# COMPUTATIONAL AND VISUAL ANALYSES OF SPATIAL INTERACTIONS: A CASE STUDY OF THE COUNTY-TO-COUNTY MIGRATION IN THE US

by

Ke Liao

Bachelor of Science
Lanzhou University, 1998

Master of Science
Northern Illinois University, 2004

---

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Geography

College of Arts and Sciences

University of South Carolina

2011

Accepted by:

Diansheng Guo, Major Professor

Susan L. Cutter, Committee Member

Michael E. Hodgson, Committee Member

Linyuan L. Lu, Committee Member

Tim Mousseau, Dean of The Graduate School





ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

# ACKNOWLEDGEMENTS

My sincerest gratitude goes to my advisor, Dr. Diansheng Guo. By being a magnificent example, he taught me what research and what doing research are about. His passion, insights, and persistence to high standards have been the resources I count on. Thank him for his great guidance throughout this dissertation process. His advising and encouragement has made this dissertation possible.

I would also like to express my thankfulness toward my great committee, Dr. Susan Cutter, Dr. Michael Hodgson, and Dr. Linyuan Lu for their help and patience. I am in debt to them for their huge support in agreeing me to serve on my committee. I thank faculty members in this department for teaching me, directly and indirectly. I am also grateful to the staff here for their wonderful assistance.

My thanks are extended to Hai Jin, Caglar Koylu, Peng Gao, and Hu Wang for their amazing help on data preparation, programming, and their inspirations.

I am thankful for my family in China, my husband, and my precious son. My husband has provided support as much as he could. My son often "delighted" me by pulling me away from my computer desk. I thank them for their love.

## ABSTRACT

Spatial interactions (SI), such as human daily movements, disease spread, and commodity flows, are among the essential forces that drive many physical and socioeconomic processes. Spatial interactions are very complex in nature. A normal SI data set often contains three different data spaces: the geographic space, the graph/network space, and the multivariate space.

The goal of this research is to address the underutilization and the underrepresentation of SI data. Currently there is a lack of powerful exploratory analytic methods that can deal with the complexity of spatial interactions, which often involve: (1) multiple data spaces, (2) various spatial constraints, (3) many variables for locations and interactions (flows), and (4) the large data size. It is unlikely that an individual method alone can fully address these challenges.

This dissertation develops an integrated computational-visual approach to examining SI data from different perspectives and synthesizing different perspective views into a holistic understanding. The contribution of this research is two-fold. *First*, it develops a graph partitioning method to discover spatially contiguous community patterns (SI regions). Evaluations with benchmark data indicate that the developed method is more effective and more computationally efficient than traditional methods.

*Second*, this research uses SI regions as a data aggregation strategy to summarize massive spatial flows. It combines the three SI data spaces in data

exploration and representation. SI regions, multivariate patterns, and geographic patterns of SI flows are analyzed simultaneously in a novel and interactive visual analytic system.

A large inter-county migration data set of the U.S. is used to assess the developed approach and implemented visual analytic system from an application perspective. The data contains over 700,000 county-to-county migration flows (i.e., origin–destination pairs). The results demonstrate that the SI regions obtained by analyzing the spatial information and network connections can unveil real-world structures such as the strong "core-suburban relationship" from a network perspective.

A focused study on income migration shows that the developed integrative approach is able to synthesize the various data spaces, address the high-dimensions, and cope with the large size of SI data. The combination of graph partition, multivariate visualization, flow mapping, and interactive interfaces creates a flexible, comprehensive, and efficient environment to explore SI data from different perspectives and obtain holistic understandings. This reported approach facilitates new and comprehensive analyses that existing research methodologies cannot support.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

Spatial interaction (SI) is a universal and important geographic phenomenon. Movements of humans or entities across the geographical space are widely observed, such as migration, flows of goods, daily travels, and exchanges of knowledge, disease spread, and animal movements. As an example, nearly half (46%) of the U.S. population changed their residences between 1995 and 2000.

The analysis and understanding of spatial interactions is crucial to a wide range of real-world practical problems and research domains, such as predicting trends of local demographic changes (Fotheringham et al. 2004, Pellegrini and Fotheringham 2002, Clark and Hunter 1992), assisting decision-making in business planning or emergency management (e.g., pandemic mitigation) (Ferguson et al. 2006, Guo 2007, Ferguson et al. 2005), and facilitating the development of public policies (Greenwood 1997). For example, analyzing place-to-place connections established by human activities can provide insights on how the virus may spread over space and time. Effective mitigation strategies can then be planned accordingly for potential pandemic outbreaks.

Existing methods for spatial interaction analysis are mostly theory-driven and model-based, e.g., various spatial interaction models (Plane and Bitter 1997, Cushing and Poot 2004). A model-based approach attempts to predict spatial interactions by fitting observed SI data to a predefined model, such as a gravity model or a regression model. The advantage of modeling approaches lies in its ability to work with incomplete data

and incorporate known theory explicitly. However, it has been noted by many researchers that it is difficult to design appropriate SI models due to our limited understanding of the complex processes that drive spatial interactions, including the spatiality of spatial interactions (i.e. distance, adjacency) (Chun and Griffith 2011, Griffith and Jones 1980, Curry et al. 1975, Sheppard et al. 1976, Mueser 1989). Some argued that the modeling approach has arrived at a stage where a breakthrough is needed by learning new knowledge from observational data so that spatial interaction models can be better designed (Roy and Thill 2004, Rae 2009, Young 2002).

Currently there is a lack of powerful exploratory analytic methods that can deal with the complexity of spatial interactions, which often involve: (1) multiple data spaces (i.e., geographic space, network space, and multivariate space), (2) various spatial constraints (e.g., travel distances, geographic contiguity, and physical barriers), (3) many variables for locations and interactions, and (4) large size: a moderate-sized dataset which involves 50-1,000 locations can easily have thousands or millions of connections.

It is unlikely that an individual method alone can fully address these challenges. For instance, computational methods have the advantage to extract patterns from large/complex data set but the extracted patterns are not directly or automatically comprehensible. Visual analytics can facilitate the understanding and communication of data and information. However, representing the data itself is not adequate or efficient enough for us to understand the patterns the data describes. This dissertation brings together different and complementary methods and develops an integrated computational-visual approach to examining SI data from different perspectives and synthesizing different perspective views into a holistic understanding.

## 1.1 Background and Motivation

Spatial interactions are very complex in nature. A spatial interaction data set usually contains a set of geographic locations, flows among them, and a set of variables associated with each location and each flow. Figure 1.1 shows the three distinctively different data spaces in a normal SI data set: (1) the geographic space, with spatial entities (e.g.. origins, destinations and flows) and measures (e.g., proximity, distance and spatial autocorrelation) (Figure 1.1a); (2) the graph/network space, with directed and weighted flow connections among locations (Figure 1.1b); (3) the multivariate space, with variables of flows (e.g. stratification of flows by income or race) and variables of locations (e.g., population, median income, and unemployment). Flows among locations are conveniently represented with a directed or undirected *N*N* matrix, where *N* stands for the number of locations (Figure 1.1c). Each cell in the matrix can be accompanied with a vector of descriptors.

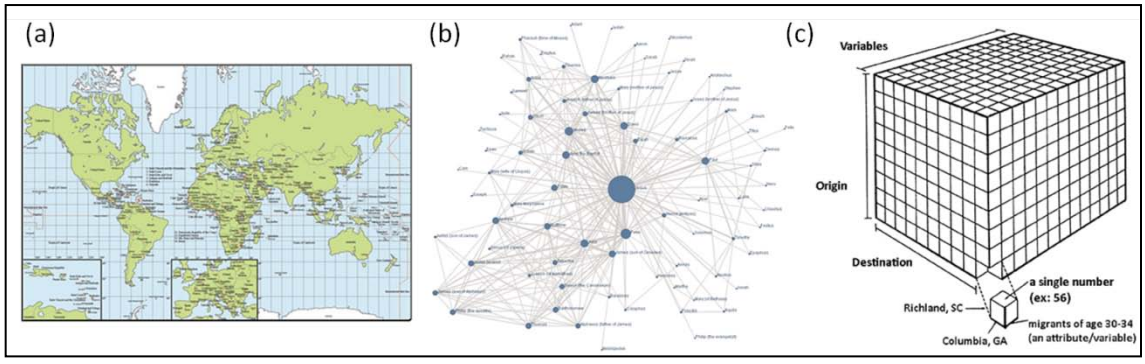


Figure 1.1 Data spaces in spatial interaction (SI) data. (a) Geographic space (source: www.webresourcesdepot.com). (b) Graph space (source: www.adrants.com). (c) Multivariate space (source: geography.uoregon.edu).

The geographic space is what distinguishes spatial interaction data from other networks such as citation networks or social networks, which do not necessarily involve

geographic locations. There are a number of visual representations for spatial interactions, including flow matrix view, arrow graph, and flow map. Each form tends to focus on a selected perspective of spatial interactions. For example, a matrix view shows location-to-location connections but ignores other spaces such as the spatial arrangement of locations and the multivariate information of locations. An arrow graph, on the other hand, groups locations/places into several classes (such as urban and rural sectors), aggregates the flows, and visualizes the flows among the different classes of locations (such as flows from urban to rural or vice versa). This visualization strategy of SI flows inevitably misses the spatial information and the multivariate structure. Since a flow map is able to present the spatial and network information simultaneously, it remains the most effective approach to visualizing spatial interactions

However, the large volume of SI data poses a great challenge for flow maps. A flow map has a severe scalability problem. It cannot work with medium- or large-sized data sets. For example, a flow map can become cluttered and difficult to read for migration flows among 48 states in the conterminous U.S. (i.e., a 48x48 flow matrix) (Tobler 2004). Nowadays, much larger data sets of spatial interactions have become available, such as county-to-county migration data in the U.S., cell phones calls among billions of mobile users (locations), among others. Although there are a number of new methods developed to improve the scalability of flow maps, such as edge bundling (Holten and Wijk 2009), subset selection (Phan et al. 2005), and edge clustering (Cui et al. 2008), there is still a clear gap between the rich information that these newly emerged bit data can potentially offer and the existing methodologies that helps us understand the data.

Spatial interaction models emphasize the influence of the multivariate space on the flow volumes. Through such a model-based approach, one may understand how the multivariate space is associated with the connections among locations. However, most spatial interactions models do not consider the graph structure hidden in SI data. Moreover, the misspecification of spatial structures has been a prominent issue of SI models (Chun and Griffith 2011, Griffith and Jones 1980, Sheppard et al. 1976). In other words, spatial interactions models focus mainly on the multivariate space.

To summarize, the geographic space, the graph space, and the multivariate space have not been combined in an analytic approach for SI data. SI data are underutilized and underrepresented. "Underutilization" refers to the lack of a powerful and comprehensive approach to analyzing the rich information lurking in large and complex SI data. "Underrepresentation" refers to the challenge related to the mapping and visualization of SI data and information. There is a lack of a powerful exploratory analytic method that can deal with the complexity of spatial interactions. One research question to address these challenges is how to integrate the multiple data spaces and deal with the large size and complexity of SI data such that useful and unknown information can be extracted, understood and communicated in a comprehensive way. In order to answer this question, this dissertation research aims to:

(1) Develop new approaches to facilitate a comprehensive analysis and understanding of SI data, combining and exploring the three data spaces (i.e., geographic, graph, and multivariate spaces);

(2) Implement an interactive, integrated, and flexible analysis environment for the exploration of SI data and the communication of SI information through visual/cartographic presentations.

The contribution of this research is two-fold. *First*, it develops a new graph partitioning method to discover spatially contiguous community patterns (SI regions) from the graph space under the spatial contiguity constraint. Evaluations with benchmark data indicate that the developed method is more effective and computationally more efficient than traditional methods. *Second*, this research combines the three SI data spaces in data exploration and representation. SI regions are extracted from the graph space under the spatial contiguity constraint and then utilized as a new aggregation scheme to summarize flow data to enable legible visualization. SI regions, multivariate patterns, and geographic patterns of SI flows are analyzed simultaneously in a novel and interactive visual analytic system. The combination of graph partition, multivariate visualization, flow mapping, and interactive interfaces creates a flexible, comprehensive, and efficient environment to explore SI data from different perspectives and obtain holistic understandings.

Note that SI data are often accompanied with temporal information. For example the Internal Revenue Service (IRS) provides annual migration data ranging from 1991 to the current. However, the temporal dimension is not discussed as a separate data space and not included in this dissertation research. The data used and the methodology developed in this research focus mainly on the spatial interaction graphs and their geographic and multivariate relations. This dissertation research does address the temporal dimension of SI data since it is already a very challenging problem to deal with

the three data spaces in SI. With that being said, the approach developed in this research is able to analyze and visualize temporal information as variables of flows, which can be meaningful and useful in some research scenarios. In the future, the approach may be extended to integrate the temporal dimension of spatial interactions (such as trajectories).

## 1.2 An Integrated Approach

This research combines the strengths of computational methods and exploratory visualization approaches. Figure 1.2 shows the methodological framework of the integrated approach, which consists of two phases. The first phase uses a computational graph partitioning method to discover and generalize spatially-embedded graph structures in SI data. Such spatial graph structures can be defined as geographically contiguous regions (or communities) that are characterized by strong internal connections (flows within regions). These spatially contiguous clusters derived from spatial interactions with special graph partitioning methods are called "SI regions" in this dissertation.

The developed graph partitioning method is evaluated with two different sets of benchmark graphs. It is shown that the method achieves satisfactory solutions and outperforms existing approaches with higher accuracy at very reasonable computation cost. Spatial contiguity is enforced to ensure the adjacency of places assigned to the same SI region. Since the graph partitioning is hierarchical, it leads to a hierarchy (or continuous sets) of SI region divisions. Each hierarchical level has a unique number of regions. Derived SI regions serve a dual purpose in the integrated approach: (1) to represent the spatiality-constrained community structures hidden in flow network; (2) to provide an aggregation means by which the SI data size can be reduced.

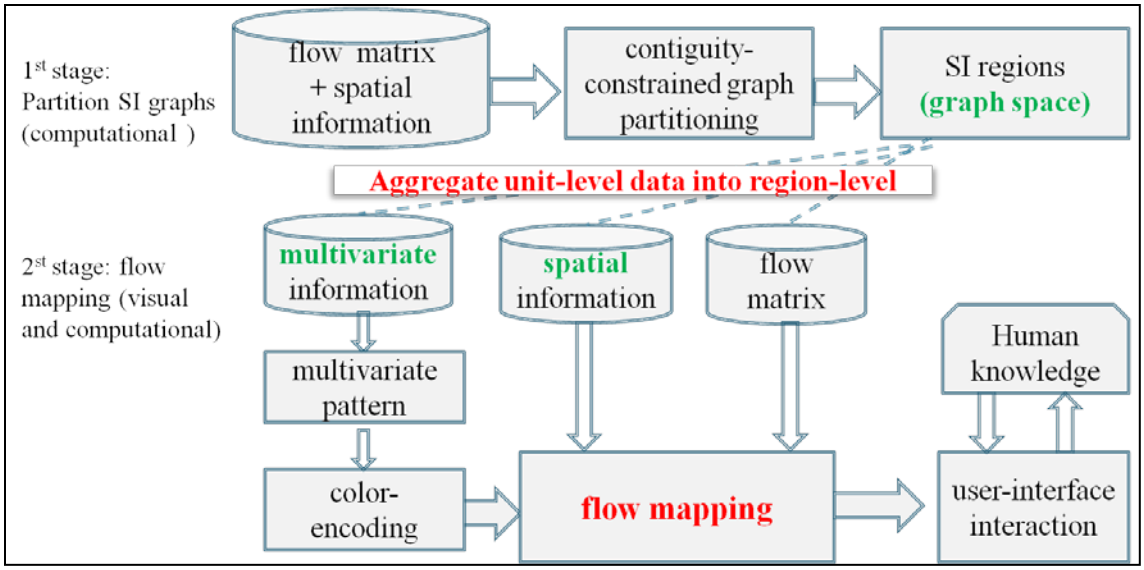


Figure 1.2 The framework of the developed approach. (1) The first phase identifies SI regions (community patterns). (2) The second phase uses SI regions to: aggregate and visualize flows, derive and present multivariate patterns of locations and flows. Computational algorithms and interactive visual environment are integrated to incorporate human's knowledge into the analysis.

The second phase of the integrated approach is to combine the three data spaces of SI data in a visual analytic system, developed and implemented in this research. SI regions are used to convert the location-based data into a region-based data to summarize spatial data, flow connections, and multivariate data. The visual framework is comprehensive and efficient because it combines the three data spaces of SI data and enables holistic visual representations of the patterns extracted from the three data spaces.

