GEOVISUAL ANALYTICS APPROACHES FOR THE INTEGRATION OF GEOGRAPHY AND SOCIAL NETWORK CONTEXTS

A Dissertation in Geography
by
Wei Luo

© 2014 Wei Luo

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

August 2014
The dissertation of Wei Luo was reviewed and approved* by the following:

Alan M. MacEachren  
Professor of Geography  
Dissertation Advisor  
Chair of Committee

Donna Peuquet  
Professor of Geography

Frank Hardisty  
Research Faculty of Geography

John Yen  
Professor of Information Sciences and Technology

Cynthia Brewer  
Professor of Geography  
Head of the Department of Geography

*Signatures are on file in the Graduate School
Spatial analysis and social network analysis typically consider social processes in their own specific contexts, either geographical or network space. Both approaches demonstrate strong conceptual overlaps. For example, actors close to each other tend to have greater similarity than those far apart; this phenomenon has different labels in geography (spatial autocorrelation) and in network science (homophily). In spite of those conceptual and observed overlaps, the integration of geography and social network context has not received the attention needed in order to develop a comprehensive understanding of their interaction or their impact on outcomes of interest, such as population health behaviors, information dissemination, or human behavior in a crisis.

In order to address this gap, this dissertation discusses the integration of geographic with social network perspectives applied to understanding social processes in place at both theoretical and methodological levels. At the theoretical level, this dissertation develops a theoretical framework to integrate geographical context, network context, and societal context (e.g., political or economic background) to understand the geo-social interaction in certain societal contexts. The framework extends the concepts of nearness and relationship in terms of the First Law of Geography as a matter of both geographical and social network distance, relationship, and interaction. At the methodological level, the integration of geography and social network contexts is framed within a new interdisciplinary field: visual analytics, in which three major application-oriented subfields (data exploration, decision-making, and predictive analysis) are used to organize discussion. In each subfield, this dissertation presents a theoretical framework first, and then reviews what has been achieved regarding geo-social visual analytics in order to identify potential future research.

This dissertation also develops two novel geo-social visual analytics tools to study the complex interaction between spatial and social relationships at different geographical levels: the regional level (e.g., country, state, or county) and the individual level (e.g., person, or
organization). The first tool, called GeoSocialApp, uses data for international trade networks among different countries to empirically study the interaction of spatial-social relationships across geographical regions at multiple levels of network hierarchy: http://www.geovista.psu.edu/GeoSocialApp/. The second tool, GS-EpiViz, takes network statistics relevant to air-borne disease transmission and control and integrates them into appropriate visualization techniques, thereby facilitating the exploration of human interaction network structures to design advanced disease control strategies. This tool also implements agent-based epidemic models to test and evaluate control strategies in a highly interactive and iterative manner. The two tools provide generic frameworks to explore spatial-social relationships at geographical scales, ranging from individual-level to national-level.

The research reported here opens a new research area: geo-social visual analytics and achieves a substantial step forward regarding the science and technology of this area. However, there is much more research that needs to be done. In the meantime, the insights generated in this research provide an initial foundation for the future scientific research and technological challenges on geo-social visual analytics.
TABLE OF CONTENTS

List of Figures ...................................................................................................................... vii
List of Tables ........................................................................................................................ xi
Acknowledgements ............................................................................................................. xii

Chapter 1 Introduction ........................................................................................................... 1
Chapter 2 Geo-Social Relationships at a Conceptual Level ............................................... 4
Chapter 3 Geo-Social Visual Analytics at a Methodological Level .................................... 12
  Data Exploration .................................................................................................................. 13
  Conceptual Framework ........................................................................................................ 13
  Geo-Social Visual Analytics in Data Exploration ............................................................... 15
  Decision-Making ................................................................................................................ 23
  Conceptual Framework ........................................................................................................ 23
  Geo-Social Visual Analytics in Decision-Making ............................................................... 25
  Predictive Analysis .......................................................................................................... 33

Chapter 4 A Geovisual Analytic Approach to Understanding Geo-Social Relationships in the International Trade Network ................................................................. 38
  Introduction ......................................................................................................................... 38
  Literature Review .............................................................................................................. 40
  Methods ............................................................................................................................... 43
    GeoSocialApp Components ............................................................................................. 44
  Data .................................................................................................................................. 47
  Results ............................................................................................................................... 49
    Spatial and Social Interaction at the First Level of CONCOR ............................................ 49
    Spatial and Social Interaction at the Second Level of CONCOR ....................................... 50
    Spatial and Social Interaction at the Third Level of CONCOR ......................................... 51
  Validation ........................................................................................................................... 53
    Degree of balkanization .................................................................................................. 54
    Pearson correlation ......................................................................................................... 55
    Validation Method .......................................................................................................... 57
    Robustness of the validation method ............................................................................... 60
  Conclusion & Contribution ............................................................................................... 62

Chapter 5 Agent-based GeoSocial Visual Analytics for Epidemic Control ....................... 68
  Introduction ......................................................................................................................... 68
  Related Research .............................................................................................................. 70
Agent-based Epidemic Models and Vaccination Strategies..........................71
Visual Analytics in Agent-Based Epidemic Models.................................72
Data .............................................................................................................73
Methodology: Agent-Based GeoSocial Epidemic Visual Analytics (GS-EpiViz) ......75
Results and Discussions..............................................................................83
Vaccination Strategies in terms of Infection Percentage..............................85
Vaccination Strategies in terms of Spatial-Social Extent of Infection............91
Conclusion and Future work ......................................................................94

Chapter 6 Discussion & Challenges..............................................................97

An Effective Interdisciplinary Integration of Spatial and Social Network Analytics.....98
Understanding the interaction of geography, network, and societal space, as
well as their respective and coupled impacts on outcomes of interest ..........98
Developing theory, methods, and tools to consider spatial and network factors
simultaneously ...............................................................................................99
Understanding the dynamics of geo-social relationships and processes ..........100
Integrating distinct applications of cognitive science to support geo-social
visual analytics ..............................................................................................101
Developing new geo-social visual analytics methods to incorporate data
exploration, decision-making, and predictive analysis as a whole .............102

The Impact of Geo-Social Relationships on Society ....................................103
Understanding the transition of daily life habits that further impact local
communities and networks because of spatial decision-making ...............103
Understanding the shift of the traditional decision-making approaches with the
emergence of social media ........................................................................104
A new concept of scale? .............................................................................105
Developing a framework to collect geo-social relationship data and assess their
fitness for different applications while also considering the potential
negative consequences for human privacy of collecting these data ..........105
References .................................................................................................107
LIST OF FIGURES

Figure 2-1: Proposed conceptual framework for geo-social relationships based on fundamental categories of embeddedness (Hess 2004). The framework consists of three kinds of embeddedness: territorial, network and societal. The paired overlaps each match with specific perspectives, as indicated and the joint overlap of all three areas suggests the extension of the First Law of Geography. ..............................................7

Figure 3-1: Feature-ID model of geovisualization; Source: Figure 8.1 in (MacEachren 1995). The model focused on identifying key components in visually-enabled reasoning about geographic phenomena and relationships. .................................................................14

Figure 3-2-a: State to state migration from 1995 to 2000. Adapted from Figure 7 in (Tobler 1987) with Flow Mapper program from CSISS.org. Figure 3-2-b: Flow mapping and multivariate visualization of large spatial interaction data. Figure courtesy of Diansheng Guo. For additional information on the research it is derived from, see (Guo 2009). ................................................................................................17

Figure 3-3: The analysis of international trade data among 192 countries in 2005 with GeoSocialApp: Node-Link (Upper Left), Dendrogram (Lower Left), Bivariate Choropleth Map (Upper Right) and Parallel Coordinate Plot (Lower Right). For a preliminary view of this tool, see Luo, et al (2011). Each node in the node-link view and the dendrogram view corresponds to one country in the bivariate choropleth map; the Parallel Coordinate Plot view is used to explore network attributes for countries (from left to right, the attributes depicted are: indegree, outdegree, closeness, eigenvector, and clustering coefficient). Red dots in the node-link and dendrogram view are periphery countries and these are highlighted on the map and Parallel Coordinate Plot with blue outlines.................................................................20

Figure 3-4: The process of decision-making for situation awareness. Situation awareness is defined as the human user's internal conceptualization of a situation, which plays a key role in effective decision making. Figure adapted from (Kohlhammer et al. 2009). Spatial data analysis and social network analysis can support decision-making from different perspectives, so the effective integration of both can enhance human user’s internal conceptualization of a situation. .................................................................25

Figure 3-5: Individual linked map-timeline and social network analysis views in TwitterHitter. The left view allows analysts to retrieve a spatiotemporal record of a suspect’s activity on twitter. The right view can uncover all potential connections among suspects through network group of twitterers within a region. They are courtesy of Jeremy D. White from Figures 3 and 4 in (White and Roth 2010). ............28

Figure 3-6: The static display of the grapevine. A main stem grows up with the increasing number of messages over geographical space. The main stem generates nodes when the number of votes to messages increases. The nodes generate buds when the number of replies (thus social connections) to messages increases. It shows the main stem in four states: (A) all features are turned off; (B) nodes are turned on; (C) the number of nodes signifies the number of votes; (D) the size of buds is according to the number of replies. Reproduced from Figure 4 in (Aguirre and Nyerges 2011). ..........32
Figure 3-7: Here I illustrate the PanViz user interface. A user has simulated an outbreak originating in Chicago and has been exploring the spread of the infectious disease over space and time. Day 21 of the simulation is depicted in the figure. The color of each county represents the percentage of the population that is ill. The plots on the right show the rate of spread (thus the rate of person-to-person connection), hospitalization, and death in user selected counties. Figure Courtesy of Ross Maciejewski from Figure 7 in [127]. For Additional Information on the research it is derived from, See (Maciejewski et al. 2011).

Figure 4-1: Dendrogram View: two layouts to visualize the hierarchical structure of CONCOR results: the left one is a tree layout and the right one is a radial layout. Slider bar is used to control the level of CONCOR results.

Figure 4-2: Dendrogram view and choropleth map view. The choropleth map depicts GDP by country. Data are divided into quintiles (5 categories with an equal number of countries in each category) depicted by 5 sequentially ordered shades of green, from low GDP (very light green) to high GDP (very dark green). Each node in the dendrogram view corresponds to one country in the choropleth map view (The highlighted nodes in blue correspond to countries with borders highlighted in blue). The first run of CONCOR process reveals two positions in the 2005 ITN.

Figure 4-3: The second run of the CONCOR process subdivides each of the first two groups. Figure 4-3A: One subgroup of economically core countries; Figure 4-3B: The other subgroup of economically core countries; Figure 4-3C: One subgroup of economic periphery countries; Figure 4-3D: The other subgroup of economic periphery countries.

Figure 4-4: The third run of the CONCOR process continues to subdivide groups. Figure 4-4A, 4-4B, and 4-4C belong to the economically core countries, whereas Figure 4-4D, 4-4E, 4-4F, and 4-4G belong to the economic periphery countries.

Figure 4-5: The degree of balkanization of all random trade networks.

Figure 4-6: The Pearson correlation values between GDP and weighted within-cluster distance of all random trade networks.

Figure 4-7: Validation results as a function of number of runs (N).

Figure 4-8: Validation results as a function of total connection numbers.

Figure 5-1. GS-EpiViz consists of four major components: display panel, control panel, xy plot, and matrix view. Display panel in the top left allows users to select different matrix displays: binary matrix, time duration matrix, reorder binary matrix, and reorder time duration matrix. Control panels allow users to design different control scenarios based on the whole population and selected population. XY plot displays the accumulated infected cases over time based on the simulation of agent-based epidemic models. The matrix view visualizes different matrix displays with class information on the bottom and on the right. Black cells in the matrix view represent
human interactions at the individual level, and white cells indicate there are no
human interactions. .......................................................................................... 76

Figure 5-2. GS-EpiViz displays the reordered binary network with new class numbers on
the bottom and on the right. .............................................................................. 77

Figure 5-3. GS-EpiViz displays the time duration network matrix with the average time
duration values in each cell............................................................................. 78

Figure 5-4. GS-EpiViz displays the reorder time duration network matrix with the
average time duration values in each cell .......................................................... 79

Figure 5-5. GS-EpiViz displays a simulated influenza infection according to the
parameters in Table 5-3 on the matrix view and XY plot. Yellow indicates that
individuals are in an exposed state, red indicates an infectious state, and the blue
indicates a recovery state. .................................................................................. 81

Figure 5-6: GS-EpiViz displays the selected population (in orange) to which the designed
control scenarios apply...................................................................................... 83

Figure 5-7-b. The design of 8 vaccination strategies with six different control rates in the
control panel. The matrix view visualizes the reordering binary matrix network,
Thursday, October 1st 2009 in (a) and Friday, October 2nd 2009 in (b). The local
control areas are highlighted in orange ............................................................. 85

Figure 5-8. The efficacy of 8 vaccination strategies in two networks, Thursday, October
1st 2009 in (a) and Friday, October 2nd 2009 in (b). The Y axis represents the percent
of infection, and the X axis represents the vaccination rates. Eight vaccination
strategies are represented by eight curves in different colors and shapes, as the
legend shows at the bottom ............................................................................... 87

Figure 5-9. The percentage of the local control success for four local vaccination
strategies within 100 simulation runs in two networks, Thursday, October 1st 2009 in
(a) and Friday, October 2nd 2009 in (b). The Y axis represents the percentage of the
local containment success, and the X axis represents the vaccination fraction. Four
local vaccination strategies are represented by four curves in different colors and
shapes, as the legend shows at the bottom .......................................................... 88

Figure 5-10. The total number of the final infected cases when the local vaccination
strategies are successful within 100 simulation runs in two networks, Thursday,
October 1st 2009 in (a) and Friday, October 2nd 2009 in (b). The Y axis represents
the total number of the final infected cases, and the X axis represents the vaccination
fraction. Four local vaccination strategies are represented by four curves in different
colors and shapes, as the legend shows at the bottom ......................................... 89

Figure 5-11. The total number of infections when the local vaccination strategies are not
successful within 100 simulation runs in two networks, Thursday, October 1st 2009
in (a) and Friday, October 2nd 2009 in (b). The Y axis represents the total number of
the infected cases, and the X axis represents the vaccination fraction. Four local
vaccination strategies are represented by four curves in different colors and shapes, as the legend shows at the bottom.

Figure 5-12. The spatial-social patterns of simulated vaccination strategies with 30% vaccination rate with the average number of infections based on 100 runs after reordering the first day network, Thursday, October 1st 2009. Figure 5-12-a is the random-based vaccination strategy, Figure 5-12-b is the degree-based vaccination strategy, Figure 5-12-c is betweenness-based vaccination strategy, and Figure 5-12-d is the strength-based vaccination strategy.

Figure 5-13. The spatial-social patterns of simulated vaccination strategies with 30% vaccination rate based on the selected areas (purple square) after reordering with the average number of infections in each cell in the first day network, Thursday, October 1st 2009. The number of infections within the selected areas is achieved through calculating the sum of infected individuals on each of the 100 runs and then dividing by 100 runs. The average number of infections outside the selected areas is calculated through dividing by the number of local control failures, because the number of infections is zero when the local control is successful. Figure 5-13- a is the random-based vaccination strategy, Figure 5-13-b is the degree-based vaccination strategy, Figure 5-13-c is betweenness-based vaccination strategy, and Figure 5-13-d is the strength-based vaccination strategy. The pie chart in each figure represents the number of local control success versus the number of local control failure with 100 simulation runs. The bar chart in each figure represents the average number of infections between the local control success and local control failure.
LIST OF TABLES

Table 4-1 Imports-exports relationship among partial countries in 2005. Flow1 means imports of importer1 from importer2 in current US millions of dollars, and flow2 means imports of importer2 from importer1 in current US millions of dollars. .............48

Table 4-2 International trade relationships among nine sample countries in a binary matrix for 0% threshold in 2005. ..............................................................48

Table 4-3: CONCOR group level attribute data. Mean GDP in 2005 for 4 groups identified at the second level of the CONCOR, mean GDP in 2005, mean distance, weighted distance by population for 7 groups at the third level of the CONCOR........57

Table 5-1. Partial node table with two attributes: Label indicates the id of each node, and classname indicates the corresponding class groups, including grade level and class number for students and “Teachers” for teachers. ....................................................74

Table 5-2. Partial edge table with 5 columns: Source indicates the id of source node, Target indicates the id of Target node, and Duration indicates that the cumulative time between two nodes measured in seconds within one day. ........................................74

Table 5-3: Basic Simulation Parameters for influenza diffusion .................................................80
ACKNOWLEDGEMENTS

I would not have been able to complete this dissertation without help from many faculty members and students in the Penn State University. First and foremost I would like to thank Alan MacEachren. I am so fortunate to have Alan MacEachren as my advisor. Recalling the Ph.D study in the past 5 years, I would say that I even could not complete my Ph.D study without Alan MacEachren. As one faculty told me if Alan MacEachren agree to be your advisor, he would never give you up. He really did never give me up even during my most depressed time period for my Ph.D study. From the bottom of my heart, I would say that he is my consummate academic father. His patient guidance, continuous encouragement, and invaluable advice make me finally achieve the academic success. His incredible time and energy devotion to research, teaching and students set a life-long high standard for me to follow. He does not only deserve much of the credit for my current achievement, but also my future success.

I also wish to thank my committee members Donna Peuquet, Frank Hardisty, and John Yen for their informative input and careful feedback at the proposal, comprehensive exam, and defense stages of my doctoral progress, as well as for the advice and opportunities they have provided to me on projects outside of the dissertation. In addition to my committee, I would like to say thanks to Brent Yarnal, Deryck W. Holdsworth, and James McCarthy for their efforts to broaden and deepen my knowledge on human, physical and human environment geography. Thanks to Cindy Brewer, the Head of the Department, for carefully reading my dissertation and providing valuable corrections.

A project of this size is not without many helping hands; I have enlisted the help of many friends and colleagues during the dissertation research. Peifeng Yin has been my partner on the GeoSocial project from its initial inception as a classroom project and subsequent extension into
one important component of my dissertation. Di Qian also has provided important contributions
to the validation of GeoSocial that deserve noting.

I also would like to say thanks to my colleagues who offer their generous help after I
failed in my candidacy exam for the first time. They are Maureen Biermann, Vanessa Massaro,
Mark Read, Raymond Tutu, Seth Baum, Greg Lankenau, and Kieron Branche. I would also like
to thank my GeoVISTA colleagues for creating an exciting atmosphere of innovation and
collaboration.

Finally, my special thanks go to my extended network of family and friends, too
countless to name, for their passionate and unconditional support.

A person cannot achieve too much without help from others. You are all my friends who
are my life-long fortune. Much love to you all.
Chapter 1

Introduction

Modern society has become an increasingly interconnected world of techno-social systems embedded with dynamic multi-scale networks (e.g., the internet, transportation). The complex interactions within and among these networks always have geographical constraints, whereas they also change or reshape the traditional notion of geographical effects (e.g., distance) (Kwan 2007). For example, small social groups are usually geographically cohesive (with the geographic span of group members a function of group size); but large social groups are less cohesive and less likely to exhibit spatial clusters (Onnela et al. 2011). Furthermore, because of advances in communication and transportation technologies over the past decade, there is a shift from networks that are both geographically and socially close (e.g., physical communities) to networks that are socially near but geographically dispersed (e.g., on-line communities) (Urry 2003, Wellman 2002). To effectively understand the complex interaction between space and techno-social networks, it is necessary to encourage interdisciplinary understanding (e.g., geography, network science) through integrating current theories and methods (e.g., spatial thinking, complex system theory) and to develop new theories and methods.

