive the territory of each individual by using a geographic distance threshold around his location at the time and then calculate the family connectedness of an individual by considering his geographic closeness, temporal overlapping and family relationship to other individuals within the territory. Figure 3.2 illustrates the individual AA's family connections that are determined by his territory (gray circle). We provide a discussion on how to determine the territory of individuals using the distance threshold in the following paragraph. While the nodes with labels illustrate connections of individual AA, empty nodes illustrate individuals that do not have any family relationship with the individual of interest. For the family tree data in this study, we define relationship as kinship and two individuals do not have a relationship if they are not members of the same family tree. For individual AA at location 2, he had five family connections, which are A, AAA, AC and AB (at two

locations, noted as $AB_1$ and $AB_2$) for the given time interval 1675-1695. Notice that, although individuals such as AD, AE, ADA and AF were from the same family with AA, they are not considered as connections because they lived outside the neighborhood buffer of AA.

The choice of the distance threshold (bandwidth) and what constitutes a connection are two important decisions for determining potential connections of an individual at a location and time. To select an appropriate bandwidth, we evaluated the distribution of move distances over time. Figure 3.3 illustrates the box-plots of distances by time intervals. Migration was highly skewed towards shorter distances, as is always the case. Over time the longest distances increased but moves at such distances were relatively rare and overall median distance is approximately 60 km. Still these distances are much greater than they were in Europe (Pooley & Turnbull, 1998) where population density was higher and people were more apt to remain in their local areas and reflect the Westward expansion of the US population.

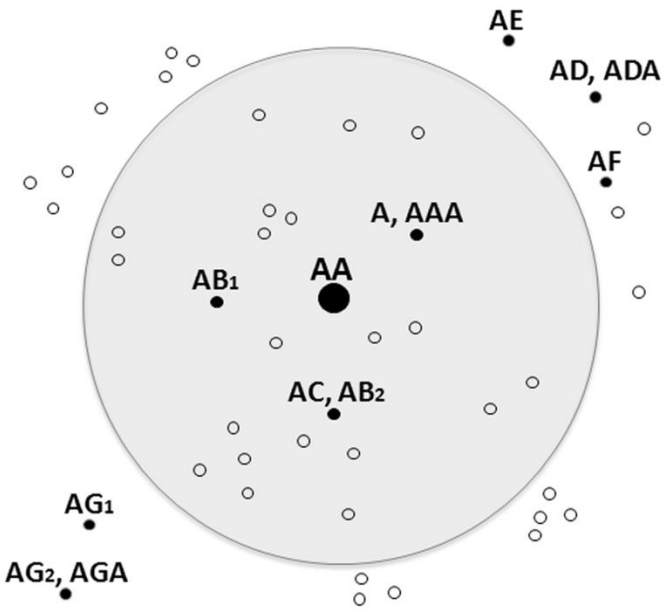



**Figure 3.2:** The potential connections of individual AA from the sample network given in Figure 3.1. The circular buffer illustrates the neighborhood of individual AA which is used to determine his/her potential connections. Nodes with labels within the neighborhood are potential connections of AA whereas empty node symbols and labeled nodes outside the neighborhood are individuals that are not connected to AA. A subscript (e.g., $AB_1$, $AB_2$) for an individual indicates his/her existence at each unique location given the time interval.

Considering the temporal resolution of the data which is composed of recorded events from the late seventeenth century till the mid-twentieth century, increasing trend of migration distance could be attributed to what transportation medium was available for the given time period. Until the mid-nineteenth century when the first railway system was built in the northeastern states, traveling was limited to the capability of horse carriages. Along the railway lines the ability to travel long distances greatly increased. However,

horse carriage remained to be a major transportation medium. On average horse quality, terrain and weather conditions, a horse carriage was able to travel 32 - 64 km a day (Bogart, 2005). Assuming that a potential for a consistent spatial interaction is possible without moving homes, we chose 60 km as a threshold distance to identify potential connections for each individual.

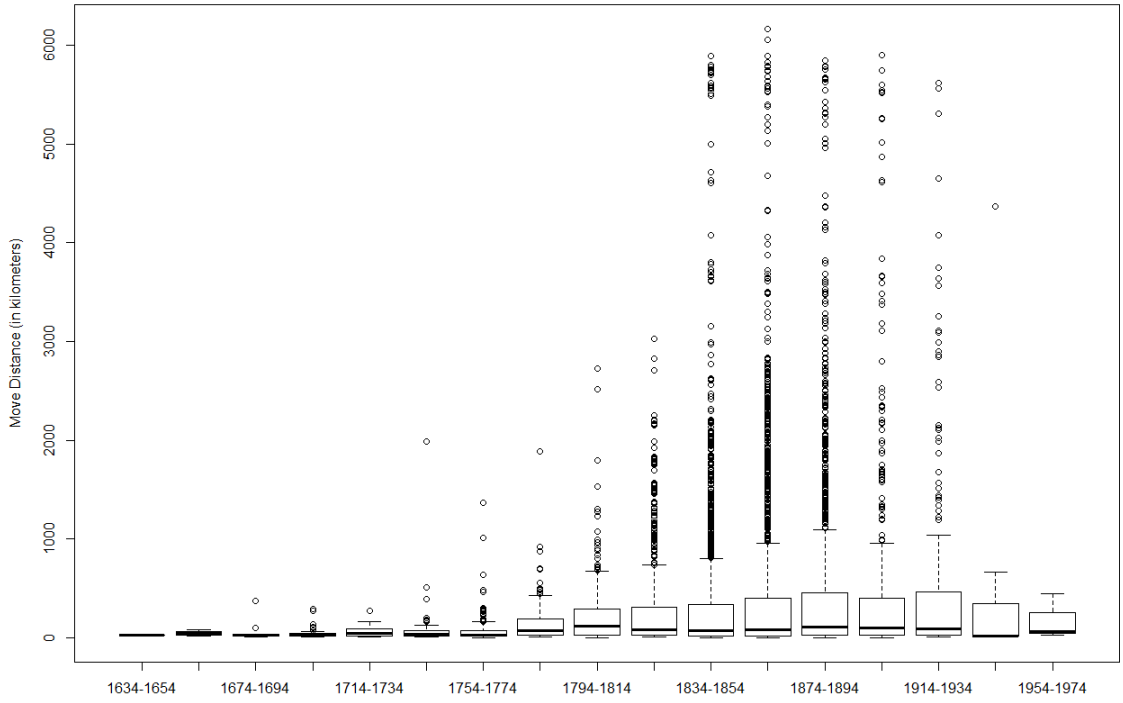



**Figure 3.3:** Distribution of move distances over time intervals. The median move distance is approximately 60km and there is an increasing trend of individuals moving greater distances over time.

The second important decision is to determine what constitutes a connection between individuals in a family network. In this study, we define connection as kinship and we assume that two individuals are connected if they are from the same family tree.

Given the connections, we use Equation 3.1 to calculate each individual's family connectedness at a specific time interval and a specific location

$$FC_{rt}(i) = \sum_{j \in N_{rt}} T_{rt}(i,j) * KP(i,j)$$

**Equation 3.1:** Family connectedness of an individual at a specific time interval and a specific location

where $FC_{rt}(i)$ is the family connectedness for individual *i* at location *r* and in time interval *t*. Individual *i* may have more than one locations (one at a time) in the time interval. $N_{rt}$ are family members within the neighborhood of the individual *i*'s location (*r*) and the time interval *t*; $T_{rt}(i,j)$ is the duration of time that individuals *i* and *j* co-existed within the neighborhood of *r* and the time interval *t*; $KP(i,j)$ is the kin proximity which describes the degree of kinship between the family members *i* and *j*.

We use consanguinity (Leutenegger et al., 2011) to quantify the degree of kinship (relation) between the members of a family, which is widely used in law and genetics. Figure 3.4 represents a family tree of four generations where A is the ancestor of all members in the family. The relation among two people is called lineal consanguinity if one is descendant from the other such as the son and the father (e.g., A-AA), or the grandfather (e.g., A-AAA), and so upwards in a direct ascending line. The degree of lineal consanguinity is directly measured by the number of lines (e.g., edges in Figure 4) between the two family members. For example, father-son relations (e.g., A-AA, AAB-AABA) are first degree; grandfather-grandson relations (e.g., A-AAA, AB-ABBB) are

second degree, and great grandfather-great grandson relations (A-AAAA, A-ABBA) are third degree.

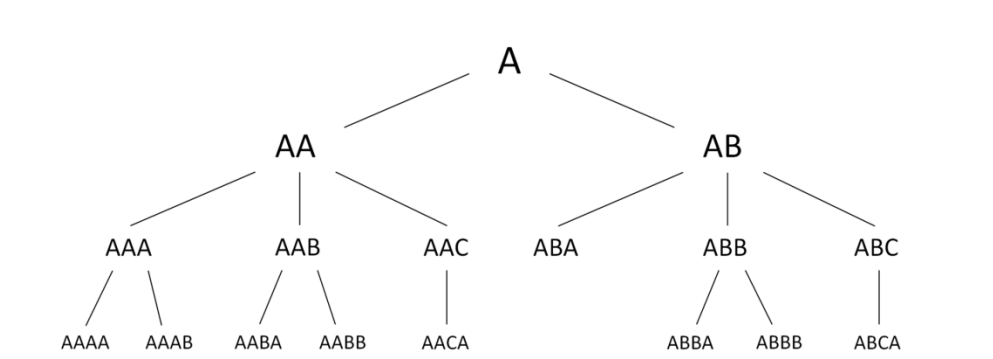



**Figure 3.4:** A sample family tree with four generations that descend from the ancestor A. The relation among two people is called lineal consanguinity if one is descendant from the other such as the son and the father (e.g., A-AA), or the grandfather (e.g., A-AAA), and so upwards in a direct ascending line. For people who descend from the same ancestor, but not from each other (e.g., cousins or uncles-nephews), the relation is called collateral consanguinity.

The relation between individuals who descend from the same ancestor, but not from each other (e.g., cousins or uncles-nephews) is called collateral consanguinity. The degree for collateral relationship is calculated by finding the common ancestor then counting the number of steps downwards to reach the two individuals. If one of the individuals is more distant (remote) to the ancestor, the number of steps to the more remote person determines the degree of consanguinity. For example, a relation between brothers (e.g., AA-AB, AAB-AAC) is considered as a first degree consanguinity since there is only one step from the father to each of them whereas an uncle-nephew relation (e.g., AA-ABA, AAB-AACA) is a second degree consanguinity because the nephew is two steps away from the common ancestor, and the rule of calculating the degree is extended to the more remote person of the collateral relationship. After determining the

degree of relation (consanguinity) between two individuals, we assign a kin proximity value to each relation by simply taking the inverse of the degree. For example, the kin proximity of a first degree relationship (e.g., father-son, brothers) is 1/1 = 1, whereas the kin proximity of a second degree relationship (e.g., grandparent-grandchildren, cousins) is 1/2 = 0.5, and a third degree relationship (e.g., great uncle/grandnephew: AA and ABBA) is 1/3 = 0.33, and so on.

### 3.5.3 Spatial Interpolation of Family Connectedness

The components of the measure, which are cumulative kinship and time for an individual at a location, are highly correlated with the presence of individuals that live within a close distance to that location. We discuss that more people living close by increases the chance of potential interactions, thus the correlation between the presence of individuals and the measure components is appropriate and does not necessitate normalization. As we are not interested in family connectedness as a cumulative measured quantity, we produce a geographically weighted average surface of family connectedness by using a spatial smoothing and interpolation method rather than a cumulative density surface of family connectedness.

Given the family connectedness for each individual at each unique location within a time window, a surface of family connectedness is produced using a smoothing and interpolation method based on inverse-distance weighting (IDW). IDW assumes that each measured value has an influence on the prediction by applying weights that are proportional to the inverse of the distance between the prediction location and the measured data point. The equation for smoothing and interpolation method is given below:

$$FC(x,t) = \sum_{i \in N_{xt}} \frac{w_i(x)\, FC_{r \in N_{xt}}(i,t)}{\sum_{j=0}^{N} w_j(x) + W_c}$$

$$w_i(x) = \frac{1}{d(x,r_i)^p}$$

**Equation 3.2:** The equation for smoothing and interpolating family connectedness

where $\boldsymbol{FC(x,t)}$ is the interpolated value of family connectedness at location *x* in time interval *t;* $\boldsymbol{N_{xt}}$ is the observations (existence of individuals at unique locations) within the neighborhood of *x* in time interval *t*; $\boldsymbol{FC_{r \in N_{xt}}(i,t)}$ is the family connectedness value for individual i at location r in time interval *t*; $\boldsymbol{w_i(x)}$ is a weighting function based on the distance $\boldsymbol{d(x,r_i)}$ from the location of the observation $\boldsymbol{r_i}$ to the unknown point *x*; *p* is a positive real number called the power parameter; $\boldsymbol{W_c}$ is a constant penalty weight added to each estimation to remove the edge effect (Lawson et al., 1999).

We determine the neighborhood for each estimation point x by using the same distance threshold (60km) we used in the previous step to identify connections between individuals. Because there are many observations (the co-existence of individuals) at the same or close by locations, the traditional IDW creates an interpolated surface which is greatly influenced by the edge effect (Figure 3.5(a)). After applying the penalty weight for locations with no or few observations the edge effect is removed (Figure 3.5(b)). A constant penalty weight does not have a significant effect on the estimation where there are many observed values by the estimation point; however, it does affect the estimation where there is a few or no observed points close by the estimation point. To balance

between over-smoothing and under-smoothing, we empirically chose a value of 0.1 for the constant weight $\boldsymbol{W_c}$.

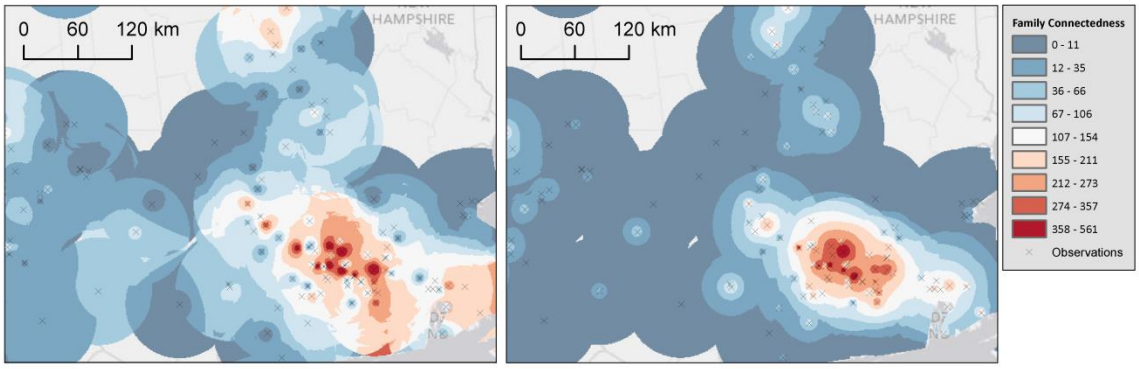



**Figure 3.5:** The comparison of the traditional IDW (a) with the modified IDW (b). The edge effect is noticeable throughout the traditional IDW surface (a). By applying additional weight that penalizes locations with no observations or few observations, the edge effect is removed in the modified IDW surface (b).

## 3.6 Results and Discussion

We produced 29 surfaces of family connectedness each of which corresponds to a 20 year time window. Each surface was produced using a constant divergent classification scheme to enable comparison between each time window. While blue hue illustrates places with low family connectedness (i.e., low potential for spatial interactions), red hue illustrates places where family connectedness is higher.

The animations of family connectedness including all families and the Chaffee family can be viewed at the following link: http://www.spatialdatamining.org/familyconnectedness (Koylu, 2013b). For the simplicity of result explanation, in this paper we only report the analysis results with the Chaffee family, which was selected over eight other genealogies on the basis of better

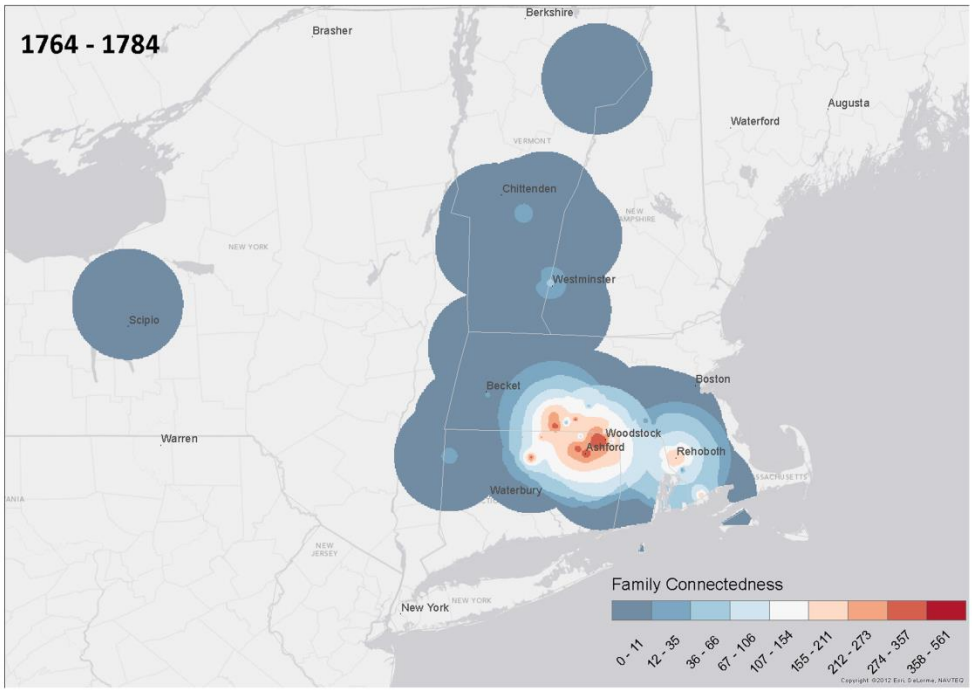
temporal resolution and information on migration. Due to the limited space we only report a small subset of the time windows that we selected based on their relevance to historical events in chronological order.

The surfaces of family connectedness from the first time window (1634-1654) to the time window of 1854-1864 illustrate the demographic and spatial expansion of a colonizing population. Colonization proceeded in spurts with a family member moving out of the settled area and then most of his descendants remaining in the new location for three generations before spawning new settlements. It takes many years in a new location for connectedness to peak.

Before the American Revolution, the earlier window (Figure 3.6(a)), there is only a few individuals in the newly settled areas such as Scipio, Warren, Chittenden, Westminster and Berkshire. Sixty years later (Figure 3.6(b)), the core of the family stayed in the area between Woodstock and Becket while new hubs started to develop in Berkshire and Scipio as the family moved North and West after the Revolution. As compared to Chittenden and Westminster there were fewer individuals in Berkshire in 1824-1844 but Berkshire became a stronger hub than Chittenden and Westminster.

When the new hubs were created, it took several generations to achieve the degree of family connectedness of the places that had been settled by the earliest generations. Due to the new births and new migrations of close kin into the area, Berkshire was able to increase its strength as a family hub in the later periods (Figure 3.7).