Figure 1.3 shows the components of the developed visual system. Spatial interaction patterns can be identified through the flow map alone or in combination with multivariate information and their spatial variations across flows, locations/regions. All components are coordinated and integrated so that a change or user action in one component can trigger updates in other components. FlowMap+, a much enhanced flow map, is the center piece of the visual framework. With the support of other visual and

computational components, FlowMap+ can visualize: (1) SI regions identified in the graph space; (2) multivariate patterns by mapping the color-encoded results of multivariate analyses; and (3) spatial patterns of flows.

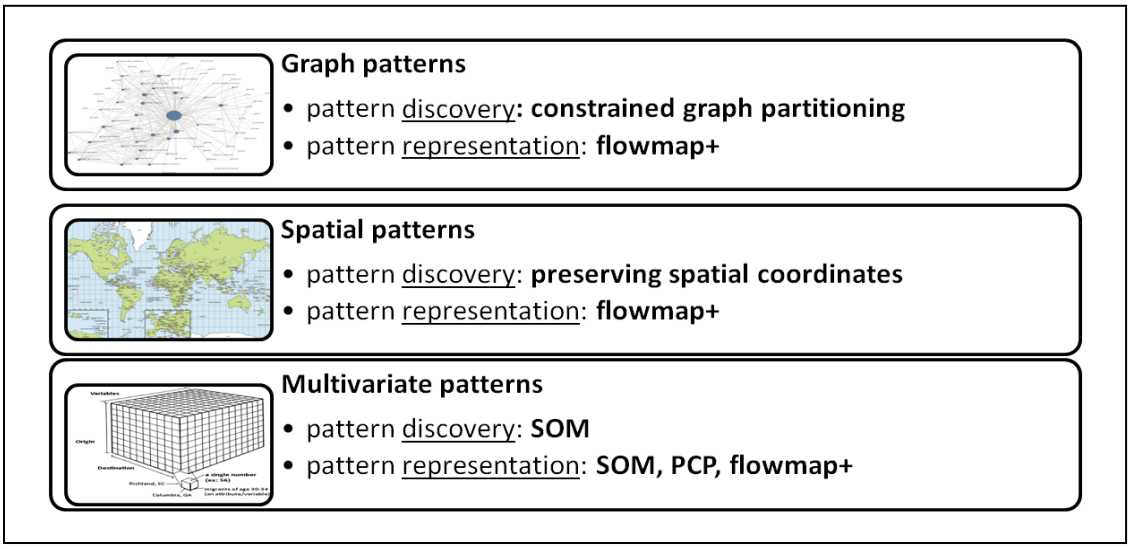


Figure 1.3 Components used to discover and visualize various patterns. All components are coordinated and integrated so that a change or user action in one component can trigger updates in other components.

The visual system supports a variety of visual configurations and exploration, including: (1) interactive selections of the scale level (i.e., # of SI regions) to examine flow patterns at different resolutions, (2) interaction techniques such as filtering, linking, and brushing, (3) highly customizable multivariate analyses, (4) interactive map operations (e.g. selection, zooming in/out, and changing symbol size and color), (5) full coordination among all visual and computational components, and (6) multiple flow measures and area-based network measures which are derived from the original data and available for analysts.

To summarize, the developed approach involves a graph partitioning method and a visual analytic framework. The former is aimed at deriving contiguity-constrained SI

regions from the flow network, which have received little attention in current spatial interaction analyses. In addition to discovering community structures in the graph space, the graph partitioning method provides a strategy to reduce the data size of data and enable effective flow visualizations. The visual analytic framework synthesizes patterns across multiple spaces into a comprehensive understanding, i.e., community structures (SI regions), spatial patterns of flows, and multivariate patterns of flows, locations, and regions.

## 1.3 Case Study

A large inter-county migration data set is used to evaluate the developed approach from an application perspective. It contains over 700,000 county-to-county migration flows in the U.S. The SI regions derived from this data set are compared with two widely used aggregation approaches in migration analyses: state divisions and urban-rural categorization. The results demonstrate that SI regions obtained by analyzing the spatial information and network connections can unveil real-world structures such as the strong "core-suburban relationship" from a network perspective. SI regions provide a meaningful and better partition and aggregation scheme for SI data than current strategies.

A focused investigation is carried out in this dissertation research to evaluate how the visual analytic system can help study the income flows induced by migration, which is a newly emerged research topic. A majority of current exploratory analyses on income flows are area-oriented. The visual system expands this line of research by examining the network structure, origin-destination flows, spatial patterns and multivariate information

simultaneously. The focused investigation presents some new findings on the demographic dimensions of flow data (e.g., the income, the age, and the education structures of flows). The demographic characteristics of migrants can help investigators better understand in-depth structures and patterns in large and complex migration data.

## 1.4 Organization of the Dissertation

The remainder of this dissertation has four chapters. Chapter 2 provides a description of the migration data used in this study. Chapter 3 focuses on the contiguity-constrained graph partitioning method. Next, the visual analytic framework and system is introduced in Chapter 4. The last chapter illustrates and assesses the developed integrative computational-visual approach from an application perspective with a large U.S. county-to-county migration data set.

# CHAPTER 2

# 2000 CENSUS COUNTY-TO-COUNTY DOMESTIC MIGRATION DATA

Early migration studies has been hindered by data scarcity and limitations (Greenwood and Hunt 2003). Often migration data were not directly available or provided with limited details. Nowadays, this data limitation is significantly relieved, with several federal agencies providing information related to migration. The data used in this analysis is a county-to-county domestic migration data set collected in 2000 by the U.S. Census Bureau. According to this data set, 46% of the population (5-years old and above in 2000) or 120 million people had a different residence location in year 2000 from year 1995 (Schachter et al. 2003); 21% of the population (or approximately 55 million) moved across county boundaries. That is, about one in two Americans changed their residence; and one in five Americans relocated to another county within 5 years.

This migration data encompasses all three data spaces: (1) the geographic space, consisting of the locations and boundaries of over 3,000 counties; (2) the graph/network space, consisting of 724,507 non-zero flows between the counties; (3) the multivariate space, consisting of demographic compositions of migration flows and a vector of variables for each county. This chapter describes the content of this data set and introduces the data processing procedures.

## 2.1 Overview of the Data

There are several providers of migration data in the U.S., including two federal sources, the Current Population Survey (CPS) and the Internal Revenue Service (IRS). Some commercial data have also been used to infer human mobility, such as moving-industry dada (Gober et al. 1996) and trajectory data of bills (Thiemann et al. 2010). Each data set has unique strength and weakness. For instance, IRS data have a better temporal resolution (once a year) but tend to under-represent poor and elder people.

The decennial Census provides high-quality migration data in terms of its population coverage, spatial resolution, geographic consistency, and its detailed information of migrants. The decennial census has been gathering migration information since the 1940's. A percentage of population was surveyed in each census with a questionnaire, asking whether they lived in the same residence or a different one five years ago. In the 2000 census, the "long-form" questionnaire containing the 5-year-ago residence questions (see Figure 2.1) was sent to 1/6 of the population. This sample size is considered adequate for ideal geographic resolutions (Isserman et al. 1982). Respondents need to be five-years of age or older on April 1st, 2000 to answer these questions.

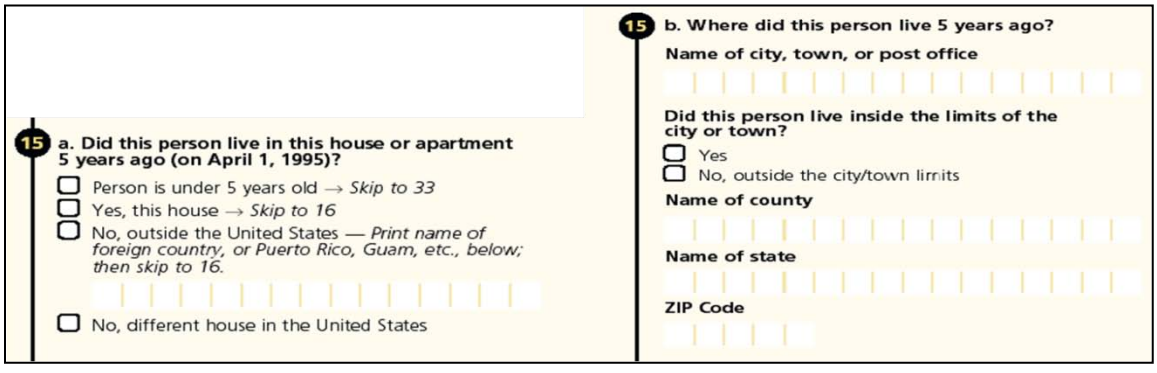


Figure 2.1 2000 Census 5-year-ago questions from the long-form questionnaire (source: U.S. Census Bureau, Population Division)

In addition to county-to-county migration flows (i.e., origin-destination county pairs), the census migration data contain a wealthy of demographic stratifications of migration flows. For example, for the migration flow from county *A* to county *B*, the data have the number of migrants for each age group, each income level, and each race. This allows for investigations on the multivariate characteristics of migration flows. Moreover, since county is a relatively stable and commonly used statistic and administrative unit, it makes it easy to compile data from different data sources and for different years. Note that it is possible to analyze the migration data from different decennial Census with the analytics developed and implemented in this research. At this time, the temporal information is not included in the analyses. This exclusion, however, shall not affect the assessment of the developed approach because the temporal information can potentially be treated as variables of flows in the SI data.

### 2.1.1. Geographic Space

Geographic information is critical for the representation and analysis of spatial patterns of flows. The boundaries of counties in this study come from the Census 2000 TIGER/Line data set[1]. Based on the county boundaries, the contiguity between every pair of counties is considered in the process of graph partitioning. In many existing migration studies, such as those studying metro and non-metro flows, fine-scaled flows are aggregated without considering the spatial adjacency (Morrill 2006, Fuguitt 1985, Fulton et al. 1997, Long and Deare 1988, Rayer and Brown 2001), which make it difficult to extract and understand the spatial structures of flows.

[1] Provided by GIS data server at University of South Carolina: http://www.cas.sc.edu/gis/dataindex.html

### 2.1.2. Graph Space

The county-level migration data of the 2000 Census was released in 2003 and became available in the "Census 2000 Migration Data DVD". The flow data were organized in two separate sets of files with similar variables: one is for the inflow and the other one is for the outflow. While outflow files indicate the flow volume with the field "Out Flow", inflow files shows flow volume with "In Flow". These two sets of files describe the same set of flows since they are only different in the directions of flows. Table 2.1 shows the variables provided in the outflow files.

The example in Table 2.1 shows 29 residents living in Independence County in Arkansas in 1995 had a new residence in Motley County in Texas in 2000. Altogether there are 735,532 unique county-to-county migration records. This research focuses on cross-county migrations occurring in the conterminous U.S. In other words, moves are excluded if the origin county and the destination county are the same. Also excluded are moves with the origin or destination county in Alaska or Hawaii.

Table 2.1 Attribute fields of the outflow data files provided by the 2000 Census

| Attribute | Explanation | Example |
|---|---|---|
| FIPS State in 1995 | 2-digit FIPS[2] code of the origin state | "05" |
| FIPS County in 1995 | 3-digit FIPS code of the origin county | "063" |
| County and State in 1995 | Names of the origin county and state | "Independence County, Arkansas" |
| FIPS State in 2000 | 2-digit FIPS code of the destination state | "48" |
| FIPS County in 2000 | 3-digit FIPS code of the destination county | "265" |
| County and State in 2000 | Names of the destination county and state | "Motley County, Texas" |
| Out Flow | count of the out-migrant population | "29" |

[2] FIPS stands for "Federal Information Processing Standard".

Migration flows between counties naturally form a network, where each node is a location (or area) and each link is a directed flow of migrants. As many other real networks (e.g., World Wide Web, scientific collaboration network) (Faloutsos et al. 1999, Newman 2001), the migration network in this research demonstrates a "power law" degree distribution (Figure 2.2). The "power-law" degree distribution refers to the variation of the frequency of nodes as a power of the node degree, which is the number of links that a node is associated with. "Power-law" degree distribution indicates two properties shared by many graphs: the cluster effect and the small average distance. The distance of a pair of nodes is the number of links in the shortest path connecting them.

Figure 2.2 is generated based on the outflow file of the migration data set. The horizontal axis (i.e., "county degree") is the degree of origin counties, which the number of counties connected to this county through links (i.e. migration flows). For instance, the county degree of Richland County in South Carolina is 944. This means from 1995 to 2000, residents of Richland in South Carolina moved to other 944 counties nationwide.

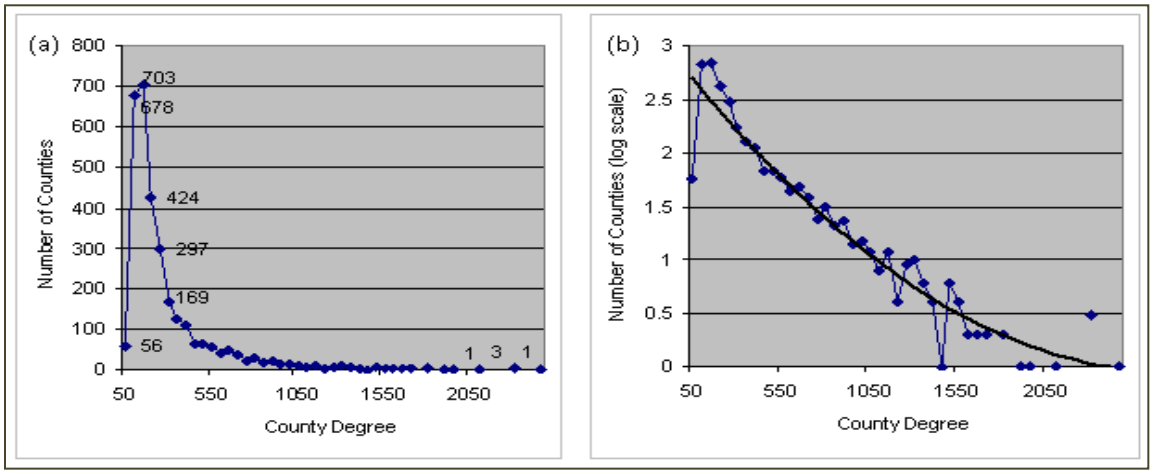


Figure 2.2 Degree distributions of counties. (a) Original degree: most counties are connected to a relatively small set of destinations and a few counties display exceptional connectivity. (b) Log of degree: exponential decrease in the frequency of counties with the county degree indicates a "power-law" distribution.

The vertical axis (i.e., "number of counties") in Figure 2.2 is the number of counties showing a certain "county degree". The long tail indicates most counties are connected to a relatively small set of destinations (100-300) while a small collection of counties demonstrates exceptional connectivity (Figure 2.2a). Los Angeles County in California displays the highest connectivity sending migrants to 2450 counties. The logarithmic line more clearly shows the exponential decrease in the frequency with the county degree (Figure 2.2b).

### 2.1.3. Multivariate Space

Multivariate data is available for both counties and flows among them. Variables for flows include characteristics of migrants, such as age, race and income. These demographic and socioeconomic features are often involved in micro-level migration models, which attempt to estimate the probability of an individual to move. The rationale is that individuals sharing certain characteristics tend to demonstrate similar moving behavior. For instance, younger people have a higher probability to move than older people in the labor force (Greenwood 1975). Within the same age group, the deterring effect of distance is less on people receiving higher education than those less educated (Schwartz 1973). The county-to-county U.S. migration data can be disaggregated by 19 demographic, socioeconomic, and household characteristics (Table 2.2).

Each characteristic has multiple categories, e.g. Sex has two categories (male and female) and Race has seven categories (including White, Black, Asian, etc.). Each category represents a variable in the analysis. The number of categories for each characteristic is shown inside the parentheses after each "characteristic" (see Table 2.2).

In this dissertation, stratifications (categories) on age, race, education, gender, and income are used in the illustration and assessment of the developed approach. In addition, two basic variables of counties are used in this study: the area and the population in year 1995. The latter variable is not directly available and needs to be inferred from other data.