Recent research in physics emphasizes the power of networks in which space becomes a background to visualize and understand network analysis results (Thiemann et al. 2010) whereas research in geography encourages the integration of spatial thinking into traditional social science through the concepts of space, place, and time (Goodchild and Janelle 2010, Goodchild et al. 2000), but often treats networks in a simplistic way. This dissertation argues that space and social networks should be considered simultaneously when framing research on human activity. This research contends that this perspective has not received enough attention and provides examples
throughout the dissertation to support this contention. This research further contends and presents evidence that the new multidisciplinary research field of visual analytics provides an approach and methods that are well suited to understanding the interaction of geographical and social network contexts. Visual analytics is defined as “the science of analytical reasoning facilitated by interactive visual interfaces” (Thomas and Cook 2005, p. 4). A core objective of research in visual analytics is to provide a framework for integration of computational analytical methods with visual interfaces to both the information of interest and the computational methods that enable human analysts to cope with large, complex, and heterogeneous data sources and complex questions that these data sources make it possible to address.

Understanding large and complex techno-social networks and their interaction with space at geographic scales requires advances in computational methods. However, computational methods alone have limits and biases because of the predefined structures they have, which greatly limit their analytical power. The process and results of any computational techniques have limited value without input from human analysts to select appropriate methods, to set parameters, to interpret results, to understand what to do next, and to draw conclusions (Andrienko et al. 2008). Visualization of data and computational processing gives users an intuitive representation, greatly promoting application of human perceptual and cognitive information processing capabilities. A simple combination of visualization with computational analysis, however, is not sufficient. Thus, the goal for visual analytics is to integrate human and computational reasoning in more fundamental ways, bringing the experts’ background knowledge, creativity, and intuition into the analysis process through an interactive visual environment, in order to combine the strengths of humans and computers to enable an insight gaining process (Keim et al. 2011).

Visual analytics provides a potential conceptual approach and set of tools to integrate the geographic and social network contexts of human processes, but the application of visual analytics to this challenge remains relatively underdeveloped in the literature. Thus, the goal of
this dissertation is to provide a base from which to develop and apply visual analytics methods to examine the interaction of both contexts and enable understanding of human processes. To achieve this goal, this dissertation address six objectives: (1) to present a theoretical framework in which geography and social network contexts can be combined and through which visual analytics methods can be developed and applied (Chapter 2) (Luo and MacEachren 2014), (2) to introduce “geo-social visual analytics” as an integrated analytical approach grounded in the conceptual framework (Chapter 3) (Luo and MacEachren 2014), (3) to review what has been achieved in relation to the integration of geographic and social network contexts in relation to three core tasks for geo-social visual analytics: data exploration, decision-making, and predictive analysis (also Chapter 3) (Luo and MacEachren 2014), (4) to develop and assess a geosocial visual analytics tool at a geographical region level (e.g., county, state, and country) to understand the role of spatial proximity in shaping international trade network (ITN) across different geographical regions (Chapter 4) (Luo et al. 2014, Luo et al. 2011), (5) to develop and assess a geosocial visual analytics tool at the individual level to allow the design and testing of advanced control scenarios in airborne disease (e.g., influenza) with individual-based epidemic models (Chapter 5), and (6) to identify potential future research challenges for advancing geo-social integration (Section 6) (Luo and MacEachren 2014). From here on, this dissertation will refer to the interaction of geographical and social relationships as geo-social relationships, and the interaction between both in terms of visual analytics as geo-social visual analytics. The key distinction between geo-social visual analytics and prior work in geo-visual analytics (Andrienko et al. 2011) is the explicit integration of social network perspectives and methods into the approach and tools.
Geo-Social Relationships at a Conceptual Level

The theoretical framework this dissertation proposes as a base upon which to develop and apply visual analytics methods that enable understanding of human processes draws upon a wide range of perspectives from human geography and social sciences more generally. One theme that connects the perspectives is that social processes take place within particular contexts. The two focused on here are spatial and social network contexts, each of which has been addressed in the past with specific methods and perspectives. Spatial analysis in social science is used to identify geographical patterns that result from social processes and to understand how space affects such processes. Most spatial analysis is based on an explicit or implicit assumption of the First Law of Geography (Tobler 1970, p. 236): “Everything is related to everything else, but near things are more related than distant things.” Social network analysis is used to understand how relationships among actors (i.e., individuals, groups, or other social collectives) within a network affect or are affected by social processes (Valente 2010). Social network analysis has an assumption, complementary to the First Law cited above, that actors with similar relations may have similar attributes/behaviors. Spatial analysis and social network analysis consider social phenomena in their own specific contexts, either geographical space or network space, but as being outlined below, considering both contexts together when they contribute simultaneously has the potential to achieve new insights about human processes.

1 Portions of this chapter have been previously published in Luo and MacEachren (2014). I led this research and, as stated in the published paper, my responsibilities and that of my co-authors in the research and the paper were as follows: Conceptualized the paper: WL. Wrote the paper: WL AMM.
Both contexts demonstrate strong conceptual overlaps. Hess (2004) proposes a geographically informed theoretical framework to understand the behaviors of social actors in geography through integrating territorial embeddedness, network embeddedness and societal embeddedness (Figure 2-1). This framework is used to study economic actions from a critical human geography perspective, but this dissertation aims to extend it into a generic framework to study geo-social relationships. The concept of embeddedness has been prominently used by geographers to understand the behaviors of social actors in specific contexts (Radil, Flint and Tita 2010). Societal embeddedness refers to societal (i.e., cultural, political, etc.) background from which actors come, in which actions of actors are influenced, and to which actors contribute. Network embeddedness refers to the importance of relational aspects (i.e., social relations, cultural relations) among social actors to shape the actors’ behaviors and of actors’ behaviors to change relations. Territorial embeddedness refers to the specific places in which the actors behave: how the places influence actors’ behaviors and attributes; how the actors’ behaviors change the territory. The overlap area in Figure 2-1 between the territorial embeddedness and the societal embeddedness fits the First Law of Geography (Tobler 1970). The overlap area between the network embeddedness and the societal embeddedness fits the homophily principle in social network analysis theory: similarity breeds connection (McPherson, Smith-Lovin and Cook 2001). The overlap area between the territorial embeddedness and the network embeddedness fits a common phenomenon: how space constrains the development of networks and how networks reshape the space. The three overlap areas in Figure 2-1 indicate that they are not mutually exclusive, but interact with each other; the emphasis is on the interaction between geographic and social network context and the impact of the interaction on outcomes of interest. The overlap area at the center suggests a possible extension of the First Law of Geography: “Everything is related to everything else, but near things are more related than distant things (Tobler 1970, p. 236).”
Nearness can be considered a matter of geographical and social network distance (Flint et al. 2009).

In addition to distance, the other two important concepts implied in the First Law of Geography are relationship and interaction: "everything is related to everything else." Flint (2002, p. 33) argues that “the nature of a place is the combination of both locations and their connections to the rest of the world.” Prager (2008) also argues that geography of relationships and interactions is complementary with geographical locations, whereas such relationships and interactions would be meaningless without the context geographical locations provide. Some of those social relationships may be constrained within the place, whereas others may stretch out to link geographical locations to wider relations and processes (Massey 1994). There is an increasing understanding of the importance of combining geographical space and networks from different sub-disciplines within geography, such as political geography (Leitner, Sheppard and Sziarto 2008), economic geography (Sheppard 2002), and geographical information science (GIS) (Bera and Claramunt 2003). Staeheli (2003, p. 160) argues that spaces become “social locations” embedded in “webs of cultural, social, economic, and political relationships.” Ashdown (Ashdown 2012) even argues that the political power is now shifting from a dominance of western culture to a collective governance at the global scale, because we have come into a new interlocked age (i.e., of complex, interconnected networks) that causes our destinies to be shared with our enemies. Andris (2014) argues that though the combination of social network analysis and GIS is necessary, social network relationships are rarely modeled with GIS environment.
Figure 2-1: Proposed conceptual framework for geo-social relationships based on fundamental categories of embeddedness (Hess 2004). The framework consists of three kinds of embeddedness: territorial, network and societal. The paired overlaps each match with specific perspectives, as indicated and the joint overlap of all three areas suggests the extension of the First Law of Geography.

Taking a perspective that complements those cited above, Massey (2004) explores relationships between identity of place (and group and individual) and responsibility of place (and group and individual) and the geographical components of each. From the perspective of this chapter, an important component of Massey’s overall conceptual argument (the details of which are well beyond our focus here) is that connections to other locations are an essential component to understanding both the identity and responsibilities of places (and the groups and individuals connected to those places). By linking individual location to group location to place, the science of social networks can explain key aspects of how observed spatial-social patterns evolve. Therefore, to know the spaces, it is very important to understand how spaces are connected geographically and socially. This contention may be increasingly true in today’s digitally hyper-connected world. We are now living life in even more complex and interrelated networks: as we
check our e-mail, make a phone call, take transportation, or update our status in Facebook (Lazer et al. 2009). The lifestyle causes the emergence of the new human geo-social relationships: people with strong emotional ties may live geographically far away (Larsen, Axhausen and Urry 2006). Christakis and Fowler (Christakis and Fowler 2009) make a list of certain rules regarding networks: people shape their network, networks shape people, their friends affect them, their friends’ friends’ friends affect them, and a network has its own life. The inherent network structures in our lives affect our ideas and behaviors (i.e., emotional, sexual, and health-related), and the interaction of such individual-level behaviors develops macrosocial phenomena observed in a spatio-temporal framework.

Massey, in a report edited by Urry, et al (2007) goes on to argue that understanding the world from a ’social’ perspective is to use ’how we are going to live together’ to motivate all social, political, and ethical questions framed in spatio-temporal frameworks. This social perspective aims to develop general social theory to address today’s highly interconnected society. Network science (Buchanan 2003) is regarded by many authors as the approach to study such phenomena (Watts 2004, Watts 2003). For example, current research projects from the Santa Fe Institute focus on studying the underlying principles and mathematical relationships concerning the evolution of human society and modern human social organizations (i.e., cities) with a network approach (Bettencourt and West 2010, Bettencourt et al. 2007). A growing literature in network science explores how everything is related to everything else (Barabasi 2002, Johnson 2012, Strogatz 2003). For example, the famous small world experiment (Milgram 1967) formalizes the notion that each person only has six degrees of separation from anyone else on earth (Kochen 1989). Sui (2004) further argues that Tobler’s First Law of Geography is a big idea for a small world, because it only takes a few steps to turn a large world into a small world. Based on all of the discussions, this dissertation argues that geo-social relationships in terms of the First Law of Geography should be extended into the notion: everything/everyone is related to
everything/everyone else, but near things/individuals are more related than distant things/individuals; Nearness and relationship can be considered a matter of geographical and social network distance, relationship, and interaction.

In addition to the exploration of geographical and social relationships and interactions, the conceptual framework (Figure 2-1) allows such relationships to be put into particular societal contexts (i.e., political, economic, cultural). On the one hand, the premise “near things are more related than distant things” in the First Law of Geography implies that certain local factors and circumstances can make geo-socially close areas different from geo-socially distant areas. For example, spatial proximity together with connections that link spatially heterogeneous groups in the population (i.e. city-wide travel of select individuals) are two major factors that determine the spatial layout and the temporal sequence of disease transmission (Mao and Bian 2010b). And, recent research by Onnela, et al (Onnela et al. 2011) suggests that the size at which spatial cohesion of groups breaks down (about 30 members) coincides with the optimal group size for cooperation in social dilemma situations. On the other hand, the premise “everything is related to everything else” indicates that there are multiple factors that make contributions to patterns and connections among areas. For example, geographical homophily has a major impact on international trade among developing countries, whereas political and cultural homophily matters the most for bilateral trade between developed and developing countries (Kali and Reyes 2007, Zhou 2013). The societal context provides the framework to explore different factors (i.e., political, economic, cultural) behind observed geo-social patterns, and how such geo-social patterns interact with those factors to generate new geo-social patterns.

In work that complements the discussion above of geo-social integrations at the conceptual level, Adams, Faust and Lovasi (2012) identify five conceptual strategies for the integration based on current geo-social relationship research: (1) spatial impacts on the development of social networks over varying spatial scales, such as offices (Sailer and McCulloh
2012, Wineman, Kabo and Davis 2009), communities (Daraganova et al. 2012, Festinger, Schachter and Kurt 1950), and so on; (2) the impact of social network on the places people select to inhabit (Verdery et al. 2012); (3) the use of peer network structures to determine neighborhood boundaries (Hipp, Faris and Boessen 2012); (4) the interactive impacts between spatial and social relationships (Preciado et al. 2012, Schaefer 2012, Lomi and Pallotti 2012); and (5) multiple context impacts on outcomes related to social, health, and other processes (Mennis and Mason 2012, Takhteyev, Gruzd and Wellman 2012, Doreian and Conti 2012). The five conceptually geo-social integrations also fit the conceptual framework (Figure 2-1). The first integration focuses more on spatial constraints; the second emphasizes the network effects on residence selection; the third and fourth stress the interaction between geographical and network relationships; and the last highlights the multiple context impacts on outcomes.

The conceptual framework introduced here builds upon ideas that derive from a critical human geography perspective. Here, this dissertation adapts it to inform the application of visual analytics to the study of geosocial processes. Previous research already provides examples of integrating critical human geography perspectives into visualization, e.g., in feminist visualization (Kwan 2002) and grounded visualization (Knigge and Cope 2006). Knigge and Cope (2006) present six connections between grounded theory and visualization, which are adopted here to integrate the proposed geo-social relationship conceptual framework and visual analytics at the theoretical level. First, the proposed conceptual framework and visual analytics are exploratory approaches which involve iterative explorations with matching mental models to construct knowledge. Second, both are iterative approaches that involve recursive processes of data collection, visualization, and analysis with critical thinking at each step. Third, both methods are enriched through connecting real-world phenomena and human experiences to broader processes. Fourth, both methods facilitate multiple interpretations and representations because of the notion that there is no single correct way to interpret and visualize data (Slocum 1999); Fifth, both
methods acknowledge the importance of situating knowledge construction into the historical, geographical, and cultural context; Lastly, both methods recognize the existence of uncertainty in the data.
Chapter 3

Geo-Social Visual Analytics at a Methodological Level

Building upon the introduction to geo-social relationships at a conceptual level, this chapter discusses how to put the geo-social relationships into practice. From an application perspective, visual analytics methods can be classified into three groups focused on support for: data exploration, decision-making, and predictive analysis. This classification is used here to organize the methodological level in terms of geo-social visual analytics. Visual analytics aims to enable the human reasoning process, so it must build on an understanding of that reasoning process (Thomas and Cook 2006). For the above three categories in visual analytics, this chapter presents reasoning frameworks related to geo-social perspectives for each group first, and then discusses corresponding geo-social visual analytics technologies in order to identify potential future research directions. In the next two chapters (Chapter 4 and Chapter 5), this dissertation develops and assesses two geosocial visual analytics tools with two application domains (i.e., international trade and public health) to address some research challenges identified in this chapter (Chapter 3).

---

2 Portions of this chapter have been previously published in Luo, et al (2011) and Luo and MacEachren (2014). I led the research and, as stated in the published papers, my responsibilities and that of my co-authors in the research and the paper were as follows. For the former one, conceived and designed the experiments: WL. Performed the experiments: WL PFY. Analyzed the data: WL. Contributed analysis tools: PFY FH WL. Wrote the paper: WL AMM PFY FH. For the latter one, conceptualized the paper WL. Wrote the paper: WL AMM.
Data Exploration

Conceptual Framework

Data exploration is a primary task in visual analytics to make sense of overwhelming amounts of disparate, conflicting, and dynamic data in a novel manner. Here, this research uses the Feature-ID model for geovisualization (Figure 3-1) (MacEachren 1995), an extension of an earlier pattern-matching model of cartographic visualization (MacEachren and Ganter 1990), as the reasoning framework for data exploration in terms of geo-social visual analytics. The pattern-matching model, in turn, draws upon a general scientific visualization perspective based on human cognition (Ganter and MacEachren 1989) to support understanding of human-display interaction in the context of map-based geovisualization.

The iterative process between human and computer interaction through “seeing”, “interpreting”, and “constructing-knowledge” shown in Figure 3-1 is an insight gaining process, which is a primary goal for visual analytics. Yi et al. (2008) characterize insight gaining as a multi-step process: provide overview, adjust, detect pattern, and match mental model. A reasonable insight gaining process starts with seeing an overview of a domain area. Having a big picture may or may not lead to direct insight, but it is a good starting point for people to make additional inquires about areas of interest to gain more knowledge. Adjust refers to a process through which people adjust the level of abstraction to explore a data subset of interest in order to make more sense of the data. Detect pattern refers to identifying interesting results that can include specific distributions, anomalies, clusters, and trends in the datasets. Match mental model (equivalent to “instantiate schema” in Figure 3-1) refers to reducing human cognitive load and amplifying recognition through providing a visual representation of data to decrease the gap between the data and user’s mental model of it, as well as the gap between the visual
representation and real-world knowledge. The whole process of visual interactive exploration is a mental model building process from knowledge development to critical breakthrough (Chang et al. 2009) that involves the iterative interaction among “seeing,” “interpreting,” and “constructing-knowledge” at each step. Therefore, geovisualization works as a larger cognitive system to support a human reasoning process.

Figure 3-1: Feature-ID model of geovisualization; Source: Figure 8.1 in (MacEachren 1995). The model focused on identifying key components in visually-enabled reasoning about geographic phenomena and relationships.
The framework provides a general model for understanding the visual and cognitive processes involved in interpreting and reasoning with geographic representations to address place-based questions. Extending the framework to geo-social analysis will require integrating: what has been learned about human perception and cognition of spatializations (Fabrikant, Monteilo and Mark 2006, Fabrikant and Skupin 2005), an understanding of how geographic scale social processes are conceptualized, and an understanding of the more complex reasoning process required by analysts attempting to understand the potentially complex relationships in geo-social processes. More generally, geo-social visual analytics aims to understand the interaction between two contexts and the impact of multiple contexts on reasoning outcomes through integrating social network space into geographic space to provide users a broader perspective. The following section illustrates the perspective through examples from existing research that each offer a step toward the objective of geo-social visual analytics to enable data exploration.

**Geo-Social Visual Analytics in Data Exploration**

Network theories and representations have not been fully considered in geographical information science (Prager 2008), but they have great potential to offer insight into complex geographical phenomena in terms of geo-social interactions. This section aims to address this gap. Given complex relationships between geographical space and social space at different spatial scales, this chapter characterizes the relationships into two groups: (1) geo-social relationships among geographical areas (e.g., nation, state, county), and (2) geo-social relationships among individuals at discrete locations (e.g., locations of mobile phone use, individual household locations, etc.).
Geo-social relationships among geographical areas cover a wide range of topics: such as migration flows at the city scale (Rae 2009), state scale (Phan et al. 2005, Tobler 2007), or country scale (Guo et al. 2012); transportation flows (Andrienko and Andrienko 2010, Demšar and Virrantaus 2010, Guo, Liu and Jin 2010); innovation diffusion as a spatial process (Hägerstrand 1968, Hägerstrand 1966); international trade among countries (Fagiolo, Reyes and Schiavo 2009, Fagiolo, Reyes and Schiavo 2008); sports competition among countries (Ahmed et al. 2010); and so on. In an early example shown in Figure 3-2-a, Tobler (1987) uses a network representation to describe the migration among different states in the U.S. In work grounded in geovisualization and geovisual analytics, Guo (2009) proposes an integrated interactive visualization framework that is used to effectively discover and visualize major flow patterns and multivariate relations from the county-to-county migration data in the U.S (Figure 3-2-b). In complementary recent research, Wood et al. (2010) propose an origins and destinations (OD) map to preserve all origin and destination locations of the spatial layout through constructing a gridded two-level spatial treemap.