(a)




(b)

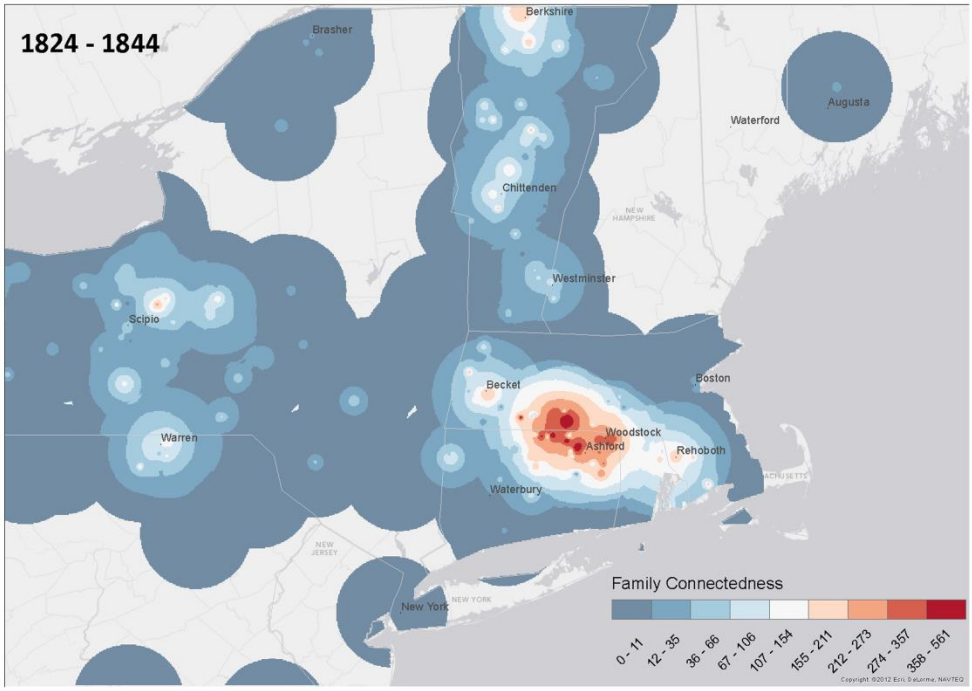



**Figure 3.6:** Family Connectedness after the American Revolution: 1764-1784 (a), 1824-1844 (b).

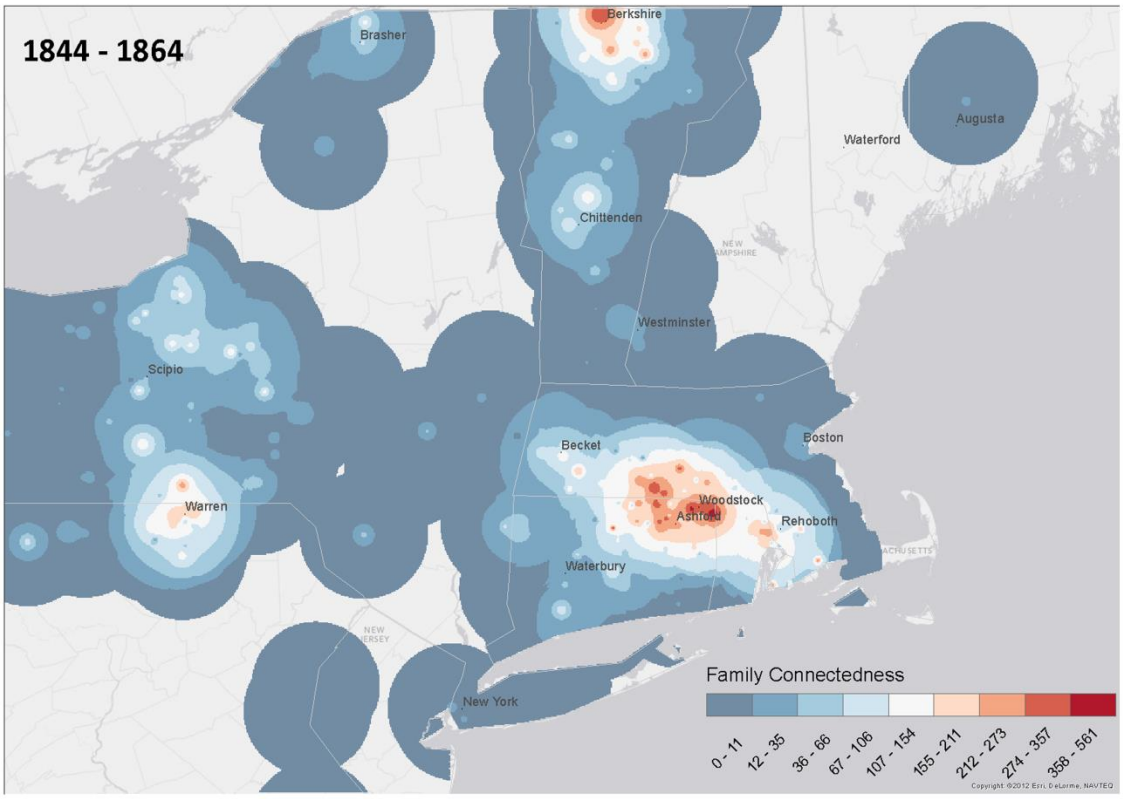



**Figure 3.7:** Family Connectedness throughout the process of urbanization (1844-1864)

Family connectedness is a composite measure of shared time (co-existence of individuals in a neighborhood) and kin proximity (e.g., the closeness of their kinship). To better understand the relationship between shared time and kin proximity, one could decompose the family connectedness of an individual at a location and time interval (Equation 3.1) into its components of total shared time (Equation 3.4) and total kin proximity (Equation 3.5). We plot these components for two distinct time intervals (i.e., 1764-1784 and 1844-1864) to capture the temporal variation of the relationship between time and kin proximity (Figure 3.8). Both components are correlated with and influenced by the presence of individuals at close by locations (i.e., 60km), thus time and kin proximity were highly correlated in both time intervals. The difference between the

intervals of 1764-1784 (Figure 3.8(a)) and 1844-1864 (Figure 3.8(b)) suggests that over time individuals spend more time in close by locations whereas the availability of kin in their territory stayed the same. This trend could partially be explained by increased co-existence of individuals with distant kin.

$$Total\ kin\ proximity_{rt}(i) = \sum_{j \in N_{rt}} KP(i,j)$$

**Equation 3.3:** Total kin proximity

$$Total\ shared\ time_{rt}(i) = \sum_{j \in N_{rt}} T_{rt}(i,j)$$

**Equation 3.4:** Total shared time

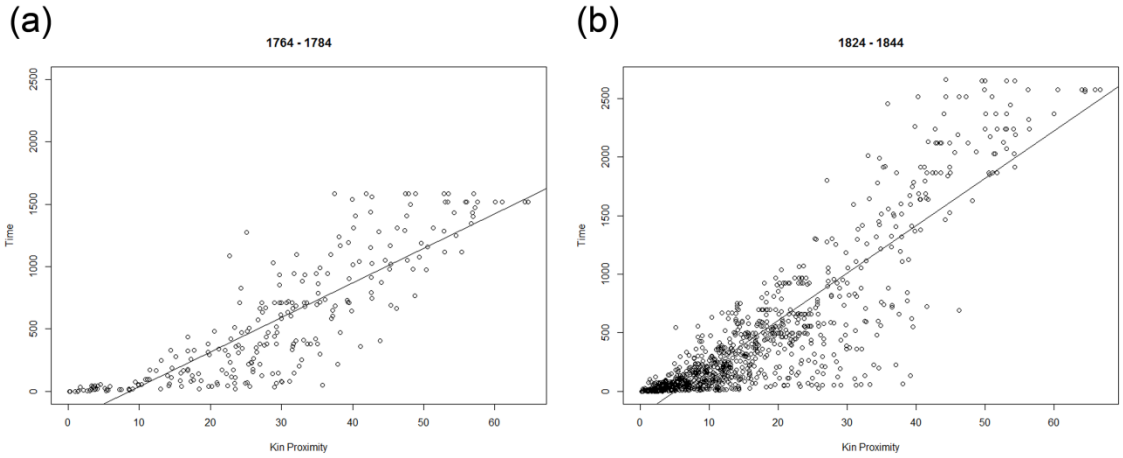



**Figure 3.8**: Relationship between the total shared time and the total kin proximity (kinship) of each individual's connections within a neighborhood in time intervals 1764-1784 (a) and 1824-1844 (b). The vertical axis represents the total shared time whereas the horizontal axis represents the total kinship.

The contribution of kin proximity and the shared time to the measure result vary across space and time. For example, an area with high family connectedness might be a result of high shared time but low kinship due to the co-existence of a large group of distant relatives (e.g., cousins, 2$^{nd}$ level cousins). In an opposite case, an area with high family connectedness might be a result of low shared time but high kinship because of the co-existence of close relatives (e.g., parent-children, siblings) in shorter periods of time.

To examine the spatial variation of the relationship between shared time and kin proximity, we performed bi-variate local indicators of spatial association (LISA) (Anselin, 1995). Bi-variate LISA examines whether local correlations between values of a variable (e.g., time) at a location and those of its neighboring values of another variable (e.g., kinship) are significantly different from what you would observe under conditions of spatial randomness. For example, a significant low-high cluster means that low values of a variable such as shared time are significantly correlated with high neighboring values of another variable such as kin proximity.

We are particularly interested in understanding contrasting patterns of kin proximity (kinship) and shared time, thus, in this article, we only report statistically significant associations with high kinship-low time, and low kinship-high time for the time intervals 1764-1784 (Figure 3.9(a)) and 1824-1844 (Figure 3.9(b)). In the early stages of the expansion (1764-1784) we observe a cluster of low time - high kinship values especially around Ashford whereas Rehoboth continued to be a location with high time and low kinship. In the later period after the American Revolution (1824-1844) the

family hubs located around Ashford still had low time but high kinship values but the spatial extent of the hubs became more dispersed.

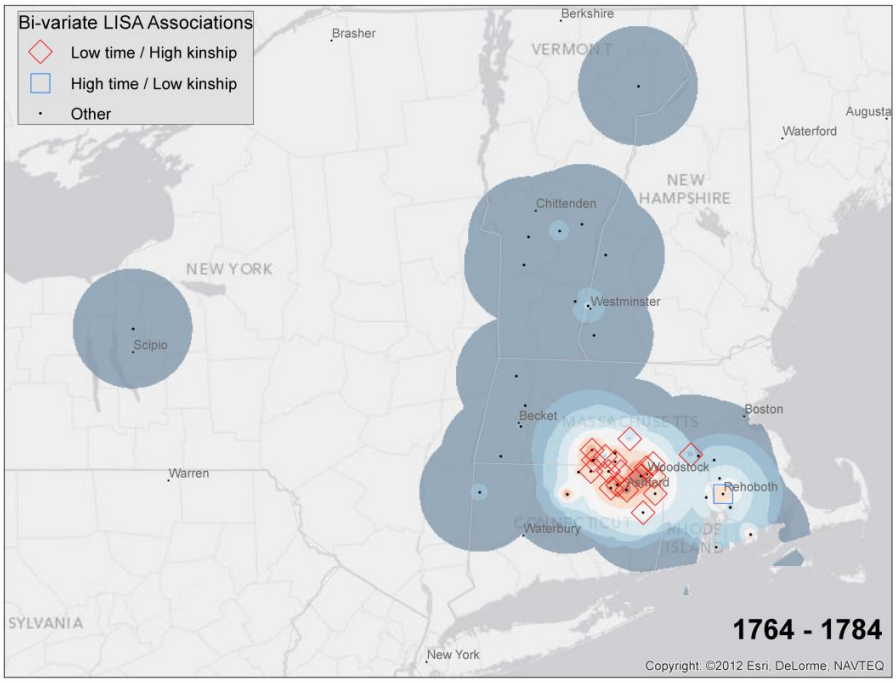
Moreover, we start to see the formation of new hubs around Scipio and Chittenden which have high time but low kinship values. Low kinship and high time associations occurred in especially Scipio and Chittenden because younger individuals which were distant relatives moved to these new hubs whereas patriarchs stayed around the old established hubs. On the contrary, we observe a contrasting pattern, high kinship - low time (red diagonals), around Berkshire in the later time period of the expansion (1824-1844). This is because Berkshire became an established hub as a result of in-migration of close relatives and high presence of patriarchs.

## 3.7 Conclusion

We introduced a measure of family connectedness that summarizes the dynamic relationships in a family network embedded in space and time. The new measure is unique because it takes into account the duration of time that each pair of individuals spend together, the distance that they live apart, and the strength of their relationship (e.g., the degree of kinship). By mapping the relational aspects of a family network across space and time, our method facilitates the discovery of hot spots (hubs) where potential for spatial interaction between individuals is relatively higher across space and time.

One aspect of social context that is often not considered by the studies that incorporate social network analysis and spatial analysis is the addition or removal of members (e.g., birth/death, entry/exit) of the network. Our work has shown that especially deaths of individuals (e.g. patriarchs) who link many others together can greatly affect family connectedness particularly in the locations where those individuals

(a)




(b)

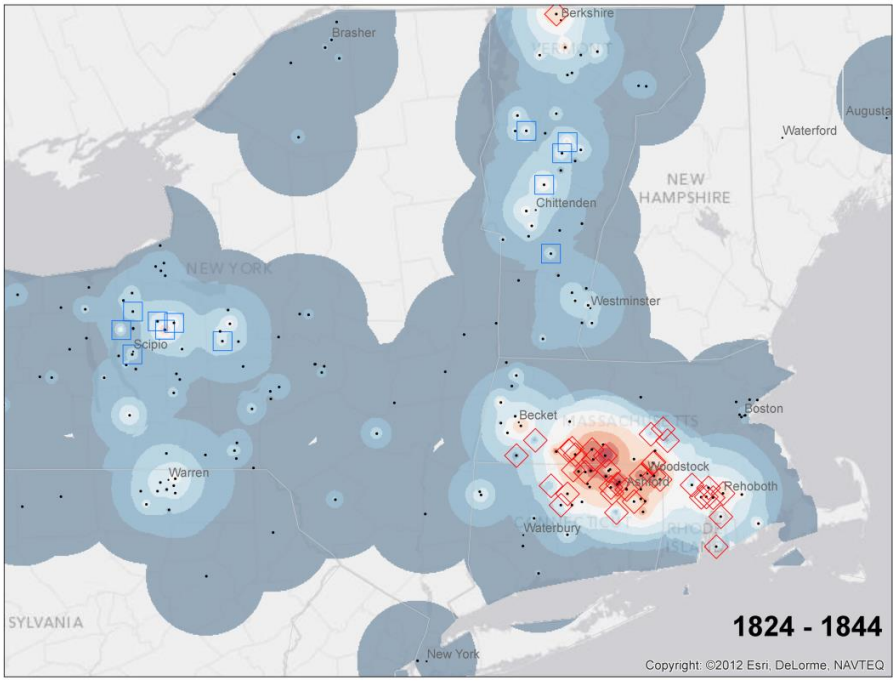



**Figure 3.9:** High-low and low-high associations of shared time and kinship in time intervals 1764-1784 (a) and 1824-1844 (b). While red diamonds represent low shared time and high kinship, blue squares represent high shared time and low kinship. The contrasting associations of cumulative shared time and kinship vary across space and time throughout the spatial and demographic expansion of the population

live. The family dataset we analyzed had a very high rate of demographic increase as was characteristic of the US North at the time. If death rates were higher, presumably there would be fewer hubs or hot spots. In other words, hot spots would disappear much more quickly with those in Europe where death rates were much higher.

We are studying potential, not actual, interaction and this is one limitation of our work. We do not have a measure of actual interaction that we could compare with the potential interaction we have described. But there are other historical data sets which do have such measures. One example would be data on witnesses to marriages or other events, which exist for several European countries in the past (Bras, 2011). Our measure could also be computed using current kinship networks and then compared with the actual interactions of family members obtained from questionnaires or by other means.

We demonstrated our approach using a family tree dataset from a population that was growing and colonizing the Northern part of the U.S. This measure has demonstrated how important migration, birth, and death of individuals to family connectedness. In this study, we define relationship as kinship and assume that two individuals have a relationship if they are from the same family. Our methodology can readily be extended to develop a measure of social connectedness using other forms of relationships, such as friendship and co-workers.

Depending upon the context of the social network, one can define the connections in any form of interaction and quantify those interactions in a variety of ways such as using the frequency of the shared content, common friends in an online social network; the number of email exchanges or meetings held together in a business network. In this regard, the voluminous data collected from social networking platforms such as Twitter,

Flickr and Foursquare and genealogy applications such as Family Search and Ancestry provide an excellent opportunity to study online social networks and family trees using our approach.

# CHAPTER 4

## UTILITY AND USABILITY EVALUATION OF FLOW MAP DESIGN[3]

[3] Koylu C. & Guo, D. (2014) Utility and Usability Evaluation of Flow Map Design. To be submitted to International Journal of Geographical Information Science.

## 4.1 ABSTRACT

Flow maps are commonly used to depict the movement of phenomena between pairs of locations (origins and destinations) for the understanding and communication of flow data and patterns, such as locational and network characteristics. Existing research on graph visualization has developed measures of readability and aesthetics to assess the utility and usability of graph layouts in presenting node and network characteristics. However, there is a lack of research for empirical evaluation of flow map designs and their utility and usability. We designed a user evaluation to obtain knowledge on how map readers perceive information presented with flow maps, how design factors such as flow line style (curved or straight) and layout characteristics may affect flow map perception and users' performance in addressing different tasks for pattern exploration. Specifically, our user testing compares four different layout settings in combination with two flow line styles, traditional straight line and Bézier curved flow line, in terms of users' performance in addressing two tasks: identifying nodes with the highest inflow and outflow. We measured correctness, response time and perceived mental effort in completing each task. The statistical analysis of the test data showed that performance (correctness and response time) and perceived mental effort of participants varied significantly depending on the factors of design, task, layout and screen resolution. The findings of this study have important implications for iterative design, interaction strategies and further user experiments on flow mapping.

Keywords: Flow mapping, usability evaluation, geovisualization, map reading

## 4.2 INTRODUCTION

Both physical and intangible phenomenon such as people, commodities and information constantly move in the geographic space and create location-to-location networks (graphs) that are often referred to as spatial interactions. In a location-to-location network, each node represents a geographic location (or area) and a link represents an interaction between a pair of locations. For example, domestic freight shipments within the U.S. form a network of state-to-state commodity flows in which there are 50 nodes (states) and thousands of links (commodity imports/exports between states).

Flow maps are commonly used and most intuitive to facilitate the understanding of flow patterns and the spatial context in a spatial interaction network. Studies in the graph visualization community (Ghoniem et al., 2005; Purchase et al., 1997) introduced measures of readability and aesthetics to assess the utility of graph layouts. There are also efforts (Alper et al., 2013; McIntire et al., 2012) to evaluate the effectiveness of weighted node-link diagrams. Such experimental studies provide important insights on how users understand visual graph drawings and have suggested various graph drawing principles such as minimizing edge crossings, maintaining large crossing angles, and obtaining symmetrical layouts.

However, there is a lack of empirical evaluation of flow maps. The heuristics learned for general graph drawing can only give very limited guidance for flow map designs. A flow map layout is constrained by the geographic coordinates of nodes, which dramatically differs from non-spatial graph drawing (where nodes can be moved freely to enhance visual clarity). Moreover, there is a lack of experimental research that assesses

the extent to which users perceive and interpret flow maps based on different design characteristics.

We introduce a user evaluation to obtain knowledge on how map readers perceive information presented with flow maps, and how design factors such as flow line style (curved or straight) and layout characteristics may affect flow map perception and users' performance in addressing different tasks for pattern exploration. Specifically, our user testing compares traditional straight line flow maps with Bézier curved flow maps with four systematically varied layout settings based on two visual tasks: identifying nodes (locations) with the highest inflow and outflow. We measured correctness, response time and perceived mental effort in completing each task. To demonstrate the application and user experiment, we used an original commodity flow dataset in the U.S. from 2007. The remainder of the paper proceeds as follows. We introduce the related work on flow mapping, and evaluation methods in Section 4.3. Then, in Section 4.4, we describe our methodology and the experiment. We finally report and discuss the results obtained from the user experiment in Section 4.5. and 4.6.

## 4.3 Related Work

### 4.3.1 Flow Mapping

Slocum et al. (2009) identifies five kinds of flow maps: distributive, network, radial, continuous and telecommunications flow maps. We focus on distributive flow maps that depict flows using abstract links between locations rather than the precise routes of flows. The term flow map is interchangeably used to refer to distributive flow map in the remaining of this paper. In the next sub-section, we discuss design issues regarding

distributive flow maps. Design issues regarding other types of flow maps can be found in (Parks 1987; Slocum et. al. 2009).

#### 4.3.1.1 DESIGN

In a flow map, a flow is often depicted as a straight or curved line connecting an origin to a destination. The color and/or width of each line can be used to represent the volume of the flow. The directionality of a flow is commonly displayed using arrows and the right-hand traffic rule that draws a flow line on the right side while the line is pointing to its destination (Guo, 2009). To reduce visual cluttering caused by overlapping arrows and lines, a number of strategies such as edge ordering, minimizing overlap with arrows, adjusting vertex positioning to optimize angular resolution and edge crossings can be employed (van de Ven, 2007). Two divergent colors can be used to distinguish the origin and destination of a flow line (Boyandin et al., 2010; Fowler & Ware, 1989). Bézier curves can also be used to draw flow lines where each line is curvy at the origin and straight on the destination end. Cognitive studies in information visualization (Ware, 2013) suggest that visual processing of line curvature is weaker than factors such as color, orientation and size. However, as compared to straight edges, the use of curved edges would lead to improved interpretation of the relational information since curvature produces wider angles between edges and the relations (connections) become more visible (Purchase et al., 2013). Xu et al. (2012) studied the impact of edge curvature on graph readability and found that uniform edge curvature had a detrimental impact on graph readability and this negative effect increased with curvature level. This paper compares traditional straight line flow maps with Bezier curved flow maps on different layout characteristics and tasks.

#### 4.3.1.2 APPLICATIONS

Tobler (1987) was the first one to develop a flow mapping application. Tobler's original software was later updated to an interactive application that included new features such as colored and scaled arrows, two-way flows and a setting to control the movement volume to be shown (W. Tobler, 2004). Yadav-Pauletti (1996) developed a migration mapping software that utilized animation with small multiples to depict migration flows over time. Similarly, Thompson and Lavin (1996) developed an application to automate the generation of animated vector field maps. Phan et al. (2005) developed a flow mapping application that bundles edges to minimize edge crossings using a hierarchical clustering method. Using node clustering and flow aggregation, Boyandin et al. (2010) introduced an interactive application to analyze temporal changes in migration flows. Boyandin et al.'s (2010) application supports user interactions such as flow and node highlighting, selection and dynamic queries for filtering out flows by their volume and length. Using multiple linked views of a flow map, a self-organizing maps and a parallel coordinate plot, Guo (2009) introduced an interactive and integrated flow mapping framework to discover community structures (natural regions), identify multivariate relations of migration flows, and examine the spatial distribution of both flow structures and multivariate patterns.

### 4.3.2 EVALUATION METHODS

#### 4.3.2.1 GEOVISUALIZATION

Traditional testing methods under controlled conditions are not suitable for evaluating geovisualization tools because of the exploratory nature of visualization and it is hard to define effectiveness or "success" for an exploratory task (Demsar, 2007). There are two

alternative approaches to evaluating a visualization tool: the insight-based approach and the objective-based (visual tasks) approach. The insight-based approach (Chang et al., 2009; North, 2006) captures and grades individual observations about the data or visualization by the participant as an insight, a unit of discovery. On the other hand, visual tasks are derivatives of basic visual operators such as identify, compare, associate, etc., and were first introduced by Wehrend and Lewis (1990). Several studies (Aufaure-Portier, 1995; Davies, 1995; Knapp, 1995; Roth, 2012) decoded the exploration process into objectives (e.g., identifying clusters in the data, finding relationships between elements, comparing values at different locations and distinguishing spatial patterns, identifying spatial positions of objects, their spatial distribution and density, etc.). Many studies (Demsar, 2007; Koua et al., 2006; C. Tobon, 2005) performed experiments to evaluate the utility and usability of geovisualization tools using the objective-based approach. Additionally, a number of studies (Fowler & Ware, 1989; Laidlaw et al., 2001; Z. Liu et al., 2012) evaluated the effectiveness of the visualization techniques for displaying particles as vector fields.

#### 4.3.2.2 Graph Visualization

Although an empirical evaluation of flow mapping has not been explicitly addressed in the geovisualization community, many studies in graph visualization (Dwyer et al., 2009; Ghoniem et al., 2005; Purchase et al., 1997; Ware et al., 2002) examined how aesthetics such as edge crossings and symmetry impact the performance of graph reading tasks in unweighted graphs. Additionally, some studies utilized eye-tracking (Huang, 2007; Körner, 2011) to understand graph perception. The findings of the experiments (Battista et al., 1999; Purchase et al., 2002) on the evaluation of graph drawing algorithms provide

important insight into users' understanding of graphs and suggest various graph drawing heuristics such as minimizing edge crossings and the ratio between the longest edge and the shortest edge; and satisfying some aesthetics criteria such as revealing symmetries. While following such heuristics ease the comprehension of graphs, flow maps can benefit from these rules in a limited manner because of fixing of nodes based on geographic coordinates.