Table 2.2 Stratifications of the 2000 Census migration data

| Characteristics | Explanation |
|---|---|
| RC_QAGE (17) | Age |
| QSEX (2) | Sex |
| RACE7(7) | Race |
| HISP_ORGN (3) | Hispanic or Latino and Race |
| POB (60) | Place of Birth – individual states/country groups |
| NAV_YR2U.S. (15) | Nativity by Year of Entry |
| ATTAINMENT(7) | Educational Attainment |
| COLL_ENROLLMENT(2) | College Enrollment |
| MIL_STATU.S.(6) | Veteran Status by Period of Service |
| QMS(5) | Marital Status |
| INCOME99(11) | Total Income in 1999 |
| LABOR_STATU.S.(4) | Labor Force Status |
| OCCUPATION(11) | Occupation types |
| INDU.S.TRY(14) | Industry category |
| HH_INCOME(6) | Household Income in 1999 |
| POV_STATU.S.(2) | Poverty Status in 1999 |
| TEN_GQ(3) | Tenure and Group Quarters |
| HHTYPE(5) | Household Type |
| MACCI(3) | Metropolitan status of residence in 2000 |

## 2.2 Data Processing

This section describes how GIS files (i.e., the boundaries and shapes of counties) are integrated with the migration data and multivariate data to make a full SI data set. The

data compilation procedure includes two steps: (1) integrating the boundary file and the migration flow files for the study area of this analysis (i.e., the conterminous U.S.), and (2) deriving the 1995 population estimates.

Figure 2.3 shows how the migration and the boundary file are integrated for the study area. The procedure starts with 3107 counties. In this study, the focus is on the conterminous 48 states and Washington D.C. Thus, counties in Hawaii and Alaska are removed, leading to 3075 counties, which form a 3075*3075 migration flow matrix.

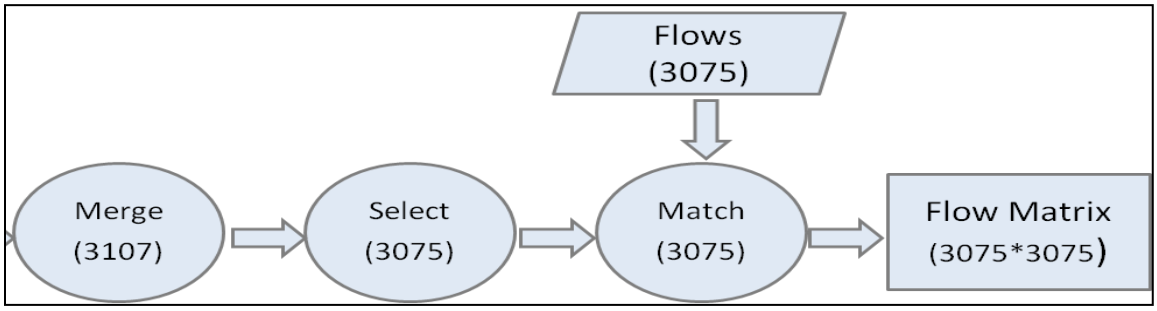


Figure 2.3 Procedures used to integrate the geographic data and the migration data

The 1995 population estimates of counties are used in this study to calculate (1) the population-weighted centroids for SI regions; and (2) various migration flow measures such as migration rates and migration expectations. Figure 2.4 shows how the 1995 population is estimated. “Non movers” refer to people who did not change their residence between 1995 and 2000. “Same-place” movers refer to movers who moved within the same county in 1995 and in 2000. “Out-migration” is the movers who had residence in different counties. The first two terms are provided in a census data file called "Gross Migration for the Population 5 Years and Over: 2000"[3], while the last term is derived from the migration flow file.

[3] Available at: http://www.census.gov/population/www/cen2000/briefs/phc-t22/index.html

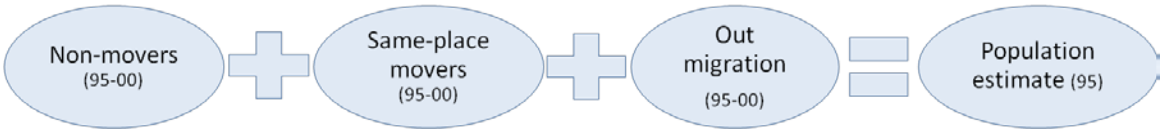


Figure 2.4 Estimates of the 1995 county populations

Eventually, four sets of files are generated for the analysis: (1) a flow file of county-to-county migrations; (2) a contiguity file generated from the geographic data, which is a list of neighboring county pairs; (3) multivariate files, containing stratifications of migration flows by various characteristics of migrants; and (4) a shape file, containing the geographic information and two descriptors (e.g. area and 1995 population estimate) of counties. The four sets of files are consistent in both the geographical scale (i.e. county) and the geographical extent (i.e. conterminous U.S.)

# CHAPTER 3

# GRAPH PARTITIONING OF SPATIAL INTERACTIONS

SI data naturally form a weighted and directed graph, where nodes represent geographic locations and edges represent flows between locations. Edges and their weights indicate the existence and strength of connections. Graph is an increasingly important representation of real systems. A property of a graph is whether its nodes exhibit a clustering structure, where there are more connections with nodes inside the same cluster than with those from the outside. Theses "clusters" are called "communities" or "modules". Communities can be identified with graph partitioning or community detection methods.

The presence of community structures in SI data is noted decades ago: "… *there is a fundamental spatial organization in the pattern of movements and suggest that the discovery of this inherent system of migration regions is the most profitable avenue of approach to the present problem*" (Ng 1969: 713). Nevertheless, the graph space in SI data has not been adequately analyzed in existing research. Efforts targeting graph structures are predominantly in the regionalization context (Poon 1997, Andresen 2009a, Slater 1984) and usually ignore other data spaces (i.e. multivariate space).

This research adopts recent advancements in graph science and incorporates the geographic contiguity constraint to develop a graph partitioning method. This chapter introduces the context, algorithm, and the evaluation of the method. First, existing

partitioning methods developed for SI graphs and relevant work for general graphs are reviewed. Next, the contiguity-constrained graph partitioning method is presented. This method is evaluated from a methodological perspective with two different sets of synthetic benchmark graphs. Results indicate that: (1) the new method achieves significantly higher quality than alternative methods; (2) the optimizing technique notably boosts the quality of the partitioning and minimizes the effect of initializations. In this chapter "region" refers to spatially contiguous "communities".

## 3.1 Review of Graph Partitioning

This section reviews two related research areas: graph partitioning methods for SI graphs and general graph partitioning that does not consider spatial information. The review on the former concentrates on flow transformations and clustering strategies developed for spatial interaction data. The review on the second area presents recent advancements in graph partitioning for non-spatial graphs such as social networks and citation networks.

### 3.1.1 Partitioning SI Graphs

One of the earliest notions of SI community is the "migration region" introduced by Ng (Ng 1969, Pandit 1994). Ng's idea is supported by the significant increase in the correlation between gross migration and population when the spatial units (e.g., provinces) are grouped into contagious areas. "Migration regions" are understood as groups of adjacent places with maximum internal migration flows and minimum external interactions (Ng 1969). Although not described in graph terminology, the concept of

"migration region" surprisingly agrees well with present ideas of graph community. More broadly, a different category of "migration region" is defined as a collection of places sharing similar origins or similar destinations (Pandit 1994). This type of clusters of places is termed "migration typologies". "Migration typologies" is not directly related to the community structure since origin/destinations are only treated as variables of places and flows among them are ignored. Hence "migration typologies" is excluded from the discussion here.

The detection of SI regions typically involves two major steps. The first step is to transform or standardize the flow matrix. The purpose is to cure the bias of flows introduced by dramatic differences in the location size (e.g., a pair of large counties tends to have much more migration between them than between a pair of small counties). The second step is to identify regions based on the transformed flows. Different approaches and strategies have been developed to address these two issues.

#### 3.1.1.1 Flow Transformations

Three major approaches have been developed to transform SI flows: (1) the ratio-based approach, (2) the iterative proportional fitting procedure (IPFP), and (3) the Intramax approach. Ratio-based transformation is the simplest, including the ratio of the flow to the out-flow total of origins (Ng 1969, Keane 1978), mobility index (i.e., the ratio of out-flow total of the origin and in-flow total of the destination) (Pandit 1994, Hollingsworth 1971), flow efficiency index as used by Morrill (1988), and trade intensity (Andresen 2009a).

The IPFP is a double constrained standardization of a flow matrix which forces the row and column totals equal to a given number after the transformation. During the IPFP process, the initial column entries and row entries are iteratively scaled by the row or column totals until the row or column totals converge to a given number (Fienberg 1970, Tyree 1973). The IPFP standardization has been utilized to detect migration regions (Slater 1980, Slater 1975, Clark 1982) or serves as an estimation strategy (Wong 1992). An IPFP-adjusted matrix has several desirable properties. First, it maintains the "spatial structure": the cross product ratios $f_{ij}f_{kl}/f_{il}f_{kj}$ remain unchanged after the transformation. Second, the adjusted matrix is a maximum entropy-estimate of the original matrix under the double constraints (Slater 2009).

A key criticism of IPFP is that it can distort relative significances of nodes (Holmes 1978, Fischer et al. 1993). The distortion effect is illuminated in Figure 3.1.

| | A | B | C | D | sum |
|---|---|---|---|---|---|
| A | 0 | 2 | 2 | 1 | 5 |
| B | 12 | 0 | 7 | 1 | 20 |
| C | 13 | 66 | 0 | 20 | 99 |
| D | 21 | 100 | 4 | 0 | 125 |
| sum | 46 | 168 | 13 | 22 | 249 |

| | A | B | C | D | sum |
|---|---|---|---|---|---|
| A | 0 | 0.13 | 0.50 | 0.38 | 1 |
| B | 0.51 | 0 | 0.41 | 0.09 | 1 |
| C | 0.16 | 0.30 | 0 | 0.53 | 1 |
| D | 0.33 | 0.57 | 0.09 | 0 | 1 |
| sum | 1 | 1 | 1 | 1 | 4 |

Figure 3.1 The distortion effect of the IPFP transformation. Flow C-B (66) is more than three times larger than flow C-D (20) in the original flow matrix. After the transformation, flow C-B (0.30) becomes smaller than flow C-D (0.53).

In this example, flow *C-B* (66) is about three times of flow *C-D* (20) in the original flow matrix. After the transformation, flow *C-B* (0.30) is around half of flow *C-D* (0.53). Such changes are due to the large variability of row/column totals. This type of distortion is not a major concern when the rows and columns in the flow matrix exhibit

small variability. However, the problem becomes severe when the matrix contains large variability. In addition, IPFP may not be a good option for sparse flow matrix containing a large number of zero entries (Fischer et al. 1993, Holmes 1978), which is often the case for fine-scale large spatial interaction data.

The third transformation approach, Intramax (Masser and Brown 1975), attempts to remove the size effect by subtracting the flow expectation from the original flow. In Figure 3.2, $I_{ij}$ is the observed flow from origin $i$ to destination $j$, $I^*_{ij}$ is the estimated flow. $I'_{ij}$ is the transformed flow between $i$ and $j$ (Figure 3.3). The expected flows are calculated as the products of the row totals of the origin and the column totals of the destination divided by the overall total (Fischer et al. 1993, Poon 1997, Mitchell and Watts 2010). By adding the transformed flows in two directions (see Figure 3.2), the Intramax transformation leads to a symmetric matrix.

| | |
|---|---|
| $I'_{ij} = (I_{ij} - I^*_{ij}) + (I_{ji} - I^*_{ji})$ | $i, j = 1, \ldots\ldots, n.$ |

Figure 3.2 Intramax transformations

| | |
|---|---|
| $I^*_{ij} = \frac{\sum_j I_{ij} \sum_i I_{ij}}{\sum_i \sum_j I_{ij}}$ | $i, j = 1, \ldots\ldots, n.$ |

Figure 3.3 The expectation term in the Intramax

Compared to IPFP, Intramax is more suitable for sparse matrices and those exhibiting a large variability of individual flows (Fischer et al. 1993). A variant of Intramax is “relative acceptance”. It divides the difference between the observed and the expected flow with the expected flow (Holmes 1978). It is argued that “Intramax” does

not remove the size effect as completely as "IPFP" (Hirst 1977). This issue, however, is less severe than the distortion introduced by the IPFP transformation (Holmes 1978).

#### 3.1.1.2 Clustering Methods

A transformed flow matrix can be treated as a similarity matrix and combined with any clustering methods to identify clusters. A widely used clustering method for spatial interaction networks is hierarchical clustering which generates communities on various abstraction levels (Pandit 1994). Hierarchical clustering is carried out in either agglomerative or divisive manner. Divisive clustering starts with a cluster containing all nodes and proceeds in a top-down manner by removing the least weighted edges successively.

Agglomerative clustering starts with each node being a cluster and proceeds in a bottom-up manner by fusing two clusters connected by the strongest edge (Andresen 2009b, Slater 1976a, Slater 1976b). Compared to divisive clustering, agglomerative clustering is more commonly used to extract SI regions. Among the three between-cluster distance definitions (i.e., single-link, average-link, and complete-link), single-link appears most favored in SI partitioning due to its simplicity, despite its relatively poor cluster quality comparing to other alternative methods. It has been coupled with IPFP transformation to detect community patterns in SI graphs (Slater 1984). A well-known deficiency of the single-link method is the chaining phenomenon: the elements which are very distant to each other may be forced to be in the same cluster.

Another agglomerative clustering method is the adjusted Ward's clustering (Masser and Brown 1975, Poon 1997, Fisher and Gopal 1994). Ward's clustering is

usually used for multi-dimensional data. It is distinguished from other agglomerative clustering by evaluating the variance of clusters during the bottom-up merging. It attempts to minimize the variance of clusters that can be formed by each merge. Masser and Brown modified Ward's method by replacing the variance metric with the Intramax measure (Masser and Brown 1975). The objective is to maximize the total "Intramax" within clusters. Further, the Intramax measure is updated after each merge since every merge causes changes to the flow matrix (i.e. row totals and column totals) that is used to derive flow expectations in the Intramax transformation.

"Intramax" transformation is always attached to "modified Ward's" (Masser and Brown 1975, Poon 1997) in the existing literature. For this reason, the partitioning method combining the "Intramax" transformation and the "modified Ward's" is referred to as the "Intramax approach" in this paper.

The Intramax approach represents an advantage over alternative methods as the clustering process is explicitly guided by an objective function (i.e. Intramax measure). The drawback of the Intramax approach is that it can be easily trapped in local optima. This is illustrated with a workable example (Figure 3.4) where the community structure is clear and predefined. This predefined graph consists of six nodes grouped in two communities of equal size. Each node is connected to the other two nodes of the same community and one in the other community. The weight is evenly 1.

Because of the uniform degree and uniform weight, the estimate weight of each edge is uniform at the beginning (i.e. $m$). The Intramax proceeds by successively merging two nodes/clusters having the largest Intramax measurement. Each merge is ensued by an update of the Intramax matrix. The first merged nodes are node $c$ and node $d$ since their

Intramax is the maximum among all pairs of nodes. Next, node *b* and *e* are merged based on the updated Intramax matrix. After four merges, Intramax assigns node *a*, *b*, and e to one community and the remaining nodes (i.e. f, c, and d) to the other community. This apparently contradicts the true pattern. A key factor responsible for the error is that this approach cannot correct or reverse local optima: two nodes are not split once merged.

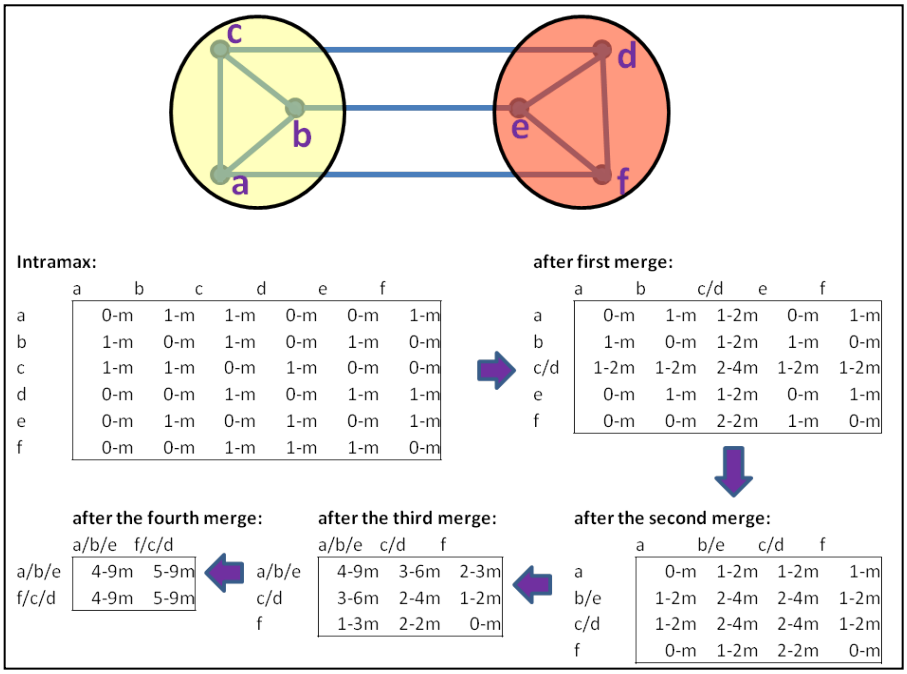


Figure 3.4 An example of local optima in the Intramax approach

### 3.1.1.3 Comparison of Existing SI Graph Partitioning Methods

Two major partitioning methods for SI data have been compared in two studies (Masser and Scheurwater 1980, Fischer et al. 1993) involving 20-100 locations (Table 3.1). The first method is a combination of the IPFP transformation and the agglomerative single-linkage clustering ("IPFP-SLK"). The second method is the Intramax approach consisting of the Intramax transformation and a modified Ward's clustering.