The above work assumes that geographic location defines the spatial-social process with explicitly spatial representation and implicitly network representation, but the assumption only holds partially true for the modern interconnected world. Such representations reflect a situation in which the current integration of network analysis in GIS only focuses on a mathematical perspective that emphasizes graph theory and topology components of networks (Curtin 2007). Such representations do not allow users to explore the relationship between geographical space and social space, because they ignore network theory behind the network representation. Miller (2004) also suggests that geographic phenomena with strong social components (e.g., infectious disease propagation (Cliff and Haggett 1998)) do not appear to follow a Euclidian metric. Thus,
geographical proximity does not necessarily mean social closeness and conversely, geographical long distance does not necessarily result in social isolation. From a cognitive perspective, explicitly spatial representation and implicitly network representation can mislead human intuition about social relationships among actors and their relationships to place. Thus, it is necessary to involve explicit network representations to consider the importance of social position, social distance, and social space.

Andris (2011) proposes five benefits to involving an explicit network representation within a geographical environment: 1) Network community structure methods can identify clusters to understand the group of interconnected places as a unit rather than as dense collocations; 2) Node measures (i.e., degree, betweenness) can show the power of places; 3) Network system measures like degree distribution, closeness distribution, and clustering coefficients can indicate the role of any connected geographic region over the whole system; 4) Multiple social flow layers can be added simultaneously like spatial overlay functions in a geographical information system (GIS) to better evaluate interaction between places; and 5)
Explicit network representation performs better to model the case in which spatial closeness does not correspond to stronger social flows between places.

Given the benefits of explicit network representations in a geographical environment, we (2011) developed a spatial-social network visualization tool, the GeoSocialApp, that supports network, geographical, and attribute spaces to allow the exploration of spatial-social networks among them (Figure 3-3). The GeoSocialApp is based on an early, less comprehensive version of the conceptual framework for geo-social relationships discussed in the Chapter 2; the conceptual framework was informed by the initial implementation and application of the GeoSocialApp. In the GeoSocialApp, different interactive and linked views for network, geographical, and attribute spaces, respectively, correspond to network embeddedness, territorial embeddedness, and societal embeddedness (see Figure 2-1).

With explicit network spaces (in a dendrogram view and node-link view), GeoSocialApp users can have an intuitive understanding of social position, social distance, and social groups directly. For example, with the international trade network among 192 countries in 2005 as a case study, two groups identified through the dendrogram view show a core-periphery structure in the node-link view in which the red nodes are in the core and the yellow nodes are in the periphery. Since each node represents one country in the map view, the results in the map also show that the countries in the world have a hierarchical structure in which red nodes in the node-link view are economic core countries without highlight and yellow nodes are economic periphery countries with highlight. The parallel coordinate plot allows users to explore the power of places based on the network measures for each country. Core countries that have a low clustering coefficient (a measure of degree to which nodes in a network tend to cluster together) have high values with the other four variables (indegree, outdegree, closeness, and eigenvector). The negative relationship implies that rich countries may benefit more from more diversified trade partners and small economies may benefit more from homogeneous trade partners.
In complementary work, Thiemann (2011) develops SPaTo Visual Explorer to allow the exploration of spatial-social networks with multiple spatial and network representations. Unlike geographical distance measures, a new shortest-path distance based on node centrality measures is implemented into SPaTo Visual Explorer (Woolley-Meza et al. 2011). This tool can easily identify the shortest social distance among different cities based on the worldwide air-transportation network. While the SPaTo Visual Explorer implements analytical methods that are relevant to the overall geo-social visual analytics perspective this dissertation presents here, the authors do not ground the work explicitly in any conceptual framework for integration of spatial with network contexts. Their related work (cited above) on the nature of borders in “spatially embedded multi-scale interaction networks,” (Thiemann et al. 2010) however, illustrates how multiscale human mobility can be understood through what this dissertation would categorize as geo-social visual analytics methods (although the authors do not use those terms). Although Themann, et al (Thiemann et al. 2010) make some use of network statistics in their analysis, this dissertation would position the analysis and its interpretation in the overlap of territorial and network embeddedness in relation to the conceptual framework outlined in chapter 2.
Figure 3-3: The analysis of international trade data among 192 countries in 2005 with GeoSocialApp: Node-Link (Upper Left), Dendrogram (Lower Left), Bivariate Choropleth Map (Upper Right) and Parallel Coordinate Plot (Lower Right). For a preliminary view of this tool, see Luo, et al (2011). Each node in the node-link view and the dendrogram view corresponds to one country in the bivariate choropleth map; the Parallel Coordinate Plot view is used to explore network attributes for countries (from left to right, the attributes depicted are: indegree, outdegree, closeness, eigenvector, and clustering coefficient). Red dots in the node-link and dendrogram view are periphery countries and these are highlighted on the map and Parallel Coordinate Plot with blue outlines.

**Geo-Social Relationships among Individual Locations**

Geo-social relationships among individual locations include ones reflected in road networks (Demšar, Špatenková and Virrantaus 2008, Bak, Omer and Schreck 2010), commuting behaviors (Kwan 1999), location-based social networks (Doytsher, Galon and Kanza 2011), and social media networks (Yau 2010). Kwan and Lee (2004) apply geovisualization methods to explore activity patterns of women in Lexington, Kentucky as they relate to the transportation network. Specifically, they use GPS data collected through a Travel Data Collection Test to
explore trips by women without children under 16 years of age. Their analysis used 3D GIS generated space-time path depictions to help find that trips by this subset of women mainly use highways and major arterials, thus illustrating that individuals within a particular demographic group can share activity patterns that use geographic networks in similar ways. In subsequent work, Kwan (Kwan 2004) applies the concept of human extensibility (the ability of individuals to utilize space-adjusting technologies, including transportation and communication, to overcome the friction of distance) to more deeply explore the complex daily interactions of individuals grounded in interlinked geographic and social networks across scales from local to global. In complementary work in the geovisual analytics domain, Shen and Ma (2008) create MobiVis which allows visual analytics of social and spatial information in a human interaction network over time, and they illustrate how easily this tool supports comparison of individual and group behavior patterns (using the MIT Reality Mining Dataset (Eagle and Pentland 2006)). Those studies, and most similar research studies, illustrate that geovisualization and geovisual analytics can reveal distinctive patterns of spatial and social behaviors of different human interaction groups in a straightforward way (Chen et al. 2011, Jia and Jiang 2012, Lee and Kwan 2011, Slingsby, Beecham and Wood 2012, Reda et al. 2009).

Complementary to the visualization advances outlined above, computational methods to explore spatial-social human interactions focus on developing quantitative representations of human movements. For example, with mobile phone data, Gonzalez et al. (2008) and Rhee et al. (2008) find that human trajectories are characterized by a regular, time independent characteristic length scale and the people are more attracted to more popular places, like home or work. With the circulation of bank notes in the United States, Brockmann et al. (2006) find that human travelling distances follow a power law distribution, and that the distribution of the time people stay in one small, spatially confined region follows algebraically long tails. Furthermore,
Chaintreau et al. (2007) and Karagiannis et al. (2010) observe that inter-contact time between mobile devices shows an approximate power law in the range of 10 minutes to 1 day.

Overall, all of the above studies suggest the existence of scale-free characteristics observed in most networks in which a small number of nodes have a high degree distribution and a large number of nodes have a small degree distribution (Barabási and Albert 1999, Albert, Jeong and Barabasi 1999) in both the spatial and temporal dimension. In other words, the number of people in terms of spatial distance or inter-contact time has a scale-free distribution: a small number of people travel a long distance or have a long inter-contact time with others, whereas a large number of people travel a short distance or have a short inter-contact time with others. The scale-free characteristics identified in the human movement data over spatial and temporal dimensions are a representation of the whole system distribution of all individuals. Representing other attributes (i.e., location, demography) of each individual in other visual analytics views and using standard linked brushing methods to connect points in the scale free graph to their matching entities in other views allows users to explore questions in terms of “who,” “where,” “when,” and “what” on each individual.

Cho, Myers and Leskovec (2011) further identify the impact of spatial and social factors on human movement: short-range travel is spatially and temporally periodic with little impact by the social network structure, whereas long-distance travel has been impacted more by social network ties. Balcan et al. (Balcan et al. 2009) develop a unified model to study the multiscale nature of human mobility and its relationship with epidemic spread, including airline traffic networks and short-range commuting interactions. Crandall et al. (2010) take work a step further to develop models that quantify how likely it is that two people know each other, if they have a very close geographic distance at approximately the same time. These results open new directions for new perspectives on not only link prediction but also network dynamics with spatial, social,
and even multiscale considerations. The topic of predictive analysis will be addressed directly in the predictive analysis section below.

**Decision-Making**

In addition to enabling an efficient insight gain from a complex dataset, another major application in visual analytics is to use the insight to support a decision-making process. To design a visual analytic tool to effectively support human decision-making related to the interactions among geographic and social contexts, it is important to understand how people process information and how people make decisions in real situations.

**Conceptual Framework**

Decision-making is a process to reduce uncertainty and doubt, enabling individuals to take a reasonable course of action facing complex decision problems, often in time pressured situations (Harris 1998). The process of decision-making consists of three steps: analyze the situation, find out relevant alternatives, and select an alternative by certain criteria (Kohlhammer, May and Hoffmann 2009). Here, this dissertation draws upon two theoretical perspectives to frame the discussion of geo-social visual analytics for decision-making: situation awareness and spatial multicriteria decision analysis.

The conceptual framework from situation awareness (SA) can represent the decision-making process from a cognitive perspective and also integrate data exploration, decision-making and predictive analysis in the context of visual analytics. SA can be defined as "the human user's internal conceptualization of a situation" (Kohlhammer and Zeltzer 2003, p. 3609). Endsley (2000) defines three levels of SA: the first level is the perception of elements in the current situation, the
second level is the comprehension of the current situation, and the third level is the projection of future status (Figure 3-4). A reasonable decision-making process should be based on an understanding of the current situation from the first two steps of SA, and also a prediction for future situations. The first two steps also match Figure 2-1 in terms of a mental model building process to comprehend the current situation.

Spatial multicriteria decision analysis aims to integrate GIS and multicriteria decision making (MCDM), and both of them can provide different techniques and methodologies to transform geographical data and the decision-maker’s preferences to obtain information and knowledge to support decision-making (Malczewski 2006a). More details regarding GIS-based MCDM (GIS-MCDM) can be found in Malczewski (1999). Spatial decision analysis is an inherently multicriteria decision process, involving economic, social, environmental, and political dimensions (Kiker et al. 2009). The territorial embeddedness and societal embeddedness in Figure 2-1 can represent the essence of multicriteria decisions in spatial decision analysis. This section below proposes adding another dimension into spatial multicriteria decision analysis, the social network.
Figure 3-4: The process of decision-making for situation awareness. Situation awareness is defined as the human user's internal conceptualization of a situation, which plays a key role in effective decision making. Figure adapted from (Kohlhammer et al. 2009). Spatial data analysis and social network analysis can support decision-making from different perspectives, so the effective integration of both can enhance human user’s internal conceptualization of a situation.

**Geo-Social Visual Analytics in Decision-Making**

Spatial data analysis and social network analysis have their independent advantages to support decision-making. As discussed in the proposed conceptual framework for geo-social relationships (Figure 2-1), both analysis approaches should be considered together. This section illustrates how integrating spatial data analysis and social network analysis frames the decision-making problem in a unique and insightful way. This research argues here that their integration can be more powerful in support of decision-making than the sum of the parts. Limited research has been carried out thus far that achieves such integration. This section proposes two categories of geo-social integrations in terms of decision-making: the first one discusses an integrative approach toward spatial and social network factors to support a decision-making process; the second one discusses how social network structures impact GIS-based MCDM.

**An Integrative Approach of Spatial and Social Network Factors**

Spatial data analysis is used to detect and visualize spatial patterns (e.g., disease, crime), and relate these patterns to salient explanatory covariates (e.g., economic and demographic factors), and then these insights are used to give decision-making support for polices (Bailey and Gatrell 1995). However, many social phenomena are complex systems that mainly grow from the bottom-up, while traditional spatial data analysis focuses on top-down methods that cannot deal with the question of how the phenomena being analyzed evolve over space and time (Holland 1996, Batty 2005). Network science, a bottom up approach, provides the potential to link
individual behaviors and interactions among individuals to the size, scale, and shape of social phenomena observed in a spatio-temporal framework (Batty 2008). For example, urban sprawl that used to be understood through deterministic methods within economic location theory (Alonso 1964) has more recently been considered as a problem of organized complexity (Batty 2013). Network science has been proposed to study both urban physical networks (e.g., transport systems, water delivery) and urban social networks (e.g., industrial ecosystem) to build theories of how cities function as complex systems (Batty 2008, Baynes 2009, Batty 2006, Ashton 2008). This dissertation argues here that, while the application of network science has achieved important insights, a problem such as urban sprawl (or any other phenomenon that is both place-based and the product of complex societal factors) can be most fully understood through conceptual approaches that integrate geographical with social context and methods that integrate spatial with network analysis. As noted elsewhere in this dissertation, the complexity of such problems is what visual analytics methods are designed to address.

The two most typical network phenomena: small-world networks (characterized by high local clustering and short average node-to-node distance) (Watts and Strogatz 1998) and scale-free networks (in which, as noted above, a small number of nodes have a high degree distribution and a large number of nodes have a small degree distribution) (Albert et al. 1999), have shown a strong relationship with space and time. For example, the famous small world experiments to study the average path length for social networks of people observed these relationships at two geographical levels: U.S. (Milgram 1967) and world (Watts 2004). Additionally, small-world and scale-free properties have been demonstrated to exist in many spatial-social networks (i.e., World Wide Web graph, power grid graph, and road networks) (Watts and Strogatz 1998, Barabási and Albert 1999, Albert, Albert and Nakarado 2004, Govindan and Tangmunarunkit 2000, Lintanakool, Schwanen and Dijst 2009, Xu and Sui 2007). Even many traditional spatial phenomena exhibit scale-free characteristics such as city and company growth (Stanley et al.
Finally, as discussed in the data exploration section, the existence of scale-free characteristics has been extended from social relationships into spatial and temporal dimensions through analyzing mobile phone and GPS data.

Geo-social visual analytic tools have the potential to directly enable decision-making that incorporates understanding of both geographic and social factors in an integrated way. One prototype of how such tools might work, TwitterHitter, was introduced by White and Roth (2010). The objective of TwitterHitter is to harvest information from Twitter.com to support the functions of crime analysis; these functions include decisions related to ongoing investigations as well as those related to deployment of personnel. TwitterHitter provides functions to plot a linked map-timeline view of the recent spatiotemporal activities of suspects on Twitter, and also can generate a directed network graph of the suspect's known associates (i.e., Twitter friends) (Figure 3-5). Some other spatial data analysis methods can also be used, such as geographically weighted regression (Fotheringham, Brunsdon and Charlton 2002), to understand the etiology (scientific analysis of the causes) of the criminal activity, with the collected tweets or their attributes as potential explanatory variables in the analysis. In complementary work focused on decisions related to disease outbreaks, Guo (2007) proposes a geo-social visual analytic approach to analyze large spatial human interaction data to support effective pandemic control measures. The approach includes two linked views: a reorderable matrix and a map view to enable pattern interpretation in a geographical context and social context simultaneously. The geo-social interaction patterns provide valuable insight toward identifying critical locations and regions to suggest hypothetical control strategies for a pandemic outbreak based on synthetic population data.
Figure 3-5: Individual linked map-timeline and social network analysis views in TwitterHitter. The left view allows analysts to retrieve a spatiotemporal record of a suspect’s activity on twitter. The right view can uncover all potential connections among suspects through network group of twitterers within a region. They are courtesy of Jeremy D. White from Figures 3 and 4 in (White and Roth 2010).

The prototype tools developed by White and Roth (2010) and by Guo (2007) illustrate that spatial data analysis and network analysis can support decision-making from different perspectives. For example, spatial analysis in crime analysis demonstrates that explanatory factors relevant to spatial clusters of crime include, but may not be limited by alcohol outlet densities (Gorman et al. 2001), single person households (Cahill and Mulligan 2007), and depression (Ross 2000). Social network analysis in crime analysis has many important implications for crime investigations, such as targeting criminal leaders (Xu and Chen 2005) and fighting organized crime proactively (McAndrew 1999). Decision-making in terms of disease control should not only require the observation of corresponding spatial patterns and driving factors behind these patterns (Wang et al. 2008), but also needs the networks through which diseases are transmitted from person to person (Bian and Liebner 2007). Epidemic models, based on human interaction networks, can be used to simulate disease transmission processes and test the effectiveness of proposed control strategies (Keeling and Eames 2005). Recent research implements human cognitive behaviors into epidemic models to consider human preventive behaviors, which results in a very good agreement with the observed influenza data (Mao and Bian 2011, Salathé and Khandelwal 2011). Similarly, decision-making related to evacuation (e.g., in response to a hurricane threat) requires both spatial analysis related to location of people,
evacuation routes, etc. as well as an understanding of how social connections impact individual evacuation decisions and behaviors (Alsnih and Stopher 2004).

The Impact of Social Network Structures on Group Decision-Making

In group/participatory settings, GIS-MCDM involves a series of activities, including defining problems, selecting evaluation criteria by group members, determining individual and collective preferences in terms of evaluation criteria and/or alternatives, sensitivity analysis with evaluation criteria and alternatives, exploring alternative combinations of individual preferences into group judgments, supporting group interaction to refine individual and group preferences, and having a final ordering of alternatives to make a compromise alternative available (Limayem and DeSanctis 2000, Malczewski 2006b). Different stakeholders can be involved in the process of GIS-MCDM to face a variety of decision-making problems, such as environmental planning, transportation, and urban planning. Social relationships among those stakeholders have significant impact on their behaviors, which further has implications for their decision abilities (Bodin and Crona 2009). However, GIS-MCDM has not taken the impact of social relationships on actors’ decision making into account.

The potential importance of social relationships for GIS-MCDM is illustrated in work by Bodin and Crona (2009). They review the role of social networks in terms of different relational patterns on governance process and outcomes: 1) high network density can facilitate collective action, reduce conflicts, and enhance knowledge development; 2) low degree of cohesiveness (i.e., clearly distinguishable subgroups) has negative effects on collaborative processes among subgroups (Granovetter 1973); 3) bonding ties among subgroups is beneficial for conflict resolution and collective action; 4) high degree of network centralization is positively correlated with collective action (Sandström and Carlsson 2008). Furthermore, Bodin and Crona’s research
shows that none of the above network characteristics has a monotonically increasing positive effect on collective actions and conflict reduction, and that increasing one characteristic may cause the reduction of another. Therefore, how to maximize the positive effects of network characteristics on the collective actions presents a key research challenge in terms of group decision-making.

The integration of visualization techniques into GIS-MCDM has received increasing attention (Andrienko and Andrienko 2003, Jankowski, Andrienko and Andrienko 2001), but most studies focus on individual decision makers rather than groups (Malczewski 2006b). Consequently, the collaborative tasks in GIS-MCDM with visualization/visual analytics have not been explored, not to mention considering the impact of social networks on decision-making.