Ghoniem et al. (2005) compared the performance of matrix-based representations and node-link diagrams on a number of tasks and showed that matrix-based visualization outperforms node-link diagrams when graphs have more than twenty nodes. However, node-link diagrams were still found to be effective in path finding tasks. Alper et al. (2013) and McIntire et al. (2012) analyzed the performance of node-link diagram and adjacency matrix for comparing two weighted graphs. Findings of their studies show that matrix representation outperforms node-links for graph comparison tasks, especially when the graph is dense or large. McGrath et al. (1997) conducted an experiment to understand how spatial properties of the graph layout affect the viewer's perception of the graph when structural features were held constant. Similarly, in order to evaluate how spatial arrangement of the layout influences graph perception, we produced four layouts from the same dataset by swapping the positions of nodes while keeping the flow structure as constant.

A graph layout may improve aesthetics, however it does not ensure understanding (Bennett et al., 2007). To address the problem, measures of mental effort and visualization efficiency have been introduced to better understand the perception of graphs using the cognitive load approach (Paas et al., 2003). Huang, Eades and Hong

(2009) further developed a measure of visualization efficiency which is the difference between cognitive cost (i.e., mental effort and response time) and cognitive gain (response accuracy). According to the definition of visual efficiency, high efficiency is gained with high accuracy and low mental effort and a short response time, whereas low efficiency occurs when low accuracy is associated with high mental effort and a long response time. In this experiment, we measure mental effort using subjective ratings of the participants in addition to response time and accuracy of each given task.

## 4.4 METHODOLOGY

This paper introduces a user evaluation to obtain knowledge on how map readers perceive information presented with flow maps. Given the large number of alternatives for flow map design and flow map reading tasks, it is challenging to choose design elements and tasks for the evaluation of flow maps. In order to finalize the experimental factors, we conducted a series of prior experiments and cognitive walkthroughs on the elements of flow map design such as line style, width, color, and arrow size; background map, and dataset (layout); and flow map reading tasks. In our final experiment which we report in this paper, we evaluate two factors of flow map design: flow line style and flow map layout based on a set of visual tasks. In the following sections, we introduce the experimental factors, research questions, participants and procedure.

### 4.4.1 EXPERIMENTAL FACTORS

#### 4.4.1.1 DESIGN

Previous studies in graph visualization showed that curved edges are easier to interpret as they produce wider angles between edges and the connections become more visible;

however, there is a lack of experimental research on how line styles such as straight and curved edges influence flow map perception. To obtain information on how different line styles facilitate the comprehension of network and node characteristics in a flow map, our particular focus is on the utility and usability of traditional straight-line flow maps with Bezier curved flow maps. We evaluated the alternatives for the symbolization of color and line width by using an interactive application: http://tinyurl.com/lbv464d (Koylu, 2014a). To encode the volume of flows, we used redundant symbolization with a sequential classed color scheme and proportional line width. We used partial arrows to delineate the direction (destination) of flow lines and each flow line follows the right-hand traffic rule: a flow line is drawn on the right side while pointing to its destination.

#### 4.4.1.2 LAYOUT

As a result of being constrained by geographic coordinates of nodes, flow map layout suffers from the visual complexity induced by the number of flows, flow lengths, and crossings. However, there is a lack of experimental research that examines the effect of layout characteristics on flow map perception. To understand how different layout characteristics influence the perception of flow maps and account for any bias that would be introduced by a particular layout and participants' knowledge of the geography, we designed the experiment with four layouts that we derived from an original commodity flow dataset: the flows of alcoholic beverages by their weight in tons, collected by the Commodity Flow Survey (CFS) in 2007. We chose this dataset because it exhibits distinctive patterns of high import and export states with few high volume links and many low and medium volume links. All commodity flow datasets can be viewed at the following link: http://tinyurl.com/lbv464d (Koylu, 2014a).

The four layouts consist of the original layout, and three fictionalized layouts created by systematically swapping coordinates of locations while using the identical network. We designed the fictitious layouts based on spatial arrangement of the most prominent nodes and larger volume flows with varying degrees of flow length and crossings. Figure 4.1 illustrates the four layouts with the two design alternatives: curved and straight flow layouts. In Figure 4.1, we ordered the layouts such that the number of edge crossings and total flow length increase from top to bottom: Layout 3, Layout 1, Layout 2 and Layout 4. Table 4.1 illustrates layout characteristics such as number of edge crossings, total flow line length and mean crossing angle. Layout 1 is the original layout of the dataset. Layout 1 suffers from edge tunneling effect (Dunne & Shneiderman, 2009) caused by overlapping flows from/to close by nodes. We designed Layout 2 so that the most prominent nodes and larger volume flows are at the periphery of the layout and there are no edge crossings among the highest-volume flows (dark purple). This layout resulted in fewer crossings of edges in general, and longer flows. Layout 3 was produced to keep the most prominent nodes at the center of their connections with no edge crossings within the highest volume class (similar to Layout 2). Layout 3 has the fewest number of edge crossings and the shortest total flow length (different from Layout 2). We designed Layout 4 so that most prominent nodes are around the periphery and there are edge crossings with the highest volume class. Layout 4 has more edge crossings and the longest flows.

#### 4.4.1.3 TASK

Flow map reading is a challenging task as it involves visual judgment of node (location or area), link (flow) and network characteristics. The basic elements of flow map reading

involve cognition of link properties such as magnitude, orientation, direction and distribution of connections (Ware, 2013). For a comprehensive evaluation of flow map reading, there is a need to construct a typology patterns and visual tasks. Similar work has been done in movement pattern analysis (Dodge, Weibel, & Lautenschutz, 2008), group level comprehension in graphs (Saket, Simonetto, & Kobourov, 2014) and task taxonomy for graph visualization (Brehmer & Munzner, 2013; Lee et al., 2006).

Given the large number of possible flow map reading tasks, it is challenging to select tasks for the evaluation of flow maps. In this study, we focus on the comprehension of location (node) prominence in flow maps. Location prominence is often described and measured by degree-based centrality measures such as degree-centrality, betweenness, closeness and eigenvector; and volume-based measures such as total flow (strength), inflow, outflow and net flow. Such measures have been widely studied in analyzing a variety of spatial networks such as migration (Koylu & Guo, 2013; Andrei Rogers & Raymer, 1998; C. Roseman & McHugh, 1982), commodity flows (Celika & Guldmann, 2007; Smith, 1970), and airline networks (O'Kelly, 1998).

We used the following criteria to select the appropriate tasks for evaluating the understanding of patterns in a flow map. First, the tasks should easily be explained to a participant without any knowledge of flow maps or graphs. Second, each task should be unique and not involve similar sub-tasks. Third, the tasks can be completed in a reasonable amount of time. Based on these criteria, we chose two tasks to identify location prominence: identifying locations with the highest total inflow, and the highest total outflow. In our prior tests we also included highest positive netflow and highest negative netflow. However, these tasks were hard to perform due to comparing inflows

and outflows; involved similar sub-tasks with inflow and outflow tasks; and resulted in excessive response times and poor response accuracy. Thus, we removed netflow tasks in the final experiment that we report in this paper. To fit the context of the commodity dataset in the experiment, we used the terms import and export rather than inflow and outflow. Below are the two tasks:

- Select the top three importers (i.e., states) with the highest total volume of imports.
- Select the top three exporters (i.e., states) with the highest total volume of exports.

A specific instruction and a hint were given to help answer each question.

#### 4.4.2 RESEARCH QUESTIONS

The overall goal of the evaluation was to obtain knowledge on map readers' perception of information presented with flow maps, how design factors such as flow line style (curved or straight) and layout characteristics; and tasks of pattern exploration influence flow map perception. Our major research questions were:

- Which type of design, curved or straight line style better assist in facilitating the comprehension of the network and node characteristics, more specifically understanding location prominence?
- How do layout characteristics such as total flow length, edge crossings, crossing angles influence the performance (correctness and response time) and perceived mental effort?
- How is user performance affected by the type of task?
- How does perceived mental effort vary across different combinations of design, layout and task?
- How do the factors of design, layout and task interact with each other and affect the performance and perceived mental effort?
- What other factors could potentially exist to explain variations in the performance and perceived mental effort?

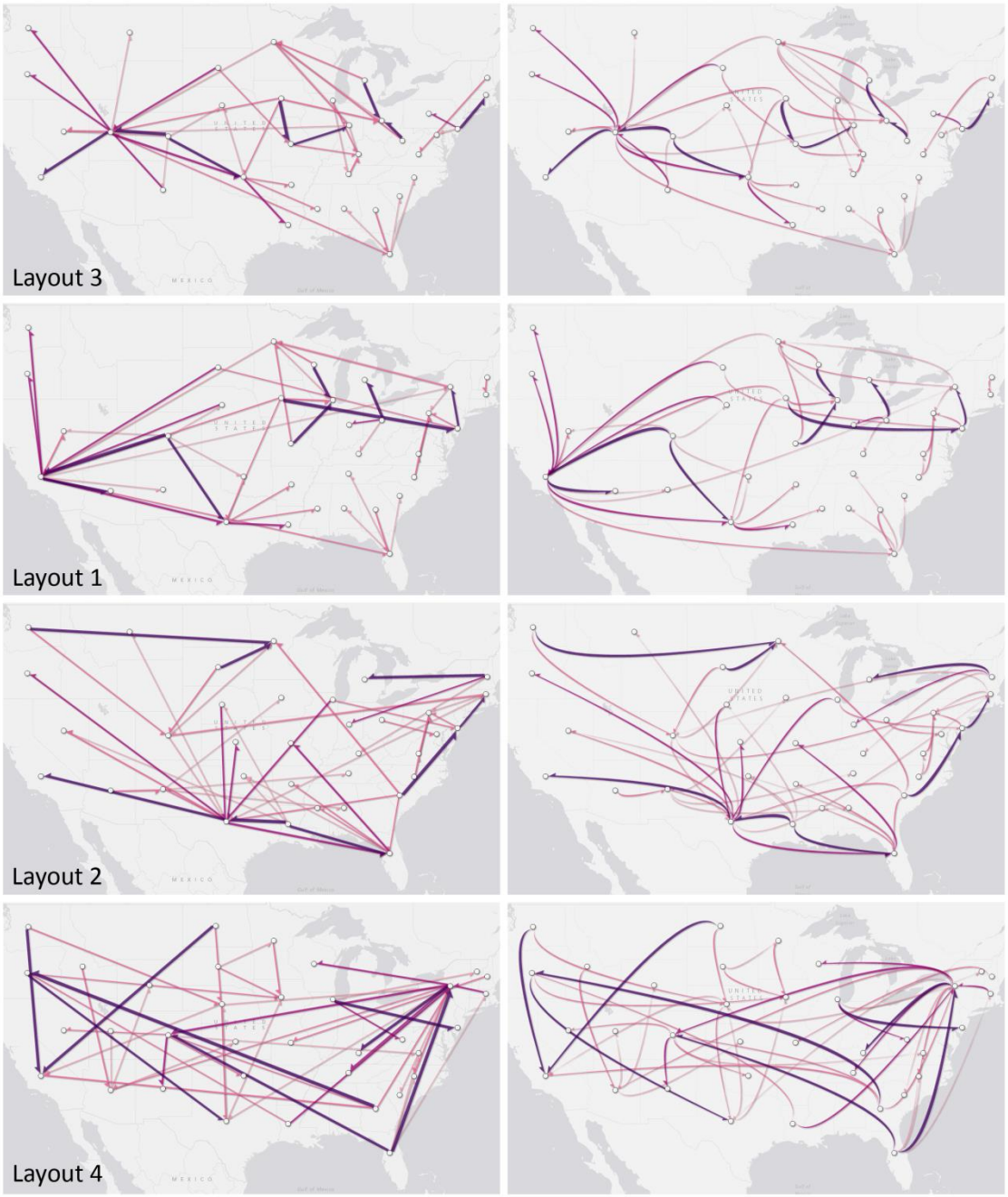



**Figure 4.1:** Flow map layouts: Straight design (left), curved design (right). Each layout displays the identical network of commodity flows. The total flow length and the number of edge crossings increase from top to bottom, whereas mean crossing angles vary between the layouts. Layout 1 is the original layout of the commodity flow dataset, whereas Layout 2, 3 and 4 were produced by swapping locations of nodes in the original network.

**Table 4.1**: Layout Characteristics

| | **Edge Crossings** (count) | | **Total Flow Length** (in pixels) | | **Mean Crossing Angles** ($0 < \alpha < 90$) | |
|---|---|---|---|---|---|---|
| | Curved | Straight | Curved | Straight | Curved | Straight |
| **Layout 1** | 52 | 32 | 12437 | 12364 | 37.69 | 45.47 |
| **Layout 2** | 120 | 93 | 15343 | 15199 | 48.11 | 42.67 |
| **Layout 3** | 43 | 32 | 10060 | 10046 | 47.2 | 43.47 |
| **Layout 4** | 149 | 136 | 18562 | 18339 | 50.25 | 44.63 |

### 4.4.3 PARTICIPANTS

To obtain a general overview of flow map comprehension, we intended to recruit participants with diverse backgrounds and expertise. Increasing number of studies has proven the usefulness of online crowdsourcing services for conducting usability experiments (Kinkeldey et al., 2013; Mason & Suri, 2012; Paolacci, Chandler, & Ipeirotis, 2010). Following this trend, we used Amazon Mechanical Turk (AMT) crowdsourcing service (https://www.mturk.com) to recruit participants. We paid each participant 50 cents to conduct the test that took 15 minutes on average. To ensure motivation, we required the participants to have greater than 5000 approved hits with a 98% hit approval rate.

202 volunteers (136 male, 66 female) participated in the test. The ages of the participants were between 19 and 69, and the average age was 33. The majority of the participants declared to have a college (41%) and graduate degree (39%), whereas there were participants with a high school degree (19%) and a degree with less than high school (1%). Most participants were from the United States (67%) and India (29%). We did not observe a significant difference between the performance and country. The majority of the participants stated that they had never seen a flow map or they did not

know what a flow map was (67%), whereas 33% said they had seen a flow map. 58 % of the participants stated (i.e., agree and strongly agree) that they understand what a flow map represents. The majority (92%) of the participants use computers more than 3 hours a day. Most of the participants (93%) use maps regularly (e.g., Google Maps) and feel comfortable about using online mapping services

#### 4.4.4 PROCEDURE

We employed a user-centered design and iteratively improved both the flow map design and experiment. We conducted three experiments with a variety of layouts, tasks and design alternatives for flow mapping. However, in this paper, we only report the results of the final experiment due to the limited space. We measured correctness of response, response time, and mental effort; collected user feedback and recorded screen resolution and interface events. The test is made available to the public using the following link: http://tinyurl.com/ksxsqvl. Participants can take the test using a personal computer at any location and time. Participants are first prompted with an instruction window that briefly describes interactive flow mapping and the online system that would be used by the participant. To reduce the cognitive load induced by instructional materials, we kept the instructions as short as possible. The test included 20 questions in which only 8 questions require participants to complete tasks using flow maps while the rest are background and follow-up questions to gauge user reactions to the test and given flow maps.

To avoid learning effects, we randomized the order of questions first by task then by layout and made sure each permutation is taken equal times before a repeating permutation is assigned to a new participant. We specifically kept the questions of the same task together in order to alleviate the confusion that may result from switching back

and forth between the different task types. The same ordering procedure was applied to both curved and straight designs separately. There was no time limit to answer any of the questions. The whole session took about 15 minutes on average.

The test interface detects screen heights narrower than 900 pixels and adjusts the zoom level of the map component to depict the whole layout on the screen which results in a smaller depiction of flow maps. Overall, 64 % of participants had smaller screen height (<900), whereas 36% had larger (>=900). Table 4.2 illustrates the number of participants that were assigned curved and straight flow maps and their screen resolutions. Participants with a smaller resolution were approximately equal for curved (64) and straight (65) designs. To account for the variation in performance and perceived mental effort, we included screen resolution as a factor in analyzing the test data.

**Table 4.2:** Number of participants by screen resolution and design type

| | Curved | Straight |
|---|---|---|
| Large | 37 | 36 |
| Small | 64 | 65 |

## 4.5 RESULTS

We organized the analysis of the test data according to the usability metrics measured in the experiment: correctness, response time, perceived mental effort and user reactions. To account for the performance variation due to screen size, we considered screen resolution (height) as a factor in our statistical analysis of the results. We conducted a detailed analysis of the test data using a mixed design analysis of variance (ANOVA) with two between-subjects factors: design (2 levels) and screen resolution (2 levels); and two

within-subjects factors: task (2 levels) and layout (4 levels). Additionally, we provide two online applications to view (1) each participant's responses: http://tinyurl.com/ppy86z7 (Koylu, 2014c); and (2) cumulative response patterns: http://tinyurl.com/p7j5nbx (Koylu, 2014b).

### 4.5.1 CORRECTNESS

To answer each flow map question, participants selected top three nodes with either the highest total inflow (import) or outflow (export). To calculate the correctness of an answer, we divide the cumulative inflow/outflow volume of the participant's selections (e.g., add up the selected three nodes' total volume of inflow or outflow based on task) by the cumulative inflow or outflow volume of the actual top three nodes (e.g., add up the actual top three nodes' total volume of inflow or outflow based on task). As a result, values of correctness range from 0 to 1.

Correctness data are illustrated using asymmetric beanplots in Figure 4.2 which allow comparison of distributions by design, layout and task types. The black and grey areas depict the density trace of each distribution whereas the lines within the density areas serve as a histogram to illustrate the frequency of observations on a particular score. Means of each distribution are displayed using red lines, and the overall mean for each layout is displayed using a dashed line. We ordered the beanplots first by layout, then by task to be able to compare the main effect of design, and how it changes depending on different choices of task and layout. The presence of a bimodal distribution indicates a major split between the participants' answers whereas a single peak shows similar answers; and a uniform distribution shows a diverse range of answers. Curved flow maps consistently resulted in negatively skewed distributions with a single peak which

highlights higher accuracy on import tasks. On the other hand, straight design produced relatively more uniform distributions on import tasks, which indicates a diverse range of performance scores. Both curved and straight designs produced bimodal distributions for export task except Layout 2 with curved design which produced a single peak.

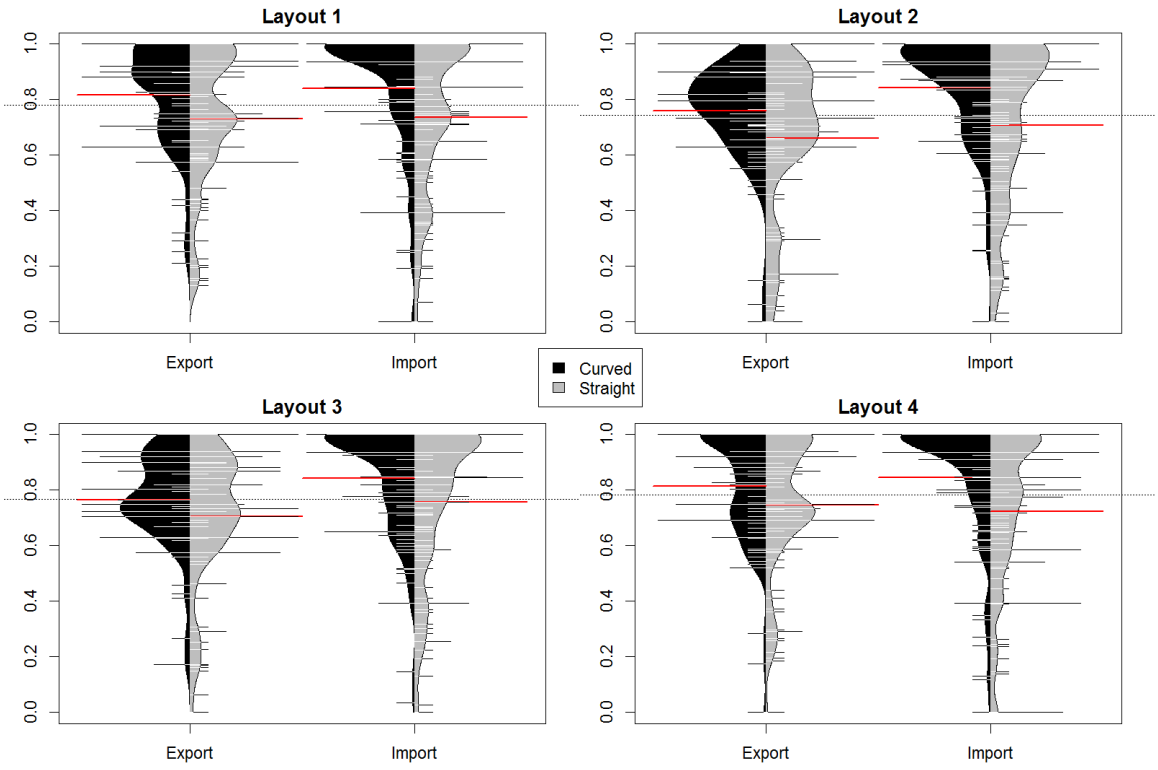



**Figure 4.2:** Asymmetric beanplots for correctness by flow map design. Red lines show the mean of each distribution whereas dashed lines illustrate the mean for each layout. The presence of a bimodal distribution indicates a major split between the participants' answers whereas a single peak shows similar answers; and a uniform distribution shows a diverse range of answers.

Further analysis of the correctness of response measure was conducted using a mixed design ANOVA with two between-subjects factors: design (2 levels) and screen resolution (2 levels); and two within-subjects factors: task (2 levels) and layout (4 levels).

As the residuals violate the assumptions of normality and homogeneity, we applied arcsine transformation to normalize the proportional data of correctness and stabilize the variance in residuals prior to the statistical analysis. For correctness, the statistical results confirmed that all four main effects, design ($p < 0.001$), screen resolution ($p < 0.01$), task ($p < 0.001$) and layout ($p < 0.001$) were statistically significant. Also, the results indicated three significant interactions: design and resolution ($p < 0.05$); task and layout ($p < 0.001$); and design and task ($p < 0.05$).