Table 3.1 Comparative analyses of IPFP-SLK and the Intramax approach

| Authors | Method | Data set | Findings |
|---|---|---|---|
| Fischer et al. 1993 | IPFP –SLK | medium-sized telecommunication | large variance in region size |
| | Intramax approach | | small variance in region size |
| Masser et al. 1980 | IPFP –SLK | medium-sized migration and journey to work | large variance in region size |
| | Intramax approach | | small variance in region size |

Similar conclusions were reached in these two studies conducted independently. Fischer's study (1993) shows that IPFP-SLK tends to generate single-node-group in late phases of clustering while the Intramax approach does not have this problem. At the same level, the regions formed by IPFP-SLK tend to have unbalanced sizes while the regions formed by the Intramax approach is more balanced. Additionally, regions obtained by the Intramax approach are more interpretable than the single-link clustering. It is found that regions obtained by Intramax approach agree well with intuitive understandings of the study area and also the existing districting of the telephone network (Fischer et al. 1993). Masser and Scheurwater (1980) believed that the Intramax is more suitable than IPFP in the search for SI regions.

These two comparative studies are common that the evaluation primarily rely on visual inspection and domain knowledge, which is reasonable but could be biased. In this dissertation research, synthetic benchmark data sets are used to evaluate the developed method in comparison with the two popular partitioning methods: IPFP-SLK and the Intramax approach.

#### 3.1.1.4 Spatial Contiguity in SI Graph Partitioning

The importance of spatial contiguity to SI regions has been recognized decades ago (Ng 1969, Masser and Scheurwater 1980). Spatial contiguity is a key property of various forms of regions (Ng 1969). Although empirical studies exist showing that SI regions tend to be contiguous even if the spatial contiguity is not considered (Thiemann et al. 2010, Slater 1984, Murtagh 1985), these observations may only be true for specific data sets and certain geographic scales.

A contiguity-constrained partitioning method for SI networks will contribute in several aspects. First it improves the understandability of detected SI regions. Community patterns are usually presented with graph drawing or dendrograms. A contiguity-constrained partitioning provides an additional option to recognize community patterns by examining the spatial boundary of regions. Second, it allows easier and more efficient exploration and visualization of the spatial patterns of flows. Despite the variety of visual methods for spatial flows, the flow map remains a convenient and important visual form of spatial interactions. The disadvantage of a flow map lies in its limited capability to handle large data sets. Contiguity-enforced SI regions provide a means to aggregate data, reduce data size, and dramatically improve a flow map's efficiency to handle large data sets. Third, the contiguity-constrained SI regions have operational or practical values. For example, in the event of an infectious disease outbreak, regions that need rapid mitigations should be contiguous for practical operations such as enforcing travel restriction or quarantine policy.

Most of existing partitioning methods for SI graphs are not constrained by spatial contiguity, including some recent ones (Thiemann et al. 2010). There are two exceptions. Once of them is implemented by Ng. It is rather primitive: after a ratio-based transformation, the rows/columns in the transformed matrix are sorted (Ng 1969). Manual adjustment is then conducted to move dominant flows into square blocks along the diagonal, which need to be contiguous to form migration regions. The intensive manual work implies a high risk of inaccuracy and poor computational efficiency. Ng (1969) applied this method to a 71-node dataset, which will become impractical for large datasets.

The other method enforcing the spatial contiguity is developed by Masser and Brown (1975). Spatial contiguity is enforced by limiting the searching space to spatially adjacent units. In addition, the spatial contiguity matrix is updated as spatially adjacent units or clusters are merged. This strategy is more accurate and computationally efficient.

### 3.1.2 Partitioning General Graphs

Community detection is a very active research area. Numerous graph-partitioning methods have been developed to identify communities in various networks. Some specialize in un-weighted networks; some are exclusively aimed at directed and weighted networks; and some others allow a node belonging to multiple communities (Fortunato 2010).

SI graphs often contain information on the strength of connections, such as the migration network where the edge weight is the magnitude of migration flows. Weights of edges provide an important clues, which can lead to different community structures

with the same graph topology (Fan et al. 2007). The discussion in this paper will concentrate on methods for weighted graphs satisfying the following requirements: (1) all nodes belong to the same type; (2) all edges belong to one category; (3) communities or SI regions do not overlap (i.e., a non-overlapping partition of space); (4) communities are defined by the strength of spatial interactions (not multivariate similarities).

Similar to the partitioning of SI graphs, general community detection consists of two key components: a definition of community and a partitioning/clustering method. The general definition of community (i.e. strong within-community connections) is intuitive but vague (Fortunato 2010). This leaves much room for a definition. For instance, "internal link" or "external link" can be formulated differently. Internal and external links can be compared in various ways (e.g., difference or ratio-based).

Community detection attempts to find an optimal partition which maximizes (or minimizes) an objective measure. The number of possible solutions grows faster than exponentially with the number of nodes (i.e. graph size) (Fortunato 2010). Very large networks, such as those with millions or even billions of nodes, make the task of partitioning extremely challenging. Exact algorithms cannot tackle NP-hard problems with affordable computational costs. Hence, researchers resort to approximation approaches for near-optimal solutions.

There are three major groups of partitioning methods for general graphs: (1) hierarchical clustering, (2) partitioning clustering (e.g. *k*-clustering methods), and (3) spectral clustering. The first group (i.e. hierarchical clustering) does not require a predefined number of communities while the latter two do. This becomes a notable advantage of hierarchical clustering since the number of communities is often unknown

beforehand. Additionally, hierarchical clustering methods are useful for data sets with an inherent hierarchical community structure.

Hierarchical clustering proceeds in either an agglomerative or divisive manner. A divisive graph partitioning method starts with all nodes being a cluster and iteratively removes inter-cluster edges to produce clusters. Newman and Girvan developed the approah to remove edges having a high "betweenness" value, which indicates the frequency of an edge being part of the shortest path between the nodes in two clusters (Newman and Girvan 2004). Such algorithms are primarily for unweighted graphs. With its current form, it is not appropriate for weighted graphs since it tends to divide nodes linked by highly weighted edges.

#### 3.1.2.1 Modularity

Modularity represents a most discussed and adopted strategy to measure community patterns. The term modularity is introduced in Newman and Girvan (2004) for unweighted and undirected graphs but its calculation is highly similar to Intramax (Masser and Brown 1975, Poon 1997). That is, both subtract the expected values from the observed. In its original form (Figure 3.5), $e_{ij}$ is a fraction of all edges between nodes in community $i$ and those in community $j$. $a_i$ is an estimate of $e_{ij}$ based upon a null model, which accommodates a number of definitions. Theoretically modularity ranges from 0 to 1. The empirical range is 0.3-0.7. Since its introduction, modularity has received enormous attention and gained remarkable popularity. It became a critical element in a diversity of partitioning methods (Fortunato 2010).

| | |
|---|---|
| $Q = \sum_i (e_{ii} - a_i^2)$ , | $a_i = \sum_j e_{ij}$ |

Figure 3.5 Original definition of modularity -- for unweighted and undirected graphs

The contribution of modularity to graph partitioning is multifold: first, modularity clearly delivers an understanding of community. In the absence of community structure (i.e. a random graph), the modularity would be or close to zero. Second, modularity provides a metric to assess the quality of community divisions. Given network data, a higher modularity indicates a better partitioning. For instance, it can be utilized in hierarchical clustering to compare the partitioning at various levels and determine the ideal level (Newman and Girvan 2004). Thus, modularity is an appropriate option of the objective function to optimize in a clustering method.

An additional and remarkable advantage of modularity is that it provides a general framework allowing considerable flexibility. The null model of modularity can be customized in various domains. The prototype can also be easily extended for weighted (Newman 2004a), directed (Leicht and Newman 2008), and weighted and directed graphs (Arenas et al. 2007) (Figure 3.6). In Figure 3.6 , $W_{ij}$ is the weight, $\delta(C_i, C_j)$ indicates whether a pair of nodes belong to the same community ($C_i$=$C_j$) or not ($C_i \neq C_j$ ).

| | |
|---|---|
| $Q = \frac{1}{2w}\sum_i \sum_j (w_{ij} - \frac{w_i^{out} - w_j^{in}}{2w})\delta(C_i, C_j),$ | $w_i^{in} = \sum_j w_{ij}$ , $w_j^{out} = \sum_i w_{ij},$ |
| $\delta(C_i, C_j) = 1 \; if \; C_i = C_j, \delta(C_i, C_j) = 1 \; if \; C_i = C_{j,}$ | $2w = \sum_i \sum_i w_{ij}$ |

Figure 3.6 Modularity for weighted and directed graphs

#### 3.1.2.2 Modularity-based Partitioning

Graph partitioning oriented on modularity represents a wide range of techniques, among which the greedy algorithm is the most popular. The simplest greedy algorithm is the agglomerative hierarchical clustering (Newman 2004b). Edges achieving the maximum increase in the modularity are added successively in the bottom-up clustering. A variety of modifications are proposed to enhance the computation efficiency of the greedy algorithm (Clauset et al. 2004), or correct the bias toward large communities (Danon et al. 2006, Schuetz and Caflisch 2008). Researchers have also attempted to improve the modularity optima through the use of intermediate configurations (Pujol et al. 2006, Du et al. 2007, Xiang et al. 2009, Ye et al. 2008). Most of these modifications are developed for unweighted graphs. Relatively few are for weighted graphs (Blondel et al. 2008).

Simulated annealing (SA) represents a different strategy used to detect community patterns (Guimer et al. 2004). It consists of local transitions (i.e. shift) and global transitions (i.e. merge and split) with the probability determined by predefined parameters. Lately SA is utilized to partition a geographically-referenced money circulation data (Thiemann et al. 2010). The drawbacks of SA include its heavy dependence on parameter configuration (e.g., initial temperature and cooling factor) and its high computational cost (Fortunato 2010). Another issue is the variation of the solutions for the same graph and same configurations due to the randomness it involves.

Extreme optimization (EO) is adopted to improve the computational efficiency and sustain the relatively high quality of SA (Boettcher and Percus 2001, Duch and Arenas 2005). The involved modularity is a sum of the fitness function of each vertex.

The fitness function is determined by the current partitioning and the property of vertex (e.g. degree). Iteration moves the vertex with the lowest fitness value to another cluster. EO has a relatively high risk of local optima due to the initial configuration (Fortunato 2010). Spectral optimization (SO) was used to optimize modularity by using eigenvectors and eigenvalues of the modularity matrix, the element of which is the modularity matrix (Newman 2006a). SO is limited to unrealistic cases where only two communities exist. There are several methods proposed to cope with this drawback (Newman 2006b, Sun et al. 2009).

#### 3.1.2.3 Evaluation with Benchmark Graphs

Two general categories of benchmark graphs have been used to evaluate partitioning methods: real graphs and synthetic graphs. Some most well-known real graph data sets include Zachary's karate club network (Zachary 1977) and the social network of bottlenose dolphins living in New Zealand (Lusseau et al. 2003). Zachary's karate club data represents a social network between 34 karate club members at a U.S. university. The dolphin social network contains 62 dolphins and 159 edges. All of these networks are unweighted and undirected. The general problem of real data sets is that the real community structure is not always clear and the data sets are usually small.

Synthetic data sets can be generated with predefined community structures, which are to be compared with the derived structures to assess the accuracy of partitioning methods. The benefit of synthetic data is obvious: the planted community structure provides a standard to test a given method. Two synthetic benchmark data sets have been introduced and utilized (Table 3.2): (1) Girvan-Newman (GN) benchmark (Newman and

Girvan 2004), and (2) Lancichinetti-Fortunato-Radicchi (LFR) benchmark (Lancichinetti et al. 2008).

GN benchmark is generated with a *l*-partition model (Condon and Karp 2001). It is characterized by its uniform degree distribution and its equal community size. In its prototype, GN graphs also have a fixed number of nodes (i.e. 128) and communities (i.e. 4). Each node has the same probability to be connected to its peers in the same community but lower probability to connect to nodes in other communities.

Table 3.2 Configurations of NG and LFR graphs

| Graph | degree distribution | community distribution | graph size (# of nodes) | community count | weighting |
|---|---|---|---|---|---|
| GN | uniform (16) | uniform | 128 | 4 | unweighted |
| LFR | power-law | power-law | varied | varied | varied |

The only parameter that varies for generating GN graphs is the out-degree (or in-degree) of nodes, which decides how strong the community structures is in the resulting graphs. Out-degree is the number of links connected with nodes outside of the community. In-degree is the number of links connected with nodes within the community. Out-degree and in-degree are dependent parameters since their total (i.e. overall degree, # of nodes/neighbors connected to a node via edges/links) is a constant (i.e. 16). A larger out-degree means weaker patterns: more between-community connections and fewer within-community connections.

LFR is more realistic and stricter than GN with its power-law distributed node degrees and community sizes, as many real graphs display (Lancichinetti et al. 2008). Different LFR graphs can be produced by varying the number of nodes, the exponent of

the power-law degree distribution, and the mixing parameter, i.e., the fraction of external degree (i.e. out-degree) to the overall degree. LFR and GN will both be used to evaluate the developed method in this research.

The mixing parameter plays a critical role in the LFR graph construction. "Mixing parameter" is the ratio of the external degree to the overall degree, which determines the strength of community patterns. A larger mixing parameter means fewer internal (i.e. in-degree) links and a weaker community pattern while a lower mixing parameter indicates a stronger community structure. Naturally, there should be a threshold of the mixing parameter distinguishing random graphs without community structures from graphs with community structures. The threshold is estimated to be: $(N-N_{max})/N$, where N is the count of nodes and $N_{max}$ is the size of the largest community (Lancichinetti and Fortunato 2009). If the mixing parameter is less than this threshold, a community structure exists in the generated graph. If not, then the generated graph is more similar to a random graph. In a random graph, the probabilities of an edge between any two nodes are approximately equal.

Evaluating a partitioning method with synthetic graphs requires a similarity measure to compare the derived partitioning solution and the true community structure. Similarity measures can be classified into three categories: pair counting, clustering matching, or information theory-based (Fortunato 2010). Among other "pair-counting" measures, Fowlkes and Mallows (FM) index (Fowlkes and Mallows 1983) is widely used though it is originally devised for hierarchical clustering comparisons.

Measures based on "clustering matching" try to identify the largest overlap between the two partitions in comparison. One problem of the "clustering matching"

score is that some clusters can be considered more than once if the overlap is large enough while some other clusters can be ignored because the overlap is too small.

Entropy measures are rooted in information theory. They consider the community labels in the two partitions to compare as two random variables. Their similarity denotes the information that is needed to infer one partition label from the other. A well cited entropy-based similarity measure is the "normalized mutual information" (Danon et al. 2005). The entropy-based similarity and the FM index will both be used in the assessment of the developed graph partitioning method (see next Section).

### 3.2 New Spatially-constrained Graph Partitioning Method

The purpose of a contiguity-constrained graph partitioning is to identify SI regions by grouping strongly connected and spatially adjacent nodes into clusters (see Figure 3.7). In addition to capturing graph structures in SI data, SI regions can provide a new and meaningful aggregation strategy of SI data.

Existing research often uses statistic or administrative regions, such as states and census regions, to aggregate flows and reduces data size. However, these predefined boundaries (regions) often vary dramatically in size and, more importantly, do not necessarily reflect the true patterns manifested by the flows. Therefore, using administrative divisions to aggregate spatial flows may miss important patterns. For example, Aiken County in South Carolina is more connected to Richmond County in Georgia than all other counties in South Carolina. According to the 2000 Census, between 1995 and 2000, the migration outflow from Aiken in South Carolina to

Richmond in Georgia is 2,386, which is about twice of the largest outflow (i.e. 1,232) from Aiken to other South Carolina counties (i.e. to Edgefield County).