Here, as a representative example of how visual analytics and GIS-MCDM can be integrated, I highlight a geo-social visual analytics tool developed by Aguirre and Nyerges (2011) to analyze public decision-making processes. I discuss the possibility to extend this tool into MCDM domains in order to consider the impact of social network structures.

Aguirre and Nyerges (2011) introduce a novel geo-social visual analytics method that they label “grapevine” (Figure 3-6) that is directed to analysis of the very complex geo-social information generated within applications of web-based public participation systems for participatory learning and decision-making. The authors applied the grapevine tool to analysis of data collected during a month-long, online and asynchronous citizen advisory activity focused on planning for transportation in Puget Sound. The analysis enabled by the tool allowed Aguirre and Nyerges to partially confirm a hypothesis about analytic–deliberative decision-making, “that decisions are better when they come from a combination of analysis and deliberation rather than from analysis alone (2011, p. 320).” But, it also allowed them to identify key challenges in supporting deliberative processes that attempt to engage a wide cross-section of the public in deliberation that includes technical information and complex problems.
The grapevine tool is intended to help researchers understand the complex geo-social activities making up technology-enabled public decision-making. There are three underlying network structures in this tool: the main stem of the grapevine connects one node to another that represents users’ posts; participants vote for each other’s posts; participants reply to each other’s posts. Aguirre and Nyerges discuss the potential to use social network analysis to understand the frequency of interactions and roles of people from a theoretical perspective, but how to use social network analysis in the real case study with the grapevine tool has not been explored. Therefore, the grapevine tool can be considered to extend from four perspectives to integrate social network perspectives into GIS-MCDM in group/participatory settings for future work. First, a group decision-making environment can be considered to add into the grapevine tool. Second, the impact of structural social networks on decision-making reviewed by Bodin and Crona (2009) can be considered. Third, although the grapevine tool is not a GIS-MCDM, it can be considered to extend to generate individual and collective alternatives for MCDM within which decision-makers can choose and negotiate to support collaborative tasks. Last, agent-based modeling, a collection of agents that assess their situations and make decisions based on certain rules (Bonabeau 2002), can be considered to add to simulate the group decision-making process.
Figure 3-6: The static display of the grapevine. A main stem grows up with the increasing number of messages over geographical space. The main stem generates nodes when the number of votes to messages increases. The nodes generate buds when the number of replies (thus social connections) to messages increases. It shows the main stem in four states: (A) all features are turned off; (B) nodes are turned on; (C) the number of nodes signifies the number of votes; (D) the size of buds is according to the number of replies. Reproduced from Figure 4 in (Aguirre and Nyerges 2011).

This section presents an argument that there are two categories of geo-social integrations in terms of decision-making: a decision-making process considering spatial and social network factors, and the impact of social network structures on GIS-based MCDM. The decision-making process is an iterative process that requires the support of a visual-interactive environment, especially in time-pressured situations (Thomas and Cook 2005). One common problem with current geo-social visual analytic tools in terms of decision-making is that they are focused on helping analysts understand the current situations from the first two steps of SA, but lack the power to support decisions relevant to the current SA or predictions for future situations.
**Predictive Analysis**

As discussed in the decision-making section, the SA model also provides the theoretical framework for predictive analysis in the context of visual analytics. Predictive analysis is not independent from the first two steps of the SA model: it requires the understanding of the past and current situations through data exploration. An internal conceptualization of a situation, aided by predictive models and human reasoning, is the key to predictive analysis.

Developing mathematical models to support predictive analysis starts with the understanding of patterns found in real-world data. For example, based on the common property: scale-free characteristics observed in many large networks (e.g., actor collaboration graph, World Wide Web graph), Barabási and Albert (1999) build a preferential attachment model to explain the development of scale-free networks in which networks tend to continue to grow with new vertices, and new vertices have a preferential attachment to vertices that are already well connected. The preferential attachment model has been used to make predictions of network growth with scale-free characteristics, but this model does not consider the impact of geographical constraints on the network growth. As discussed in the data exploration section, scale-free characteristics have been extended into spatial and temporal dimensions in terms of human mobility. Lee et al. (2009a) develop a new mobility model called SLAW (Self-similar Least Action Walk) that can capture all human mobility features reviewed in the data exploration section (p. 21-22), including the Lévy flight travel patterns (Brockmann et al. 2006), spatial heterogeneously bounded mobility (Gonzalez et al. 2008), power-law inter-contact times (ICTs) (Karagiannis et al. 2010, Chaintreau et al. 2007), and fractal waypoints (Rhee et al. 2008). However, none of those mathematical models in terms of geo-social relationships have been implemented into geo-social visual analytics to empower prediction.
Recent studies have shown that mathematical models have a better prediction performance when they consider multiple components of context information \(i.e.,\) spatial, network, societal in Figure 2-1) rather than just one. For example, Andris, Halverson, and Hardisty (2011) develop a new model considering physical and social space for predicting future migration, and for U.S. migration flows among major cities; the model outperforms a gravity model considering physical space alone. In complementary work, Takhteyev et al. (2012) find that pre-existing ties \(i.e.,\) frequency of air travel) between places and people is the best predictor of Twitter ties compared to three other spatial and social factors including geographic distance, national boundaries, and language. A related study shows that using place-based attributes \(i.e.,\) social, economic and ecological context) can successfully predict community membership more than 70% of the time in a large-scale social network of cell phone towers (Caughlin et al.).

Geo-social visual analytics designed to support predictive analysis implements mathematical models in a visual-interactive environment to allow users to select appropriate methods, to set parameters, to interpret results, to understand what to do next, and to draw conclusions based on different scenarios. For example, Brigantic et al. (2010) introduce a visual analytic tool (PanViz) with metapopulation-based epidemic models to rapidly assess alternative mitigation strategies in terms of pandemic influenza to give decision makers support. The PanViz system was subsequently extended and deployed in the Indiana State Department of Health Planning (Figure 3-7) to support analysis of potential epidemic control strategies (Maciejewski et al. 2011). While PanViz demonstrates the potential of geovisual analytics, it does not explicitly include capabilities to incorporate social network information into the analysis. Bisset and Marathe (2009) have developed a similar tool that does include such capabilities: EPISIMS with individual-based epidemic models to simulate the dynamics of millions of individuals, traffic of entire cities, and disease spread, respectively. In related research, Broeck et al. (2011) present
“GLEaMviz” available to the public that allows the user to set a variety of parameters to simulate the human-to-human infectious disease spread across the world.

Figure 3-7: Here I illustrate the PanViz user interface. A user has simulated an outbreak originating in Chicago and has been exploring the spread of the infectious disease over space and time. Day 21 of the simulation is depicted in the figure. The color of each county represents the percentage of the population that is ill. The plots on the right show the rate of spread (thus the rate of person-to-person connection), hospitalization, and death in user selected counties. Figure Courtesy of Ross Maciejewski from Figure 7 in [127]. For Additional Information on the research it is derived from, See (Maciejewski et al. 2011).

One big challenge in terms of predictive analysis is that geo-social systems are highly sensitive to social adaptive behaviors (Vespignani 2009). In crisis situations (e.g., pandemics, natural disasters), the geo-social systems behave abnormally which is still a challenge to predict. Vespignani (2009) proposes three challenges for future work to predict social adaptive behaviors in a period of crisis: collecting data on information spread and social reactions in the period of crisis; developing models to quantify the effect of risk perception and awareness phenomena of individuals; and deploying monitoring infrastructures capable of informing computational models
in real time. The three suggestions also need geo-social visual analytics to support methods to address all three challenges in an integrative system.

In research directly focused on predictive analytics for crisis events, Bengtsson et al. (2011) focus on displacement after the 2010 Haiti Earthquake. Specifically, they track population with phone call data from six weeks before the disaster to five months after. Their estimates in terms of the number, the timing, and geographical distribution of population movements correspond well with a retrospective survey. A follow-up study identifies that the destinations in which people stayed had significant social bonds and the time that the displaced population stays outside the city follows a skewed, fat-tailed distribution (Lu, Bengtsson and Holme 2012).

Bengtsson and colleagues have taken the first step to demonstrate that the prediction of population movements with the use of phone data in disaster response is possible, but substantial future work is still required before common usage. For example, cell phone data availability during natural disasters is still limited (Bengtsson et al. 2011); data coverage varies over space, time, and different groups of people (Gething and Tatem 2011); and data privacy is always a big concern (Landwehr et al. 2012).

Above all, the process of predictive analysis is an iterative process that needs human interpretation, control, experiences, and imagination, especially in crisis periods. As discussed here, current geo-social visual analytic tools in terms of decision-making are good at enabling perception of elements in and comprehension of the current situations, thus the first two steps of SA, but they lack the power to make predictions for future situations. Meanwhile, there is a gap between the insight gained in the process of data exploration and predictive analysis. For example, the current predictive epidemic visual analytic tools implement epidemic models to simulate disease transmission and design corresponding control strategies without flexible approaches analyzing human spatial-social interactive clusters as knowledge input to improve
control strategy design (like those illustrated by Guo (2007) in Section: An Integrative Approach of Spatial and Social Network Factors).

Based on the above discussion, multiple geo-social visual analytics research questions need to be addressed in relation to data exploration, decision-making, and predictive analysis. In addition, data exploration, decision-making, and predictive analysis have inner connections: knowledge development in the data exploration process can support decision-making and predictive analysis, which sometimes leads decision-making and predictive analysis to develop a new knowledge construction process. Therefore, the overarching goal of visual analytics is to involve human ability with the whole complex analytical process rather than each step separately (Andrienko et al. 2007).

The following two chapters will demonstrate two novel geo-social visual analytics tools: one is at a geographical region level (e.g., county, state, and country) and the other one is at an individual level, and their applications to two research domains: international trade and disease transmission and control, respectively.
Chapter 4

A Geovisual Analytic Approach to Understanding Geo-Social Relationships in the International Trade Network

Introduction

As mentioned at the end of the last chapter, this chapter is building on the base of theory and review of the literature in the previous chapters. This chapter presents the development of the GeoSocialApp and its application to the international trade network (ITN).

The world has become an increasingly interconnected system with multi-scale geographically embedded networks (i.e., transportation, internet). Spatial analysis aims to understand such systems in terms of spatial patterns, relationships, processes, and change within and among geographical spaces (Bailey and Gatrell 1995). Social network analysis has been used to understand how systems emerge through the interaction of individual agents (i.e., humans, companies). Both approaches have advantages and limitations as methods through which to understand the complex geo-social interconnected world. Many geo-social interconnected systems mainly grow from the bottom-up, but traditional spatial analysis is a top-down approach that cannot deal with the evolution of the systems over space and time (Batty 2003, Holland 1996). Social network analysis, a bottom-up approach, can link individual-level behaviors and

---

3 Portions of this chapter have been previously published in Luo, et al (2014). I led the research and, as stated in the published paper, my responsibilities and that of my co-authors in the research and the paper were as follows. Conceived and designed the experiments: WL QD. Performed the experiments: WL PFY QD. Analyzed the data: WL QD. Contributed analysis tools: PFY FH WL. Wrote the paper: WL AMM QD PFY FH. To provide a logical self-contained discussion, this chapter includes a few small sections that repeat or summarize material presented in earlier chapters (e.g., the first paragraph shares ideas with the section “An Integrative Approach of Spatial and Social Network Factors” above) and this chapter has its own literature review tailored to the application developed.
interactions to the emergence of social phenomena (Batty 2008), but the approach typically ignores geographical constraints (Onnela et al. 2011). An effective integration of both approaches has the potential to aid understanding of geo-social systems from a more comprehensive perspective. For example, the integration of spatial consideration into a social network approach enables understanding of why and how an air-borne disease diffuses within an urban area in a manner that can generate disease hot spots as well as cold spots (Mao and Bian 2010b). The integration of spatial analysis and social network analysis has the potential to link individual-level behaviors and interactions (i.e., human, vehicle, organization) to understand urban sprawl over space and time (Batty 2008). Although spatial analysis and social network analysis have the potential to complement each other, the formal integration of two approaches remains relatively underdeveloped in the literature (Adams et al. 2012).

This chapter therefore integrates spatial analysis and social network analysis into a unified framework through a geovisual analytics approach. Geovisual analytics tools integrate computational methods with interactive visualization, in order to enable insights on large and complex geospatial datasets (Andrienko and Andrienko 2012, MacEachren et al. 2011a, Andrienko et al. 2010, Guo et al. 2006). Specifically, this research presents and applies a geovisual analytics tool, GeoSocialApp (Luo et al. 2011), that consists of three major analytical “spaces” implemented as linked components: a geographic space, a network space, and an attribute space. Each performs a specific task and can coordinate with other components to facilitate a process through which insights are enabled. This chapter illustrates how the GeoSocialApp facilitates development of hypotheses, with the international trade network (ITN) as a case study. The explicit geographical and network representations in the GeoSocialApp facilitate and enable insight in terms of different roles that spatial and social relationships have in the ITN across geographical regions with network hierarchies at different scales. One major goal of geovisual analytics is to develop hypotheses on how space matters based on the patterns
identified from geo-spatial data (Andrienko et al. 2011); but the validation of geovisual analytics results is still regarded as a challenge (Keim et al. 2011). Here, this research proposes a Monte-Carlo approach as a statistical validation to support the hypothesis developed through visual-computational exploration of spatial and social interaction in the ITN.

The chapter begins below by reviewing the development of geo-social visual analytics methods in geography and network domains. This chapter then presents an overview of the methods and the international trade network data used in this study. The results obtained through applying the methods to the data provide insights on the different roles that spatial and social relationships play in relation to trade across geographical regions. This chapter next introduces the Monte-Carlo approach as a statistical validation to support the insights discussed as above. Finally, this chapter presents conclusions and an outlook for future research.

**Literature Review**

Current geo-social visual analytics tools can be classified into two major groups: the first group, rooted in geography, focuses on geographical analysis explicitly with a network representation; the second group, rooted in social network science, has an explicitly network representation with geography as a background to visualize the results. This section reviews the geo-social visual analytics tools from geography and social network science domains, and argues for a more balanced approach that emphasizes spatial relationships and social networks simultaneously.

Spatial interactions/flows associated with topics such as human migration and disease transmission are major research domains for integrating network representation into geovisual analytics. For example, Andrienko and Andrienko (2010) develop a spatial generalization method to transform trajectories with common origins and destinations into aggregated flows maintaining
essential characteristics of the movement between areas. In complementary research, Guo (2009) proposes an integrated interactive visualization framework that is applied to county-to-county migration data in the U.S. in order to visualize and discover network structures, multivariate relations, and their geographic patterns simultaneously. Additional relevant research can be found in recent papers by Andrienko et al. (2009), Demšar and Virrantaus (2010), Guo, Liu and Jin (2010), and Wood, Dykes and Slingsby (2010).

All of the above studies consider the geo-social processes from a primarily geographical perspective. Spatial interactions/flows in research taking this perspective are typically visualized on maps, which provide important information on spatial context. The observed spatial patterns can be related to the spatial context (e.g., big cities tend to be hotspots for human interaction). The methods for geo-social interaction discussed so far assume that geographic locations define the geo-social process, but new communication and transportation technologies clearly spread social networks beyond traditional geographical constraints (i.e., distance) (Larsen et al. 2006). Therefore, understanding the social meaning behind the geo-social processes is equally important.

Geo-social visual analytics from a social network science perspective tends to have an explicit network context with an implicitly geographical representation. Ahmed et al. (2010) introduce new visual analysis methods with dynamic network views (e.g., wheel layout, radial layout, and hierarchical layout) to explore the 2006 International Federation of Association Football (FIFA) World Cup competition in which countries are clustered based on their geographical locations in the dynamic graph representation. The visual analysis methods allow users to analyze and compare each country’s performance within the geo-social context. The explicit network representation and implicitly geographical representation require analysts to relate the explicit network representation to his or her unrepresented geographic background knowledge in the visually interactive process (Andrienko and Andrienko 2012). Thiemann (2011) developed the SPaTo Visual Explorer, which implements multiple explicitly geographical and
network representations. Using a case study focused on global air flight networks, they illustrate how SPaTo can allow users to develop hypotheses about the interaction between geographical distance and social network distance. For example, he derives evidence showing that geographical proximity of cities corresponds with short social distance among the cities. Beyond the above, four additional research efforts have focused on specific components of methods to involve explicitly geographical representations into a traditional social network approach: 1) a spatial point pattern exploration approach (e.g., kernel density) can be used to understand spatial impacts on the development of social networks (Verdery et al. 2012); 2) a spatial autocorrelation coefficient (e.g., Moran’s I) has been applied to social networks to measure the statistical similarity of individuals (Mercken et al. 2009); 3) explicitly spatial representations facilitate practical implementation of decision-making in certain social network application domains (e.g., infectious disease control) (Mao and Bian 2010a); and 4) certain geo-social systems (e.g., human migration, international trade network) can be better understood or predicted through mathematical models considering physical and social space (Andris et al. 2011, Zhou and Park 2012).

As discussed above, understanding geo-social systems requires consideration of both geographical relationships and social network relationships. Therefore, it is necessary to involve explicitly geographical and social network representations. Andris (2011) lists five benefits to having an explicit network representation within a geo-spatial framework: 1) the group of connected geographical regions can be studied as a unit with social closeness based on a network community detection approach; 2) the social power of places can be represented by node measures (i.e., degree, betweenness); 3) the social role of interconnected places over the whole system can be represented by network system measures (i.e., degree distribution, betweenness distribution); 4) the complex social interaction between places can be understood through adding multiple social flow layers on Geographical Information System (GIS); and 5) the geo-social
systems in which spatial closeness and social closeness do not match can be better modeled with an explicit network representation.

The above discussion illustrates that there is the lack of explicitly spatial and social network representations in current geovisual analytics and the importance of such representations to understand geo-social systems (Luo and MacEachren 2014). It is also still a challenge to statistically support hypotheses developed through visual exploration (Cusumano-Towner 2009), particularly the hypotheses directed to geo-social interaction. To fill the gap, this chapter introduces the GeoSocialApp with the 2005 international trade network as a case study to understand the interaction between spatial and social relationships, and introduces the use of a Monte-Carlo approach to validate the hypothesis developed in our geo-social visual exploration.