The interaction between screen resolution and design (Figure 4.3) showed that participants' performances were significantly lower when they were assigned straight flow maps on a small screen resolution (screen height < 900). On the other hand, performance of curved flow maps did not have a significant difference between small and large screen resolutions. We could attribute consistently high performance of curved design regardless of the screen resolution to the use of two visual clues (line curvature and arrows) for direction and more clear separation of flow lines between every pair of nodes. The interaction between layout and task (Figure 4.4) suggests that participants performed well with Layout 1 and Layout 4 for both tasks, whereas Layout 2 and Layout 3 led to decreasing accuracy on export task. We further analyze the potential reasoning behind this interaction in section 4.5.4. Finally, the interaction between design and task indicates a substantial increase in accuracy when participants were assigned an import task (as opposed to an export task) on a curved design.

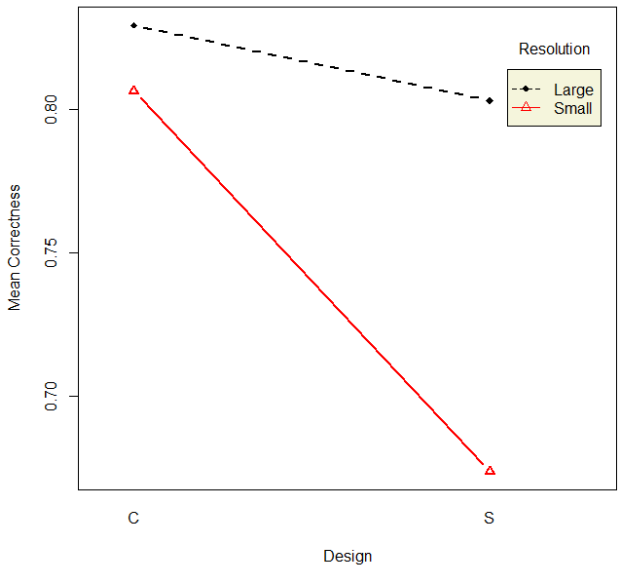



**Figure 4.3:** Significant interaction between design and screen resolution showed a clear association between small screen resolution and lower accuracy when using straight flow maps.

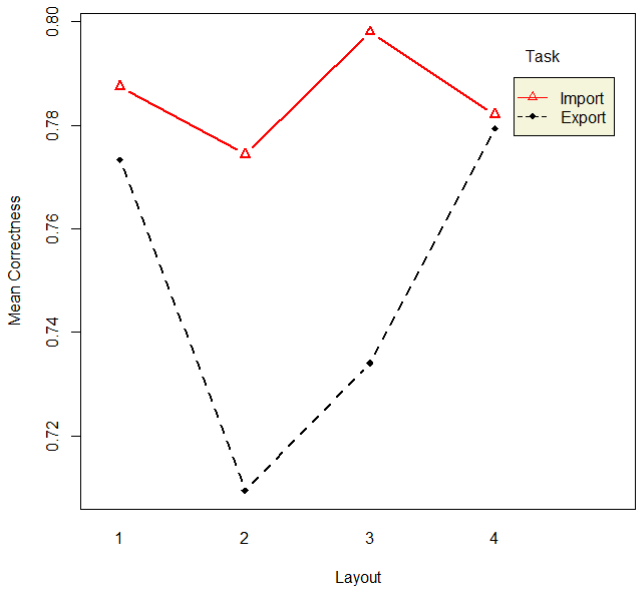



**Figure 4.4**: Significant interaction between layout and task for correctness. Both import and export task resulted in similar accuracy when the users performed tasks on Layout 1 and Layout 4, whereas accuracy was lower when participants were given an export task on Layout 2 and Layout 3.

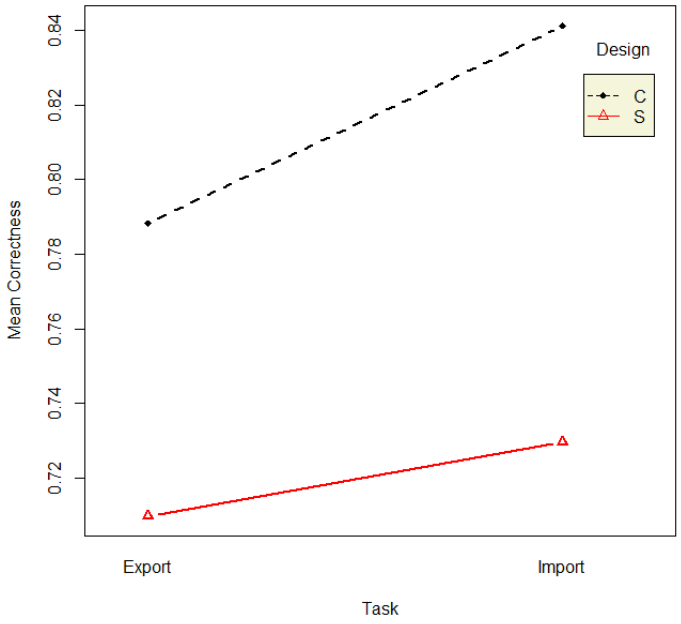



**Figure 4.5:** Significant interaction between task and design for correctness. Although the difference in the accuracy of curved and straight design is an outcome of the screen resolution, accuracy substantially increased when participants were assigned an import task (as opposed to an export task) on a curved design.

### 4.5.2 Response Time

Participants' response times are illustrated using asymmetric beanplots ordered by design, layout and task (Figure 4.6). Unlike the correctness of response, response times for curved and straight flow maps were similar, whereas participants performed export tasks substantially faster than import tasks. To meet the assumptions on the normality and homogeneity of residuals, we applied logarithmic transformation prior to conducting the ANOVA. The statistical results of the mixed model ANOVA confirmed significant main effects of task ($p < 0.001$) and layout ($p < 0.001$), whereas design ($p = 0.828$) and resolution ($p = 0.459$) were not found to be significant. Layout and task was found to be

the only significant ($p < 0.05$) interaction effect (Figure 4.7). Import tasks required more time than export tasks, and the average time spent on an import task varied depending on the type of layout.

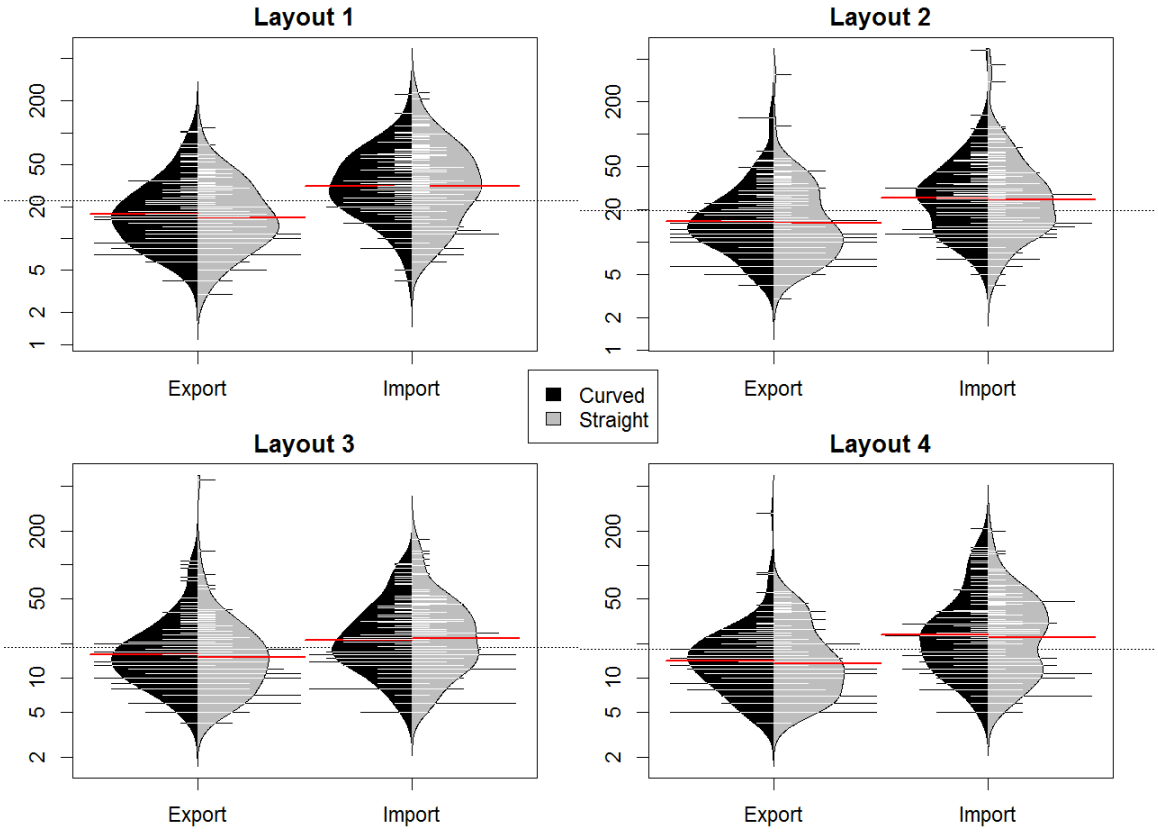



**Figure 4.6:** Asymmetric beanplots for response time (in seconds) by flow map design. Red lines show the mean response time for each distribution whereas dashed lines illustrate the mean response time for each layout. Unlike the correctness, response times for curved and straight flow maps are similar, whereas an export task is performed significantly faster than an import task.

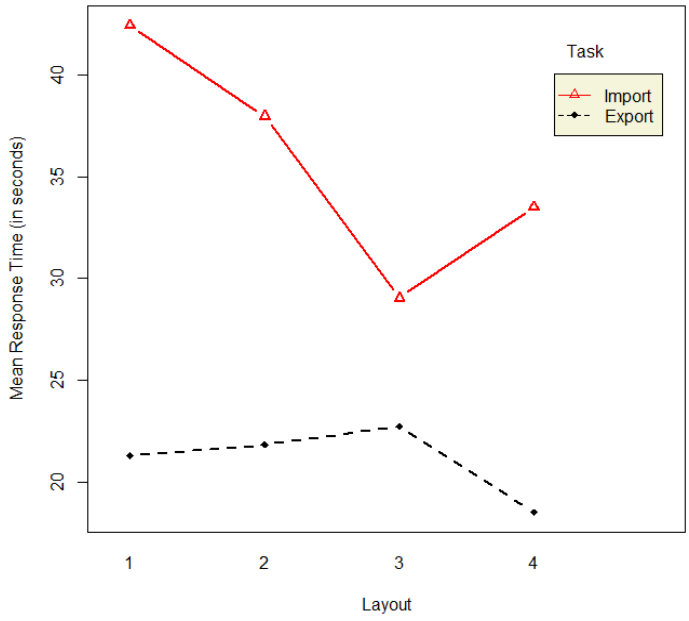



**Figure 4.7**: Significant interaction between layout and task for response time (in seconds). Import tasks required more time than export tasks, and the average time spent on an import task varied depending on the type of layout.

### 4.5.3 MENTAL EFFORT

Participants reported their perceived mental effort after each flow map question using a 9-point Likert scale which we treated as a continuous variable. The results of the ANOVA indicated that the differences among large and small screen resolution was highly significant ($p < 0.001$) whereas the factors of design, layout and task were not found to have a significant effect on perceived mental effort. None of the interactions between the factors were significant.

To gain more insight into perceived mental effort, we illustrate the interactions between task and design, and layout and task. Task and design interaction in Figure 4.8 shows a substantial decrease in perceived mental effort when participants' completed an

import task on a curved flow map. Layout and task interaction in Figure 4.9 indicates similar mental effort for import tasks and export tasks on Layout 1, and 2, whereas Layout 3 and 4 were ranked as less challenging on import tasks. Additionally, we expected Layout 4 to have a relatively higher mental effort due to its longer flows with relatively more edge crossings; however, participants rated Layout 4 as the least challenging layout when they were given an import task.

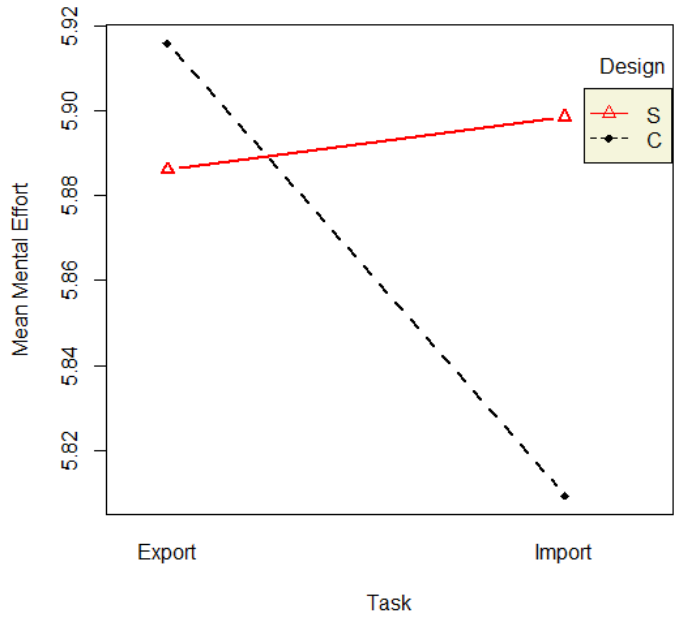



**Figure 4.8**: Task and design interaction for perceived mental effort. Participants found import tasks less challenging when they were given curved flow maps.

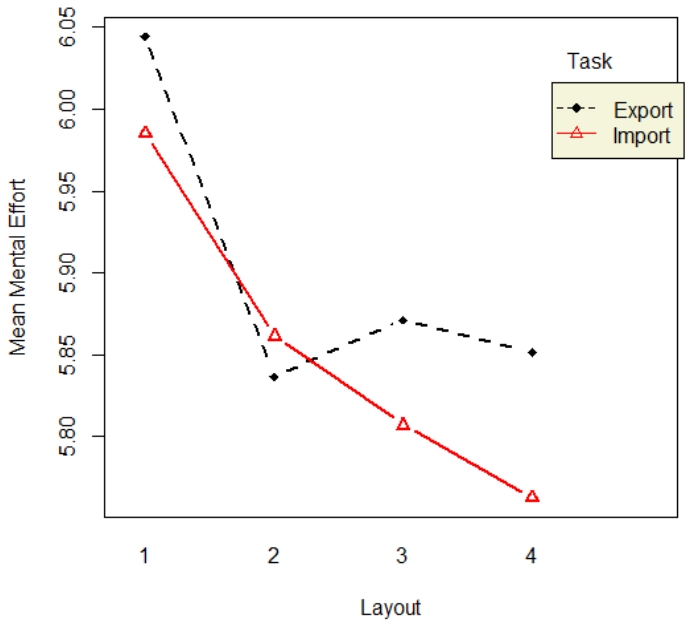



**Figure 4.9:** Design and layout interaction for perceived mental effort. Although we expected Layout 4 to have a higher mental effort due to its longer flows with more edge crossings; participants rated Layout 4 as the least challenging layout when they were assigned an import task.

### 4.5.4 Task and Layout Interaction

To capture cumulative response patterns and explain the potential causes in the variation of performance and perceived mental effort, we developed an interactive application: http://tinyurl.com/p7j5nbx (Koylu, 2014b) that displays the frequency of each node's appearance in participant's top three choices given a design, task and layout combination. Figure 4.10 illustrates the average rate of appearances for the top four nodes (the true four highest importers and exporters) among participants' answers (top three selections) by task and layout. On average, approximately 80 % of the participants picked the first ranked importer and exporter among their top three selections. For the second rank, the rate of appearance (60 %) for import task was consistent across all layouts whereas the rate for export task decreased greatly on Layout 2 and 3. Similarly, the rate for the third

rank was substantially lower for export task on Layout 3. Only smaller differences were observed for the fourth rank.

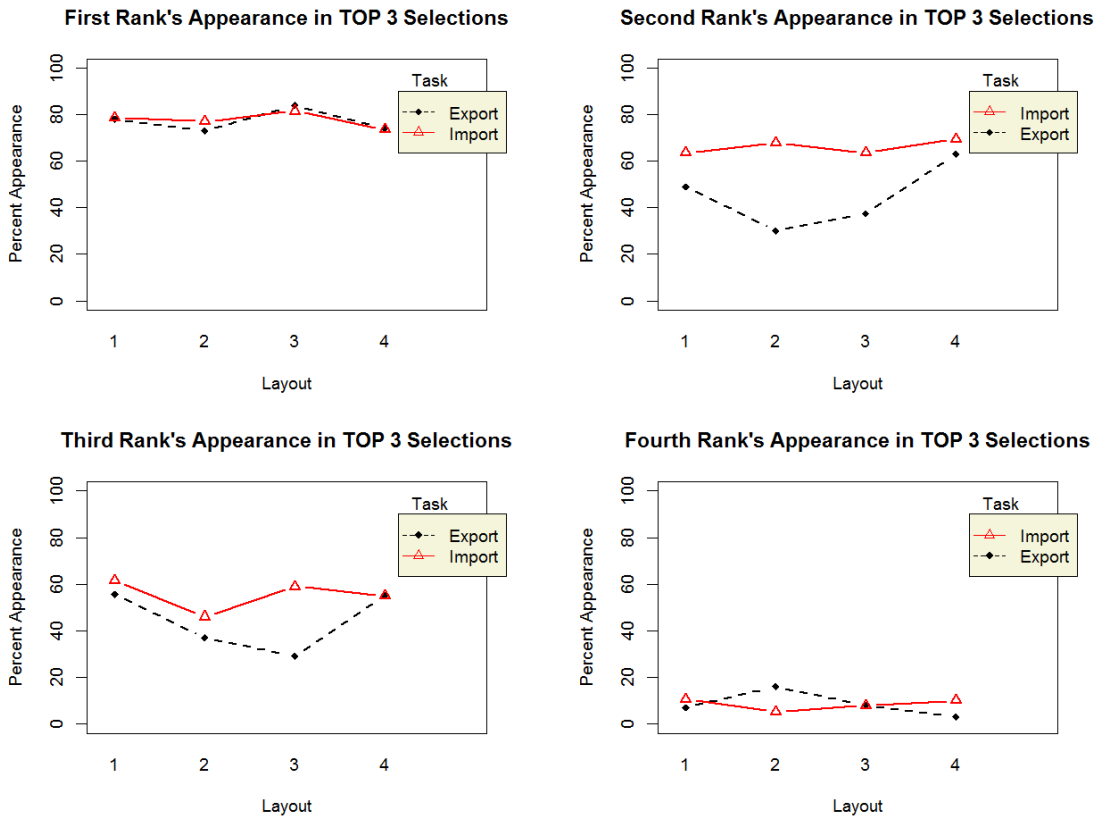



**Figure 4.10**: Average rate of appearances for the actual top 4 nodes in participants' top 3 selections. Average rate of appearances for the second and third ranks suggest substantial performance drop for export tasks when layout 3 was used.

Figure 4.11 illustrates a comparison of the frequency of selections on export tasks for Layout 3 and 4. In Figure 4.11, the actual top three nodes with the highest export are colored yellow with a blue outline in which darker tons of blue indicate higher rank. In Layout 3, the actual top three nodes were selected 151 times (92+32+27) whereas the same nodes were selected 203 times (79+65+59) in Layout 4. The second and third rank

nodes received substantially less number of hits in Layout 3 than in Layout 4. Although there were substantially more edge crossings in Layout 4 (149) than Layout 3 (43), on average participants received higher correctness scores on Layout 4 (see Figure 4.4). This finding agrees with the hypothesis that edge crossings and the graph layout do not have a significant influence on the perception of node importance (Huang, Hong, & Eades, 2006).

As large map symbols attract more attention than small ones (Alan M. MacEachren, 1995), in a flow map, we expected to observe salience bias (Mitchell, Ware, & Kelley, 2009) towards flows of higher magnitude and/or longer length. We hypothesize that decreasing accuracy on Layout 3 was caused by participants' increased tendency for selecting alternative nodes (incorrect choices) that were visually salient as a result of longer length and clear depiction of their flows. On the other hand, in Layout 4, second and third rank nodes were visually more salient than their alternatives as a result of the extended length of their outflows (especially the highest volume class with dark purple color). Thus, participants selected the correct nodes as they had visually dominant (darker color, thicker and longer) flows.