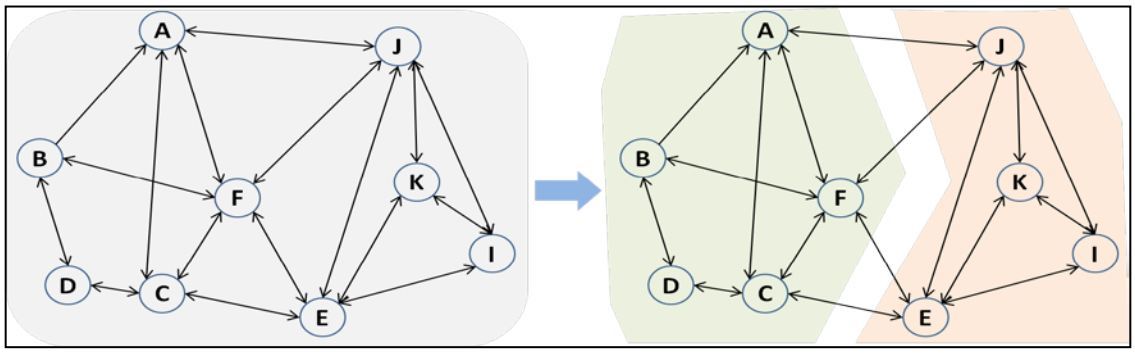


Figure 3.7 Contiguity-constrained graph partitioning. Spatially adjacent nodes are grouped into strongly connected sub-graphs (i.e., clusters or communities).

Graph partitioning methods developed in the SI context and in the general context share some common elements: (1) the formulation of "community" (i.e., the "modularity" defined in the general context and the "Intramax" formulated in the SI context); and (2) the use of clustering approaches to group nodes/places. Despite these similarities, partitioning methods developed for general graphs cannot be applied directly to SI graphs. This is because they do not consider the spatial contiguity--a key condition of a "region" in geography and a distinguishing property of spatial phenomena. A graph partitioning that considers spatial information would facilitate the understanding the spatial patterns of flows and make a connection between the graph space and the geographic space of SI data. However, existing partitioning methods for SI graphs are not effective to avoid local optima and thus the derived regions are often not of good quality.

The research in this dissertation develops a partitioning method that is able to derive contiguous communities and at the same time more effectively optimize the solution. This section starts with an outline of the developed method. Then the

modularity measure used in the method is introduced. Last, the optimization strategy used to avoid local optima is presented.

### 3.2.1 Overview of the Partitioning Method

The new graph partitioning method takes the following as inputs: (1) an $n*n$ flow matrix, representing the connection strength of paired locations; and (2) an $n*n$ contiguity matrix, specifying the contiguity relationship among locations. The overall procedure of the partitioning method is outlined in Table 3.3.

Phase 1 is a bottom-up process that performs an initial hierarchical clustering of the nodes (locations) in the SI data. Step 1 converts the flow matrix into a modularity matrix to partially remove the size effects of locations (counties) on flows. Each element in the modularity matrix is the modularity between a pair of places in two directions. Note that, since modularity is the difference between the actual flow and the expectation, a larger modularity represents a stronger connection. Based on the modularity matrix, a hierarchy of clusters of nodes is constructed using a contiguity constrained agglomerative clustering algorithm (Step 2), where a cluster is a set of strongly connected and spatially contiguous nodes (locations).

Table 3.3 Procedures of the contiguity-constrained graph partitioning method

| | | |
|---|---|---|
| Phase 1 | Step 1 | Calculate the modularity for each pair of nodes |
| | Step 2 | Construct an initial hierarchical tree by agglomerative clustering |
| Phase 2 | Step 3 | Cut the tree into two sub-trees that maximize modularity |
| | Step 4 | Modify and improve the two regions (sub-trees) by a Tabu search |
| Phase 3 | Step 5 | Repeat steps from 1 and 4 within each region to generate more regions until the needed abstraction level is reached |

Four clustering methods can be used to initialize the clustering: single linkage (SLK) clustering, complete-linkage (CLK) clustering, average-linkage (ALK) clustering, and the Ward's method. They are different in their definition of inter-cluster distance. ALK calculates the distance as the average of all distances between the nodes in one cluster and those in the other cluster. SLK takes the minimum among all distances between the nodes in two clusters, while CLK takes the maximum as the inter-cluster distance. The Ward's method calculates the inter-cluster distance as a modularity measure in this research (see the next subsection 3.2.2). The Ward's method updates the modularity matrix after each merge since each merge alters the flow matrix (cluster-based).

No matter which clustering method is used, only spatially contiguous nodes/clusters can be merged. The contiguity matrix is dynamically modified to reflect the topological change due to each merge. For example, two unconnected nodes *A* and *B* can become neighbors if *A* is merged to a node *C* that is adjacent to *B*.

The optimization (Phase 2) first cuts the tree into two sub-trees that maximize modularity (Step 3). Then the cut is modified to improve the two regions (sub-trees) by a Tabu search (Step 4), which is aimed at identifying the optima or optima cut and avoid local optima. Phase 1 and Phase 2 continues until a desired level (i.e., # of clusters/regions) is obtained (Step 5).

### 3.2.2 Extended Formulations of Modularity

Modularity calculation is a critical element in the developed method. Modularity is used in this research to: (1) to construct a distance metric for clustering (Steps 1 and 2); (2) to serve as the objective function during the optimizing phase (Step 3 and 4).

Modularity is defined as the difference between the observed flow and the expected flow between two nodes or two clusters of nodes. In this research, the calculation is adjusted to: (1) account for non-zero diagonal flows (i.e., no internal flows of within locations); and (2) allow hierarchical expectations to obtain more accurate flow estimates during the partitioning.

The between-node expectation and modularity are calculated as shown in Figure 3.8. Using node-based modularity as a similarity measure in the hierarchical clustering (phase 1) can partially remove the size bias, i.e., large flows associated with large nodes (in terms of population, for example).

| | |
|---|---|
| $e_{ij} = f_{i.} * f_{.j} * F_S/({F_S}^2 - \sum_{i \subseteq S}(F_{i.} * F_{.i})),\ if\ i \neq j$<br>$e_{ij} = 0, \quad if\ i = j$ | $i \subseteq \boldsymbol{S}, j \subseteq S$ , $F_S = \sum_i \sum_j f_{ij}$<br>$f_{i.} = \sum_{j=1}^{n} f_{ij},\ f_{.j} = \sum_{i=1}^{n} f_{ij}$ |
| $Q_{ij} = f_{ij} - e_{ij}$ | |

Figure 3.8 Adjusted flow-based expectation and the modularity between nodes

In Figure 3.8 , $e_{ij}$ is the expected flow between node *i* and node *j*. This flow-based estimation is considered as the null model in the formulation of modularity, which assumes that flows between locations are proportional to the total outflow of the origin location and the total inflow of the destination location. Within-location flows are not considered and thus the modularity within the same location (node) is set to zero. This formulation ensures that the total expected flow is equal to the total of actual flows. This flow-based estimation is similar to the one introduced in Leicht and Newman (2008), except that it does not normalize the difference of the observed flow and the expected flows with the flow volume. This flow-based estimation does not have such

normalization to avoid the suppression of communities associated with large volumes of interactions. The metric introduced by Leicht and Newman (2008) would favor places associated with small flows since small-sized nodes tend to have a large normalized modularity.

Another extension of the modularity lies in the adjustment for zero-valued diagonal elements in the flow matrix. In Figure 3.8, the nominator of flow estimates is the product of the row-marginal of the origin, the column-marginal of the destination, and the grand total flow *Fs.* This part of this expectation is similar to the Intramax measure. It is distinguished from Intramax by its denominator, which is adjusted to account for diagonal flows, for which the actual flows are not considered (i.e. enforced to be zeros) and the expectations are set to zeros.

Without this adjustment, estimates of diagonal elements (within-location flows) tend to dominate the flow matrix and exert considerable impacts on the estimates of other flows, since they are usually much larger than between-location flows. As a result, the denominator needs to be adjusted to keep the total of the observed flow and the total of the expectation consistent. Otherwise, the formulation would lead to non-zero and large estimates of within-location flows.

Figure 3.9 describes between-region modularity, which is similar to the between-node formulation. Within-region modularity (Figure 3.10) is a special case of between-region modularity when the two involved regions are the same. If adjustments for within-location flow estimates are not necessary, the formulation of between-node and between-region modularity is simpler (Figure 3.11and Figure 3.12).

| | |
|---|---|
| $E(A,B) = F_{A.} * F_{.B} * F_S/({F_S}^2 - \sum_{i\subseteq S}(F_{i.}F_{.i}))$, | $A \subseteq \boldsymbol{S}, B\subseteq$S, A$\cap$B$= \emptyset$ |
| $Q_{AB} = F(A,B) - E(A,B) = \sum_{i\in A}\sum_{j\in B} f_{ij} - E(A,B)$ | $F_{A.} = \sum_{i\in A} fi.,\ F_{.B} = \sum_{i\in B} f_{.j}$ |

Figure 3.9 Adjusted flow-based expectation and the modularity between regions.

| | |
|---|---|
| $E(A,A) = (F_{A.}F_{.A} - \sum_{j\subseteq A}(F_{j.}F_{.j})) * F_S/({F_S}^2 - \sum_{i\subseteq S}(F_{i.}F_{.i}))$, | $F_{A.} = \sum_{i\in A} fi.$, $F_{.A} = \sum_{i\in A} f_{.j}$ |
| $Q_{AA} = F(A,A) - E(A,A) = \sum_{i\in A}\sum_{j\in A} f_{ij} - E(A,A)$ | |

Figure 3.10 Adjusted flow-based expectation and the modularity within regions.

| | |
|---|---|
| $e_{ij} = f_{i.}\, f_{.j}/F_S$ | $i \subseteq \boldsymbol{S}, j\subseteq$S, $f_{i.} = \sum_{j=1}^{n} f_{ij}$ |
| $Q_{ij} = f_{ij} - e_{ij}$ | $f_{.j} = \sum_{i=1}^{n} f_{ij}, F_S = \sum_i \sum_j f_{ij}$ |

Figure 3.11 Unadjusted flow-based expectation and modularity between nodes.

| | |
|---|---|
| $E(A,B) = F_{A.}\, F_{.B}/F_S$ | $A \subseteq \boldsymbol{S}, B\subseteq$S. |
| $Q(A,B) = F_{AB} - E(A,B) = \sum_{i\in A}\sum_{j\in B} f_{ij} - E(A,B)$ | $F_{A.} = \sum_{i\in A} fi.,\ F_{.B} = \sum_{i\in B} f_{.j}$ |

Figure 3.12 Unadjusted flow- based expectation and modularity between regions.

The above adjustments to the original modularity formulation are crucial. First, the adoption of the flow-based SI model makes the partitioning results more meaningful and easier to interpret in the SI context. Second, the adjustment for within-location flows avoids their undesirable impacts.

### 3.2.3 Hierarchical Partitioning and Hierarchical Modularity

The reported partitioning method is a hierarchical approach. In order to generate a *K*-region partition, (*K-1*) edges must be removed. Each edge removal is equivalent to a bisection cut. That is, removing the first edge leads to a 2-region partition. Removing the second edge within one of the two regions will then generate 3 regions. The overall objective of the top-down partitioning is to remove the edge that gives the smallest between-modularity (Figure 3.9) or the largest within-modularity (Figure 3.10). The partitioning proceeds by successively removing edges.

Hierarchical partitioning and hierarchical modularity are closely coupled in the developed method. The latter means that the flow expectation in the modularity is recalculated after each partition. Specifically, after the original set of locations is divided into two regions, the modularity between nodes and between clusters are calculated using the connections/links within regions only (i.e., treating each new region as a separate data set). For example, suppose the data are divided in two regions *A* and *B*, then a new flow matrix and expectation matrix are constructed within A and B, separately. In other words, further partitions in A will not be influenced by the links in B. Hierarchical expectation makes the presented approach a dynamic one. A static approach would never alter the expectation term in the entire partitioning process. Dynamically updating measures has been shown an effective strategy in graph partitioning (Newman and Girvan 2004).

### 3.2.4 Tabu-based Optimization

The hierarchical clustering introduced above can give a good partition of the SI graph under the contiguity constraint. However, there is still significant room to improve

since the clustering process does not optimize the objective function (i.e., within region modularity) directly. Therefore, this research applies a Tabu-based heuristic algorithm (Figure 3.13) to further improve the cluster partition. It is integrated with the top-down partitioning procedure (Step 4 in Table 3.3) to improve each split.

Input: an initial two-way partition of a graph G into two regions: A and B;
A spatial contiguity matrix of the nodes in G;
n = the number of nodes in G;
Output: an improved partition of G into two new regions A' and B'

Algorithm:
Set the Tabu list T=Ø, the number of moves M=0; the best plan so far P = (A, B);

Substep1: Find best location move on the boundary of region A and region B
(1) Set U=Ø
(2) For each location u in A or B
If *u* is on the boundary of A and B (i.e. the neighbor list of u include at least one from A and one from B )
If move u to B (if *u* ∈A) or to A (if u ∈ B), the new formed A and B remain contiguous
Insert *u* into U
(3) Find the location *u** from U which results in the best improvement of least decrease in the modularity when moved

Substep2: Update the Tabu list T and the best plan so far:
(1) Move *u** to B (if *u** ∈ A) or to A (if *u** ∈ B) to have a new partition (A' and B')
(2) If (A',B') is better than P, then P = (A',B');
(3) If the size of T is less than a threshold (e.g., 0.08*n, as used in this research) and *u** is not in the T
Add this move (*u**) to the Tabu list and increase M by 1
(4) If the size of T is equal to the threshold and *u** is not in the T
Remove the last element in T
Add this move (*u**) to the Tabu list and increase M by 1;

Substep3: If M is less than a threshold (e.g., 3*n)
Repeat Substeps 1-2
Record the best modularity value and its partition A and B;

Substep4: Restart the Tabu procedure:
(1) Set T=Ø;
(2) Repeat Substeps 1-3 until no further improvements is possible;
(3) Report the best partition P.

Figure 3.13 The procedures of Tabu optimization.

### 3.2.5 Hierarchical SI Regions

SI regions are organized in a hierarchy formed during the graph partitioning. The region hierarchy can be visualized with a dendrogram. The county-to-county U.S. migration data as described in Chapter 2 can be partitioned into a hierarchy of regions with the contiguity-constrained graph partitioning method (ALK initial configuration) (see Chapter 3). Figure 3.14a presents the dendrogram for the top 10 regions to show the general organization of the derived regions (i.e. contiguity communities). The regions are labeled with numbers following its hierarchical order. At each level, a region is partitioned into two to produce the next level. In other words, the next level has one more region than the previous (top) level.

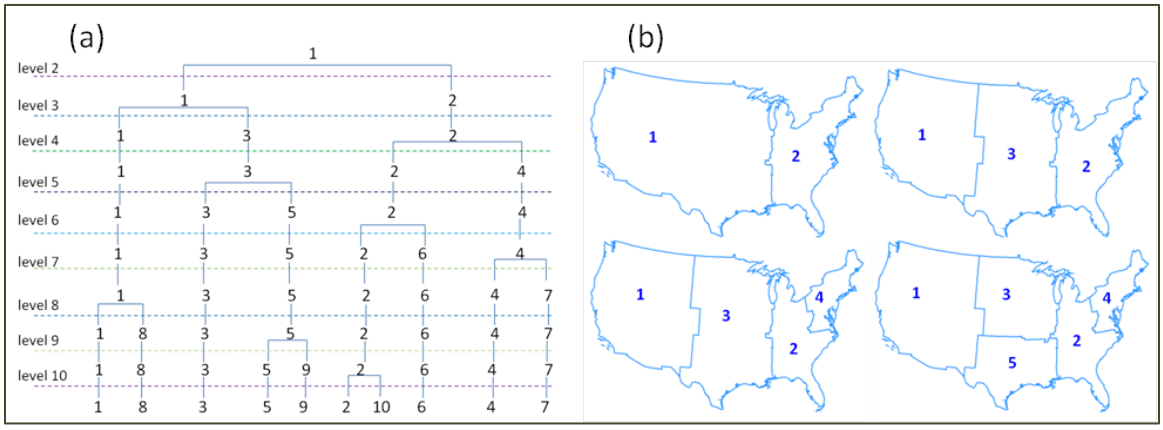


Figure 3.14 The hierarchy of SI regions derived from the migration data. (a) The hierarchy the first 10 regions. (b) Maps of the top 5 regions: each region, at any hierarchical level, is spatially contiguous.