Methods

This research extends and applies the GeoSocialApp, a geovisual analytics tool initially introduced in preliminary form in Luo et al. (2011). The GeoSocialApp implements traditional network analysis methods within the context of an environment that links explicitly spatial and social representations to understand the interaction of spatial and social relationships in the ITN. The GeoSocialApp is an extension of the GeoViz Toolkit (GVT) developed in the GeoVISTA Center at Penn State (Hardisty and Robinson 2010). The research presented here makes use of the existing choropleth mapping capabilities of GVT to support geographical analysis as well as the component coordination methods that enable dynamic linking and brushing across views, and adds a dendrogram component that supports multiple graph-based views to represent a varying network hierarchy. Details about other GVT components that could be used to extend the analysis presented here can be found in http://www.geovista.psu.edu/GeoSocialApp/.
GeoSocialApp Components

As noted above, this research uses two components in the GeoSocialApp: a dendrogram view and a choropleth view. The dendrogram view implements the convergence of the iterated correlations (CONCOR) algorithm (Breiger, Boorman and Arabie 1975, Wasserman and Faust 1994) to group nodes with equivalent positions in a single network or multiple social networks together. Equivalent positions refer to collections of actors that have similar ties to and from all other actors in the network. The implication of actors having equivalent positions is that they play similar social roles in a relational network. The relational network can be described by an adjacency matrix A, which can generate a position similarity matrix R to measure the equivalent positions, whose element value $r_{ij}$ is defined as:

$$r_{ij} = \frac{\sum (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) + \sum (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum (x_{ki} + \bar{x}_i)^2 + \sum (x_{ik} + \bar{x}_i)^2} \sqrt{\sum (x_{kj} + \bar{x}_j)^2 + \sum (x_{jk} + \bar{x}_j)^2}}$$ ...........(1)

where $\bar{x}_i$ ($\bar{x}_j$) is the mean of the values in row i (j) of the matrix A and $\bar{x}_i$ ($\bar{x}_j$) is the mean of the values in column i (j) of the matrix A. At the initial level of analysis, CONCOR performs the above equation calculations iteratively on the position similarity matrix R until all values converge to either 1 or -1, resulting in all nodes being grouped into one of two categories. Two groups can be too generalized for some studies, so hierarchical structures can be achieved by running CONCOR on each subgroup. In this way, CONCOR can continue to split nodes into successively smaller groups: two become four, four become eight, and so on. Although this algorithm was developed originally for application to social networks of individuals, it has been demonstrated to be an effective method to empirically locate structural positions in terms of the ITN (Smith and White 1991, Luo et al. 2011).
Equivalent positions in terms of the ITN refer to collections of countries that have similar import and export trade relationships with all other countries (Breiger 1981). The implication of countries having equivalent positions is that they play similar social roles in the ITN. According to world system theory, the economic development of different countries is affected by their structural positions: core, semi-periphery, and periphery through unequal economic exchanges among them (Wallerstein 1974). Core countries focus on capital-intensive production, periphery countries provide low-skill labor and raw materials, and semi-periphery countries are the industrializing countries positioned between the periphery and core countries. The CONCOR algorithm can classify the ITN into these three structural equivalence positions (Snyder and Kick 1979, Nemeth and Smith 1985).

A tree layout and a radial layout are implemented in the dendrogram view to visualize the hierarchical structure of CONCOR results (Figure 4-1). The tree layout organizes the graph in a hierarchical way by placing child nodes under their common ancestors. An informationally equivalent radial view can be transformed from the tree by putting child nodes in the enclosing circle of their common ancestors (Jeong and Pang 1998, Carriere and Kazman 1995). The dendrogram view in the GeoSocialApp also provides a slider to control the hierarchical level of CONCOR results.
Figure 4-1: Dendrogram View: two layouts to visualize the hierarchical structure of CONCOR results: the left one is a tree layout and the right one is a radial layout. Slider bar is used to control the level of CONCOR results.

The dendrogram view of social space is dynamically linked to a choropleth map view used for visual exploration in geographical space. Each node in the dendrogram view corresponds to a geographical unit (i.e., states, countries) in the choropleth map. The choropleth map allows users to choose the number of classes, the classification method (i.e., equal intervals, quantiles), the variable to display, and the ColorBrewer palette (Harrower and Brewer 2003) for color selection. Thus, the linked dendrogram and map views allow exploration of social positions and social groups and their corresponding spatial positions and spatial groups simultaneously. With the hierarchical level control in the dendrogram view, the linked views further support the explicit exploration of interaction between social space and geographical space and its impact on
outcomes of interest at different network hierarchy (Figure 4-2). This capability will be illustrated in the case study presented below, after the data used in that case study are first described.

Figure 4-2: Dendrogram view and choropleth map view. The choropleth map depicts GDP by country. Data are divided into quintiles (5 categories with an equal number of countries in each category) depicted by 5 sequentially ordered shades of green, from low GDP (very light green) to high GDP (very dark green). Each node in the dendrogram view corresponds to one country in the choropleth map view (The highlighted nodes in blue correspond to countries with borders highlighted in blue). The first run of CONCOR process reveals two positions in the 2005 ITN.

**Data**

Our analysis of the interaction between spatial and social relationships in the ITN is based upon import and export data among 192 countries in 2005. These data were extracted from the CorrelatesOfWar (COW) Database and include volume of imports and exports in current U.S. dollars (Barbieri, Keshk and Pollins 2008). 2005 ITN data is converted into a directed network in which countries are the nodes of the network and an import/export trading relationship is represented by a link between two countries. This research then organizes the data into a binary matrix form to fit the CONCOR algorithm with columns as exporting countries and rows as importing countries. As an illustration, Table 4-1 is the original import and export data among sample countries in 2005, and Table 4-2 is the binary matrix for the first 10 countries in our data; “1” represents presence of import/export trade between countries, “0” represents no trade. A
binary matrix is used rather than a weighted matrix for the reason: one basic idea of the CONCOR algorithm is that the primary indicator of a relationship is the absence of links between individuals rather than the occurrence of the links (White, Boorman and Breiger 1976); given this idea, the past research in international trade has typically used the binary matrix with the CONCOR algorithm to identify three structural equivalence positions: core, semi-periphery, and periphery (Snyder and Kick 1979, Nemeth and Smith 1985, Cassi, Morrison and Ter Wal 2012).

Table 4-1 Imports-exports relationship among partial countries in 2005. Flow1 means imports of importer1 from importer2 in current US millions of dollars, and flow2 means imports of importer2 from importer1 in current US millions of dollars.

<table>
<thead>
<tr>
<th>Year</th>
<th>importer1</th>
<th>importer2</th>
<th>flow1</th>
<th>flow2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>United States of America</td>
<td>Canada</td>
<td>291944</td>
<td>195151</td>
</tr>
<tr>
<td>2005</td>
<td>United States of America</td>
<td>Bahamas</td>
<td>726.3</td>
<td>1945.79</td>
</tr>
<tr>
<td>2005</td>
<td>United States of America</td>
<td>Cuba</td>
<td>0</td>
<td>397.87</td>
</tr>
<tr>
<td>2005</td>
<td>United States of America</td>
<td>Haiti</td>
<td>458.5</td>
<td>756.91</td>
</tr>
<tr>
<td>2005</td>
<td>United States of America</td>
<td>Dominican Republic</td>
<td>4721.4</td>
<td>5179.24</td>
</tr>
<tr>
<td>2005</td>
<td>United States of America</td>
<td>Jamaica</td>
<td>410.9</td>
<td>1962.2</td>
</tr>
<tr>
<td>2005</td>
<td>United States of America</td>
<td>Trinidad and Tobago</td>
<td>8342.2</td>
<td>1583.01</td>
</tr>
<tr>
<td>2005</td>
<td>United States of America</td>
<td>Barbados</td>
<td>33.4</td>
<td>595.28</td>
</tr>
<tr>
<td>2005</td>
<td>United States of America</td>
<td>Dominica</td>
<td>3.8</td>
<td>67.43</td>
</tr>
</tbody>
</table>

Table 4-2 International trade relationships among nine sample countries in a binary matrix for 0% threshold in 2005.
This research uses three additional data variables: GDP, population, and geographical distance, to validate the hypothesis developed through visual exploration using the GeoSocialApp. This research downloaded 2005 GDP and population data for each country from the World Bank website (http://data.worldbank.org/), and calculated the linear distance between national capitals to measure the geographical distance between countries with ArcGIS. This measure of between-country distance is picked over others (e.g., distance between country centroids, distance between the nearest points of country borders, etc.), because gravity models used in other international trade network studies use the same distance measure (Zhou 2011).

Results

Spatial and Social Interaction at the First Level of CONCOR

This research uses the dendrogram view in the GeoSocialApp to explore Table 4-2 to identify social relationships among all countries, and the univariate choropleth map to visualize the spatial distribution of GDP for all countries (Figure 4-2). Comparing the dendrogram view and the map view, and using the dynamic linking between them to explore specific details for individual and groups of countries, can provide insight about spatial and social interactions within the ITN.

Initially, the network data is divided into two groups using the dendrogram view. After highlighting one group (blue nodes in the dendrogram view and blue outlines in the map view), this study finds that most countries in the highlighted group are economic periphery countries (i.e., most countries in Central America and Africa) and most countries in the other group are economically core countries (i.e., North America and European Union). The univariate
choropleth map depicts GDP for each country. The sequential colors reinforce this classification: economically less-productive countries are indicated by light green, whereas other, more economically productive countries are indicated by dark green. The two classifications identified by CONCOR imply that economically core countries tend to have similar international trade partners, and economic periphery countries tend to have similar trade partners. This study focuses on the interaction between spatial and social relationships in the ITN. At the first level of CONCOR in Figure 4-2, all countries with close social relationships tend to exhibit spatial proximity.

**Spatial and Social Interaction at the Second Level of CONCOR**

The second application of CONCOR to the ITN subdivides the first two categories, resulting in a total of four groups as shown by Figure 4-3. The core countries and the periphery countries are partitioned into four new geographies, which further indicate a core–periphery arrangement: the mean GDP for each geography is sorted in Table 4-3. Figure 4-3A mainly includes more developed countries in the economically core group: North America, most countries in Europe, Australia, South Africa, and economically more-important countries in Asia (i.e., China, India), whereas Figure 4-3B mainly consists of less developed countries in the economically core groups: Russia, most countries in South America, and a small number of countries in Europe. Figure 4-3C mainly includes more developed countries in the economic periphery group: Central America, and a few countries from Eurasia (i.e., Vietnam, Iran), whereas Figure 4-3D mainly consists of the less developed countries in the economic periphery group: countries from Africa and some countries from Asia (e.g., Mongolia). In terms of spatial and social interaction identified by the second level of CONCOR, economically core countries in

---

A list of countries for each group is in File S1.
Figure 4-3A and Figure 4-3B (i.e., North America, Europe), as well as more developed periphery countries in Figure 4-3C exhibit regional patterns (i.e., Central America, Central Asia) that also fall into the same social groups across the globe. It suggests that international trade partners for those countries are related to both spatial proximity and similar economic development level (Figure 4-3A, 4-3B, and 4-3C). Economic periphery countries in Figure 4-3D have one major cluster (i.e., Africa). Compared to 4-3A, 4-3B, and 4-3C, Figure 4-3D suggests that spatial proximity has a stronger impact on the least developed countries in terms of international trade partners they have.

Spatial and Social Interaction at the Third Level of CONCOR

The third run of CONCOR applied to the ITN again subdivides the previously identified groups into seven different subgroups (Figure 4-4). At this level the geographies are considerably more complex but this research highlights three features. First, only seven new

---

5 A list of countries for each group is in File S1.
subgroups are identified in this level: CONCOR does not divide countries depicted in Figure 4-3A any further, resulting in the same group of countries in Figure 4-4A, because economically core countries in this group have highly similar import and export trade partners. Second, some groups of countries at this level further confirm a core-periphery hierarchical structure in terms of the ITN: the top economically core countries in Figure 4-4A; a clear distinction between east African countries (the second least developing places) in Figure 4-4F and west African countries (the least developing regions) in Figure 4-4G. Third, the role that spatial and social relationships play in terms of the ITN identified by the third level of CONCOR becomes more noticeable. Core countries in Figure 4-4A, Figure 4-4B, and Figure 4-4C have their own distinct geographical regions (i.e., North America, Europe), but social relationships to connect different regions are also strong. Figure 4-4D and Figure 4-4E identify two distinct geographical regions (Central America and Central Asia) compared to Figure 4-3C that put both into the same social group. The distinct geographical regions suggest that spatial constraints are stronger than social connections between the two regions at this network level. Comparing the two distinct geographical regions identified in Figure 4-4D and Figure 4-4E to distinct geographical regions (i.e., North America, Europe, and Australia) in Figure 4-4A confirms the above finding: spatial constraints have less impact on economically core countries and more impact on economic periphery countries to determine the international trade partners they have.
Figure 4-4: The third run of the CONCOR process continues to subdivide groups. Figure 4-4A 4-4B, and 4-4C belong to the economically core countries, whereas Figure 4-4D, 4-4E, 4-4F, and 4-4G belong to the economic periphery countries.

Validation

As outlined above, using an interactive visual approach, this research found that developing countries with structural equivalence tend to exhibit a pattern of geographical proximity, and developed countries with structural equivalence tend to exhibit a pattern in which geographical proximity remains a factor, but one that is overcome by some connections to distant places. Based on the patterns, this research develops the two-part hypothesis that: international trade network clusters with structural equivalence are strongly 'balkanized' (spatially fragmented).
according to geography of trading partners, and the geographical distance within each network cluster has a positive relationship with the development level of countries. The next step, reported in this section, is to verify this visual finding with a robust statistical evaluation. This research has two steps to verify the hypotheses. The first step introduces two indicators (degree of balkanization and Pearson correlation) to quantify the observed patterns, and the second step uses a Monte-Carlo method to measure the statistical level of the two indicators. It is also important to note that these two linked parts of the analytic process (visual hypothesis generation and confirmatory analysis) provide an iterative means of arriving at stronger conclusions.

**Degree of balkanization**

The first part of the hypothesis is that the network cluster with structural equivalence is strongly 'balkanized'. Based upon the eight clusters identified by the CONCOR process, distances among all pairs of countries are calculated and Balkanization (B) is measured as the difference of (a) mean within-cluster distances and (b) mean between-cluster distances. Let $C_i$ denote clusters in the international trade network after applying the CONCOR algorithm. Each cluster is composed of several countries \( \{X_i, X_{i+1}, \ldots, X_m\} \), and all countries \( X \) are classified into those mutually exclusive clusters.

\[
X = \{X_1, X_2, \ldots, X_N\}; \text{ X: country; There are } N \text{ countries in the international network.}
\]

\[
C = \{C_1, C_2, \ldots, C_K\}; \text{ There are } K \text{ clusters after applying CONCOR algorithm.}
\]

\[
C_i = \{X_{m_1}, X_{m_2}, \ldots, X_{m_i}\}; \text{ The } i^{\text{th}} \text{ cluster after applying CONCOR algorithm, which has } m_i \text{ countries. } \forall i, j \leq K \text{ and } i \neq j, C_i \cap C_j = \emptyset; \text{ Clusters are mutually exclusive. } D(X_i,X_j):
\]

geographical distance between two countries.
A positive value of $B$ means that countries that belong to the same trade cluster are geographically dispersed; the higher the positive value, the higher the degree of balkanization. If $B$ is equal to zero, the countries from the same cluster have no geographic proximity at all and display a random geographic distribution. A negative value of $B$ indicates that countries from the same trade cluster are geographically grouped. The degree of balkanization of 2005 international trade data set is denoted as $\tilde{B}$, with value of 2774.008 km. The absolute value indicates little about the degree of balkanization unless it is compared to some benchmark. The Monte-Carlo method can provide such a benchmark and produce a statistical significance measure of the absolute result, which will be discussed after describing our approach to measuring the relationship between GDP and distance by network cluster.

**Pearson correlation**

This research uses Pearson correlation (Rodgers and Nicewander 1988) to measure the positive relationship between geographical distance within each network cluster and the development level of countries, which is determined by GDP in this chapter.

$$P_{X,Y} = \frac{1}{n} \sum_{i=1}^{n} (G_i - \bar{G})(D_i - \bar{D}) \over \sigma_G \sigma_D$$

$G_i$ is the average GDP of each cluster. $D_i$ is the average within-cluster distance of each cluster. $\sigma_G$ is the standard deviation in terms of average GDP of each cluster. $\sigma_D$ is the standard deviation in terms of average within-cluster distance of each cluster. $P$ ranges from -1 to 1. A
positive P value implies that there is a positive relationship between geographical distance within each network cluster and GDP. A negative P value implies that geographical distance increases as GDP decreases. If P is around zero, it means that the geographic factor of each network cluster is independent from GDP.

When the average within-cluster distance is calculated, this research gives more weight to the countries that are more populous by weighting the distance by the population. The reason for this is explained below. The Pearson correlation between the average within-cluster distance without weight and GDP is only 0.13; this does not reflect the strong relationship that is apparent between the two variables as observed visually from the GeoSocialApp. I checked the GeoSocialApp again in order to figure out the reason behind this initial result, and found that simply calculating the average distance between any pair of countries may introduce some noise. For example, island countries in the middle Pacific (Figure 4-4F) that are far away from any other countries may raise the average within-cluster distance. The cluster in Figure 4-4F includes mainly developing countries in North Africa and the Mideast, as well as some island countries (e.g. Solomon Islands, Vanuatu). These islands only represent 1.5% of the population and 3.8% of the GDP for the cluster, but increase the within-group distance by 47.71%. Such a dramatic rise of within-group distance makes the distance-GDP nexus indistinct and brings down the Pearson correlation. This study tests the impact of those islands on the Pearson correlation through removing those islands in Figure 4-4F, which raises the correlation to 0.36. Given the similar issue existing in some of the other clusters (i.e., Figure 4-4D, 4-4E), this study weights the distance between all countries proportionally to their population without removing any island countries (Table 3-3). Following from these preliminary results, this research refines the hypothesis into: the geographical distance weighted by population within each network cluster has a positive relationship with the development level of countries. The 2005 international trade
data set’s Pearson correlation (\(\bar{P}\)) between average GDP per cluster and population weighted within-cluster distance is determined to be 0.97.

Table 4-3: CONCOR group level attribute data. Mean GDP in 2005 for 4 groups identified at the second level of the CONCOR, mean GDP in 2005, mean distance, weighted distance by population for 7 groups at the third level of the CONCOR.

<table>
<thead>
<tr>
<th>FigureID</th>
<th>Mean GDP (billions of dollars)</th>
<th>Mean GDP (billions of dollars)</th>
<th>Mean Distance (km)</th>
<th>Weighted Distance</th>
<th>Mean GDP (billions of dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3A</td>
<td>912.00</td>
<td>4A</td>
<td>912.00</td>
<td>6664</td>
<td>5.55E+19</td>
</tr>
<tr>
<td>3B</td>
<td>250.00</td>
<td>4C</td>
<td>116.00</td>
<td>8086</td>
<td>5.15E+18</td>
</tr>
<tr>
<td>3B</td>
<td>384.00</td>
<td>4B</td>
<td>7146</td>
<td>9.76E+18</td>
<td>384.00</td>
</tr>
<tr>
<td>3B</td>
<td>250.00</td>
<td>4C</td>
<td>116.00</td>
<td>8086</td>
<td>5.15E+18</td>
</tr>
<tr>
<td>3C</td>
<td>48.10</td>
<td>4D</td>
<td>3403</td>
<td>1.84E+17</td>
<td>48.10</td>
</tr>
<tr>
<td>3C</td>
<td>34.90</td>
<td>4E</td>
<td>21.80</td>
<td>5125</td>
<td>4.67E+17</td>
</tr>
<tr>
<td>3D</td>
<td>13.70</td>
<td>4F</td>
<td>8838</td>
<td>1.19E+18</td>
<td>13.70</td>
</tr>
<tr>
<td>3D</td>
<td>12.30</td>
<td>4G</td>
<td>10.90</td>
<td>5833</td>
<td>6.37E+17</td>
</tr>
</tbody>
</table>

**Validation Method**

Here, this study uses a Monte-Carlo method to assess the hypothesis generated from visual-computational exploration. Monte-Carlo methods are a set of mathematical tools that use randomly generated data to evaluate mathematical expressions or to achieve the distribution of some desired variables (Gentle 2003). Results that are generated from the random inputs serve as benchmarks to determine whether the phenomenon that have been observed exhibits a statistically significant difference from that generated by a random process, thus whether the phenomenon is unlikely to have occurred by chance.