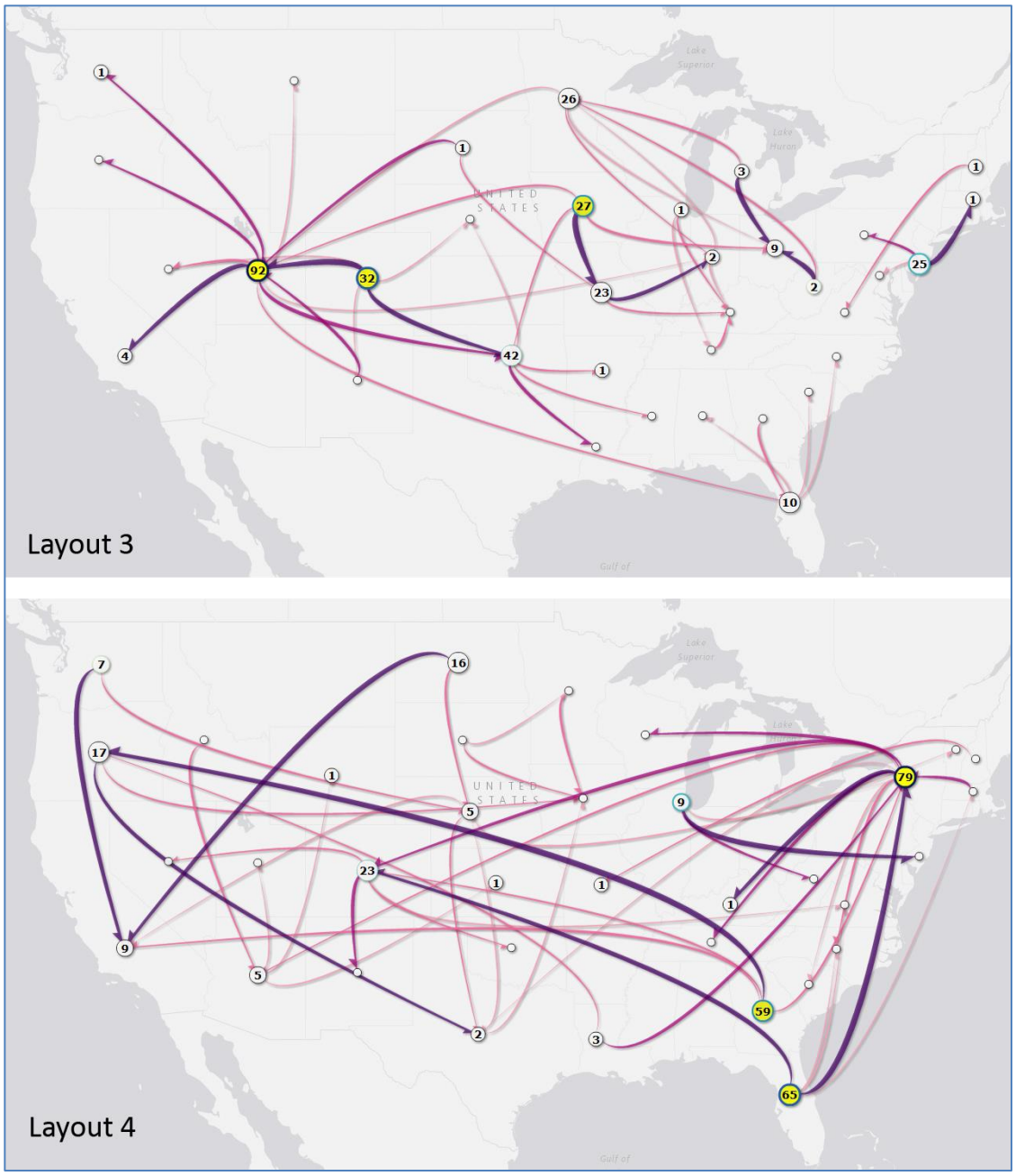



**Figure 4.11**: Frequency of hits on export task: Top-Layout 3, Bottom-Layout 4. Although the network is identical in both layouts, the second and third rank nodes received substantially less number of hits (compare 62 and 57 to 30 and 26) when Layout 3 was used. We hypothesize that decreasing accuracy on Layout 3 was caused by participants' increased tendency for selecting alternative nodes (incorrect choices) that were visually salient as a result of longer length and clear depiction of their flows.

### 4.5.5 User Reactions

We did not observe significant difference among the performance of participants and their demographics such as age, gender and education level. Approximately 30 % of participants provided additional feedback on the flow map design and issues related to the interactive features of the flow mapping and testing environment. Feedback left by each participant can be viewed at the application link that displays test entries: http://tinyurl.com/ppy86z7 (Koylu, 2014c). A statistical analysis on whether the between-subjects factors of design and screen resolution had an influence on leaving a complaint or feedback showed no significant effect. The majority of the complaints and constructive feedback was about the difficulty with determining the direction of flows due to the overlapping flow lines and arrows. A number of participants suggested larger arrows; and more variation and contrasting hue for color scheme. Some participants suggested adjusting (increase/decrease) the thickness of the flow lines when using the zoom function.

## 4.6 Discussion and Conclusion

The analysis of correctness, response time and perceived mental effort revealed interesting patterns. Curved flow maps facilitated consistently high accuracy regardless of screen resolution whereas correctness scores significantly dropped when straight flow maps were used with a small screen resolution (screen height < 900). Screen resolution was found to be the only significant effect that influenced perceived mental effort, and participants with smaller screens rated tasks as more challenging. In contrast to its effect on correctness, flow line style did not have a significant effect on either response time or perceived mental effort. The effectiveness and potentially easier perception of curved

design could be attributed to a number of factors such as using two visual clues (curvature and arrow) while the straight line uses only arrow for depicting direction; less line occlusion and edge tunneling effect (Dunne & Shneiderman, 2009) and wider angles that make links and arrows more visible.

The statistical analysis of the test data showed that performance (correctness and response time) of participants varied significantly depending on task and layout combination. Participants were less accurate when they performed an export task on layouts (Layout 2 and 3) that alternative nodes (incorrect choices) were visually salient as a result of the length and clear depiction of their flows. On the other hand, participants were significantly more accurate on import tasks; their performance was consistent across all layouts; and they rated the tasks with substantially less perceived mental effort; whereas they took significantly more time to complete. Also, the average time spent on an import task significantly varied depending on the layout. Further analysis is needed to understand the cognitive processes that result in performance variation such as consistent and higher accuracy, and higher response time when a participant is assigned an import task. We believe that an eye tracking experiment could help gain insight into the cognitive processes and sequence of flow map reading when completing import and export tasks.

We observed that edge crossings did not matter for the difference in performance; however, further studies are needed to empirically test both the effect of flow salience (longer, darker color and thicker flows) and edge crossings. We would like to acknowledge that our findings are limited by the experimental parameters, and some of the conclusions may not apply to the general comparison of flow map designs.

Since the participants were from Amazon Mechanical Turk with a certain level of computer skills the findings of the study are not necessarily representative of a broader population with diverse backgrounds. Because participants were given the freedom to conduct the test without any time limit and the test is not administered by an experimental facilitator, confounding factors related to the test environment and multi-tasking is expected. To account for this factor, we administered a pilot test on 36 undergraduate students in a computer lab at the University of South Carolina. We did not observe a significant difference between the performance of administered test takers and AMT users. We excluded the results of the administered test not to introduce bias that would be produced by the lab environment.

In our evaluation of flow map design, we used two popular design alternatives: curved and straight flow lines with partial arrows; four alternative layouts of an identical network; and two tasks which we derived from a simple typology (i.e., high total inflow, outflow) that emphasize the importance of locations. Given the large number of possible flow map reading tasks, it is challenging to select tasks to evaluate the effectiveness and efficiency of flow maps. Based on the idea that comprehension of location prominence involves essential flow map reading tasks such as identifying volume and direction of flows, distinguish in and out connections of nodes, and compare cumulative inflow or outflow volumes of multiple nodes; we used perception of location (node) prominence as a way to capture the general perception of flow maps. For a more extensive evaluation of flow map reading, there is a need to construct a comprehensive typology patterns and visual tasks. For future work, we plan to include more complex tasks such as identifying spatial regions, network structure (clustering), and flow patterns.

The findings of this study have important implications for iterative design, interaction strategies and further user experiments on flow mapping. Future work is needed to improve both flow mapping and experiment. We plan to integrate interactions such as highlighting, isolation and animation to help reduce the cognitive load associated with effects such as edge tunneling, edge crossings and crossing angles. Alternative designs for line style with varying curvature, arrow size, and color schemes could be implemented. An insight-based approach would be useful to ensure that users are able to generate insights into flow data and visualization. Also, we believe that an eye movement analysis would be greatly helpful to study cognitive processes and behaviors linked to flow map reading.

# CHAPTER 5

## CONCLUSION

Our view of the world has drastically changed with an increasing focus on the flows of both physical and intangible phenomenon such as people, commodities, flights, money, information, ideas and innovation. This dissertation uses the concept of geo-social networks, which connect places by the flows of physical and intangible phenomenon, to describe and study the complex system of flows from an integrated perspective of GIScience and network science. This dissertation made the following contributions to the theory and methodologies that aim at understanding complex geo-social data by integrating methods of computation, visualization and usability evaluation.

Chapter 2 introduced a new network-based smoothing approach to calculating and mapping locational (graph) measures in spatial interaction networks. The new approach introduces a generic framework that can be used to smooth various graph measures and is the first attempt that truly considers the flow structure in implementing spatial kernel smoothing in a spatially embedded network. The approach helps overcome spurious data variations and unstable graph measures that exist as a result of the size-difference and small area problems in spatial interaction datasets. The demonstration of the approach in smoothing net migration rate and entropy measures in county-to-county migration data in the U.S. helped discover natural regions of attraction (or depletion) and other structural

characteristics that the original (unsmoothed) measures failed to reveal. Furthermore, with the new approach, one can also smooth spatial interactions within sub-populations (e.g., multivariate components such as different age groups), which are often sparse and impossible to derive meaningful measures if not properly smoothed. Moreover, the results of the case study in migration dataset also highlighted the effectiveness of the approach in discovering patterns at multiple scales (e.g., national, regional and local).

Chapter 3 introduced a novel approach to discover spatial and structural patterns among individual locations of a dynamic geo-social network embedded in space and time. A measure of connectedness was introduced to summarize the dynamic relationships in a point-based geo-social network by taking into account the distance (how far individuals live apart), time (the duration of individuals' coexistence within a neighborhood), and the relationship (kinship or kin proximity) between each pair of individuals. The new approach facilitates the discovery of hot spots (hubs) where potential for spatial interaction between individuals is relatively higher across space and time. The approach was demonstrated using a family tree dataset and the results highlighted the formation of family hubs that change across space and time as a result of demographic processes such as migration and population change.

Flow mapping is commonly used to visualize geo-social networks. Flow maps heavily rely on viewers' comprehension of flow patterns and the spatial context in a geo-social network; yet very little is known about how users interpret and use flow maps. Chapter 4 introduced a user evaluation to obtain knowledge on how map readers perceive information presented with flow maps, and how design factors such as flow line style (curved or straight) and layout characteristics may affect flow map perception and users'

performance in addressing different tasks for pattern exploration. The analysis of the test data showed that performance (correctness and response time) and perceived mental effort of participants varied significantly depending on the factors of design, task, layout and screen resolution. Both the analysis of usability metrics such as correctness, response time and mental effort and user feedback provided important implications for iterative design, interaction strategies and further user experiments on flow mapping.

In the remainder of this chapter, broader impacts and future research directions are discussed.

## 5.1 Broader Impacts

The wide use of social networking and media applications has led to a revolution that brings together large numbers of citizen sensors who engage in the creation of voluminous geographic data. A variety of applications have been developed to analyze and understand the movement of phenomenon such as information, diseases, innovations, protests and activities across space and time by examining volunteered geographic information (VGI) collected from cyber-space and social media. Examples of those applications can be found in various fields such as public health (Ghosh & Guha, 2013), finance (Rao & Srivastava, 2012), linguistics (Graham & Zook, 2013), disaster management (Heinzelman & Waters, 2010) and urban geography (Hollenstein & Purves, 2013). A broader impact of this dissertation is to bring cross-disciplinary studies of geo-social data using a web-based platform and share new theories, computational and visual tools, and research findings. Currently, the findings, and methodologies developed in this dissertation are publically available at www.geo-social.org. Online and freely available tools will be included in this platform to analyze large, complex and diverse geo-social

data and promote education and training of not only geographers and spatial scientists but also social scientists. Sharing of research findings from diverse subject areas such as migration, spatial analysis of social networks, disease spread and analysis of social media data will help foster multiple perspective thinking and build a community that is particularly interested in combining theories and methodologies of diverse fields.

## 5.2 Future Directions

This dissertation introduced a network-based smoothing approach that considers spatial and network structure in discovering location characteristics in spatial interaction data (Chapter 2) and space-time visualization of connectedness that takes into account spatial, temporal and relational dimensions of a location based social network (Chapter 3). Both of these approach are similar in that they integrate spatial and social (network) factors together. One future direction should consider the development of visual analytic approaches that consider the dimensions of geography, network and time simultaneously.

The ultimate goal for developing computational and visualization methods is to gain insight into the underlying processes that form complex patterns of geo-social networks, and help decision-making, hypothesis generation and testing. However, little is known about cognitive aspects and usability issues related to visualization of geo-social networks. Chapter 4 introduces an evaluation study to gain insight into how map readers perceive information presented with flow maps, and how design factors such as flow line style (curved or straight) and layout characteristics; and different tasks for pattern exploration influence flow map perception. Future work is needed to improve both flow mapping and the experiment. Interaction techniques such as highlighting, isolation and animation could be integrated into flow map design to help reduce the cognitive load

associated with effects such as edge tunneling, edge crossings and crossing angles. Alternative designs for line style with varying curvature, arrow size, and color schemes could be implemented. An insight-based approach would be useful to evaluate how users generate insights into flow data and visualization. Also, an eye movement analysis would be greatly helpful to study cognitive processes and behaviors linked to flow map reading.

The three manuscripts presented in this dissertation highlight the need for a theoretical foundation with a comprehensive pattern typology to guide the design, development and evaluation of computational and visualization tools for understanding the complex patterns of geo-social networks. Examples of comprehensive pattern typologies can be found in trajectory analysis (N. Andrienko & Andrienko, 2007; Dodge et al., 2008), information visualization (Munzner, 2009), temporal visualization (X. Li, 2010) and graph visualization (Brehmer & Munzner, 2013; B. Lee et al., 2006; Saket et al., 2014). A pattern typology is useful for two reasons. First, to guide designers what type of patterns need to be perceived or detected with the use of the tool so that the users can be instructed to detect the types of patterns the tool is oriented to. Second, the theory provides a framework to evaluate the utility and usability of the proposed method.

A data model (characterization) is essential for building a typology of patterns. Peuquet (1994) introduced a triad model that provides a conceptual basis for characterizing spatiotemporal data using the perspectives (questions) of location (where), time (when) and attribute (what). Mennis et al. (2000) further extended the triad model into a pyramid model to include "object" as a knowledge dimension on a higher level. The knowledge component in the pyramid model concerns the representation of derived objects, their classification and inter-relationships. Li (2010) adopted the pyramid model

and used the knowledge component to describe the existence (appearance and disappearance) of objects through time. To describe social network data embedded in geographic space and time, Ma (2012) further added a new dimension of social network elements (i.e., node, link, sub-network) using the object component. While Ma's (2012) framework is useful for formalizing social network tasks, however, the framework does not adequately consider the interactions between the components of network, geography and time. Future work is needed to build a data model that takes into account geography, time and network dimensions as well as the interactions between those dimensions.

Growing use of social media and networking applications and increasing volume of the collaborative user generated content (UGC) through crowd sourcing provide a great potential to obtain and analyze large and complex geo-social data and address the problems concerning the environment and society, and the interaction among them, such as disasters, public health, security, migration, and transportation. Because the user generated content is unstructured and often imprecise, it is crucial to understand the validity, accuracy, representativeness and uncertainty of the data in order to imply the causal and behavioral characteristics of the users (L. Li, Goodchild, & Xu, 2013). Aside from the concerns for the quality, credibility and representativeness of the data, there are major challenges regarding the handling, analyzing and communicating the information mined from such large amount of user generated data:

- Limits on the size of the datasets require integrating distributed storage and computing technologies for data handling and processing.
- Due to the variety of target domains (e.g., public health, emergency response) for analysis, and mediums (e.g., Twitter, Flickr) for collecting the data it is

challenging to develop automated or semi-automated processes to clean and interpret the content.

- Although the content generated by users' posts spread through a network of users, very little research has looked at the relational (network) dimension of the data. This is mainly because it is difficult to analyze massive quantities of data which form complex networks with rich semantics.
- Because of the size and dynamic nature of the data, it is challenging to develop computational and visual methods for in depth analysis of spatial, temporal and relational (network) aspects of the user generated content.

## REFERENCES


Abello, J., van Ham, F., & Krishnan, N. (2006). ASK-GraphView: A large scale graph visualization system. *Ieee Transactions on Visualization and Computer Graphics, 12*(5), 669-676.

Abramson, I. S. (1982). On Bandwidth Variation in Kernel Estimates-A Square Root Law. *The Annals of Statistics, 10*(4), 1217-1223.

Adams, J. W., & Kasakoff, A. B. (1984). Migration and the family in colonial New England: The view from genealogies. *Journal of Family History, 9*(1), 24-43.

Agarwal, P., & Skupin, A. (2008). *Self-organising maps: Applications in geographic information science*: Wiley.

Aigner, W., Miksch, S., Schumann, H., & Tominski, C. (2011). *Visualization of time-oriented data*: Springer.

Alper, B., Bach, B., Henry Riche, N., Isenberg, T., & Fekete, J.-D. (2013). *Weighted graph comparison techniques for brain connectivity analysis.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., . . . Tominski, C. (2010). Space, time and visual analytics. *International Journal of Geographical Information Science, 24*(10), 1577-1600.

Andrienko, G., Andrienko, N., & Wrobel, S. (2007). Visual Analytics Tools for Analysis of Movement Data. *SIGKDD Explorations, 9*(2), 38-46.

Andrienko, N., & Andrienko, G. (2007). Designing visual analytics methods for massive collections of movement data. *Cartographica, 42*.

Andrienko, N., Andrienko, G., Voss, H., Bernardo, F., Hipolito, J., & Kretchmer, U. (2002). Testing the Usability of Interactive Maps in CommonGIS. *Cartography and Geographic Information Science, 29*, 325-342. doi: 10.1559/152304002782008369

Andris, C. (2011). *Metrics and methods for social distance.* Massachusetts Institute of Technology.

Anselin, L. (1995). Local indicators of spatial association--LISA. *Geographical Analysis, 27*(2), 93-115.

Atkinson, R. D. (1998). Technological change and cities. *Cityscape*, 129-170.

Aufaure-Portier, M.-A. (1995). Definition of Visual Language for GIS. In T. L. Nyerges, et al. (Ed.), *Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems* (pp. 163-178): Kluwer Academic Publishers.

Backstrom, L., Sun, E., & Marlow, C. (2010). *Find me if you can: improving geographical prediction with social and spatial proximity.* Paper presented at the Proceedings of the 19th international conference on World wide web.

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science, 286*, 509-512.

Battista, G. D., Eades, P., Tamassia, R., & Tollis, I. G. (1999). *Graph drawing: algorithms for the visualization of graphs*: Prentice-Hall.
Bender-deMoll, S., & McFarland, D. A. (2006). The art and science of dynamic network visualization. *Journal of Social Structure, 7*(2), 1-38.
Bennett, C., Ryall, J., Spalteholz, L., & Gooch, A. (2007). *The Aesthetics of Graph Visualization.* Paper presented at the Computational Aesthetics.
Bogart, D. (2005). Turnpike trusts and the transportation revolution in 18th century England. *Explorations in Economic History, 42*(4), 479-508.
Bonacich, P. (1987). Power and Centrality - A Family of Measures. *American Journal of Sociology, 92*(5), 1170-1182.
Borgatti, S. P. (2005). Centrality and network flow. *Social Networks, 27*(1), 55-71. doi: 10.1016/j.socnet.2004.11.008
Borruso, G., & Schoier, G. (2004). Density Analysis on Large Geographical Databases. Search for an Index of Centrality of Services at Urban Scale. In A. Laganá, M. Gavrilova, V. Kumar, Y. Mun, C. Tan, & O. Gervasi (Eds.), *Computational Science and Its Applications – ICCSA 2004* (Vol. 3044, pp. 1009-1015): Springer Berlin / Heidelberg.
Bors, A., & Nasios, N. (2009). Bayesian Estimation of Kernel Bandwidth for Nonparametric Modelling. In C. Alippi, M. Polycarpou, C. Panayiotou, & G. Ellinas (Eds.), *Artificial Neural Networks – ICANN 2009* (Vol. 5769, pp. 245-254): Springer Berlin / Heidelberg.
Boyandin, I., Bertini, E., & Lalanne, D. (2010). *Using flow maps to explore migrations over time.* Paper presented at the Geospatial Visual Analytics Workshop in conjunction with The 13th AGILE International Conference on Geographic Information Science.
Bras, H. (2011). Intensification of family relations? Changes in the choice of marriage witnesses in the Netherlands, 1830-1950.
Brehmer, M., & Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *Visualization and Computer Graphics, IEEE Transactions on, 19*(12), 2376-2385.
Buchin, K., Speckmann, B., & Verbeek, K. (2011). Flow map layout via spiral trees. *Visualization and Computer Graphics, IEEE Transactions on, 17*(12), 2536-2544.
Buja, A., Cook, D., & Swayne, D. F. (1996). Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics, 5*(1), 78-99.
Butts, C. T., Acton, R. M., Hipp, J. R., & Nagle, N. N. (2012). Geographical variability and network structure. *Social Networks, 34*(1), 82-100.
Cadwallader, M. T. (1992). *Migration and residential mobility : macro and micro approaches*. Madison, Wis.: University of Wisconsin Press.
Cairncross, F. (2001). *The death of distance: How the communications revolution is changing our lives*: Harvard Business Press.
Carlos, H., Shi, X., Sargent, J., Tanski, S., & Berke, E. (2010). Density estimation and adaptive bandwidths: A primer for public health practitioners. *International Journal of Health Geographics, 9*(1), 1-8. doi: 10.1186/1476-072x-9-39
Castells, M. (1996). The rise of the network society. Vol. 1 of The information age: Economy, society and culture. *Massachusetts and Oxford: Blackwell*.