Figure 3.14b presents four maps of the regions at consecutive levels (i.e., 2-, 3-, 4- and 5-region level). Within this hierarchy of regions, each region at a lower level is completely contained by a region at a higher level. Evidently, each new region is fully contained by an existing region. Other regions in the previous level remain unchanged.

Figure 3.15 shows two more hierarchical levels. SI regions are delineated with black lines. Each region is shaded based on the population density. The population density of counties are also shown but without the boundary. The two partition levels also suggest the relationship among the hierarchical region levels: the pattern discovered at lower resolutions (fewer regions) is contained at higher resolutions (more regions). That is, higher-resolution patterns unfold gradually as the number of regions increases (i.e., moving down the hierarchy).

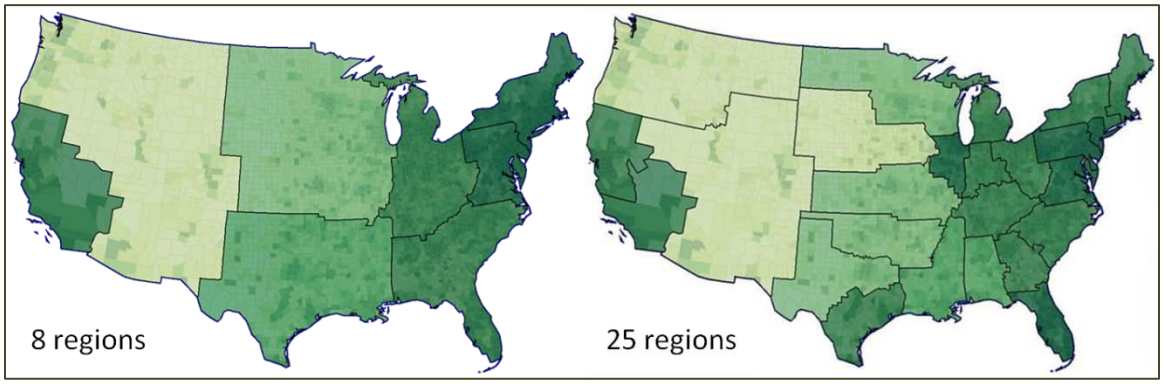


Figure 3.15 The hierarchy of regions derived from the migration data (8- and 25-level).

## 3.3 Evaluations and Comparisons

Evaluation experiments are carried out to compare and test two the visual system from two perspectives. The first is the overall performance and accuracy of the reported method and other existing methods to discover community patterns from SI data. The second aspect is the sensitivity and robustness of the developed approach and other methods in detecting patterns.

### 3.3.1 Overview of the Evaluation Design

The developed partitioning method can have four variants--by combining the optimizing strategy with four alternative hierarchical clustering methods (see Section 3.2.1 and Table 3.4). These variants are compared against two existing SI partitioning

methods: (1) IPFP-SLK (IPFP transformation combined with the SLK clustering) (Slater 1975), and (2) the Intramax approach (the Intramax transformation coupled with a modified Ward's clustering) (Masser and Brown 1975). Therefore, six different methods are compared in this section, out of which four methods are developed in this research. Note the difference between the reported methods with these two existing methods is that the existing methods do not address the local optimal with an optimization strategy. The difference between the reported methods with methods developed for general graphs is that general graph partitioning methods do not enforce the spatial contiguity on the resulting communities.

Table 3.4 Graph partitioning methods for SI data considered in the evaluation

| *Method* | *IPFP-SLK* | *Intramax* | *Developed Methods* | | | |
|---|---|---|---|---|---|---|
| | | | *optiALK* | *optiCLK* | *optiSLK* | *optiWard* |
| *Initial transformation* | IPFP | Intramax | Modularity | | | |
| *Clustering method* | SLK | Ward | ALK | CLK | SLK | Ward |
| *Objective function* | none | Intramax | Modularity | | | |
| *Optimization* | none | none | Tabu | | | |

The evaluation uses two different types of synthetic graphs with known patterns (i.e., community structures): the GN-type and the LFR-type. Each method is applied to detect communities in each of the evaluation data sets. The detected communities are then compared with the true (known) community patterns in the data. Specifically, two similarity measures are calculated respectively to quantify how the detected community structure matches the true structure. A higher similarity suggests higher accuracy and better performance of the method. The two similarity measures used in the evaluation are

FW score (Fowlkes and Mallows 1983) and the entropy-based "normalized mutual information" score (Danon et al. 2005).

FM index (see Figure 3.16) constructs a connection matrix ($C$) for each partition (derived or the true structure), the element of which is $C_{ij}$. $C_{ij}$ *i*s 1 if element (or node) $i$ and element (or node) $j$ are in the same community and 0 if not. $C^A$ and $C^B$ represent the two partitions to compare. The numerator counts the pairs of nodes that belong to the same community in both partitions. The denominator is the square root of the number of node pairs within the same cluster in partition $A$ times the number of node pairs within the same cluster in partition $B$. The Rand index (Rand 1971) is another "pair-counting" measures. It has the similar form as FM index except that the numerator is the count of node pairs within the same communities and different communities in the two solutions and the denominator is simply the count of all node pairs.

$$FM\ (A,B) = \sum_{i=0}^{n}\sum_{j=0}^{n} {C^A}_{ij}{C^B}_{ij} \Bigg/ \sqrt{\left(\sum_{i=0}^{n}\sum_{j=0}^{n} {C^A}_{ij} * {C^A}_{ij}\right) * \left(\sum_{i=0}^{n}\sum_{j=0}^{n} {C^B}_{ij} * {C^B}_{ij}\right)}$$

Figure 3.16 Fowlkes and Mallows (FM) similarity Index

Entropy-based "normalized mutual information" measure (see Figure 3.17) is constructed from a "confusion matrix" $N$ in which the rows represent the communities/clusters in partition $A$ and the columns represent the communities/clusters in partition $B$. The element of $N$ *(i.e.* $N_{ij}$) is the count of nodes which appear in cluster $i$ of partition $A$ and belong to cluster $j$ of partition $B$. $N_{i.}$ is the sum of $N_{ij}$ over $j$ and $N_{.j}$ is the sum of $N_{ij}$ over $i$, $C_A$ and $C_B$ are the number of clusters in $A$ and $B$ respectively. This

measure reaches its maximum if two partitions are exactly the same (i.e. 1). If the two sets of community labels are completely independent, the measure has its minimum 0.

$$I(\mathrm{A},\mathrm{B}) = \left(-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij}\log\left(N_{ij}N/N_{i.}N_{.j}\right)\right) \Bigg/ \left(\sum_{i=1}^{C_A} N_{i.}\log\left(N_{i.}/N\right) + \sum_{j=1}^{C_B} N_{.j}\log\left(N_{.j}/N\right)\right)$$

Figure 3.17 Entropy-based "normalized mutual information" similarity measure

### 3.3.2 Evaluation Data Sets

Two synthetic benchmark data sets are used in the evaluation: (1) Girvan-Newman (GN) benchmark (Newman and Girvan 2004), and (2) Lancichinetti-Fortunato-Radicchi (LFR) benchmark (Lancichinetti et al. 2008) (see section 3.1.2.3 for more details). Two small LFR graphs are generated to illustrate the community structures planted in the synthetic graphs. Since it is unrealistic to show a graph involving 1000 nodes on a paper, the two examples have a reduced size (i.e. 200). Nodes belonging to the same planted community are represented by dots in the same color (Figure 3.18).

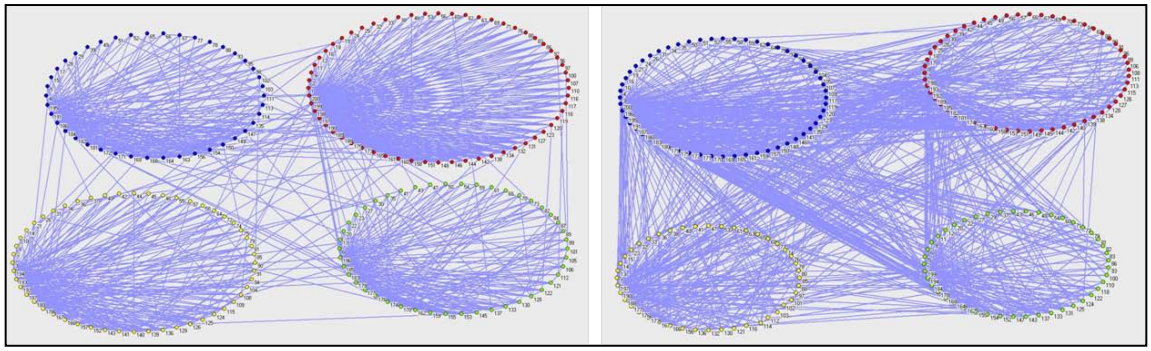

Figure 3.18 Example of synthetic LFR graphs. (a) Mixing parameter=0.1, relatively strong community patterns. (b) Mixing parameter=0.5, relatively weak community patterns.

The configurations of the two small LFR graphs are presented in the first row of Table 3.5. These two examples share the same setting except for the mixing parameter. Mixing parameter is the ratio of the external degree to the overall degree. It determines the strength of community patterns. A low mixing parameter indicates a strong community structure and a large mixing parameter means a weak community pattern.

Table 3.5 Configurations of example and evaluation graphs

| | Node | | | | Edge | Community | | |
|---|---|---|---|---|---|---|---|---|
| Name | *Count* | *Avg. degree* | *Max degree* | *Degree dist.* | *Mixing parameter* | *Size dist.* | Min. Size | Max. Size |
| LFR | 200 | 10 | 100 | PL (-1) | 0.1,0.5 | PL (-2) | 20 | 50 |
| GN | 1000 | 50 | 50 | even | 0.3,0.4, 0.5, 0.6,0.7 | even | 100 | 100 |
| LFR | 1000 | 57 | 500 | PL (-1) | 0.2,0.3, 0.4, 0.5,0.6 | PL (-2) | 100 | 300 |

Figure 3.5a shows that the graph generated with a smaller mixing parameter (i.e. 0.1) has considerable links within communities, exhibiting clear community patterns. Figure 3.5b shows the graph generated with a larger mixing parameter (i.e. 0.5). It has more between-community (external) links and the community pattern is obscure (Figure 3.5b). Since there is no geographic information, the layouts of the two graphs are uniformly circular, which is only a simplified representation of real geo-referenced SI graphs.

Two sets of synthetic GN and LFR graphs are generated with a freely available toolkit[4] (Lancichinetti 2008). The graphs are undirected and unweighted. The toolkit is able to generate both GN and LFR graphs with proper settings. But the toolkit is not

[4] Available at: http://sites.google.com/site/santofortunato/inthepress2

effective in generating weighted graphs since it does not consider weights in the construction of the community patterns (Lancichinetti et al. 2008, Lancichinetti and Fortunato 2009). Table 3.6 presents the configuration of the two sets of evaluation graphs on the 2nd and 3rd rows. Each of the evaluation graphs has 1000 nodes. The GN graph contains ten equal-sized communities and the degree of each node is uniformly 50. The threshold of the mixing parameter (to ensure the presence of community patterns) is 0.9 (see section 3.1.2.3). The mixing parameter of the GN set ranges from 0.3 to 0.7.

An LFR graph is different from a GN graph in that it has a power-law distributed community size and node degrees. The number in the parentheses in the column "Degree dist." and the column "Size dist." is the negative of the exponent of the (node) degree distribution and the (community) size distribution. The average degree of LFR is set at 57 such that the resulting graph can have approximately the same graph density (i.e., the ratio of the number of edges to the number of possible edges) as the migration data used in this research. The range of community sizes is set at 100--300. The mixing parameter needs to be less than 0.7 (i.e. (1000-300)/1000) to ensure the presence of community patterns. The mixing parameter of the LFR set ranges from 0.2 to 0.6.

### 3.3.3 Evaluation Results

Using the setting in Table 3.5, five sets of GN and five sets of LFR benchmark graphs are generated with varied mixing parameters. Each set contains ten graphs. The evaluation tests are conducted on machine of 3.16 GHz CPU and 3.25 GB of RAM. The quality of the derived partitioning result for each graph is assessed with two similarity measures, which quantify how well the partitioning can discover the true patterns.

Figure 3.19 and Figure 3.20 present the evaluation results for the GN and LFR benchmark graphs respectively. Each curve corresponds to one of the methods and points on the curve represent the average similarity scores (vertical axis, ranging from 0 to 1) at a level of the mixing parameter ($\mu$) (horizontal axis) (see Table 3.5 for the configuration of other parameters of the graphs). As mentioned earlier, the mixing parameter controls the strength of community patterns, where higher mixing values lead to weaker community structures. A high similarity suggests a good quality.

The evaluation results show that the IPFP-SLK does not work well regardless of the strength of community patterns or the types of graphs. This may be partially related to the distortion introduced by the IPFP transformation and the chaining effect of the SLK clustering. The FM scores of IPFP-SLK are higher than its Entropy scores because the FM measure only counts pairs of nodes correctly assigned to the same communities but does not punish incorrect community assignments. The Intramax approach is significantly better than IPFP-SLK. Overall, the methods developed in this research outperform both the IPFP-SLK and Intramax. Particularly, the developed methods discover the planted community structure with the best fidelity when the planted pattern is strong.

Not surprisingly, weaker structures are more difficult to detect. The new methods consistently work better than the two existing methods at different mixing parameter values (i.e., strengths of community structures). The superior performance of the reported methods over the Intramax approach is due to the Tabu optimization embedded in the former.

Moreover, the performance of the presented methods is also stable with different initiations. In other words, the four variants of the presented method are very similar in performance, indicating that the optimization strategy effectively reduces the sensitivity to the initial configuration. Among the four variants, the one combined with the Ward's clustering (optiWard) generates slightly better solution. This may be related to the fact that the optiWard uses the same measure (i.e., modularity) in both the clustering and the optimization process.

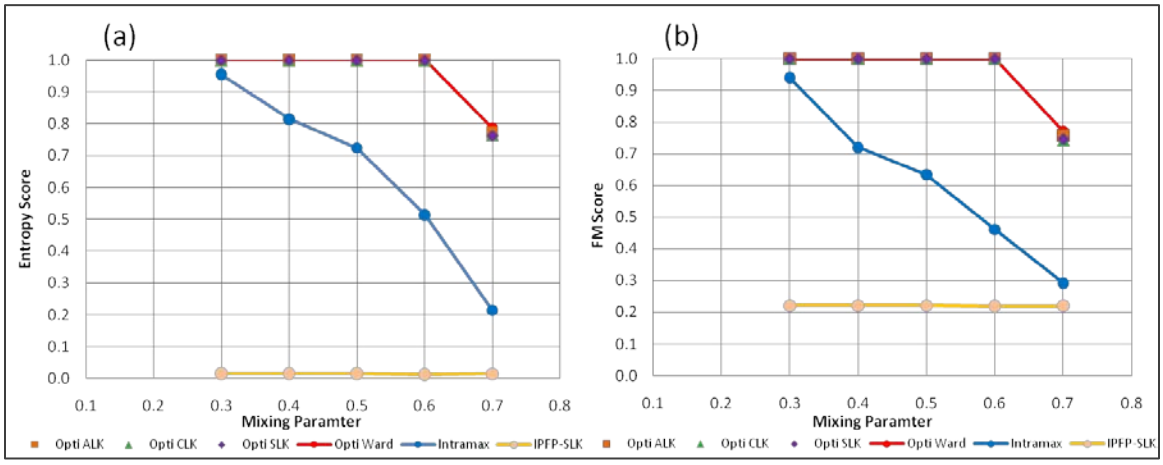


Figure 3.19 Similarity scores of methods tested on the GN benchmark graphs. (a) Entropy-based scores. (b) FM scores

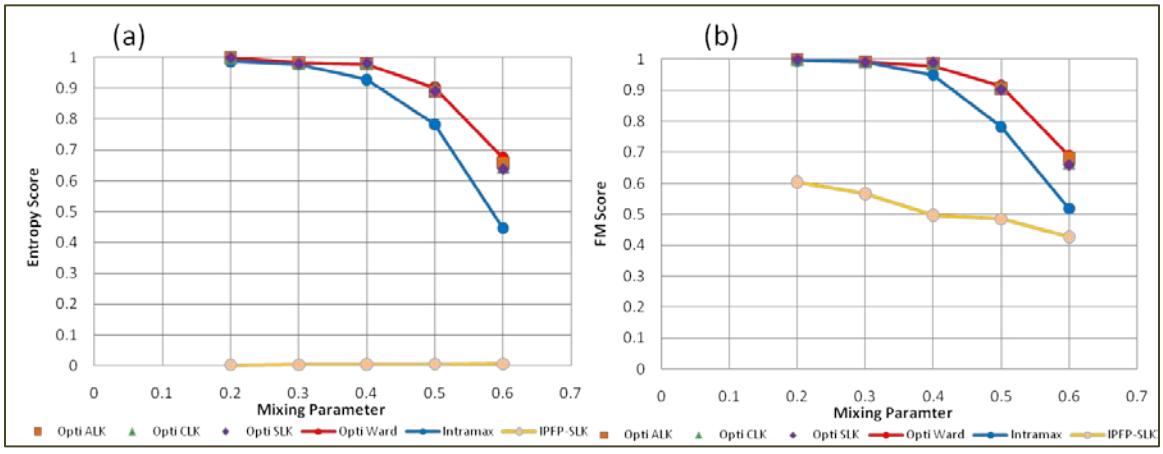


Figure 3.20 Similarity scores of methods tested on the LFR benchmark graphs. (a) Entropy-based scores. (b) FM scores.