To start, this study generates 10,000 random international trade networks. The basic idea of this data simulation process is to create trade networks with equal numbers of nodes and links, but to connect the nodes randomly. This keeps the number of nodes and links constant to make clustering results from random trade networks comparable to results from the actual ITN data.
For each random network, the degree of balkanization $B$ and Pearson correlation $P$ are calculated after performing the CONCOR algorithm. The 10,000 results offer a numerical approach to calculate the statistical significance of the original degree of balkanization and Pearson correlation by counting the percentage of random networks that have an equal or larger degree of balkanization or Pearson correlation. For the 2005 international trade data set, the degree of balkanization ($\bar{B}$) and the statistical significance (p value) of the Pearson correlation ($\bar{P}$) is calculated as follows:

$$p_B = \frac{\text{Number of random networks with } B \geq B}{\text{Total number of random networks}}$$

$$p_P = \frac{\text{Number of random networks with } P \geq P}{\text{Total number of random networks}}$$

For this analysis, we set the confidence level for $p$ at 0.05. Figure 4-5 shows the histogram of the degree of balkanization ($B$) based on all of the random trade networks. This figure shows an imperfect bell-shaped curve, culminating around 0. Its average mean is slightly less than 0, specifically -0.54. An intuitive explanation is that the countries that belong to the same cluster have a random geographic distribution for most random trade networks. The p value of $\bar{B}$ is <0.0001, which means that less than one trade network within every 10,000 random trade networks has a clustering structure that equals or exceeds that of the 2005 international trade network. In other words, the observed high degree of balkanization within the 2005 trade data is unlikely to be a randomly produced result. Thus, the network cluster with structural equivalence exhibits statistically significant geographical clustering.
Figure 4-5: The degree of balkanization of all random trade networks.

The Pearson correlation values calculated between the average GDP and the weighted within-cluster distance for all random trade networks are displayed in Figure 4-6. Unlike the previous result in Figure 4-5, the distribution of Pearson correlation values is an irregular with one peak around 0.1 and another mini-peak around 0.9. That the majority of results are associated with the peak around 0.1 can be interpreted to mean that if trade networks were random, the relationship between GDP and the weighted within-cluster distance would be irrelevant or have a very weak positive or negative relationship. The bi-modal distribution could be caused by a combination of clusters of countries with similar GDPs and the weighting procedure used. A nearly perfect correspondence between trade clusters and GDP is possible, but if trade links are broken, the patterns rapidly decohere into the default slight positive correlation. Only a small portion of random trade networks exhibit a strong positive relationship between these two variables. The p value is 0.0171, which is significant at the 0.05 confidence level. It indicates that less than 2 of every 100 random trade networks display a stronger correlation between GDP and weighted within-cluster distance than found in the actual 2005 ITN data. In other words, the observed strong positive relationship from the visual exploration is unlikely to occur randomly,
and the positive relationship between weighted geographical distance within each network cluster and the development level of countries is statistically significant.

Figure 4-6: The Pearson correlation values between GDP and weighted within-cluster distance of all random trade networks.

Robustness of the validation method

This study uses two approaches to test the robustness of the validation results. The first approach is to change the number of runs for each Monte-Carlo validation. The second approach is to create random trade networks with different total connection numbers. For both approaches, we keep the number of nodes constant to make clustering results from random trade networks comparable to original results. If two tests exhibit consistent results with minor fluctuations, such results support a contention that our validation method is robust against these kinds of changes. Similar test approaches have been used in other fields, such as meteorology (Anderson 2012).

The first approach examines whether the number of runs in each Monte-Carlo validation influences the final results. If results are robust, validation results will converge as the number of runs increases. Figure 4-7 displays the results in which the number of runs (N) is 1,000, 2,000, 5,000 and 10,000. When N is small, such as 1,000, the results display some reasonable
fluctuations. As the number of runs rises, those results are smoothed and finally converge (as shown by the turquoise line on each plot representing 10,000 runs).

Figure 4-7: Validation results as a function of number of runs (N).

The second approach uses different numbers of connections among nodes to test the robustness of the validation. We examine the robustness with 50%, 75%, 100%, 150%, and 200% of the original connection number and rerun the validation methods. Figure 8 shows that the distributions of degree of balkanization and Pearson correlation are largely consistent based on the five different scenarios.

Figure 4-8: Validation results as a function of total connection numbers.
This section applies Monte-Carlo methods to validate the hypotheses developed from the GeoSocialApp-based visual-computational exploration of the 2005 ITN. Monte-Carlo simulation produces many randomized pseudo-networks, calculates statistical indicators, and compares the results with those from the original ITN. The results from the 2005 ITN analysis are shown to be statistically significant. In other words, the Monte-Carlo method verifies that the patterns observed from the GeoSocialApp are unlikely to have resulted from random processes. Moreover, this research tests the robustness of the validation methods by changing the number of runs and the number of connections. In both scenarios, the Monte-Carlo method produces consistent results, which provides evidence that our validation method is robust.

**Conclusion & Contribution**

This chapter presents the GeoSocialApp, a visual analytics application that supports exploration of the complex interaction between spatial and social network relationships and demonstrates its capabilities by investigating the ITN across geographical regions at different levels of the network hierarchy. The explicit focus of the GeoSocialApp on both geographical and social representations enables a process that generates insight related to the different roles that spatial and social relationships have within the varying network hierarchy levels. To address the network relationships, the GeoSocialApp implements the CONCOR algorithm that has been used in many past studies of the ITN. Although this algorithm has known limitations (Clark 2010), the focus here is on demonstrating the potential of a geovisual analytics approach that integrates spatial and network analysis methods, not on developing novel methods to measure structural equivalence in networks. In addition, the CONCOR algorithm is still frequently used to measure structural equivalence of the ITN in recent research (Cassi et al. 2012, Zhou and Park 2012). Thus, relying on a method with a long history was appropriate. The first run of CONCOR applied
to our ITN data suggests a complex interaction between spatial and social relationships for the ITN, but also obscures the separate roles that each relationship has. The second and third run of CONCOR, identifying successively more homogeneous clusters, makes it clear that spatial constraints exist for all groups, but suggests that they are more influential for groups that include economic periphery countries.

Developing hypotheses about phenomena through visual-computational exploration is one major goal of visual analytics; but recent research recognizes that a weakness of many visual analytics methods developed thus far is that they lack mechanisms to validate the hypotheses that are generated (Keim et al. 2011, Cusumano-Towner 2009). This research develops two indicators to quantitatively assess the patterns identified through visual-computational analysis and then uses a Monte-Carlo method with robustness tests to support the hypothesis with statistical evidence. In addition to using this method to test the hypothesis, this research also uses the feedback of the first statistical analysis, as discussed in the validation section, to refine the hypotheses. This research proposes that the approach outlined here may open a new research direction to support iterative hypothesis development, testing and refinement through combined visual-computational exploration and statistical validation.

A future goal for the GeoSocialApp specifically is to integrate this validation method directly within the tools. Monte-Carlo methods are suitable to validate the statistical significance of patterns identified through visual analytics for two reasons: a) patterns revealed through visual analytics tend to be complex and at the same time knowledge about their statistical distributions is absent in most situations (e.g., no existing theory can tell us what statistical distributions international trade networks follow, so we assume that every country has equal possibility of being connected in the process of generating 10000 networks); and b) one goal of Monte-Carlo methods is to achieve the distribution of some desired variables with randomly generated data (Gentle 2003). To effectively integrate Monte-Carlo methods into the visual analytics tools, there
are two major challenges: a) how to generate random data to provide baseline distributions based on different applications; and b) Monte-Carlo methods are time-consuming processes because they need to generate a sufficiently large number, e.g., 10,000, of new random data and then calculate the distribution of the desired variables. To address the first challenge, one solution is to understand the process of pattern revelation theoretically and mathematically, and to design Monte-Carlo methods accordingly. To address the second challenge, since each Monte-Carlo realization is completely independent, one solution is to design parallel Monte-Carlo methods, and apply them within a parallel computing environment, e.g., cluster computing frameworks (Liu 2008).

In addition to integrating the validation method within the application, another future goal for the GeoSocialApp is to convey more information with novel visual designs to improve the process of hypothesis generation. For example, in the radial graphical view, more information (e.g., the distance or GDP distribution within each cluster) could have been symbolized. For the map view, one potentially useful addition might be a paired distance histogram (with 5-7 bins of short to long distance) that summarizes the distribution of between country distances for any selected cluster. In this way, more attribute information can be visualized on the map and network views to understand the interaction between geographical space and social network space.

Social network approaches have been widely applied to study the ITN, with a focus on the importance of network positions and relationships (Shutters and Muneepeerakul 2012, Ercsey-Ravasz et al. 2012, Kali and Reyes 2007, De Benedictis and Tajoli 2011). Fagiolo et al. (2009) argue that the role of geographical proximity in shaping the structure of the ITN has not been explored, especially across geographical regions. To fill this gap, recent research integrates two important approaches in the study of global trade: social network analysis and the gravity model (Zhou and Park 2012, Fagiolo 2010). The researchers add network parameters into gravity models to represent the impact of the global trade network on bilateral trade, but those models are
still not complex enough to consider both relationships across different geographical regions at varying levels. The hypothesis developed here through visual-computational exploration and then assessed through statistical validation can be considered as another effort toward future international trade models that consider more fully the complex geo-social interactions that occur across different geographical regions at varying levels. The next step will extend the analysis to the temporal domain in order to understand how such geo-social patterns do change over a longer time period (e.g., from 1989 to 2009).

This study applies the proposed indicator: degree of balkanization to all eight clusters identified through the CONCOR process. Based upon the 8 clusters identified by the CONCOR process, distances among all pairs of countries are calculated and Balkanization is measured as the difference of (a) mean within-cluster distances and (b) mean between-cluster distances. In the future, it would be insightful to compare this overall Balkanization metric to the same metric calculated on a cluster-by-cluster basis. In other words, this study will apply the indicator to eight clusters one by one to get an idea whether each cluster supports the first hypothesis (international trade network clusters with structural equivalence are strongly 'balkanized' (spatially fragmented) according to geography of trading partners), followed by the application of the indicator to all eight clusters together. Given that the clusters in terms of core countries (which tend to be spatially grouped and spatially dispersed at the same time) demonstrate greater balkanization than periphery country clusters (which tend to be only spatially grouped), this study will consider using the concept of a small-world network (Watts and Strogatz 1998) to provide an another metric to quantify international trade network. The small-world phenomenon characterizes highly local clustering and short global separation (Watts 1999). Given that the international trade network clusters identified through the CONCOR algorithm exhibit what can be described as a similar small-world phenomenon (global distribution of local trade clusters), it is possible to
explore the potential extension of small-world phenomenon in the international trade network from a geographical perspective.

Given that Pearson correlation is sensitive to the sample size, the high correlation of 0.97 between geographic proximity weighted by population and the development level of countries should be interpreted with caution. However, the goal of this chapter is not to produce the definitive analysis of the ITN but to demonstrate the value of applying a geovisual analytics approach as a method to account for both geographic and social network factors in complex processes. Application of the visual-computational methods was able to generate hypotheses about the interaction between level of economic development for countries and relative proximity of international trading partners and the statistical analysis (of which the Pearson correlation is a part) was used to provide support for the hypotheses. The positive relations are further validated statistically and robustly through application of a Monte-Carlo method. The future work will consider using a Wilcoxon rank sum test (Rosner 2011) and other similar non-parametric methods to complement the results from Pearson correlation for three reasons: a Wilcoxon rank sum test works well even if sample size is small; a Wilcoxon rank sum test conducts a formal statistical test and computes a p-value, which provides quantitative information in comparison with descriptive methods like Pearson correlation; and non-parametric methods have fewer assumptions and are applicable to more general situations.

The combination of spatial and social network context supports exploration of the interaction between these components and consideration of their impact on outcomes of interest (Adams et al. 2012), but the combination has not received enough attention generally, not just with respect to the ITN. The GeoSocialApp provides generic frameworks to explore any analysis contexts that include spatial and social relationships among geographical regions (e.g., human migrants among different states in the U.S., war conflicts among different countries in the world, vector borne disease propagation, or the impact of social media on behavior in the world). To our
knowledge, this is the first tool to allow users to explore the interconnections of spatial and social relationships at a geographical region level.
Chapter 5

Agent-based GeoSocial Visual Analytics for Epidemic Control

Introduction

Following Chapter 4, this chapter is also building on the base of theory and review of the literature in the previous chapters. This chapter presents the development of the Agent-based GeoSocial Visual Analytics and its application to the epidemic control.

Epidemics cause a huge cost to society. The 1918 influenza pandemic infected one third of the world’s population and caused 50 million deaths worldwide (Taubenberger and Morens 2006). Severe Acute Respiratory Syndrome (SARS) and Swine/H1N1 Influenza have had a dramatic impact over most of the world in the 21st century (Leung et al. 2004, Fraser et al. 2009). Although the world’s public health system has made tremendous efforts to detect, prepare and control epidemics, epidemic outbreaks of novel infection (i.e., Middle East Respiratory Syndrome) will continue to occur, exacerbated by the increasing urbanization and the mobility of contemporary society (Anderson et al. 2004). Therefore, how to effectively control epidemic transmission is still an open question.

The spread of airborne diseases occurs via the network of physical contacts among individuals through a population over space and time (Meyers et al. 2005). Therefore, human contact patterns and movement patterns over space and time play important roles to determine spread of diseases such as SARS (McLean 2005, Lloyd-Smith et al. 2005), influenza (Halloran et al. 2008, Longini Jr et al. 2005), and many others. Previous research has demonstrated that a

---

To provide a logical self-contained discussion, this chapter includes a few small sections that repeat or summarize material presented in earlier chapters and this chapter has its own literature review tailored to the application developed.
better understanding of the underlying network structure gives insight into disease dynamics and control strategies (Read, Eames and Edmunds 2008, Salathé et al. 2010, Meyers et al. 2003, Eubank et al. 2006). However, the complexity of human interaction network structures, with spatial and temporal considerations, restricts human ability to analyze the structures, discover useful patterns, and facilitate the pandemic decision-making process (Guo 2007).

Researchers use agent-based epidemic models to simulate disease transmission and assess control scenarios over a human contact network. Disease transmission occurs over spatial and social spaces simultaneously, but little work has been done to effectively analyze network structure from an integrative spatial-social perspective before the simulation (Luo and MacEachren 2014). The lack of effective network structure analysis over spatial and social space can limit the potential of simulation models to enable understanding of disease transmission and to support the design of advanced control scenarios. Therefore, this research proposes a new concept in terms of an effective disease control that should start from an understanding of the network structure over spatial and social spaces, designing effective control measures accordingly, and evaluating the efficacy of different control measures. Such requirements pose considerable challenges for epidemic models in terms of disease control design.

To address these challenges, this research proposes a novel visual analytics tool to implement the new concept in terms of an effective disease control strategy. The proposed visual analytics tool aims to achieve the following linked research objectives: 1) Develop visual analytics methods and tools to represent complex human interaction data as geo-social forms that can facilitate the discovery of useful patterns in terms of disease spread and transmission control; 2) Develop methods to transform the useful patterns into reliable knowledge to support decision-making processes in epidemic control. The visual analytics tool consists of three components: geo-social network data mining, agent-based epidemic models, and combined visualization methods. The geo-social network data mining allows users to analyze network structures and
identify useful patterns in terms of disease transmission and control. The combined visual-computational methods allow users to transform the useful patterns into knowledge to design advanced control scenarios. The agent-based epidemic models allow users to evaluate the efficacy of the advanced control scenarios.

The paper begins below by reviewing agent-based epidemic models and relevant work in visual analytics (Section 2). This study then presents human interaction data from a French primary school collected and made available for research purposes by Stehle et al. (2011) (Section 3) and an overview of the proposed method: agent-based geosocial epidemic visual analytics (Section 4). The results obtained through applying the method to the data provide insights on the effective control scenarios with spatial and social considerations of network structures (Section 5). Finally, this paper presents conclusions and an outlook for future research (Section 6).

**Related Research**

This section reviews the state of art in agent-based epidemic models and relevant work in visual analytics and discusses how this research can make contributions in both areas. In terms of epidemic control, this study addresses decisions about vaccination, and implements various immunization strategies into the geosocial visual analytics tool to allow the design and testing of advanced control scenarios. In addition to vaccination control strategies, a comprehensive review about all of the control strategies used in the previous research can be found in Lee et al. (2009b). In addition to the integration of agent-based epidemic models into visual analytics, there is also other research that integrates population-based epidemic models into visual analytics (Broeck et al. 2011, Maciejewski et al. 2011).
Agent-based Epidemic Models and Vaccination Strategies

Agent-based epidemic models are based on network structures, in which each node is regarded as an agent and links between nodes represent possible infection channels between individuals (Bian and Liebner 2007). Contact networks are built by a series of nodes with social or spatial locations and links between those nodes, which are fundamentally linked to the spatial spread of infectious disease (Meyers 2007). Knowledge of the network structure allows models to simulate infection dissemination in a population based on individual-level behaviors (Keeling and Eames 2005). At this point, network topology plays a significant role in the speed and extent of epidemic dynamics within a population (Koopman 2004). Therefore, epidemiologists currently use agent-based epidemic models to simulate disease transmission and corresponding control scenarios on different network structures.

Given the limited supply of vaccines, vaccinations aim to achieve the highest efficacy through immunizing a fraction of the population (Longini and Halloran 2005, Emanuel and Wertheimer 2010). Current vaccination strategies identify the targeted population with combinations of different network structures (Mao and Bian 2010a, Salathé et al. 2010, Masuda 2009, Zanette and Kuperman 2002, Carrat et al. 2006). The basic idea of targeted strategies is first to rank the importance of nodes and then remove (vaccinate/isolate) the nodes from highest importance to lowest. Such important measures include degree centrality (the number of links connected directly by others) (Wasserman and Faust 1994), Eigenvector centrality (individuals who are connected to many well-connected peers are more central than those who are connected to an identical number of poorly-connected peers) (Bonacich 1972), betweenness centrality (capturing the extent to which a particular node lies on the bridge among different communities) (Freeman 1978), closeness centrality (assigning scores based on the mean distance to each other vertex) (Freeman 1978), and travel distance (assigning scores based on the daily geographical
distance each individual travels). For example, targeting high degree individuals for immunization is an effective control strategy (Pastor-Satorras and Vespignani 2002, Zanette and Kuperman 2002), when community structure is not strong. When community structure becomes stronger, targeting individuals bridging communities becomes more effective than targeting individuals with high degree (Salathé and Jones 2010).

One major drawback in the above vaccination strategies is that those strategies assume that the disease outbreak occurs within the entire population. According to the assumption, those strategies remove a fraction of individuals based on the whole network. Previous research shows that disease transmission (i.e., influenza) starts with a local growth followed by a long distance transmission to the whole population (Mao and Bian 2010b), which results in a specific research question: whether applying the vaccination strategies to the local network can achieve better control efficacy than applying them to the whole network. This research question brings a new challenge: how to effectively analyze network structures to identify local human interaction patterns to which epidemiologists can apply the vaccination strategies? To address this challenge, this research proposes a new visual analytics tool that will be discussed in the Methodology section.