Celika, H. M., & Guldmann, J.-M. (2007). Spatial interaction modeling of interregional commodity flows. *Socio-Economic Planning Sciences, 41*(2), 147-162.
Chaffee, W. (1909). The Chaffee Genealogy. *New York: Privately printed*.
Chang, R., Ziemkiewicz, C., Green, T. M., & Ribarsky, W. (2009). Defining insight for visual analytics. *Computer Graphics and Applications, IEEE, 29*(2), 14-17.
Chen, J., Shaw, S.-L., Yu, H., Lu, F., Chai, Y., & Jia, Q. (2011). Exploratory data analysis of activity diary data: a space–time GIS approach. *Journal of Transport Geography, 19*(3), 394-404.
Cho, E., Myers, S. A., & Leskovec, J. (2011). *Friendship and mobility: user movement in location-based social networks.* Paper presented at the Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.
Clark, G. L. (1982). Volatility in the geographical structure of short-run US interstate migration. *Environment and Planning A, 14*(2), 145-167.
Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*(6). doi: 10.1103/PhysRevE.70.066111
Cohen, J. D. (1997). Drawing graphs to convey proximity: an incremental arrangement method. *ACM Transactions on Computer-Human Interaction, 4*(3), 197-229.
Costa, L. D., Rodrigues, F. A., Travieso, G., & Boas, P. R. V. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics, 56*(1), 167-242. doi: 10.1080/00018730601170527
Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., & Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences, 107*(52), 22436-22441. doi: 10.1073/pnas.1006155107
D'Amico, M., & Ferrigno, G. (1990). Technique for the evaluation of derivatives from noisy biomechanical displacement data using a model-based bandwidth-selection procedure. *Medical and Biological Engineering and Computing, 28*(5), 407-415. doi: 10.1007/bf02441963
Danese, M., Lazzari, M., & Murgante, B. (2008). Kernel Density Estimation Methods for a Geostatistical Approach in Seismic Risk Analysis: The Case Study of Potenza Hilltop Town (Southern Italy). In O. Gervasi, B. Murgante, A. Laganà, D. Taniar, Y. Mun, & M. Gavrilova (Eds.), *Computational Science and Its Applications – ICCSA 2008* (Vol. 5072, pp. 415-429): Springer Berlin / Heidelberg.
Daraganova, G., Pattison, P., Koskinen, J., Mitchell, B., Bill, A., Watts, M., & Baum, S. (2012). Networks and geography: Modelling community network structures as the outcome of both spatial and network processes. *Social Networks, 34*(1), 6-17. doi: 10.1016/j.socnet.2010.12.001
Davies, C. (1995). Tasks and Task Descriptions for GIS. In T. L. Nyerges, et al. (Ed.), *Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems* (pp. 327-341): Kluwer Academic Publishers.
Davies, T., M., & Hazelton, M., L. (2010). Adaptive kernel estimation of spatial relative risk. *Statistics in Medicine, 29*(23), 2423-2437. doi: 10.1002/sim.3995
Davies, T. M., & Hazelton, M. L. (2010). Adaptive kernel estimation of spatial relative risk. *Statistics in Medicine, 29*(23), 2423-2437. doi: 10.1002/sim.3995

De Montis, A., Barthelemy, M., Chessa, A., & Vespignani, A. (2007). The structure of interurban traffic: A weighted network analysis. *Environment and Planning B-Planning & Design, 34*(5), 905-924.
De Waard, D., & Studiecentrum, V. (1996). *The measurement of drivers' mental workload*: Groningen University, Traffic Research Center.
Demsar, U. (2007). Investigating visual exploration of geospatial data: An exploratory usability experiment for visual data mining. *Computers, Environment and Urban Systems, 31*(5), 551-571. doi: DOI: 10.1016/j.compenvurbsys.2007.08.006
Demšar, U., & Virrantaus, K. (2010). Space–time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science, 24*(10), 1527-1542. doi: 10.1080/13658816.2010.511223
Dodge, S., Weibel, R., & Lautenschutz, A.-K. (2008). Towards a taxonomy of movement patterns. *Inf Visualization, 7*(3-4), 240-252.
Doreian, P., & Conti, N. (2012). Social context, spatial structure and social network structure. *Social networks, 34*(1), 32-46.
Dorigo, G., & Tobler, W. (1983). Push-Pull Migration Laws. *Annals of the Association of American Geographers, 73*(1), 1-17.
Dunne, C., & Shneiderman, B. (2009). Improving graph drawing readability by incorporating readability metrics: A software tool for network analysts. *University of Maryland, HCIL Tech Report HCIL-2009-13*.
Dwyer, T., Lee, B., Fisher, D., Quinn, K. I., Isenberg, P., Robertson, G., & North, C. (2009). A comparison of user-generated and automatic graph layouts. *Visualization and Computer Graphics, IEEE Transactions on, 15*(6), 961-968.
Egerbladh, I., Kasakoff, A. B., & Adams, J. W. (2007). Gender Differences in the Dispersal of Children in Northern Sweden and the Northern USA in 1850. *The History of the Family, 12*(1), 2-18.
Estrada, E., & Bodin, O. (2008). Using network centrality measures to manage landscape connectivity. *Ecological Applications, 18*(7), 1810-1825.
Estrada, E., Hatano, N., & Gutierrez, A. (2008). 'Clumpiness' mixing in complex networks. *Journal of Statistical Mechanics-Theory and Experiment*. doi: P03008 10.1088/1742-5468/2008/03/p03008
Fagiolo, G., Reyes, J., & Schiavo, S. (2009). World-trade web: Topological properties, dynamics, and evolution. *Physical Review E, 79*(3). doi: 036115 10.1103/PhysRevE.79.036115
Faust, K., Entwisle, B., Rindfuss, R. R., Walsh, S. J., & Sawangdee, Y. (2000). Spatial arrangement of social and economic networks among villages in Nang Rong District, Thailand. *Social Networks, 21*(4), 311-337. doi: 10.1016/S0378-8733(99)00014-3
Faust, K., & Lovasi, G. S. (2012). Capturing context: Integrating spatial and social network analyses. *Social networks, 34*(1), 1-5.
Festinger, L., Schachter, S., & Back, K. W. (1963). *Social pressures in informal groups : a study of human factors in housing*. Stanford, Calif.: Stanford University Press.
Fischer, M. M., Essletzbichler, J., Gassler, H., & Trichtl, G. (1993). Telephone Communication Patterns in Austria: A Comparison of the IPFP-based Graph-

Theoretic and the Intramax Approaches. *Geographical Analysis, 25*(3), 224-233. doi: 10.1111/j.1538-4632.1993.tb00293.x
Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: the analysis of spatially varying relationships*. New York: John Wiley & Sons.
Fowler, D., & Ware, C. (1989). *Strokes for representing univariate vector field maps.* Paper presented at the Proceedings of Graphics Interface.
Freeman, L. C. (1977). Set of Measures of Centrality Based on Betweenness. *Sociometry, 40*(1), 35-41.
Frey, W. H. (2005). Metropolitan America in the New Century. *Social Science Quarterly*.
Frey, W. H., Liaw, K., Xie, Y., & Carlson, M. J. (1995). Interstate Migration of the US Poverty Population: Immigration "Pushes" and Welfare Magnet "Pulls". Ann Arbor: Population Studies Center University of Michigan.
Fyfe, D. A., Holdsworth, D. W., & Weaver, C. (2009). Historical GIS and Visualization Insights From Three Hotel Guest Registers in Central Pennsylvania, 1888—1897. *Social Science Computer Review, 27*(3), 348-362.
Gansner, E., Koren, Y., & North, S. (2004). Graph Drawing by Stress Majorization. *Proceedings of 12th Int. Symp. Graph Drawing (GD'04), Lecture Notes in Computer Science, Springer Verlag, 3383*, 239--250.
Gansner, E. R., Koren, Y., & North, S. C. (2005). Topological fisheye views for visualizing large graphs. *IEEE Trans Vis Comput Graph, 11*(4), 457-468.
Gao, S., Liu, Y., Wang, Y., & Ma, X. (2013). Discovering spatial interaction communities from mobile phone data. *Transactions in GIS, 17*(3), 463-481.
Gastner, M. T., & Newman, M. E. J. (2006). The spatial structure of networks. *The European Physical Journal B - Condensed Matter and Complex Systems, 49*(2), 247-252.
Ghoniem, M., Fekete, J.-D., & Castagliola, P. (2005). On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization, 4*(2), 114-135.
Ghosh, D., & Guha, R. (2013). What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science, 40*(2), 90-102. doi: 10.1080/15230406.2013.776210
Goodchild, M. F., Anselin, L., Appelbaum, R. P., & Harthorn, B. H. (2000). Toward spatially integrated social science. *International Regional Science Review, 23*(2), 139-159.
Gotz, D., & Zhou, M. X. (2008). Characterizing users' visual analytic activity for insight provenance. *Inf Visualization, 8*(1), 42-55.
Graham, M., & Zook, M. (2013). Augmented realities and uneven geographies: exploring the geolinguistic contours of the web. *Environment and Planning A, 45*(1), 77-99.
Guo, D. (2009). Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data (Vol. 15, pp. 1041-1048). IEEE Transactions on Visualization and Computer Graphics: IEEE.
Guo, D., & Zhu, X. (2014). Origin-Destination Flow Data Smoothing and Mapping. *Visualization and Computer Graphics, IEEE Transactions on, PP*(99), 1-1. doi: 10.1109/TVCG.2014.2346271

Heinzelman, J., & Waters, C. (2010). Crowdsourcing crisis information in disaster-affected Haiti.
Hipp, J. R., Faris, R. W., & Boessen, A. (2012). Measuring 'neighborhood': Constructing network neighborhoods. *Social networks, 34*(1), 128-140.
Hollenstein, L., & Purves, R. (2013). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*(1), 21-48.
Holmes, J. (1978). Transformation of Flow Matrices to Eliminate the Effects of Differing Sizes of Origin-Destination Units: A Further Comment. *Systems, Man and Cybernetics, IEEE Transactions on, 8*(4), 325-332.
Holten, D., & van Wijk, J. J. (2009). Force-Directed Edge Bundling for Graph Visualization. *Computer Graphics Forum, 28*(3), 983-990.
Huang, W. (2007). *Using eye tracking to investigate graph layout effects.* Paper presented at the Visualization, 2007. APVIS'07. 2007 6th International Asia-Pacific Symposium on.
Huang, W., Eades, P., & Hong, S.-H. (2009). Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization, 8*(3), 139-152.
Huang, W., Hong, S.-H., & Eades, P. (2006). *How people read sociograms: a questionnaire study.* Paper presented at the Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation-Volume 60.
Hughes, H. L. (1993). Metropolitan Structure and The Suburban Hierarchy. *American Sociological Review, 58*(3), 417-433.
Hägerstrand, T. (1976). Geography and the study of interaction between nature and society. *Geoforum, 7*(5), 329-334.
Irwin, M. D., & Hughes, H. L. (1992). Centrality and The Structure of Urban Interaction - Measures, Concepts, and Applications. *Social Forces, 71*(1), 17-51.
Kafadar, K. (1994). Choosing among two-dimensional smoothers in practice. *Computational statistics & data analysis, 18*(4), 419-439.
Kasakoff, A. B., & Adams, J. W. (2000). The effects of migration, place, and occupation on adult mortality in the American North, 1740–1880. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 33*(2), 115-130.
Kato, M., Nagasaki, M., Doi, A., & Miyano, S. (2005). Automatic drawing of biological networks using cross cost and subcomponent data. *Genome Inform, 16*(2), 22-31.
Keim, D. A. (2002). Information visualization and visual data mining. *Visualization and Computer Graphics, IEEE Transactions on, 8*(1), 1-8.
Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2001). *An online algorithm for segmenting time series.* Paper presented at the Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on.
Kinkeldey, C., Mason, J., Klippel, A., & Schiewe, J. (2013). *Assessing the impact of design decisions on the usability of uncertainty visualization: noise annotation lines for the visual representation of attribute uncertainty.* Paper presented at the Proceedings of the 26th international cartographic conference.
Knapp, L. (1995). A Task Analysis Approach to the Visualization of Geographic Data. In T. L. Nyerges, et al. (Ed.), *Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems* (pp. 355-371): Kluwer Academic Publishers.

Kolaczyk, E. D., Chua, D. B., & Barthelemy, M. (2009). Group betweenness and co-betweenness: Inter-related notions of coalition centrality. *Social Networks, 31*(3), 190-203. doi: 10.1016/j.socnet.2009.02.003

Koua, E. L., & Kraak, M.-J. (2004). *A Usability Framework for the Design and Evaluation of an Exploratory Geovisualization Environment*. Paper presented at the Proceedings of the Information Visualisation, Eighth International Conference.

Koua, E. L., Maceachren, A., & Kraak, M. J. (2006). Evaluating the usability of visualization methods in an exploratory geovisualization environment. *International Journal of Geographical Information Science, 20*(4), 425-448. doi: 10.1080/13658810600607550

Koylu, C. (Producer). (2013a, 5/5). Animation of Family Migration. Retrieved from http://129.252.37.169:8400/flowvis/trajectories/index.html

Koylu, C. (2013b, 10/22). Family Connectedness. from http://www.spatialdatamining.org/familyconnectedness

Koylu, C. (2014a, 6/15). Commodity Flow Mapper. from http://129.252.37.169:8400/flowvis/commodity/index.html

Koylu, C. (2014b). Flow Map Test Analytics. Retrieved 10/14, 2014, from http://129.252.37.169:8400/flowvis/testanalytics/index.html

Koylu, C. (2014c). Flow Map Test Responses. Retrieved 10/14, 2014, from http://129.252.37.169:8400/flowvis/testentries/index.html

Koylu, C., & Guo, D. (2013). Smoothing locational measures in spatial interaction networks. *Computers Environment and Urban Systems, 41*, 12-25. doi: 10.1016/j.compenvurbsys.2013.03.001

Koylu, C., Guo, D., Kasakoff, A., & Adams, J. W. (2014). Mapping family connectedness across space and time. *Cartography and Geographic Information Science, 41*(1), 14-26.

Kumar, S., Morstatter, F., & Liu, H. (2014). *Twitter Data Analytics*: Springer.

Körner, C. (2011). Eye movements reveal distinct search and reasoning processes in comprehension of complex graphs. *Applied Cognitive Psychology, 25*(6), 893-905.

Laidlaw, D. H., Davidson, J. S., Miller, T. S., da Silva, M., Kirby, R., Warren, W. H., & Tarr, M. (2001). *Quantitative comparative evaluation of 2D vector field visualization methods.* Paper presented at the Proceedings of the conference on Visualization'01.

Lambert, A., Bourqui, R., & Auber, D. (2010). *3D Edge Bundling for Geographical Data Visualization*. Paper presented at the Proceedings of the 2010 14th International Conference Information Visualisation.

Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F., & Bertollini, R. (1999). *Disease mapping and risk assessment for public health*: John Wiley & Sons.

Lee, B., Plaisant, C., Parr, C. S., Fekete, J.-D., & Henry, N. (2006). *Task taxonomy for graph visualization.* Paper presented at the Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization.

Lee, J. Y., & Kwan, M. p. (2011). Visualisation Of Socio-Spatial Isolation Based On Human Activity Patterns And Social Networks In Space-Time. *Tijdschrift voor economische en sociale geografie, 102*(4), 468-485.

Leutenegger, A. L., Sahbatou, M., Gazal, S., Cann, H., & Génin, E. (2011). Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us. *European Journal of Human Genetics, 19*(5), 583-587.

Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences, 109*(1), 68-72. doi: 10.1073/pnas.1109739109

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science, 40*(2), 61-77. doi: 10.1080/15230406.2013.777139

Li, X. (2010). *The time wave in time space : a visual exploration environment for spatio-temporal data.* (Proefschrift Enschede), s.n.], [S.l. Retrieved from http://edepot.wur.nl/199249

Limtanakool, N., Schwanen, T., & Dijst, M. (2009). Developments in the Dutch Urban System on the Basis of Flows. *Regional Studies, 43*(2), 179-196. doi: 10.1080/00343400701808832

Liu, L. (1995). PPFLOW: An interactive visualization system for the exploratory analysis of migration flows. *Geographic Information Sciences, 1*(2), 118-123.

Liu, Z., Cai, S., Swan, J. E., Moorhead, R. J., Martin, J. P., & Jankun-Kelly, T. (2012). A 2D flow visualization user study using explicit flow synthesis and implicit task design. *Visualization and Computer Graphics, IEEE Transactions on, 18*(5), 783-796.

Lomi, A., & Pallotti, F. (2012). Relational collaboration among spatial multipoint competitors. *Social networks, 34*(1), 101-111.

Long, L. (1988). *Migration and residential mobility in the United States*: Russell Sage Foundation.

Long, L., E. (1988). *Migration and residential mobility in the United States*. New York: Russell Sage Foundation.

Long, L. E., & National Committee for Research on the 1980 Census. (1988). *Migration and residential mobility in the United States*. New York: Russell Sage Foundation.

Luo, W., & MacEachren, A. M. (2014). Geo-social visual analytics. *Journal of spatial information science*, 27-66.

Luo, W., MacEachren, A. M., Yin, P., & Hardisty, F. (2011). *Spatial-social network visualization for exploratory data analysis*. Paper presented at the Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Chicago, Illinois.

Luo, W., Yin, P., Di, Q., Hardisty, F., & MacEachren, A. M. (2014). A Geovisual Analytic Approach to Understanding Geo-Social Relationships in the International Trade Network. *PloS one, 9*(2), e88666.

Ma, D. (2012). *Visualization of social media data: mapping changing social networks.* (Master of Science), University of Twente.

MacEachren, A. M. (1995). *How maps work: representation, visualization, and design*: New York; London: The Guilford Press.

MacEachren, A. M., Wachowicz, M., Edsall, R., Haug, D., & Masters, R. (1999). Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science, 13*(4), 311-334.
Maggioni, M. A., Nosvelli, M., & Uberti, T. E. (2007). Space versus networks in the geography of innovation: A European analysis*. *Papers in Regional Science, 86*(3), 471-493.
Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods, 44*(1), 1-23.
McGrath, C., Blythe, J., & Krackhardt, D. (1997). The effect of spatial arrangement on judgments and errors in interpreting graphs. *Social Networks, 19*(3), 223-242.
McIntire, J. P., Osesina, O. I., Bartley, C., Tudoreanu, M. E., Havig, P. R., & Geiselman, E. E. (2012). *Visualizing weighted networks: a performance comparison of adjacency matrices versus node-link diagrams.* Paper presented at the SPIE Defense, Security, and Sensing.
Mennis, J., & Mason, M. J. (2012). Social and geographic contexts of adolescent substance use: The moderating effects of age and gender. *Social networks, 34*(1), 150-157.
Mennis, J. L., Peuquet, D. J., & Qian, L. (2000). A conceptual framework for incorporating cognitive principles into geographical database representation. *International Journal of Geographical Information Science, 14*(6), 501-520.
Michelson, W. M. (1970). *Man and his urban environment: a sociological approach*. Reading, Mass.,: Addison-Wesley Pub. Co.
Mitchell, P., Ware, C., & Kelley, J. (2009). *Investigating flow visualizations using interactive design space hill climbing.* Paper presented at the Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on.
Moody, J., McFarland, D., & Bender-deMoll, S. (2005). Dynamic network visualization. *American Journal of Sociology, 110*(4), 1206-1241.
Morrill, R. L. (1988). Migration Regions and Population Redistribution. *Growth and Change, 19*(1), 43-60. doi: 10.1111/j.1468-2257.1988.tb00461.x
Munzner, T. (2009). A nested model for visualization design and validation. *Visualization and Computer Graphics, IEEE Transactions on, 15*(6), 921-928.
Nag, M. (2009). Mapping Networks: A New Method for Integrating Spatial and Network Data. *Unpublished manuscript. Princeton University, Department of Sociology*.
Nakaya, T., & Yano, K. (2010). Visualising Crime Clusters in a Space-time Cube: An Exploratory Data-analysis Approach Using Space-time Kernel Density Estimation and Scan Statistics. *Transactions in GIS, 14*(3), 223-239.
Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters, 89*(20), 4. doi: 208701
10.1103/PhysRevLett.89.208701
Newman, M. E. J. (2003). The structure and function of complex networks. *Siam Review, 45*(2), 167-256.
Newman, M. E. J., & Leicht, E. A. (2007). Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences of the United States of America, 104*(23), 9564-9569. doi: 10.1073/pnas.0610537104
Nielsen, J. (1993). *Usability engineering*. Boston: Academic Press.

North, C. (2006). Toward Measuring Visualization Insight, *26*, 6-9.
O'Kelly, M. E. (1998). A geographer's analysis of hub-and-spoke networks. *Journal of Transport Geography, 6*(3), 171-186. doi: http://dx.doi.org/10.1016/S0966-6923(98)00010-6
Ohmae, K. (1990). The borderless world. *New York*.
Onnela, J.-P., Arbesman, S., González, M. C., Barabási, A.-L., & Christakis, N. A. (2011). Geographic Constraints on Social Network Groups. *PLoS ONE, 6*(4), e16939. doi: 10.1371/journal.pone.0016939
Openshaw, S. (1983). *The modifiable areal unit problem* (Vol. 38): Geo Books Norwich.
Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist, 38*(1), 63-71.
Pandit, K. (1994). Differentiating Between Subsystems and Typologies in the Analysis of Migration Regions: A U.S. Example. *The Professional Geographer, 46*(3), 331 - 345.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making, 5*(5), 411-419.
Parks, M. J. (1987). *American Flow Mapping: A Survey of the Flow Maps Found in Twentieth Century Geography Textbooks, Including a Classification of the Various Flow Map Designs*: Georgia State University.
Patil, D. J. (Producer). (2011, 08/19/2012 11:00 pm). Visualize your LinkedIn network with InMaps. Retrieved from http://www.youtube.com/watch?v=PC99Nw2JX8w&feature=player_embedded
Peuquet, D. J. (1994). It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of american Geographers, 84*(3), 441-461.
Phan, D., Xiao, L., Yeh, R., & Hanrahan, P. (2005). *Flow map layout.* Paper presented at the IEEE Symposium on Information Visualization.
Phithakkitnukoon, S., Calabrese, F., Smoreda, Z., & Ratti, C. (2011). *Out of Sight Out of Mind--How Our Mobile Social Network Changes during Migration.* Paper presented at the Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom).
Plane, D., A., & Rogerson, P., A. (1991). Tracking the baby boom, the baby bust, and the echo generations: how age composition regulates US migration. *The Professional Geographer, 43*(4), 416 - 430.
Plane, D. A., & Heins, F. (2003). Age articulation of US inter-metropolitan migration flows. *Annals of Regional Science, 37*(1), 107-130.
Plane, D. A., & Jurjevich, J. R. (2009). Ties That No Longer Bind? The Patterns and Repercussions of Age-Articulated Migration. *The Professional Geographer, 61*(1), 4 - 20.
Plane, D. A., & Mulligan, G. F. (1997). Measuring spatial focusing in a migration system. *Demography, 34*(2), 251-262.
Pooley, C. G. (1979). Residential Mobility in the Victorian City. *Transactions of the Institute of British Geographers, 4*(2), 258-277.