Table 3.6 and Table 3.7 represent the time cost of the methods tested on the GN and LFR benchmark graphs respectively. As the real community structure becomes weaker, the computational cost of the method reported in this research increases. Given the optimization procedures embedded in the reported method, it is understandable that the new reported methods takes longer time to derive a partition then the other two methods. Given the significant improvement the partition that the new methods generate, the extra computational cost is worthwhile and affordable. The largest difference (i.e. 87 seconds) is between optiWard and IPFP-SLK when they are tested on the most difficult graph in both types of graphs.

Table 3.6 The time cost of methods tested on the GN benchmark graphs (in seconds)

| Method | GN(0.3[5]) | GN(0.4) | GN(0.5) | GN(0.6) | GN(0.7) |
|---|---|---|---|---|---|
| IPFP-SLK | 20 | 22 | 22 | 17 | 19 |
| Intramax | 24 | 25 | 26 | 26 | 25 |
| OptiALK | 53 | 46 | 53 | 60 | 77 |
| OptiCLK | 36 | 55 | 67 | 80 | 66 |
| OptiSLK | 40 | 75 | 59 | 76 | 104 |
| OptiWard | 47 | 69 | 76 | 61 | 51 |

Table 3.7 The time cost of methods tested on the LFR benchmark graphs (in seconds)

| Method | LFR(0.2) | LFR(0.3) | LFR(0.4) | LFR(0.5) | LFR(0.6) |
|---|---|---|---|---|---|
| *IPFP-SLK* | 27 | 15 | 15 | 12 | 19 |
| *Intramax* | 37 | 32 | 42 | 40 | 40 |
| *OptiALK* | 63 | 48 | 65 | 65 | 72 |
| *OptiCLK* | 49 | 41 | 46 | 40 | 59 |
| *OptiSLK* | 71 | 43 | 48 | 62 | 54 |
| *OptiWard* | 51 | 59 | 55 | 61 | 106 |

### 3.4 Summary and Discussions

The presented partitioning method for SI graphs consists of two stages: the bottom-top clustering and the top-down partitioning and optimization. The modularity

[5] The mixing parameter of the graph that the methods are tested on.

measure or objective function is embedded in both stages. The main features of the new partitioning method include: (1) explicit enforcement of spatial contiguity; (2) consistent use of modularity in the partitioning process; (3) Tabu-based optimizing strategy that significantly improves the partition quality; and (4) superior and robust performance. Experiments results with benchmark graphs show that the presented partitioning method is not sensitive to the initialization clustering method and can better detect community structures in graphs than existing methods.

In order to partition a SI graph/network, one needs to decide which spatial characteristics (e.g. distance, contiguity, and barriers) to consider and how to incorporate it in the graph partitioning. This research chooses to enforce spatial contiguity and pursue the direction of contiguity-constrained SI regions introduced by Ng (Ng 1969), since spatial contiguity is a key property of various region definitions in geography.

In order to obtain reliable and interpretable community structures, we need first define a community structure. The presented approach adopts the modularity-based definition. Moreover, several extensions are made to better account for special characteristics in spatial interaction graphs. The consistent use of modularity facilitates the production of reliable and meaningful partition results. In addition, with the Tabu-based optimization, the method can better escape local optima and minimize the influence of initial configuration and clustering.

The presented method is evaluated with two types of synthetic benchmark graphs and compared with two existing graph partitioning methods: the IPFP-SLK approach (Slater 1975, Clark 1982, Slater 1984) and the Intramax approach (Masser and Brown 1975, Poon 1997). The evaluation results show that: (1) when the community structure is

relatively strong, the Intramax approach and the presented approach both performs very well and much better than the IPFP-SLK approach; and (2) for weak community structures, the presented method outperforms both the Intramax approach and the IPFP approach. The new method is computationally efficient. It takes nearly 2 minutes on a machine of 3.16 GHz CPU and 3.25 GB of RAM to partition 3,075 counties into 100 regions.

In its current form, the presented approach cannot automatically determine the number of communities to be detected. In other words, the methods presents a hierarchy of communities, with each hierarchical level corresponds to a specific number of communities. For example, the synthetic data sets that we use for the evaluation only have flat structures (i.e., each graph has a specific number of communities, which do not form a hierarchy), the new method still gives a hierarchy of communities. Therefore, it is up to the user to determine which level is the best solution.

This limitation will be addressed in future research. It is possible to determine the best hierarchical level based on the modularity gain at each partition and/or the trend of modularity values at each level. It is also likely that community patterns may exist at different abstraction levels. In this case, the method needs to first determine the levels at which community patterns exist. Then the community pattern at each of the levels needs to be identified. With the above expansions, the presented method would become an even more useful and powerful tool to extract community patterns from SI data.

# CHAPTER 4

# VISUAL EXPLORATION OF FLOW PATTERNS

"Knowledge is always gained by the orderly loss of information." -- Kenneth Boulding (Boulding 1970: 2)

Confirmatory and exploratory analyses are two different and complementary approaches to analyzing data. Confirmatory modeling has a long tradition in migration analysis. For example, modeling analyses on the determinants of migration has been the focus of 65% publications in five leading regional science journals between 1955 and 1994 (Plane and Bitter 1997). The modeling approach is especially favored by regional scientists, who have made substantial contributions to migration studies (Cushing and Poot 2004).

In sharp contrast to the plethora of SI models, there is much less research on the exploratory analysis of SI data (Greenwood and Hunt 2003). This unbalanced development was noted a few decades ago:"... *papers have placed a premium on the development and testing of new hypotheses rather than on descriptions of facts and their collation*" (Haenszel 1967: 260). The lagging development of exploratory analysis became a restraint of SI modeling. SI modeling has arrived at a stage where a breakthrough is needed to obtain new knowledge about spatial interactions (Roy and Thill 2004), which is insufficiently represented in current modeling analyses (Chun and Griffith 2011, Griffith and Jones 1980, Curry et al. 1975, Sheppard et al. 1976).

Our knowledge on SI has been severely hindered by our capability to analyze SI data and obtain unknown/unexpected information (Rae 2009, Young 2002). This chapter focuses on a visual analytic system developed for SI analysis, which can facilitate systematic and flexible investigations of spatial interaction patterns. "Systematic" means that multiple data spaces of SI can be linked and examined simultaneously. "Flexibility" refers to the capability to view the data at varying resolutions, from different perspectives, and in multiple coordinated visual forms.

Hierarchical SI regions derived with the spatially-constrained graph partitioning method (see Chapter 3) are used in the visual analytic system to present patterns at different geographic scales. This chapter will show how the visual analytics approach: (1) visualizes discovered community patterns, (2) reveals spatial patterns of region-to-region flows; (3) extracts multivariate patterns across flow/regions/locations. This chapter is organized into three sections. The first section reviews existing methods for the visualization and mapping of spatial interactions. The second section introduces the components of the developed visual system and the coordination among different components. The third section presents an analysis example to illustrate the main features of the exploratory analysis environment.

## 4.1 Mapping SI Flows: Related Work

A flow map is a common exploratory approach in SI analysis, where origins, destinations, and flows are of primary concerns. The goal of flow mapping is to visualize georeferenced flow data (Figure 4.1) that consist of: (1) spatial information of origins/destinations and (2) a data cube (or data table) representing the multivariate

information of flows (e.g., types of flows, compositions of flows). For example, the multivariate information of migration flows can be the category of migrations: seasonal, return, or chain migration, voluntary or involuntary. Stratifications of migration flows based on social, economic, demographic, and household characteristics of migrants are also typically available. A desired flow mapping should be able to show the spatial patterns and the associated multivariate patterns simultaneously.

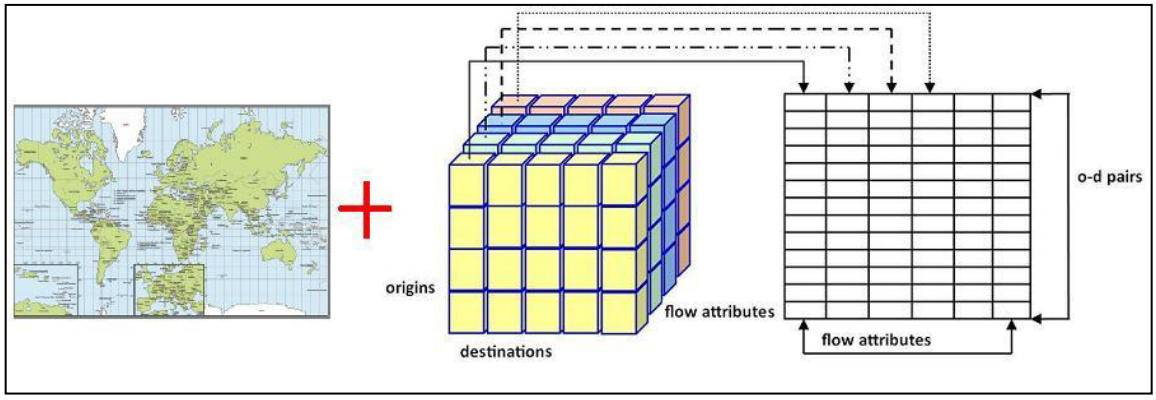


Figure 4.1 Geo-referenced flow data: consisting of spatial information and a flow data tube or flow data table (right) (based on: Yan and Thill 2009)

In order to visualize SI data, three major issues need to be addressed: (1) the primary visualization form(s); (2) a strategy to abstract patterns from voluminous flow data; and (3) a method to transform original flows to visualize the flow magnitude.

Flow map, flow matrix, and arrow graph are among the main visual forms for geo-referenced flows. Due to the growing size and complexity of SI data, direct depiction or mapping of individual origin-destination pairs becomes unrealistic in many scenarios. Researchers therefore often use data summarization or computational methods to extract certain features from the data (Andrienko et al. 2008) to enable legible flow visualizations. Moreover, due to the varying sizes of spatial units, it is usually not meaningful to directly map the raw flow volume. Instead, the raw flow volume should be

normalized or transformed. For example, instead of mapping the raw migration among the U.S. states, it is more meaningful to map net migrations, flow efficiencies, etc.

### 4.1.1 Flow Maps

A flow map is an intuitive and commonly used approach in visualizing SI data. One of the earliest uses of flow map can be found in the systematic migration studies a century ago (Ravenstein 1885). Flow maps connect origins and destinations using flow lines. The width of flow lines may signify the magnitude of flows (Tobler 1981). A number of alternative flow symbols have been proposed (Tobler 1987). For instance, a flow symbol can be a rectangular band with the width or shading representing the flow magnitude.

Trajectory data are a special category of SI data where the route of movements should be mapped. Therefore, a flow map for trajectory data not only shows origin and destination but also locations on the route. Figure 4.2 provides two examples of flow map. The first one (see Figure 4.2a) is a well known example visualization of trajectory data. It is based on Minard's graph of Napoleon's Russian campaign (Tufte 1986). The second example (see Figure 4.2b) maps the telecommunication traffic in Europe.

For a long time, the flow map was a desirable exploratory tool of SI data. As the two examples indicates, it is simple to implement and easy to understand. Flow map remains a popular choice for visualizing SI flows nowadays. However, flow map has a serious scalability problem due to the quadratic growth of flows as the number of involved locations increases. A flow table involving 50 places can have 2,450 flow lines, which are already too large for a flow map to present without suffering severe visual

cluttering problem. Another limitation of existing flow maps is its limited capability to represent multivariate information of locations and flows. Traditional flow map may be able to show multiple properties of flows and places only when the number of flows and variables is small, such as the map shown in Figure 4.2a.

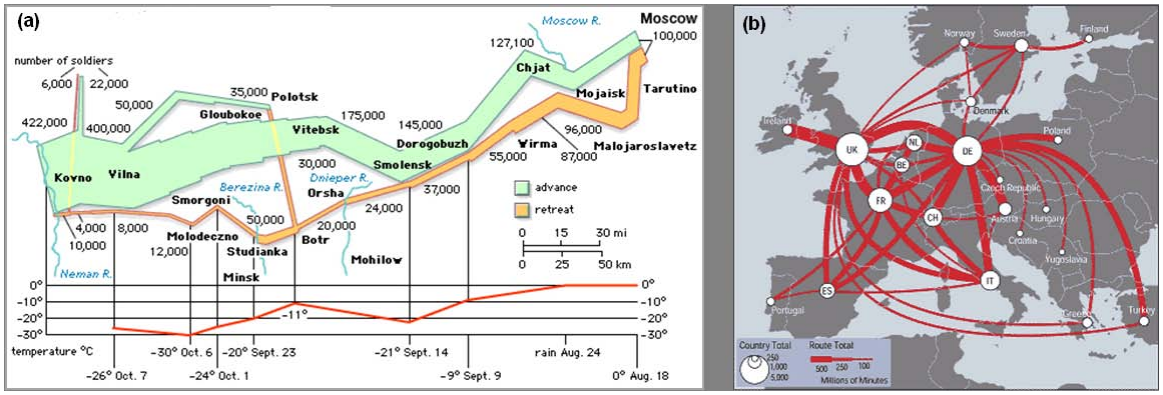


Figure 4.2 Examples of flow maps. (a) Napoleon's 1812 Russian campaign (source: **www.emersonkent.com**): temporal information (i.e. date) and a variable of locations (i.e. temperature) are depicted, in addition to the geographic locations, routes, and the changing number of soldiers along the route. (b) Telecommunication traffic in Europe (source: **www.mundi.net**): circles placed in each country represent the total traffic and curves connecting two countries represent the traffic between them.

In order to address the scalability problem of flow map, a number of solutions have been proposed. Interaction techniques, such as the filtering and focusing used in (Rae 2009), can allow the user to choose a small subset of data to map. However, mapping a selected subset cannot provide a comprehensive view of the data. Another group of solutions groups flows based on their geometric similarities or densities and/or adjusts flow graphics to partially resolve cluttering (Holten and Wijk 2009). A straight forward attempt would be using a large physical displays (e.g., multiple monitors) (Abello et al. 1999) but this approach still cannot handle large data sets. One may also reduce the data size by summarizing or extracting certain patterns from the data prior to

the mapping. These solutions are briefly reviewed below as two separate groups: those focusing on manipulating the graphic displays and those involving data aggregations.

#### 4.1.1.1 Strategies Focusing on Graphics

Strategies focusing on graphics attempt to handle large data volumes by modifying flow symbols to minimize the visual cluttering problem. A variety of approaches have been proposed to visualize large networks in the field of graph visualizations. Most of them, however, are not directly useful for mapping SI data since they either require certain features of data, such as a hierarchy or multi levels (Eades et al. 1996, Schaffer et al. 1996), or do not consider fixed locations of nodes (e.g., geographic locations).

Edge bundling is an approach that is applicable to SI data. It attempts to adjust the shape of flow lines and combine them to alleviate the overall congestion problem by bundling lines in the map (see Figure 4.3). Specifically, portions of edges can be merged while the node locations remain unchanged (Holten and Wijk 2009). This strategy first divides each edge into segments and then move the end points of the segments based on a force-directed model. This method works well when the majority of flows are long-distance. But it cannot significantly reduce the congestion of short-distance flows.