**Visual Analytics in Agent-Based Epidemic Models**

Visual analytics is an emerging research area that aims to leverage the power of human reasoning and computational analysis through visual interfaces that enable analysts to turn complex data into useful information and knowledge (Thomas and Cook 2005). Several studies have been done to integrate agent-based epidemic models with visual analytics tools, in order to allow users to enable analysts to set up parameters to simulate disease transmission and design control scenarios. The Epi-Fast tool allows for a disease transmission and public health
intervention simulation based on the explicit representation of social contact networks among individuals (Bisset et al. 2009, Bisset and Marathe 2009). The epidemic models and interventions are pre-configured into the tool, so it does not allow users to explore the social contact networks to identify human interaction patterns to design advanced control scenarios. Guo (2007) develops a visual analytics tool to allow the identification of human interaction patterns to support the design of pandemic control scenarios. The tool implements a new algorithm to identify human interaction patterns, but it does not include agent-based epidemic models to allow users to design and evaluate the efficacy of control scenarios considering the patterns identified through the tool.

The above discussion illustrates that vaccination strategies with agent-based epidemic models typically consider the whole network structure rather than the local human interaction patterns. In terms of visual analytics with agent-based epidemic models, there is a gap to transform human interaction patterns into knowledge to design and evaluate the efficacy of control scenarios. This research aims to implement a new agent-based epidemic visual analytics tool to fill this gap.

Data

This study uses data on face-to-face interactions among 242 individuals including 232 children and 10 teachers, across 10 classes over two days (Thursday, October 1st 2009 and Friday, October 2nd 2009) in a French primary school collected by Stehle et al. (2011) (The data set is available at http://www.sociopatterns.org/datasets/primary-school-cumulative-networks/). Stehle et al. (2011) used a proximity-sensing infrastructure based on radiofrequency identification devices (RFID) (Cattuto et al. 2010) to measure the contact network of the primary school, in order to capture interaction information for the study of infectious disease transmission. Each node represents individuals and edges are face-to-face interactions. Each node has the attribute of
classname that indicates the corresponding grade level and class number, and the teachers are assigned as the class of “Teachers” (Table 5-1). Edges between two nodes use the attribute “duration” to describe the cumulative time between two nodes in face-to-face proximity within one day (Table 5-2). The cumulative time is measured over an interval of 20 seconds that allows RFID to assess the proximity of two individuals with a probability over 99% (Cattuto et al. 2010).

This study uses the data measured on the two days. This study converts the weighted edge table into the weighted networks among different individuals. Each node has the attribute of classname to distinguish nodes from different communities.

Table 5-1. Partial node table with two attributes: Label indicates the id of each node, and classname indicates the corresponding class groups, including grade level and class number for students and “Teachers” for teachers.

<table>
<thead>
<tr>
<th>Label</th>
<th>Classname</th>
</tr>
</thead>
<tbody>
<tr>
<td>1538</td>
<td>4A</td>
</tr>
<tr>
<td>1539</td>
<td>4A</td>
</tr>
<tr>
<td>1551</td>
<td>3B</td>
</tr>
<tr>
<td>1552</td>
<td>3B</td>
</tr>
<tr>
<td>1650</td>
<td>Teachers</td>
</tr>
<tr>
<td>1653</td>
<td>Teachers</td>
</tr>
<tr>
<td>1656</td>
<td>1B</td>
</tr>
</tbody>
</table>

Table 5-2. Partial edge table with 5 columns: Source indicates the id of source node, Target indicates the id of Target node, and Duration indicates that the cumulative time between two nodes measured in seconds within one day.

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1538</td>
<td>1539</td>
<td>260</td>
</tr>
<tr>
<td>1538</td>
<td>1545</td>
<td>120</td>
</tr>
<tr>
<td>1538</td>
<td>1546</td>
<td>660</td>
</tr>
<tr>
<td>1538</td>
<td>1548</td>
<td>60</td>
</tr>
<tr>
<td>1538</td>
<td>1549</td>
<td>40</td>
</tr>
<tr>
<td>1538</td>
<td>1618</td>
<td>360</td>
</tr>
<tr>
<td>1538</td>
<td>1653</td>
<td>420</td>
</tr>
</tbody>
</table>
Methodology: Agent-Based GeoSocial Epidemic Visual Analytics (GS-EpiViz)

This study treats each class as the basic local human interaction unit for two reasons: the network density for within class interactions is substantially higher than between class interactions; it is practical to implement control strategies based on spatial confinement (e.g., class, household, school). Based on the basic unit, the average time duration, which is equal to total time duration/(the basic time measure unit of 20 seconds*all potential connections), between two communities in the network is used to measure the social connection strength among communities. Communities with strong social connections can result in a high probability with which infections can spread. According to the disease transmission characteristics, communities with strong social connections can develop higher level human interaction patterns from one community to another. The disease transmission characteristics at the individual-level within each class and at the community level among different classes can transform the network into a hierarchical structure.

Visualizing the hierarchical structure of the social network in an appropriate way allows users to design vaccination strategies with spatial-social interaction patterns according to the position of infected cases and connection strength among different communities. Network visualization has a rich history (Di Battista et al. 1998, Herman, Melançon and Marshall 2000) that has generated many variants on two primary categories of network visualization methods: node-link visualization and matrix-based visualization (Henry, Fekete and McGuffin 2007). This study applies the matrix view to represent our network data (Figure 5-1) for two reasons: a matrix view has the advantage to exhibit high-level structures (relationships between different communities) by finding the proper ordering of rows and columns (Bertin and Barbut 1967); the proper ordering of rows and columns in this application is determined by the communities in which infected cases are located. For example, the matrix view on the right in the Figure 5-1
displays the interaction data with rows and columns organized by grade and class from 1st grade class A at the top left to 5th grade class B in the lower right. The first infected individual in the scenario is in grade/class 5A. Figure 5-2 reorders all classes in the social space: it puts group 5A, with the first infected case, on the top left followed by other groups according to the social connection strength from the highest to the lowest. Matrix views in Figure 5-3 and Figure 5-4 show the network densities among different communities. Using this ability to sort the aggregate entities (classes) based on social connection strength, users can identify the communities with the strongest connections to the community having the first infected case and then focus vaccination or other preventative measures within those communities rather than applying the response uniformly to all communities. This targeted response is potentially more efficient (in overall use of resources) and more effective (in minimizing the proportion of individuals who are infected).

Figure 5-1. GS-EpiViz consists of four major components: display panel, control panel, xy plot, and matrix view. Display panel in the top left allows users to select different matrix displays:
binary matrix, time duration matrix, reorder binary matrix, and reorder time duration matrix. Control panels allow users to design different control scenarios based on the whole population and selected population. XY plot displays the accumulated infected cases over time based on the simulation of agent-based epidemic models. The matrix view visualizes different matrix displays with class information on the bottom and on the right. Black cells in the matrix view represent human interactions at the individual level, and white cells indicate there are no human interactions.

Figure 5-2. GS-EpiViz displays the reordered binary network with new class numbers on the bottom and on the right.
Figure 5-3. GS-EpiViz displays the time duration network matrix with the average time duration values in each cell.
Figure 5.4. GS-EpiViz displays the reorder time duration network matrix with the average time duration values in each cell.

In addition to representing human interaction network data in the different displays to support control strategy design, this tool also implements agent-based epidemic models to simulate disease transmission and different control scenarios. Specifically, each individual in the population is assigned to a disease state (susceptible, exposed, infectious, recovered). A single influenza infection is randomly introduced into the network with all other initially susceptible individuals. Influenza is chosen as an exemplar because it is a common infectious disease with reasonably well known transmission characteristics. This study assumes that transmission can only occur during the day time, and only on weekdays (thus when the individuals involved are at the school). Though this simplifying assumption is not realistic, it allows an analyst to analyze the disease spread and design control scenarios starting from one single infected case without considering multiple introductions of infected cases.
After coming in contact with an infection, a susceptible individual has a transmission probability 0.0015 per 20 seconds of contact (basic duration measurement unit) to be infected. This value has been chosen because it generates values of $R_0$ (basic productive number) consistent with observed $R_0$ of influenza (0.9 to 2.1) in previous studies (Mills, Robins and Lipsitch 2004, Ferguson et al. 2005). $R_0$ is defined as the average number of secondary cases generated in a susceptible population (Diekmann, Heesterbeek and Metz 1990). The calculation of $R_0$ in this case study follows the steps: randomly generate 1 new infected case 100 times, simulate disease transmissions 100 times for each new infected case within the network according to parameters from the Table 3, and then calculate average $R_0$ for each simulation. Both of the derived $R_0$ based on the two networks respectively are approximately equal to 1.8~1.9 which fall into the observed $R_0$ influenza (0.9 to 2.1). Upon infection, the individual enters into the exposed period (infected but not infectious). The mean exposed days, 3 days, will be used in this simulation (Heymann 2004). After the exposed period, an exposed individual will become symptomatic and infectious. The infectious period used in this project is 7 days, the mean days for patients who recovered (Heymann 2004). This study assumes that individuals cannot be infected again after recovery.

Figure 5-5 shows a simulated influenza infection according to the parameters in the Table 5-3.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Default Value</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total humans in simulation</td>
<td>242</td>
<td>Stehle et al. (2011)</td>
</tr>
<tr>
<td>Length of exposed period</td>
<td>3 days</td>
<td>Heymann (2004)</td>
</tr>
<tr>
<td>Length of infectious period</td>
<td>7 days</td>
<td>Heymann (2004)</td>
</tr>
<tr>
<td>$R_0$</td>
<td>0.9 to 2.1</td>
<td>Mills et al.(2004) and Ferguson et al. (2005)</td>
</tr>
<tr>
<td>Infection probability</td>
<td>0.0015</td>
<td>Measured based on $R_0$.</td>
</tr>
</tbody>
</table>
Figure 5.5. GS-EpiViz displays a simulated influenza infection according to the parameters in Table 5-3 on the matrix view and XY plot. Yellow indicates that individuals are in an exposed state, red indicates an infectious state, and the blue indicates a recovery state.

In terms of control strategies, GS-EpiViz allows users to design four vaccination strategies: random-based, degree-based, betweenness-based, and strength-based vaccination strategies; these are selected because they are the most typical ones used to compare effectiveness of vaccination strategies based on different network structures (Mao and Bian 2010a, Salathé and Jones 2010, Salathé et al. 2010). The random-based vaccination strategy randomly identifies a fraction of the population to vaccinate. The degree-based vaccination strategy identifies individuals with a large number of direct contacts for vaccination first. The betweenness-based vaccination strategy prioritizes individuals who connect between communities. The strength-based vaccination strategy prioritizes individuals whose total time exposed to others is long.

Given that the vaccination results are sensitive to vaccination rates, the tool provides a variety of options in terms of the percentage of the population vaccinated: 5%, 10%, 15%, 20%, 25%, and
30%. To compare the effectiveness of vaccination strategies based on the whole population and local human interactions as reviewed in the section of Related Research, this tool provides two panels on the left of the tool: *Whole Population Control Design* and *Selected Population Control Design*. The former panel allows users to select control strategies and control rates based on the whole population, whereas the latter panel allows users to select control strategies and control rates based on the selected areas (orange area in the Figure 5-6).

Vaccination strategies are applied at the beginning of the spread of influenza in the network. The percent of the infected population and the spatial-social extent of infection are used to evaluate the efficacy of those strategies. There are two networks, two different control regions, four strategies, six vaccination rates, yielding 96 (2*4*6) combinations to simulate. The efficacy of vaccination strategies for each combination is estimated for 100 simulation runs, resulting in a total of 9,600 epidemic simulation runs.
Results and Discussions

This study aims to compare the efficacy of vaccination strategies between the local selected areas and the whole areas based on the real world networks provided by the school interaction data. The better vaccination strategies are expected to generate a lower number of infections. Another measure of how well we can contain the epidemic locally is the number of infected cases occurring outside the selected areas. If this number is zero, the local control scenarios with the selected areas are considered successfully. Otherwise, the local control scenarios are considered as failures. Figure 5-7 shows the design of eight vaccination strategies with six different control rates in the two networks, Thursday, October 1st 2009 in (a) and Friday, October 2nd 2009 in (b). Two networks have been reordered according to the strength of social connections to the classroom with the first infected case. The local control areas are highlighted in orange. The efficacy of eight vaccination strategies in the two networks is evaluated through infection percentage and the spatial-social extent of infection as described below.
The first infected individual is indicated by the red cell here.
Figure 5-7-b. The design of 8 vaccination strategies with six different control rates in the control panel. The matrix view visualizes the reordering binary matrix network, Thursday, October 1st 2009 in (a) and Friday, October 2nd 2009 in (b). The local control areas are highlighted in orange.

Vaccination Strategies in terms of Infection Percentage

The better vaccination strategies in terms of infection percentage are expected to generate a smaller number of infections. Figure 5-8-a and Figure 5-8-b show that all of the eight vaccination strategies can produce a decreasing number of infections in proportion to the increasing vaccination fractions. The random-based vaccination strategy produces the highest number of infections, followed by the random-based vaccination strategy with the local selected
control areas in Figure 5-7-a and Figure 5-7-b. The other three pairs: degree-based, betweenness-based, and strength-based vaccination strategies exhibit the same pattern: the three strategies with the local selected control areas outperform those strategies with the whole area, respectively. The explanation for the pattern is illustrated in Figure 5-9 and Figure 5-10. Figure 5-9 displays the percentage of the local control success for the four local vaccination strategies within 100 simulation runs. As the vaccination fractions increase, all of the four local vaccination strategies can produce a higher percentage of the local control success, but to different degrees. In addition, when the local control scenarios are successful, they can produce a significantly lower number of infected cases (Figure 5-10). Those results show that when the vaccination fraction reaches a relevantly high level (i.e., 30%) with the selected control areas, the disease transmission can be confined locally at a very high percentage (i.e., 90% to 95%), resulting in only a small number of infections (i.e., 2 to 5). Figure 5-9 and Figure 5-10 show that the former can stop the disease transmission locally at a certain probability (Figure 5-9), which also results in the lower number of infections (Figure 5-10).

Figure 5-9 shows that the disease transmission cannot be confined locally at 100% percentage, which also results in the high number of infections (Figure 5-11). For example, when the local vaccination fraction in terms of degree-based, betweenness-based and local strength-based control strategies reaches to 30%, a high number of individuals (i.e., 25 to 50) in Figure 5-11 are infected at a very low percentage (5% to 15%) in Figure 5-9. Those results suggest that containing disease outbreaks locally should be highly recommended, but the follow-up control strategies are needed if the local control strategies are not successful.
Figure 5-8-a

Figure 5-8-b

Figure 5-8. The efficacy of 8 vaccination strategies in two networks, Thursday, October 1st 2009 in (a) and Friday, October 2nd 2009 in (b). The Y axis represents the percent of infection, and the X axis represents the vaccination rates. Eight vaccination strategies are represented by eight curves in different colors and shapes, as the legend shows at the bottom.
Figure 5-9-a

Figure 5-9-b

Figure 5-9. The percentage of the local control success for four local vaccination strategies within 100 simulation runs in two networks, Thursday, October 1st 2009 in (a) and Friday, October 2nd 2009 in (b). The Y axis represents the percentage of the local containment success, and the X axis represents the vaccination fraction. Four local vaccination strategies are represented by four curves in different colors and shapes, as the legend shows at the bottom.
Figure 5-10. The total number of the final infected cases when the local vaccination strategies are successful within 100 simulation runs in two networks, Thursday, October 1\textsuperscript{st} 2009 in (a) and Friday, October 2\textsuperscript{nd} 2009 in (b). The Y axis represents the total number of the final infected cases, and the X axis represents the vaccination fraction. Four local vaccination strategies are represented by four curves in different colors and shapes, as the legend shows at the bottom.
Figure 5-11. The total number of infections when the local vaccination strategies are not successful within 100 simulation runs in two networks, Thursday, October 1st 2009 in (a) and Friday, October 2nd 2009 in (b). The Y axis represents the total number of the infected cases, and the X axis represents the vaccination fraction. Four local vaccination strategies are represented by four curves in different colors and shapes, as the legend shows at the bottom.
Vaccination Strategies in terms of Spatial-Social Extent of Infection

From a spatial-social perspective, effective vaccination strategies with the selected areas are expected to confine the disease outbreak locally. This section only displays the simulation results with the first day, Thursday, October 1\textsuperscript{st} 2009, because the second day’s results will generate the same conclusion as below. Figure 5-12 and Figure 5-13 compare spatial-social extent of affected areas between the whole area and the selected area control scenarios with 30\% vaccination rate. The average number of infections based on the simulation results is displayed on each cell in each matrix. Within the selected areas in each matrix (in purple square), local control scenarios can produce a much lower number of infections (Figure 5-13) compared to the whole area control scenarios (Figure 5-12). This is because placing the same amount of vaccines within a smaller spatial-social extent of areas would cause a lower number of infections locally. However, outside the selected areas, there is a much larger number of infections (Figure 5-13) compared to the whole area control scenarios (Figure 5-12). Each pie chart in Figure 5-13 shows the number of local control successes versus the number of local control failures with 100 simulation runs. Each bar chart in Figure 5-13 represents the average number of infections between the local control success and local control failure. It shows that there is a high probability (see pie chart) to contain the disease outbreak locally with 30\% vaccination rate, but a larger disease outbreak would occur (see bar chart) if the local control fails at a low probability. Those results suggest that if the local control scenarios fail, new control strategies have to be implemented to avoid disease outbreaks outside the selected control areas. Therefore, the efficacy of vaccination strategies in terms of spatial-social extent of infection also suggests that local vaccination control would be a good control strategy. On the other hand, this strategy should be complemented with other control strategies if the local ones are not successful at a low likelihood.
Figure 5-12. The spatial-social patterns of simulated vaccination strategies with 30% vaccination rate with the average number of infections based on 100 runs after reordering the first day network, Thursday, October 1st 2009. Figure 5-12-a is the random-based vaccination strategy, Figure 5-12-b is the degree-based vaccination strategy, Figure 5-12-c is betweenness-based vaccination strategy, and Figure 5-12-d is the strength-based vaccination strategy.
Figure 5-13. The spatial-social patterns of simulated vaccination strategies with 30% vaccination rate based on the selected areas (purple square) after reordering with the average number of infections in each cell in the first day network, Thursday, October 1st 2009. The number of infections within the selected areas is achieved through calculating the sum of infected individuals on each of the 100 runs and then dividing by 100 runs. The average number of infections outside the selected areas is calculated through dividing by the number of local control failures, because the number of infections is zero when the local control is successful. Figure 5-13-a is the random-based vaccination strategy, Figure 5-13-b is the degree-based vaccination strategy, Figure 5-13-c is betweenness-based vaccination strategy, and Figure 5-13-d is the strength-based vaccination strategy. The pie chart in each figure represents the number of local control success versus the number of local control failure with 100 simulation runs. The bar chart in each figure represents the average number of infections between the local control success and local control failure.

**Conclusion and Future work**

This research proposes a new concept in terms of effective disease control that starts from the understanding of social network structure over space and time, is followed by designing effective control measures accordingly, and then evaluates the efficacy of different control measures. This concept is used to frame design of a new visual analytic tool: GS-EpiViz. This tool first proposes a geo-social cluster identification method applicable to infectious disease transmission and control, then implements the method and agent-based epidemic models into a visually interactive environment. With real world human interaction data as a case study, this research compares the efficacy of vaccination strategies between the local selected areas and the whole areas. The simulation results show that locally targeted vaccination has the potential to keep infection to a small number and prevent spread to other regions. At some small probability, the local control strategies will fail; in these cases other control strategies will be needed. The case study shows how GS-EpiViz does support the design and testing of advanced control scenarios in airborne disease (e.g., influenza). The spatial-social patterns identified through exploring human interaction data provide useful knowledge to support decision-making processes in epidemic control.
When the local control strategies fail at a small probability, other control strategies are needed to deal with infections outside the containment areas. The integration of agent-based epidemic models into visual analytics approaches allows the adjustment of control scenarios to specifically deal with those new infections. An extension of GS-EpiViz will allow the users to add/adjust control strategies in the process of simulations in the future work.