Pooley, C. G., & Turnbull, J. (1998). *Migration and mobility in Britain since the eighteenth century*. London ; Bristol, Penn.: UCL Press.
Poon, J. P. (1997). The Cosmopolitanization of Trade Regions: Global Trends and Implications, 1965–1990*. *Economic Geography, 73*(4), 390-404.
Porta, S., Latora, V., Wang, F., Strano, E., Cardillo, A., Scellato, S., . . . Messora, R. (2009). Street centrality and densities of retail and services in Bologna, Italy. *Environment and Planning B: Planning and Design, 36*(3), 450-465.
Purchase, H. C., Carrington, D., & Allder, J.-A. (2002). Empirical evaluation of aesthetics-based graph layout. *Empirical Software Engineering, 7*(3), 233-255.
Purchase, H. C., Cohen, R. F., & James, M. I. (1997). An experimental study of the basis for graph drawing algorithms. *Journal of Experimental Algorithmics (JEA), 2*, 4.
Purchase, H. C., Hamer, J., Nöllenburg, M., & Kobourov, S. G. (2013). *On the usability of Lombardi graph drawings.* Paper presented at the Graph Drawing.
Radil, S. M., Flint, C., & Tita, G. E. (2010). 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, 100*(2), 307-326.
Rana, S., & Dykes, J. (2003). A framework for augmenting the visualization of dynamic raster surfaces. *Information Visualization, 2*(2), 126-139.
Rao, T., & Srivastava, S. (2012). *Analyzing Stock Market Movements Using Twitter Sentiment Analysis*. Paper presented at the Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012).
Robertson, G., Fernandez, R., Fisher, D., Lee, B., & Stasko, J. (2008). Effectiveness of Animation in Trend Visualization. *IEEE Transactions on Visualization and Computer Graphics, 14*(6), 1325-1332. doi: 10.1109/tvcg.2008.125
Robinson, A. H. (1967). The thematic maps of Charles Joseph Minard∗.
Rogers, A. (1992). Heterogeneity, spatial population dynamics, and the migration rate. *Environment and Planning A, 24*(6), 775-791.
Rogers, A., & Raymer, J. (1998). The spatial focus of US interstate migration flows. *International Journal of Population Geography, 4*(1), 63-80.
Rogers, A., & Sweeney, S. (1998). Measuring the Spatial Focus of Migration Patterns. *The Professional Geographer, 50*(2), 232 - 242.
Roseman, C., & McHugh, K. (1982). Metropolitan areas as redistributors of population. *Urban Geography, 3*(1), 22-33.
Roseman, C. C. (1977). Changing Migration Patterns Within the United States. In R. P. f. C. Geography (Ed.), (Vol. 2, pp. 44): Association of American Geographers, Washington, D.C. Commission on College Geography.
Roth, R. E. (2012). Cartographic interaction primitives: Framework and synthesis. *Cartographic Journal, The, 49*(4), 376-395.
Roy, J. R., & Thill, J. C. (2004). Spatial interaction modelling. *Papers in Regional Science, 83*(1), 339-361. doi: 10.1007/s10110-003-0189-4
Rubin, J., Chisnell, D. (2008). *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests* (2 ed.): Wiley Publishing, Inc.
Sailer, K., & McCulloh, I. (2012). Social networks and spatial configuration—How office layouts drive social interaction. *Social networks, 34*(1), 47-58.

Sain, S. R., & Scott, D. W. (1996). On Locally Adaptive Density Estimation. *Journal of the American Statistical Association, 91*(436), 1525-1534.
Saket, B., Simonetto, P., & Kobourov, S. (2014). Group-Level Graph Visualization Taxonomy. *arXiv preprint arXiv:1403.7421*.
Saraiya, P., North, C., & Duca, K. (2004). *An Evaluation of Microarray Visualization Tools for Biological Insight*. Paper presented at the Proceedings of the IEEE Symposium on Information Visualization.
Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011). Socio-spatial properties of online location-based social networks. *Proceedings of ICWSM, 11*, 329-336.
Schaefer, D. R. (2012). Youth co-offending networks: An investigation of social and spatial effects. *Social networks, 34*(1), 141-149.
Schloegel, K., Karypis, G., & Kumar, V. (2000). Parallel multilevel algorithms for multi-constraint graph partitioning. *Euro-Par 2000 Parallel Processing, Proceedings, 1900*, 296-310.
Scott, J. (2000). *Social Network Analysis: A Handbook* (2nd Ed. ed.). London: SAGE.
Shaw, S. L., Yu, H., & Bombom, L. S. (2008). A Space-Time GIS Approach to Exploring Large Individual-based Spatiotemporal Datasets. *Transactions in GIS, 12*(4), 425-441.
Shepherd, I. (1995). Putting time on the map: Dynamic displays in data visualization and GIS. *Innovations in GIS, 2*(2), 169-187.
Shi, X. (2009). A Geocomputational Process for Characterizing the Spatial Pattern of Lung Cancer Incidence in New Hampshire. *Annals of the Association of American Geographers, 99*(3), 521-533. doi: 10.1080/00045600902931801
Shi, X. (2010). Selection of bandwidth type and adjustment side in kernel density estimation over inhomogeneous backgrounds. *International Journal of Geographical Information Science, 24*(5), 643-660.
Shi, X., Duell, E., Demidenko, E., Onega, T., Wilson, B., & Hoftiezer, D. (2007). A polygon-based locally-weighted-average method for smoothing disease rates of small units. *Epidemiology, 18*(5), 523.
Shneiderman, B. (1996). *The eyes have it: A task by data type taxonomy for information visualizations.* Paper presented at the Visual Languages, 1996. Proceedings., IEEE Symposium on.
Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London ; New York: Chapman and Hall.
Slater, P., B. (1975). Hierarchical regionalization of RSFSR administrative units using 1966-69 migration data. *Soviet Geography Review and Translation, 16*(7), 453-465.
Slater, P. B. (1976). Hierarchical regionalization of Japanese prefectures using 1972 inter-prefectural migration flows. *Regional Studies, 10*(1), 123-132.
Slater, P. B. (1976). The use of state-to-state college migration data in developing a hierarchy of higher educational regions. *Research in Higher Education, 4*(4), 305-315. doi: 10.1007/bf00991624
Slater, P. B. (1984). A partial hierarchical regionalization of 3140 US counties on the basis of 1965 - 1970 intercounty migration. *Environment and Planning A, 16*(4), 545-550.

Slocum, T. A., McMaster, R. B., Kessler, F. C., & Howard, H. H. (2009). *Thematic cartography and geovisualization*: Pearson Prentice Hall Upper Saddle River, NJ.
Smith, R. H. T. (1970). Concepts and Methods in Commodity Flow Analysis. *Economic Geography, 46*, 404-416. doi: 10.2307/143153
Sohn, K., & Kim, D. (2010). Zonal centrality measures and the neighborhood effect. *Transportation Research Part A: Policy and Practice, 44*(9), 733-743. doi: 10.1016/j.tra.2010.07.006
Thompson, W., & Lavin, S. (1996). Automatic generation of animated migration maps. *Cartographica: The International Journal for Geographic Information and Geovisualization, 33*(2), 17-28.
Tobler, W. (2004). Movement mapping.
Tobler, W. R. (1976). Spatial interaction patterns. *Journal of Environmental Systems, 6*, 271-301.
Tobler, W. R. (1987). Experiments in migration mapping by computer. *American Cartographer, 14*, 155–163.
Tobon, C. (2005). Evaluating geographic visualization tools and methods: an approach and experiment based upon user tasks. In J. Dykes, A. M. MacEachren, & M.-J. Kraak (Eds.), *Exploring Geovisualization* (pp. 645-666): Amsterdam: Elsevier.
Tobon, C. m. (2002). Usablity testing for improving interactive geovisualization techniques (pp. 24 b:CountryRegion London).
Todo, Y., Yadate, D. M., Matous, P., & Takahashi, R. (2011). *Effects of geography and social networks on diffusion and adoption of agricultural technology: Evidence from rural Ethiopia.* Paper presented at the CSAE 25th Anniversary Conference.
van de Ven, B. (2007). *Algorithms for flow maps.* Technische Universiteit Eindhoven.
Viry, G. (2012). Residential mobility and the spatial dispersion of personal networks: Effects on social support. *Social networks, 34*(1), 59-72.
Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing* (1st ed.). London ; New York: Chapman & Hall.
Ware, C. (2013). *Information visualization: perception for design*: Elsevier.
Ware, C., Purchase, H., Colpoys, L., & McGill, M. (2002). Cognitive measurements of graph aesthetics. *Information Visualization, 1*(2), 103-110.
Warren Liao, T. (2005). Clustering of time series data—a survey. *Pattern Recognition, 38*(11), 1857-1874. doi: 10.1016/j.patcog.2005.01.025
Wasserman, S., & Faust, K. (1994). *Social network analysis : methods and applications*. Cambridge [England] ; New York: Cambridge University Press.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440-442.
Wehrend, S., & Lewis, C. (1990). *A problem-oriented classification of visualization techniques*. Paper presented at the Proceedings of the 1st conference on Visualization '90, San Francisco, California.
Whisler, R. L., Waldorf, B. S., Mulligan, G. F., & Plane, D. A. (2008). Quality of Life and the Migration of the College-Educated: A Life-Course Approach. *Growth and Change, 39*(1), 58-94. doi: 10.1111/j.1468-2257.2007.00405.x
Wierwille, W. W., & Casali, J. G. (1983). *A validated rating scale for global mental workload measurement applications.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Wise, S. M., Haining, R. P., & Ma, J. (1997). Regionalization tools for the exploratory spatial analysis of health data. In M. Fischer & A. Getis (Eds.), *Recent developments in spatial analysis: spatial statistics, behavioural modelling and neuro-computing*. Berlin: Springer-Verlag.

Xu, K., Rooney, C., Passmore, P., Ham, D.-H., & Nguyen, P. H. (2012). A user study on curved edges in graph visualization. *Visualization and Computer Graphics, IEEE Transactions on, 18*(12), 2449-2456.

Yadav-Pauletti, S. (1996). *MigMap, a Data Exploration Application for Visualizing US Census Migration Data.* University of Kansas, Geography.

Yan, J., & Thill, J. C. (2009). Visual data mining in spatial interaction analysis with self-organizing maps. *Environment and Planning B-Planning & Design, 36*(3), 466-486. doi: 10.1068/b34019

Yang, B. S., Luan, X. C., & Li, Q. Q. (2010). An adaptive method for identifying the spatial patterns in road networks. *Computers Environment and Urban Systems, 34*(1), 40-48. doi: 10.1016/j.compenvurbsys.2009.10.002

Young, D. A. (2002). A New Space-Time Computer Simulation Method for Human Migration. *American anthropologist, 104*(1), 138-158.

Zhou, S., & Mondragon, R. J. (2004). The rich-club phenomenon in the Internet topology. *Ieee Communications Letters, 8*(3), 180-182. doi: 10.1109/lcomm.2004.823426

Zijlstra, F. R. H. (1993). *Efficiency in work behaviour: a design approach for modern tools.* Retrieved from http://resolver.tudelft.nl/uuid:d97a028b-c3dc-4930-b2ab-a7877993a17f

# APPENDIX A – ADDITIONAL DISCUSSION AND RESULTS FOR CHAPTER 2

Due to the copyright agreement of the published material, Appendix A is attributed to further analyses and discussion on the results of the methodology introduced in Chapter 2. In an area-based spatial interaction network such as the county-to-county migration data, spatial units (counties) vary greatly not only by their population, but also their areal extent. For example, rural counties or counties in the Western U.S. have substantially larger areas than urban counties and counties in the Eastern U.S. The variation in the area of counties results in two major challenges. First, contiguous counties with varying area cause a bias for neighborhood selection. For example, a county with a large area could potentially have neighbors that span an extent with greater distances, whereas a county with a small area surrounded by smaller counties could form neighborhoods in much shorter distances. One potential solution to address this problem is using areal interpolation to consider counties partially in selecting a neighborhood, and estimate flows based on the included portion of each county within the neighborhood. Second, on a measure map result, a county with larger area is visually more dominant even though its population or capacity to generate flows is not high. Using a cartogram based on each county's population or capacity to generate flows is a potential solution to eliminate such misleading patterns.

On a smoothed measure map such as the net migration rate, the color of a spatial unit (county) illustrates the measure calculated considering the flows from/to that county's neighborhood rather than just the flows to/from that particular county.

Such information is hard to convey on a static measure map. A dynamic map with user controls to visualize each county's neighborhood and flows associated with each neighborhood would help users better understand the dynamics and effects of parameter selection in the smoothing approach.

In the final part of Appendix A, a series of figures that illustrate smoothed net migration rate for all age groups are given below. These figures demonstrate the effectiveness of the smoothing approach in identifying spatially and structurally distinct migration patterns of different age groups. It is possible to summarize the general overview of the migration of age groups into three distinct patterns. The first group consists of the age groups 30-34 (Figure A.1), 35-39 (Figure A.2), 40-44 (Figure A.3), 45-49 (Figure A.4) and 50-54 (Figure A.5) that represent the families with kids (younger than 20-25) who prefer living in suburbs surrounding the metropolitan areas. The second group consists of the age groups 55-59 (Figure A.6), 60-64 (Figure A.7), 65-69 (Figure A.8), 70-74 (Figure A.9) and 75-79 (Figure A.10) that represent the retirees who leave metropolitan areas and target recreational places close to forests and coastal areas; and with warmer temperatures. Within the second group, there is also an increasing tendency to choose places with warmer temperature (e.g., Florida, Arizona and Coastal Carolinas) as the age increases. The third group consists of the age groups 80-84 (Figure A.11) and 85 and above (Figure A.12) that represent individuals who need nursing care. The distinct patterns that the third group highlights places with higher availability of nursing homes and children who potentially take care of their elderly parents.