As mentioned in the section of related research, there are many other control strategies besides vaccination. For example, antiviral prophylaxis seeks to treat infected cases to prevent infection (Halloran et al. 2008), quarantine seeks to isolate contacts of infected cases effectively to prevent spread (Longini Jr et al. 2005), and so on. Extensions to GS-EpiViz will implement different control strategies to allow the design of combination strategies, because previous research shows that combination strategies are more effective than individual strategies (Ferguson et al. 2005).

In this study, vaccination strategies are applied at the beginning of the spread of influenza in the network. In realistic situations, infections, especially novel ones, cannot always be recognized at the beginning of an outbreak. If the first infected case assumed is not really the first case, it would be important to test effectiveness of local control scenarios with “hidden” cases based on GS-EpiViz. For example, I can run the experiment with random a number of “hidden” cases in the same classroom of the first assumed infected case, within the boundary of local control scenarios, or outside that boundary in order to test the effectiveness of the proposed spatial-social control scenarios. If novel infections have been reported from multiple locations before the execution of the control strategies, an extension of GS-EpiViz will allow the users to design local control strategies to deal with infections from multiple locations in the future work.

This paper applies GS-EpiViz to a relatively small set of human interaction data from a high school. The future development of GS-EpiViz needs to scale-up this tool to include much
larger and more complex human interaction data at urban, state, and even national scales. To achieve this goal, there are a series of research questions I will address. For example, what are the characteristics of human dynamics in the context of epidemic disease transmission at larger scale? How can the characteristics of human dynamics in the context of epidemic disease transmission help experts design advanced control scenarios?
Chapter 6

Discussion & Challenges

Geo-social data do not make any sense when abstracted from their appropriate contexts (Abbott 1997). Geo-social contexts do not only demonstrate conceptual and observed overlaps, but also shed light on data from different perspectives. Downs and DeSouza (2006) argue that spatial thinking serves three purposes: 1) a descriptive function, 2) an analytic function, and 3) an inferential function. Social thinking describes how human interactions impact their ideas, emotions, and behaviors (Winner 2002). A social network approach provides a means to describe human interactions and analyze and infer how such interactions impact the social thinking process. This dissertation follows Downs and DeSouza’s lead and proposes that the three purposes that spatial thinking serves can be applied to social thinking to develop integrative spatial-social thinking for people to acquire knowledge. The three purposes also match the primary goal of visual analytics: interactive visual interfaces should support human analytical reasoning in an efficient and effective way. Supporting spatial and social thinking to address the three purposes outlined is a prototypical example of meeting this goal.

This dissertation first develops a theoretical model of geosocial context (Chapter 2), and then discusses multiple geo-social visual analytics research questions that need to be addressed in relation to data exploration, decision-making, and predictive analysis (Chapter 3). It also implements the theoretical model (Chapter 2) through visual analytics approaches that account for geographical, social network, and conceptual dimensions of that context (Chapter 4 and 5).

---

7 Portions of this chapter have been previously published in Luo and MacEachren (2014). I led the research and, as stated in the published papers, my responsibilities and that of my co-authors in the research and the paper were as follows. Conceptualized the paper: WL. Wrote the paper: WL AMM.
The theoretical model of geosocial context and its practical implementation through visual analytics tools, as well as our analysis of the literature provides several potential research directions for further investigation of geo-social visual analytics. This dissertation concludes by highlighting nine core challenges that will require interdisciplinary efforts to meet. The challenges can be classified into two categories: the first category focuses on further integration of spatial and social network analytics from an interdisciplinary perspective, and the second focuses on understanding the impact of geo-social relationships on society.

An Effective Interdisciplinary Integration of Spatial and Social Network Analytics

The entire dissertation argues the importance of a more unified integration of spatial and social network analytics. Though this dissertation makes substantial progress towards the science of the integration of spatial and social network analytics, there is still much work that needs to be done. This subsection identifies five future research challenges that need to be addressed.

Understanding the interaction of geography, network, and societal space, as well as their respective and coupled impacts on outcomes of interest.

This challenge underlies the entire dissertation. This dissertation applies much more work from social network analysis than from spatial statistics to geo-social systems. One future direction should consider including current spatial statistics (e.g., Moran’s I, Getis’ G) into the geo-social systems (i.e., GeoSocialApp, GS-EpiViz). A comprehensive discussion of those spatial statistics can be found in Anselin (1995), Bailey and Gatrell (1995), and Rogerson and Yamada (2009). The potential insight of applying spatial statistics to geo-social systems can be found in the section of Literature Review in the Chapter 4 (p.42). Based on the dissertation it seems clear that understanding how spatial proximity and network relationships interact for outcomes of
interest is at an early stage (Takhteyev et al. 2012, Verdery et al. 2012). Thus, further research is required to investigate how geographical and social relationships operate explicitly in different geo-social systems. In addition to spatial analysis and social network analysis, multivariate analyses methods, that conceptually fit the societal context in Figure 2-1, can be implemented into geovisual analytics tools to understand the relationships between variables and their relevance to the spatial-social patterns being studied. This dissertation primarily focuses on the spatial-social relationships and methods to understand them; complementary research on multivariate analysis and geovisualization can be found in the following papers (Guo et al. 2006, Guo et al. 2005, Guo 2003, Chen, MacEachren and Guo 2008, Chen and MacEachren 2008, Chen et al. 2006).

**Developing theory, methods, and tools to consider spatial and network factors simultaneously.**

Most geo-social analytical approaches use independent traditional spatial analysis and social network analysis methods simultaneously to explore the same datasets (Emch et al. 2012). It is necessary to develop new theory and methods that integrate spatial and social factors together. For example, Radil et al. (2010) propose a new method that borrows the concept of social position to explore an actor’s position in a spatial contiguity matrix simultaneously with his or her position in social networks. The proposed method can identify statistically significant violence patterns that cannot be captured by the classical spatial autocorrelation method, Global Moran's I (Moran 1948, Tita and Radil 2011). Radil and his colleagues’ work is only one attempt to explore new geo-social theory and methods. As discussed in this dissertation, spatial analysis and social network analysis exhibit strong conceptual and observed overlaps, so much more future work is needed.
Understanding the dynamics of geo-social relationships and processes.

Geo-social relationships are not static but dynamic, so it is important to understand the change of the relationships over time and compare the dynamic change to the static understanding (Adams et al. 2012) as well as to investigate the linked geographic and social process that drives the change. Hess (2004) also discusses the need to involve a temporal concept into the proposed categories of embeddedness (Figure 1) through taking into account developments over time and changes in the spatial configuration of networks at different scales. This dissertation argues that three models can be developed to represent temporal geo-social relationships among geographical or individual units. The first model has fixed nodes with unchanging geographical relationships and varying social relationships. The model is based on the geo-social phenomena in which geographical relationships are fixed because of their relative geographical locations (e.g., cities, states, and nations), and social relationships change over time. Those geo-social phenomena include the dynamic trade network at a country-to-country scale (Zhou and Park 2012, Fagiolo 2010), human migration in the U.S. at a county-to-county scale (Guo 2009), and technology adoption (i.e., Twitter) at a city-to-city scale (Toole, Cha and González 2012). The second model has fixed nodes with changing geographical relationships and relatively fixed social relationships (i.e., human movement, mobile-based social media). The last model can support more complex geo-social dynamic behaviors in which networks can expand and recede (Batty 2005). Recent developments that apply the concept of “rendezvous” (bringing sensors close to one another in space or time (Honicky 2011a) to shed light on human mobility characteristics (Honicky 2011b)) provides data and methods to support the last two models.

Another important model: Hägerstrand’s space-time cube from time geography (Hägerstrand 1970), can also be considered to extend for understanding geo-social relationships. The space-time cube is like a three-dimensional “aquarium” with two-dimensional plane
representing physical areas and a vertical axis representing time. Human movement in space and time can be described as a line between the starting locations at the starting time and the ending locations at the ending time. Miller (1991) demonstrated how the space-time cube could be applied to modern GIS transportation networks. In addition, Kwan (1998) demonstrated how the space-time cube could show disparities in gender accessibility. This dissertation argues that four kinds of social networks in terms of their space-time relationships can be considered productively in future research from a space-time cube perspective: social networks at the same space and time (e.g., social interactions at a group meeting), social networks at the same space and different time (e.g., alumni networks), social networks at different space and same time (e.g., cell phone call networks at the same time), and social networks at different space and time (e.g., the evolution of friend networks).

Integrating distinct applications of cognitive science to support geo-social visual analytics.

Cognitive science provides theoretical frameworks for the design of geovisualization tools (MacEachren 2004), it provides a conceptual approach (e.g., distributed cognition) to understand human reasoning as enabled by visual tools, and it also offers fundamental theories and approaches to understand and model human behaviors in network science, as discussed in the decision-making section. For example, the Organizational Risk Analyzer (ORA) uses both network theory and social psychology to model human behaviors, and ORA has been used to analyze 1500 videos made by insurgents in Iraq and effectively reduce sniper activity by 70% (Bohannon 2009a). From a social cognitive perspective, human behaviors result from an interaction between human internal cognition and externally environmental effects (Bandura 2001). The external effects also fit the three spaces in the proposed conceptual framework for geo-social relationships (Figure 1), because the external effects include human socioeconomic
status (societal embeddedness), their relationships with others (network embeddedness), and materials provided by a specific location in which the person is located (territorial embeddedness). Involving social cognitive theory in modelling human behaviors may make contributions to predict geo-social systems in crisis situations discussed in the predictive analysis section. Therefore, cognitive science should not only be used to design visualization tools (Fisher, Green and Arias-Hernández 2011), but to support geo-social analytic models as well.

**Developing new geo-social visual analytics methods to incorporate data exploration, decision-making, and predictive analysis as a whole.**

As discussed in chapter 3, most geo-social visual analytics methods only support one step, so visual analytics cannot effectively transform knowledge through visual exploration into complex analytical strategies directly. One possible solution is to improve inter-disciplinary cooperation through understanding the human analytical reasoning of real decision-makers to design visual analytic tools accordingly (Andrienko et al. 2011), such as has been attempted for maritime anomaly detection (Riveiro 2011), bridge management system analysis (Wang et al. 2010), and other application domains. In addition, visual analytics have not synergistically integrated computational methods to maximize human conceptual, perceptual and reasoning capabilities in the whole scientific and problem-solving process. One possible theoretical framework to link the whole complex analytical process can be found in Gahegan (2005), situating human reasoning, concretized representation, conceptual structures, visual representation, and mathematical models into the whole science process. To support a full range of applications of geo-social visual analytics, however, the approach must be generalized beyond the context of scientific research, which was the target of Gahegan’s model.
The Impact of Geo-Social Relationships on Society

In addition to geo-social relationships everywhere in our daily life (i.e., social media, and mobile phone), such relationships can influence our daily behaviors (i.e., where to work or live) that have an impact on our environment/society. To understand the geo-social relationships better, it is necessary to put such relationships in the broader context of human environment interaction. This subsection identifies four future research challenges that need to be addressed.

Understanding the transition of daily life habits that further impact local communities and networks because of spatial decision-making.

As reviewed in the decision-making section, GIS-MCDM includes broad application domains (e.g., transportation, urban planning). Decision-making in terms of those domains involves interests among different stakeholders, but it also has a corresponding influence on the practices of everyday life. The influence will further impact local community interaction and network structures. For example, people who used to live in hutongs (traditional alleyway neighbourhoods) in Beijing report a substantial disruption of the high quality and frequency of local interaction they had in the hutong compared to that after they were relocated to mega-block high rise apartment complexes on the city periphery when their neighbourhood underwent urban renewal (Rock 2012). Tita et al. (2011) point to a similar impact of urban redevelopment on social networks in their argument that the clear north-south geographical division in the gang rivalry networks in a section of Los Angeles is due to a landscape feature: the San Bernadino Freeway. Although spatial decision making has a significant role in shifting local community network structures, research in GIS-MCDM has not taken such shifts into account.
Understanding the shift of the traditional decision-making approaches with the emergence of social media.

The emergence of social media (i.e., Facebook, Twitter, LinkedIn) changes the world via collective power through on-line social networks. One person can communicate with hundreds or even more people about products, news, cultures, and any information. The communication occurs in a smaller world than in the pre-internet era; this is illustrated by Kwak et al. (2010) who find that the average path length of a Twitter network is 4.12 compared to “six degrees of separation” (Milgram 1967) in the real world. The impact of people-to-people communication has greatly changed the traditional sense of decision-making, because social media based conversations help people to be accountable and occur outside of the direct control of decision-makers (Mangold and Faulds 2009). For example, social media is playing an increasing role in the most recent anti-government protests, including the Arab Spring, Occupy Wall Street, and the London Riots (Tsou and Yang 2012). How to use social media to offer strategies for disaster and emergency management has received substantial research attention in recent years, such as for flood (MacEachren et al. 2011b, Dashti et al. 2014), hurricane (Caragea et al. 2014), and disrupted infrastructure (Al-Akkad et al. 2014). While social media have a transformative impact on traditional decision-making approaches, strategies through which responding organizations can successfully leverage these technologies are just beginning to be considered (Yates and Paquette 2011). Hiltz, Kushma and Plotnick (2014) identify practical barriers to the use of social media in the context of disaster and emergency management through interviewing U.S. public sector emergency managers. Before effective use of social media in decision-making can be achieved, there are many unexplored research questions. For example, how does information diffuse geographically and socially via social media? How do social media change human behaviors in normal and crisis situations? How do social media transform individual voices into collective power to be accountable?
A new concept of scale?

Scale is a fundamental concept in geography. As discussed above, many spatial-social phenomena show scale-free characteristics observed across different spatial scales (Song, Havlin and Makse 2005). Barabási (2009) claims that although there are diverse dynamical processes on networks (i.e., the spread of viruses and ideas on physical space network and the flow of information over cyberspace network), it is possible that these dynamical processes share some common characteristics. Based on this argument and the relationships between geography and networks discussed in this dissertation, this research proposes that it is important to ask what the spatial factors are behind those common characteristics; and are those spatial factors scale-dependent? Through understanding geo-social relationships, is it possible to reconceptualize the concept of spatial scale as constructed by relational spatial units (i.e., human, company, city, country)? Does the cyber world bring a new life to the concept of scale? What are the common and different characteristics between the scale in the physical world and the cyber world? How does information spread in a cyberspace network interact with information spread in the physical space network?

Developing a framework to collect geo-social relationship data and assess their fitness for different applications while also considering the potential negative consequences for human privacy of collecting these data.

The potential of geo-social visual analytics, especially during natural disasters, disease outbreaks, and similar events that put people and property at risk, provides additional motivation for future collection of network and spatial data. Most popular geo-social network data collection methods include surveys (Read et al. 2008), crawling social media sites (Salathé and Khandelwal 2011), collecting data from mobile devices such as cell phones (Eagle, Pentland and Lazer 2009),
and leveraging wireless sensor technology (Kazandjieva et al. 2010). Details about each method and their pros and cons can be found in Salathé et al (2010). A key problem here is that there is no theory/framework to assess whether the collected data are suitable to study different applications. For example, Salathé and Jones (2010) study disease transmission at individual-level through building social interaction networks. Nodes represent individuals and edges consider both friendships in Facebook and physical proximity in real world (i.e., the same dorm, the same class). However, the demographics of social media users is a biased sample of the whole population (in relation to age, gender, race, etc) and such networks are still at a rather coarse resolution for the study of disease transmission. As more complete data sets become available, individual-level network data with spatial and temporal information may make it possible to predict human behaviors better, but collection of individual level data raises a range of privacy concerns (Bohannon 2009b).

To sum up, geo-social visual analytics is based on the conceptual extension of the First Law of Geography: Everything/everyone is related to everything/everyone else, but near things/individuals are more related than distant things/individuals (Tobler 1970); Nearness and relationship can be considered a matter of geographical and social network distance, relationship and interaction. The observed social phenomena in a spatio-temporal framework motivate the development of social, political, and ethical research questions to finally develop the general geo-social theory to understand the world. Thus, at the methodological level, geo-social visual analytics should facilitate the integration of computational methods with human reasoning abilities to answer research and application questions in the context of data exploration, decision-making, and predictive analysis. The resulting methods will, from an integrative perspective of spatial thinking and social science, enable research to understand the geo-social mechanisms and processes that underlie human behavior.


Ashdown, P. 2012. The global power shift. TED.


Emanuel, E. J. & A. Wertheimer. 2010. Who should get influenza vaccine when not all can? : LIT Verlag Münster.


--- (1968) Innovation diffusion as a spatial process. *Innovation diffusion as a spatial process*.


Malczewski, J. 1999. GIS and multicriteria decision analysis. JOHN WILEY Wiley & SONS, INC.


---. 2006b. Multicriteria Decision Analysis for Collaborative GIS. In Collaborative geographic information systems, eds. S. Balram & S. Dragićević. IDEA GROUP PUBLISHING.


Valente, T. 2010. *Social Networks and Health: Models, Methods, and Applications*. Oxford Univ Pr.


Winner, M. G. 2002. Thinking about you thinking about me. Michelle Garcia Winner San Jose, CA.


Yau, N. 2010. Facebook worldwide friendships mapped.


VITA:
Wei Luo

Wei Luo is an educator and researcher interested in the topics of GIScience, Geovisualization, and Geovisual Analytics. Wei completed his doctoral studies in The Pennsylvania State Department of Geography (2009-2014) under the direction of Dr. Alan MacEachren. While at Penn State, Wei was a research assistant at the Penn State GeoVISTA Center. Wei completed his Bachelors (2003-2007) studies at Northwest University Department of Urban and Environmental Science and Masters (2007-2009) studies at the University at Buffalo Department of Geography under the direction of Dr. Ling Bian. Starting Fall 2014, Wei will go to Arizona State University School of Computing, Informatics, and Decision Systems Engineering as a Postdoctoral Research Assistant.

Wei is broadly trained in the discipline of Geography with specialization in the sub-discipline of GIScience. Wei’s core research areas within Cartography, Geovisualization, Geo-Social Visual Analytics, Spatial Analysis, Network Analysis, Complex Systems, Machine Learning, GEOINT, GeoInformatics, and Epidemic Modeling with applications to the geographic components of Big Data, Economics, Social Media, Urban Dynamics, Spatial Epidemiology/Public Health, International Trade, Crime Analysis, and Crisis Management.

Wei is an active contributor to the GIScience community, coordinating multiple university and national activities and participating in a variety of academic and professional societies. At the time of submitting the dissertation, Wei served as Program Committee of IEEE International Conference on Big Data and 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN 2014), as well as Education Committee of The International Association of Chinese Professionals in Geographical Information Sciences (CPGIS). Wei also is an active member of Association of American Geographers (AAG) and International Cartographic Association (ICA) Commission on Cognitive Visualization. Among other awards and distinctions, Wei was named recipient of the 2004-2006 Northwest University Scholarship, 2010 GIScience Best Paper Award led by Dr. Rob Roth, and the 2012 E. Willard Miller Award in Geography 1st Place.

Wei was born in XIAN, SHAANXI (CHINA) and grew up in XIAN, SHAANXI (CHINA). Wei currently resides in STATE COLLEGE, PA.