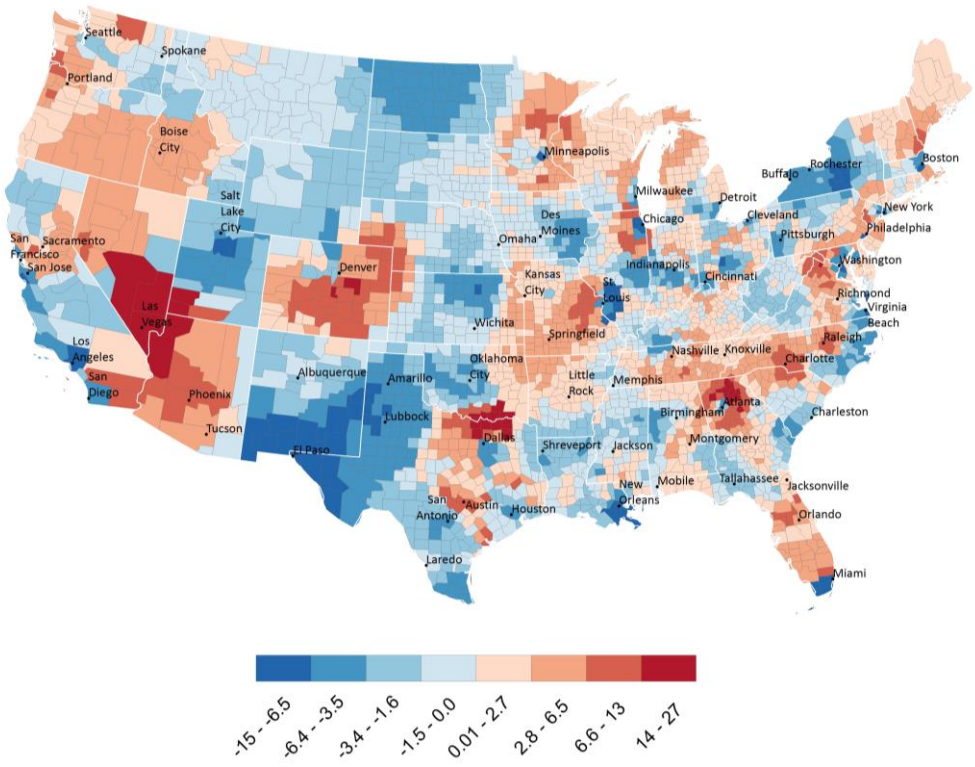



Figure A.1: Smoothed Net Migration Rate for age group 30-34

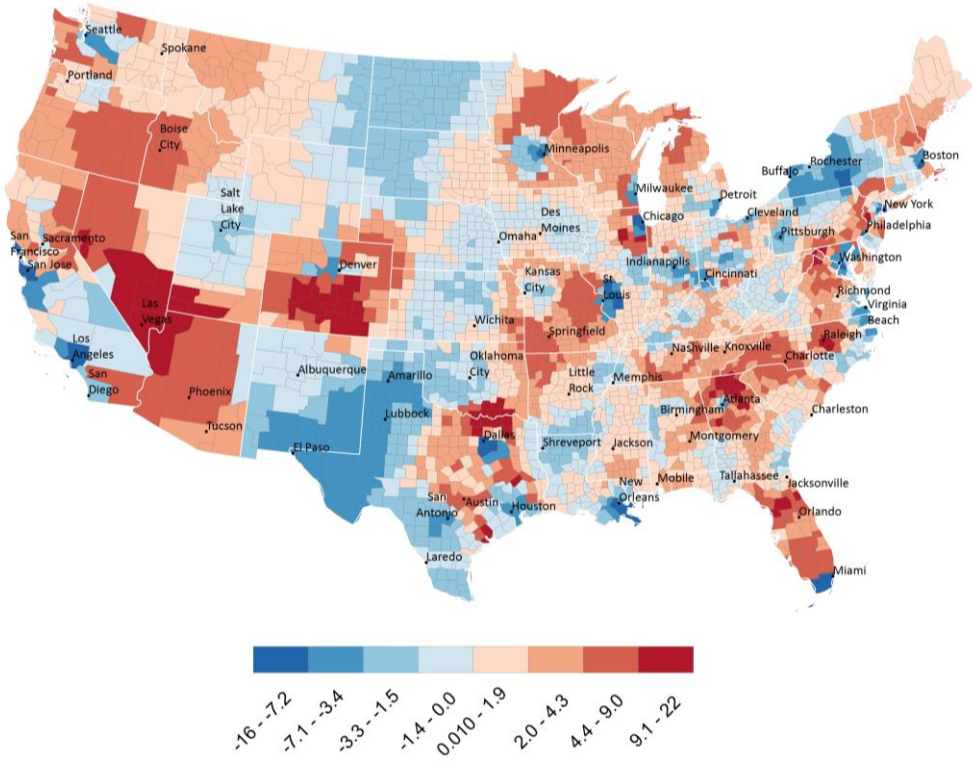



Figure A.2: Smoothed Net Migration Rate for age group 35-39

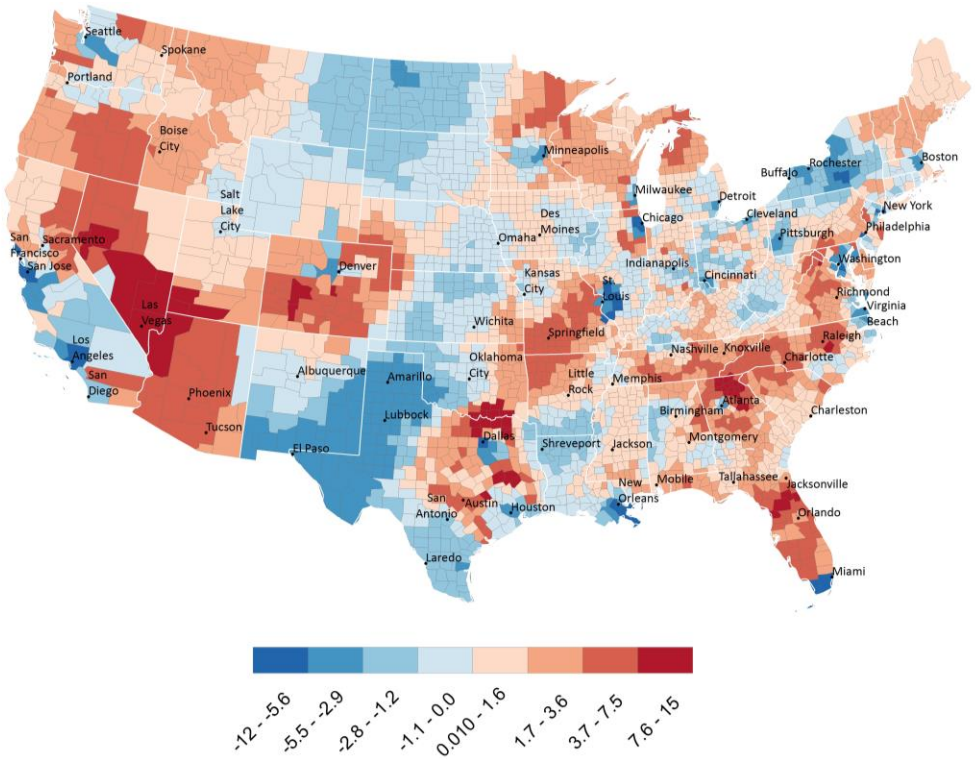



Figure A.3: Smoothed Net Migration Rate for age group 40-44

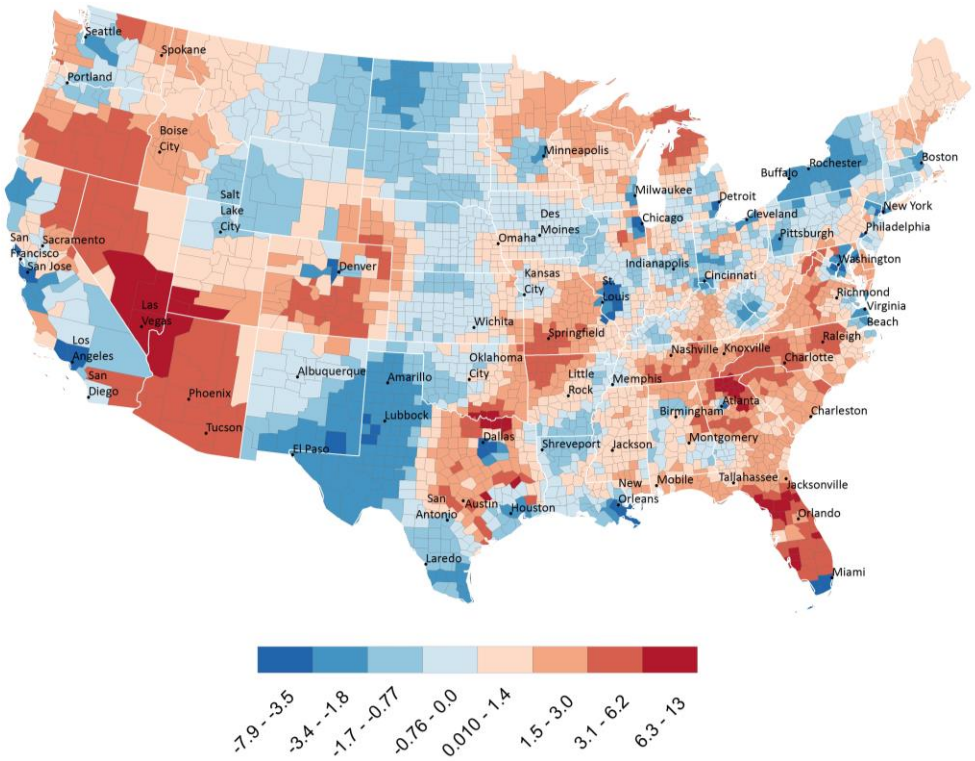



Figure A.4: Smoothed Net Migration Rate for age group 45-49

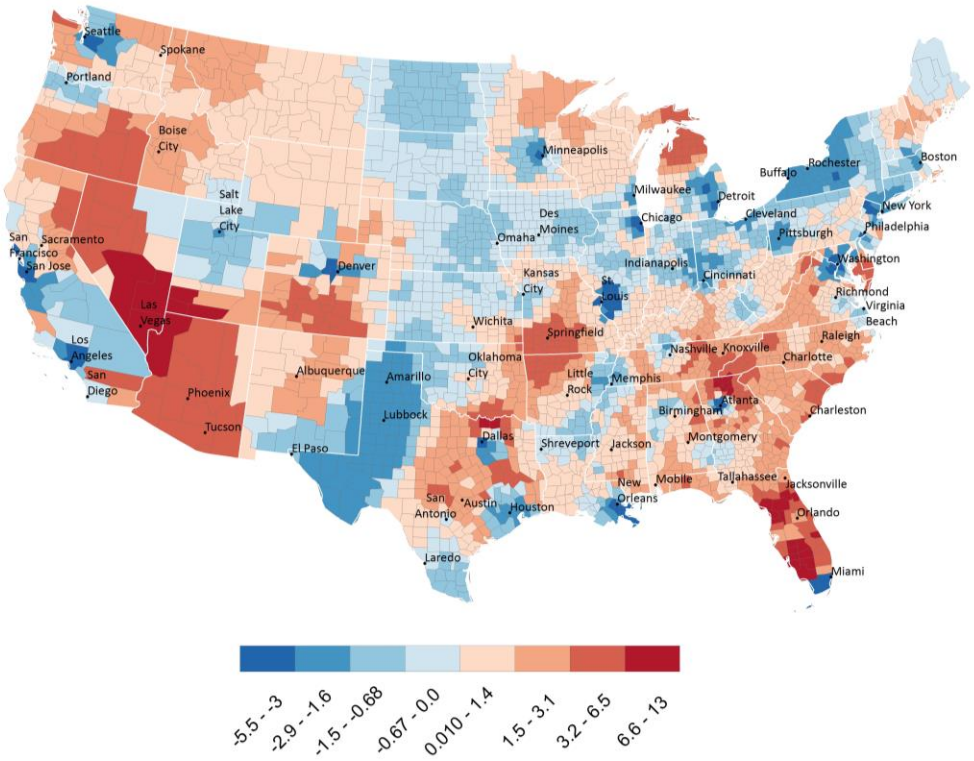



Figure A.5: Smoothed Net Migration Rate for age group 50-54

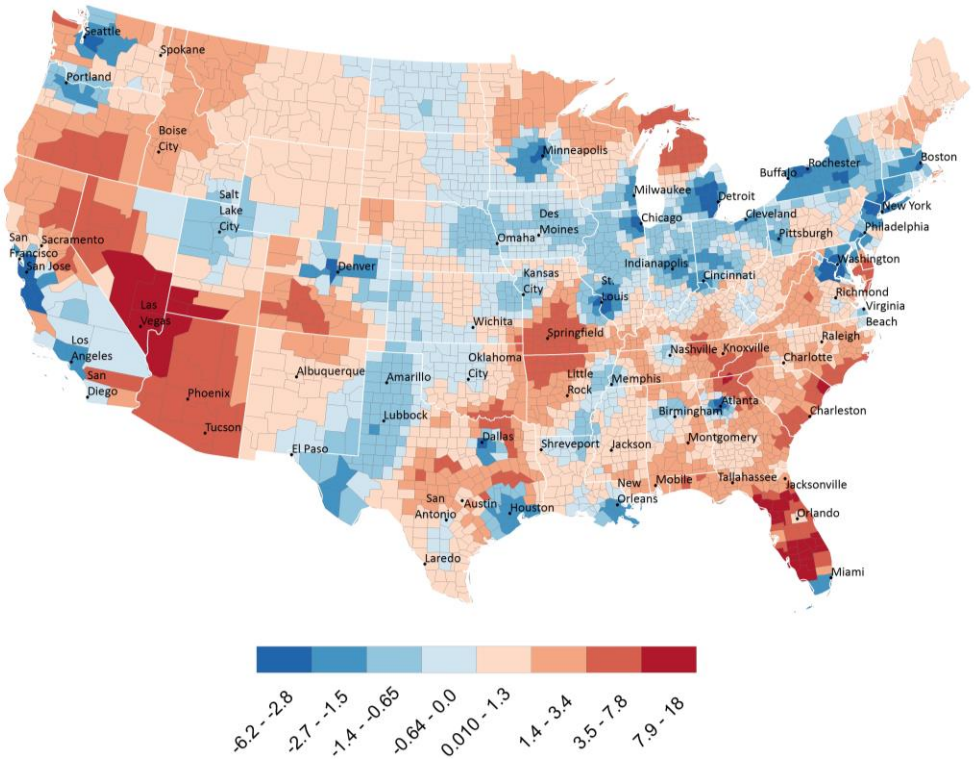



Figure A.6: Smoothed Net Migration Rate for age group 55-59

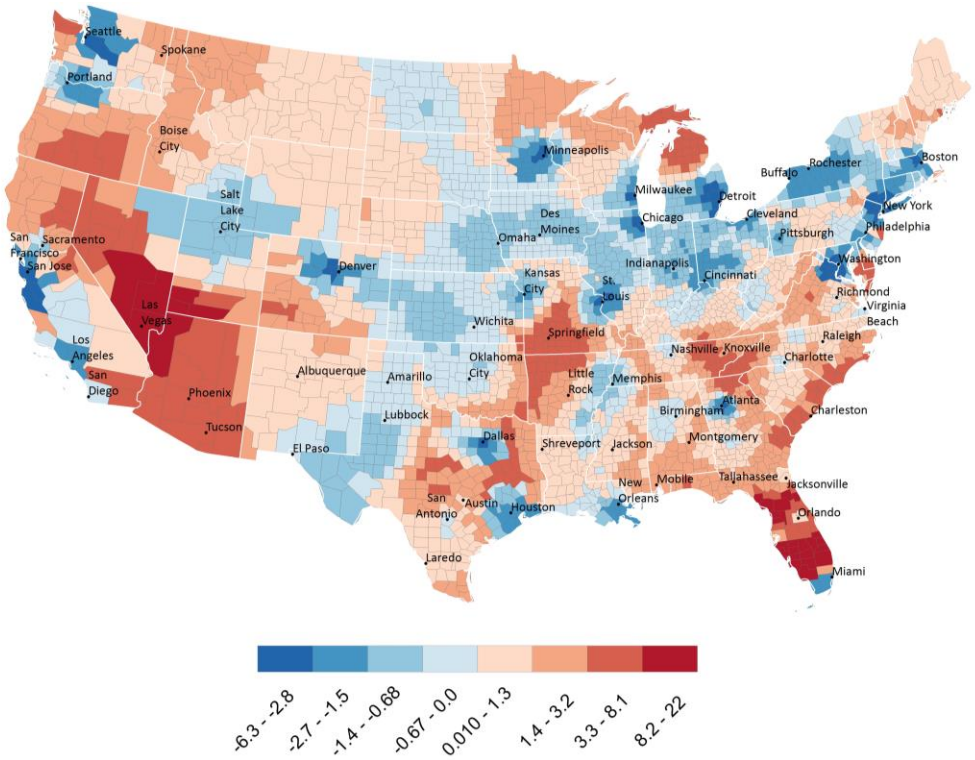



Figure A.7: Smoothed Net Migration Rate for age group 60-64

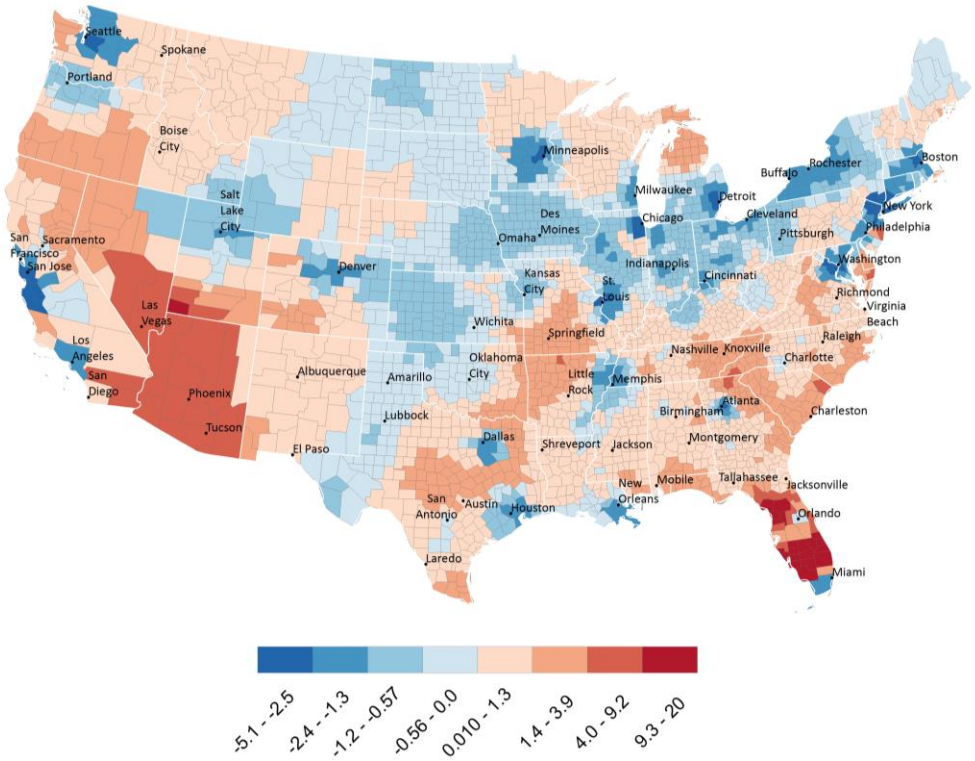



Figure A.8: Smoothed Net Migration Rate for age group 65-69

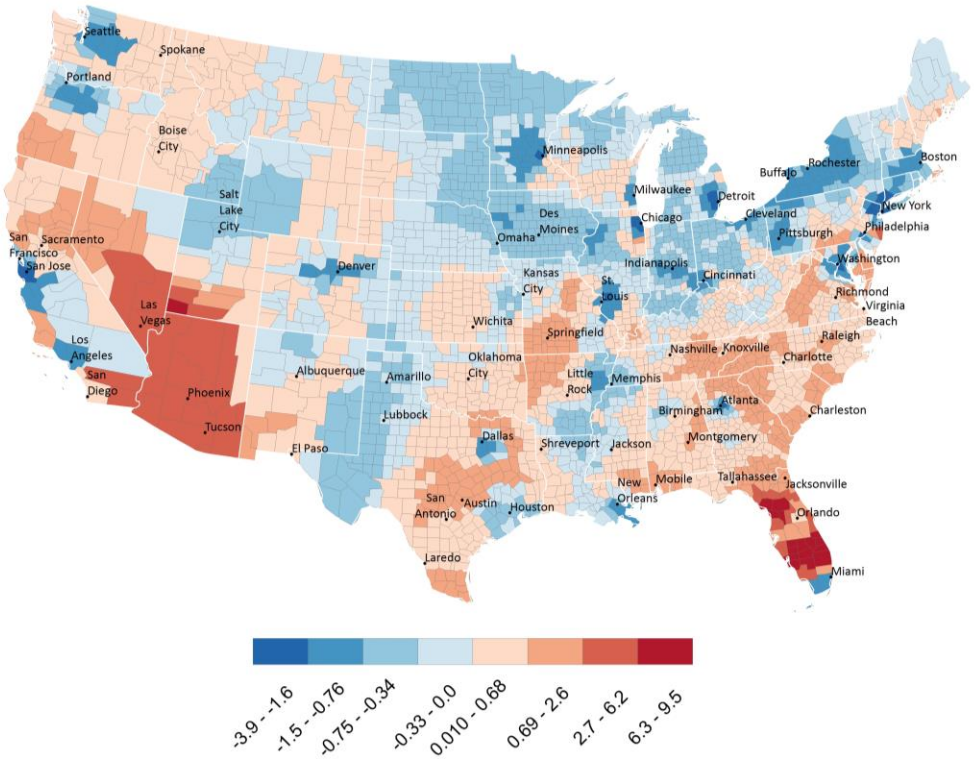



Figure A.9: Smoothed Net Migration Rate for age group 70-74

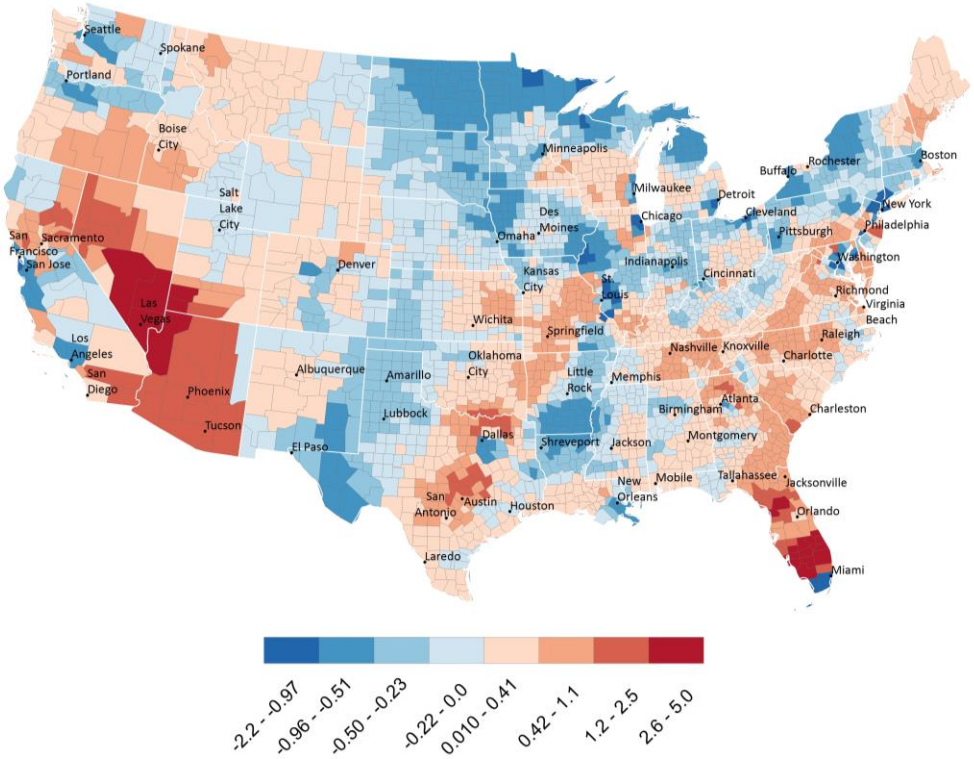



Figure A.10: Smoothed Net Migration Rate for age group 75-79

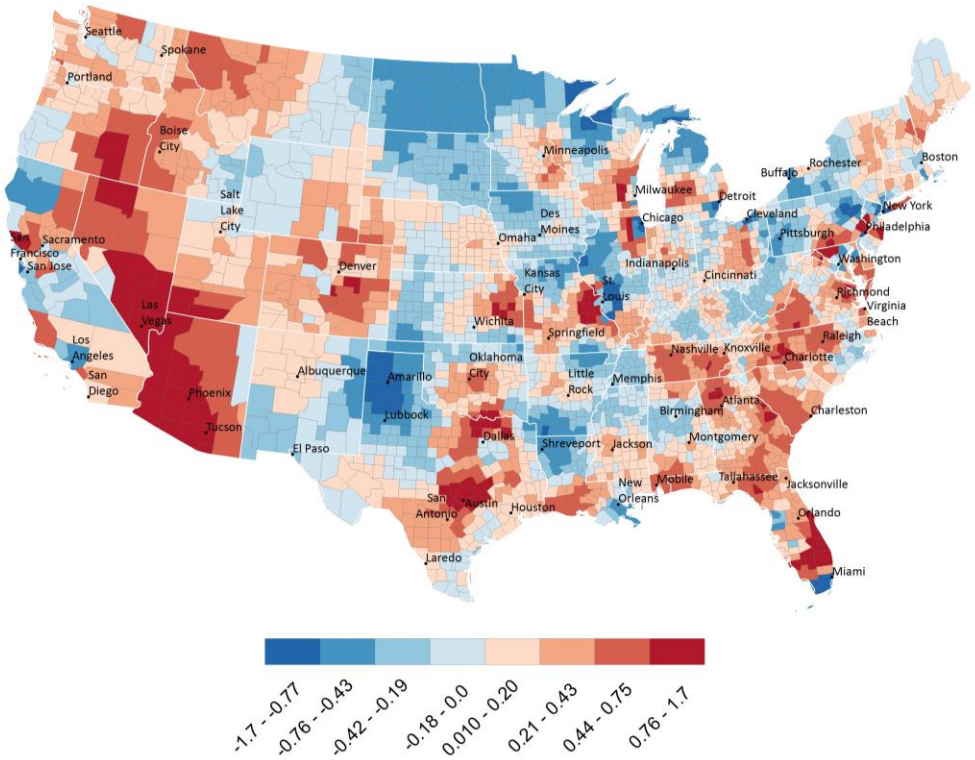



Figure A.11: Smoothed Net Migration Rate for age group 80-84

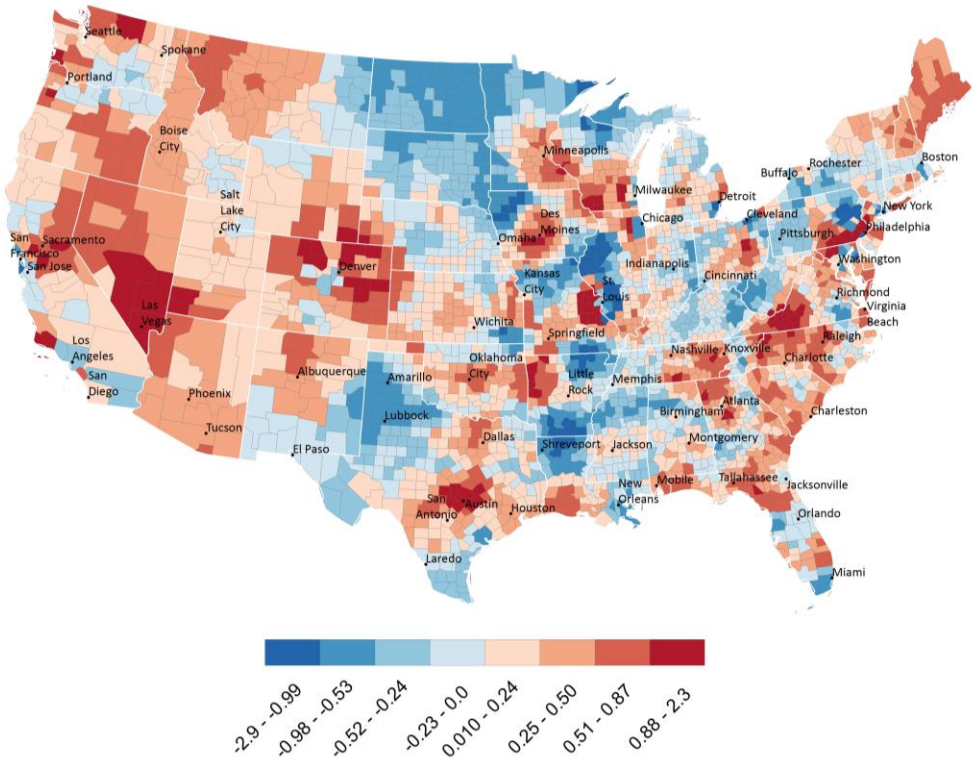



Figure A.12: Smoothed Net Migration Rate for age group >= 85

# APPENDIX B - COPYRIGHT PERMISSIONS



**ELSEVIER LICENSE**
**TERMS AND CONDITIONS**

Nov 24, 2014

This is a License Agreement between Caglar Koylu ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

**All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.**

| | |
|---|---|
| Supplier | Elsevier Limited<br>The Boulevard,Langford Lane<br>Kidlington,Oxford,OX5 1GB,UK |
| Registered Company Number | 1982084 |
| Customer name | Caglar Koylu |
| Customer address | |
| License number | 3515470638534 |
| License date | Nov 24, 2014 |
| Licensed content publisher | Elsevier |
| Licensed content publication | Computers, Environment and Urban Systems |
| Licensed content title | Smoothing locational measures in spatial interaction networks |
| Licensed content author | Caglar Koylu,Diansheng Guo |
| Licensed content date | September 2013 |
| Licensed content volume number | 41 |
| Licensed content issue number | n/a |
| Number of pages | 14 |
| Start Page | 12 |
| End Page | 25 |
| Type of Use | reuse in a thesis/dissertation |
| Intended publisher of new work | other |
| Portion | full article |
| Format | both print and electronic |
| Are you the author of this Elsevier article? | Yes |
| Will you be translating? | No |
| Title of your | Understanding Geo-Social Network Patterns: Computation, |

| | |
|---|---|
| thesis/dissertation | Visualization, and Usability |
| Expected completion date | Dec 2014 |
| Estimated size (number of pages) | 160 |
| Elsevier VAT number | GB 494 6272 12 |
| Permissions price | 0.00 USD |
| VAT/Local Sales Tax | 0.00 USD / 0.00 GBP |
| Total | 0.00 USD |