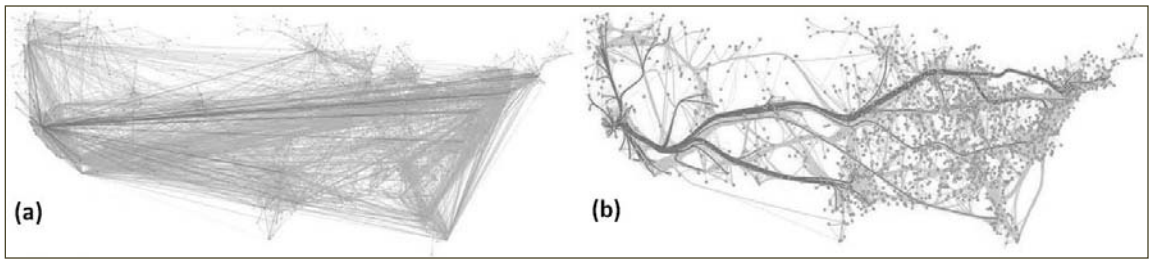


Figure 4.3 The use of edge bundling in mapping U.S. migration flow. (a) Without edge bunding. (b) With bundling (source: (Holten and Wijk 2009)).

Figure 4.3 shows an application of the edge bundling strategy with a migration data set with 1,715 nodes and 9,780 edges. It can be seen that edge bundling emphasizes long-distance edges while short-distance flows are difficult to follow, such as flows in the highly urbanized Northwest. Moreover, edge bundling is based on pure geometric operations and does not consider the real meaning of the "bundled lines" or flows. For example, the migration map with this method (see Figure 4.3b) can be misleading since a migrant from New York to California does not have to "go" through Minnesota and Arizona. Another problem with this approach is that it is difficult to follow a specific flow. For example, after the bundling, it is not clear how many migrants moved from New York to California since they are not directly connected and mapped.

Another strategy similar to edge bundling is to rearrange node positions and merge routes (Phan et al. 2005). This approach renders easily distinguishable flow patterns but can only map flows attached to one or a small set of locations. Therefore it cannot map a large data set. The geometry-based edge clustering proposed by Cui requires a construction of control mesh (Cui et al. 2008). This approach leads to considerable variation in curvature (Holten and Wijk 2009).

#### 4.1.1.2 Strategies Involving Data Abstractions

Data abstraction or generalization is to reduce the number of flows to a level that traditional flow map can handle. Often large SI data are aggregated on predefined regions such as administrative units (e.g. states) or statistic units (e.g., census regions). Statistical economic regions are commonly used in studying the flows between the urban and the rural sectors, such as the 26-region divisions from the State Economic Area system

(Fuguitt and Beale 1993). Other than administrative or census units, regular grids may also be used aggregate flow lines through each grid (Rae 2009). The grids are shaded based on the intensity of flow lines. The resulting line-density map is easy to understand but it is sensitive to the arbitrary grid partition. Moreover, the origin and destination of flows are missing in the representation.

Another type of strategy to reduce data size is to use computational methods to summarize data and extract. For example, multivariate flow data tube (see Figure 4.1) can be condensed into abstract flow patterns with factor analyses (Berry 1966). Yan and Thill use self-organizing map (SOM) to cluster flow data and facilitate informative selection of flows (Yan and Thill 2009). Further, the use of SOM offers a view of the multivariate structures of flows. Its main limitation is that clustering results cannot be used to alleviate the congestion caused by dense flow symbols since the clustering only considers the multivariate space of SI data (not the geographic space). Flows belonging to a multivariate cluster may scatter over the study area.

This study combines the pattern extraction strategy and the data generalization to enable effective and meaningful flow mapping. First, network patterns are extracted directly from spatial interaction graphs with the contiguity-constrained graph partitioning method. The detected SI regions are then used as the generalization/aggregation scheme to reduce the number of flows to visualize.

The combination of SI regions and the data aggregation is different from existing methods since existing data aggregation strategies (as reviewed above) are not based on the flow network. The newly developed method in this research aggregates flows based

on network structures in the SI data and thus is able to render more meaningful data abstractions at different scales by moving up or down the region hierarchy.

### 4.1.2 Other Visual Representations: Flow Matrix and Arrow Graph

Flow matrix and arrow graph are two most popular alternatives to flow maps for visualizing directed flows. A matrix view is applicable to general graph data. Compared to a 3-d matrix, a 2-d matrix view is more conceivable and more commonly used (Wood et al. 2010). A 2-d flow matrix has the rows/columns representing the nodes and the cells indicating the presences or intensities of edges. The order of matrix columns and/or rows can be rearranged to highlight interesting patterns. An appropriate ordering can lead to a dramatically improved effect (Guo 2007). Various techniques have been proposed to achieve optimal ordering (Makinen and Siirtola 2000, Siirtola and Makinen 2005, Guo and Gahegan 2006). Friendly and Kwan describe two sets of "effect-ordering" techniques for frequency data and multivariate data (Friendly and Kwan 2003).

Systematic comparisons were conducted between matrix and node-link representations (e.g. force-based graph layout, flow map) of general network data (Ghoniem et al. 2004, Ghoniem et al. 2005). The conclusion is that "*node-link diagrams are well suited for small graphs, and matrices are suitable for large or dense graphs*" (Ghoniem et al. 2004: 23). Here a "small graph" refers to networks with 20 or fewer nodes. This conclusion is based upon the performance of the node-link diagrams and the matrices on a set of seven generic graph visualization problems. However, since the evaluation problems are not designed for georeferenced graphs, the conclusion is

questionable for SI data. A set of evaluation tasks concerning the spatiality of georeferenced graphs would lead to more convincing conclusions for SI data.

When applied to SI data, matrix visualization does not preserve the spatiality of locations and flows, although its layout indeed facilitates cognitive coupling of origins and destinations. However, the rows/columns of a matrix view can be reordered to keep spatially contiguous origins/destinations neighboring to each other as much as possible in the matrix (Guo and Gahegan 2006, Andrienko and Andrienko 2008, Marble et al. 1997). This strategy is utilized in this research in a supplemental matrix visualization of flows.

Arrow graph is another visualization form for graph data (Figure 4.4) but it does not focus on the spatial dimensions. Instead, it aggregates flows based on a pre-defined categorization of places, often based on the attribute information of locations. The predefined categories of locations are represented by parallel and horizontal lines in the arrow graph. Arrows are drawn vertically between the horizontal category lines to show the flows among these categories. Typically, the color and width of an arrow indicate the flow direction and the flow volume.

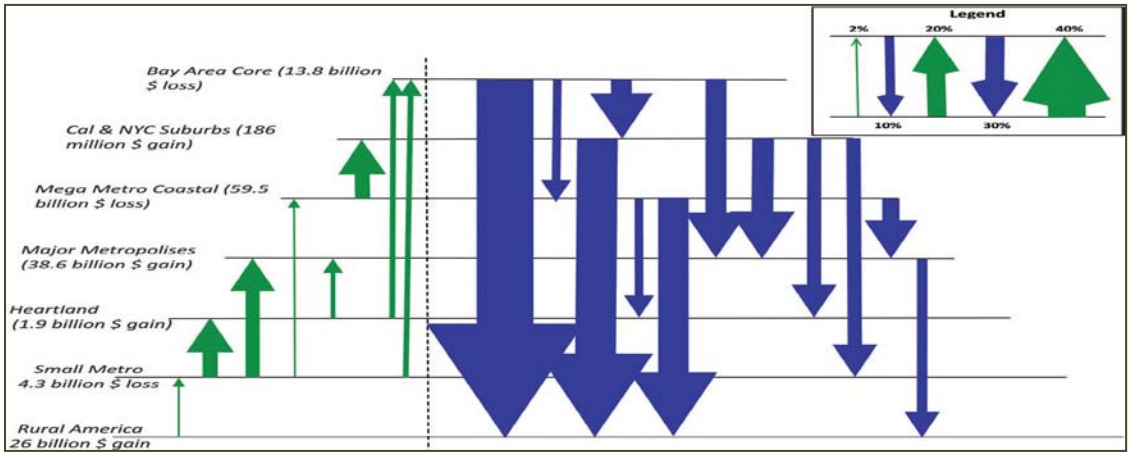


Figure 4.4 An arrow graph showing SI flows (source: Shumway and Otterstrom 2010)

The arrow graph is sensitive to the classification of places. Urban-rural spectrum is usually utilized in arrow graphs for migration analysis (Plane et al. 2005, Shumway and Otterstrom 2010). Flows along the urban-rural hierarchy can be clearly displayed with arrow graphs. The limitation of arrow graphs lies in: (1) the absence of spatial information regarding specific origin-destination flows; (2) different classification schemes can result in dramatically different visualizations.

.

### 4.1.3 Flow Measures

After choosing a visual form, the next issue of flow mapping is to decide the flow measure. Measures such as gross flow and net flow are widely used and easy to understand. The size effect, however, is often present in these absolute terms. For instance, a populous place tends to have more migration.

Two general strategies have been used to offset the size effect. One strategy is to estimate flows with relevant factors and deduct the estimate from the observed value. The idea is conceptually similar to the modularity measure discussed in Chapter 3 of this paper. Various models involving different factors have been devised to estimate flows. In the early history of migration studies, distance is often involved in gravity models for migration estimation (Zipf 1946, Stouffer 1940). It is argued that geographic or transportation is an inaccurate estimator without corrections (Plane 1984a, Anderson 1955). Further, distance-oriented estimates cannot capture the interference of natural or administrative barriers. For instance, it is shown that state boundary may restrict migrations, which cannot be represented by distance (Anderson 1955).

Estimations based on population are conducted in a number of works, including the Chi-statistics (Wood et al. 2010) which divides the difference between the estimate and the actual flow (i.e. flow modularity) by the square root of the estimate. Another transformation with the population of locations is to normalize flow rates by the population of the origin or the cross product of populations in the origin and the destinations (Haenszel 1967).

The other strategy used to remove size effect does not count on predictors. For instance, flow efficiency counters the size impact by taking the ratio of the flow difference to their total on both directions, i.e. net migration to the gross migration (Podolák 1995, Plane 1984b).

In addition to flow measures, there are also area-based measurements for locations/nodes devised from a network perspective. For instance, the GINI index (Duncan 1957, White 1986, Plane and Mulligan 1997) and the coefficient of variation (CV) (Allison 1978, Rogers and Sweeney 1998) reflect the spatial concentration of flows. "Migration drift" measures the average net directionality and the distance moved (Plane 1999b). Plane used this measure to compare the moving behavior exhibited by various migrant groups. Flow efficiency is a measure which can be easily extended for locations. Flow efficiency of locations is simply the ratio of the sum of net flows to the sum of gross flows. According to the classification scheme of measures used in migration studies (Bell et al. 2002), net flow, gross flow, and flow modularity measure the "intensity of migration", flow efficiency may represent the "effect of migration", while network measures indicate "migration connectivity".

Absolute terms and relative terms offer different and complementary perspectives. For instance, ratio-based flow measurements evaluate the relative significance of flows. However, they cannot distinguish flows when the ratios are close. As an example, the (migration) flow efficiency of Los Angeles and Dayton-Springfield in Ohio are close (-5.25% and -5.57% respectively) (Newbold and Peterson 2001) but the net out-migration of Los Angeles is almost 10 times of Dayton-Springfield. Therefore, investigating both the absolute and relative terms of flows can lead to a more comprehensive understanding of the flow patterns.

## 4.2 An Integrated and Interactive Analysis Environment

### 4.2.1 Overview of the Framework

The visual analytic framework developed in this research is described in Figure 4.5. The analysis starts with a hierarchy of SI regions derived with the graph partitioning method and the three data spaces of SI data. The SI regions are used to aggregate location-to-location flows to obtain region-to-region flows. It is also used to convert the location-based variables into region-based variables.

Then the analyst can choose the region level (or abstraction level) to aggregate flow data and a flow measure for the aggregated flows. A threshold can be set to filter flows, i.e., only show flows that have a flow measure higher than the specified threshold. This filtering function helps analysts focus on major flows. If no multivariate information is selected (see the purple arrow in Figure 4.5), flows and SI regions are visualized in a flow map (FlowMap+) and complementary visualizations (e.g., flow matrix and/or data/flow table). If one or more variables are selected for univariate/multivariate analyses

(orange arrows in Figure 4.5), a multivariate clustering or a univariate classification is performed. The multivariate clustering or univariate classification results are represented by colors in the flow map, with the assistance of two other visual components: self-organizing map (SOM) and parallel coordinate plot (PCP).

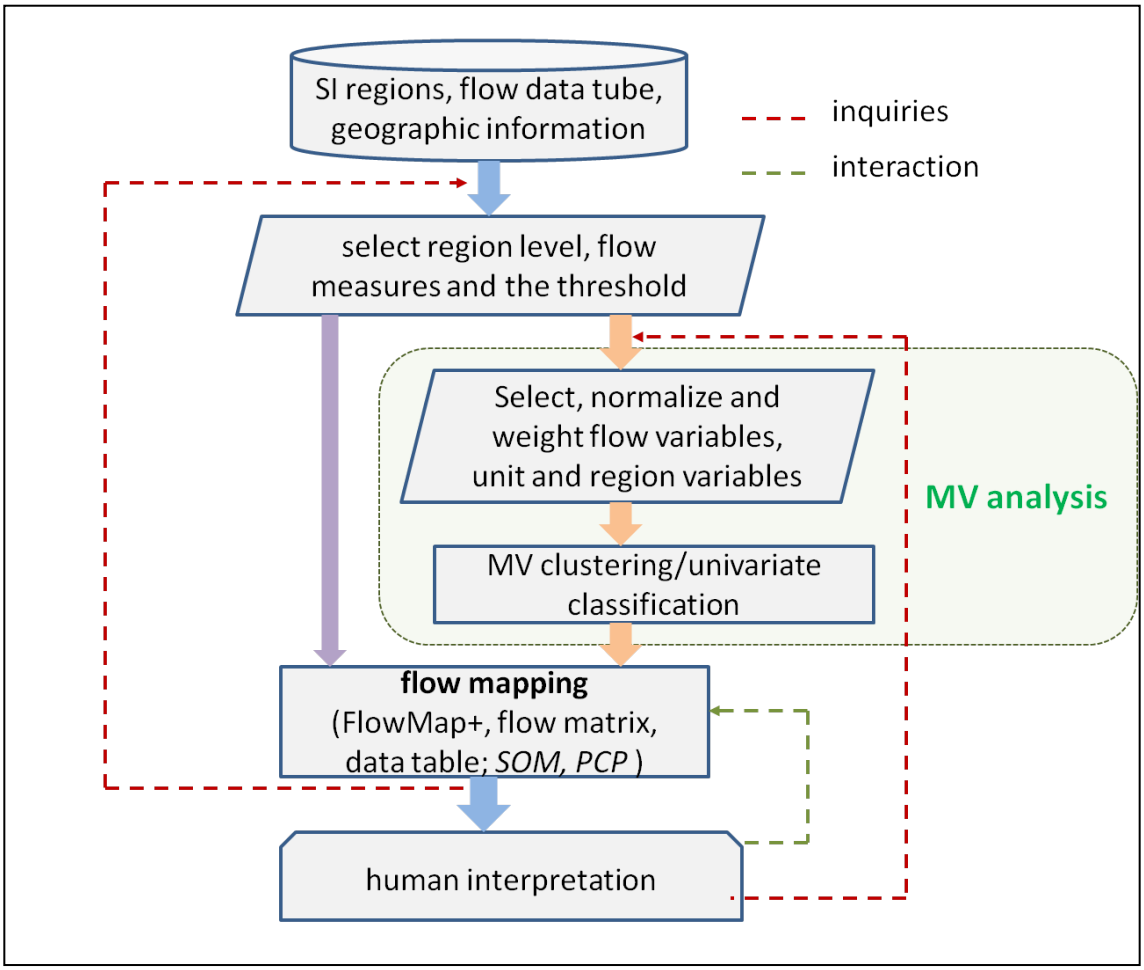


Figure 4.5 The visual analytic framework. Flow data and multivariate data are aggregated by SI regions first; flows are visualized with FlowMap+. If variables are selected for multivariate analysis, the clustering or classification results are presented in FlowMap+ and other relevant components.

The system supports an iterative analysis process, where the user can change the setting of the system (e.g. different region numbers, different flow thresholds, or different sets of variables) and examine the changing patterns (see the red dotted lines in Figure

4.5). All visual components, including the flow map and other chosen forms, are updated automatically upon user interactions. A rich set of interaction techniques is implemented to facilitate dynamic and flexible flow mapping.

The visual system developed and implemented in this research consists of multiple components. Figure 4.6 gives an example layout of these components and their functions are listed in Table 4.1.

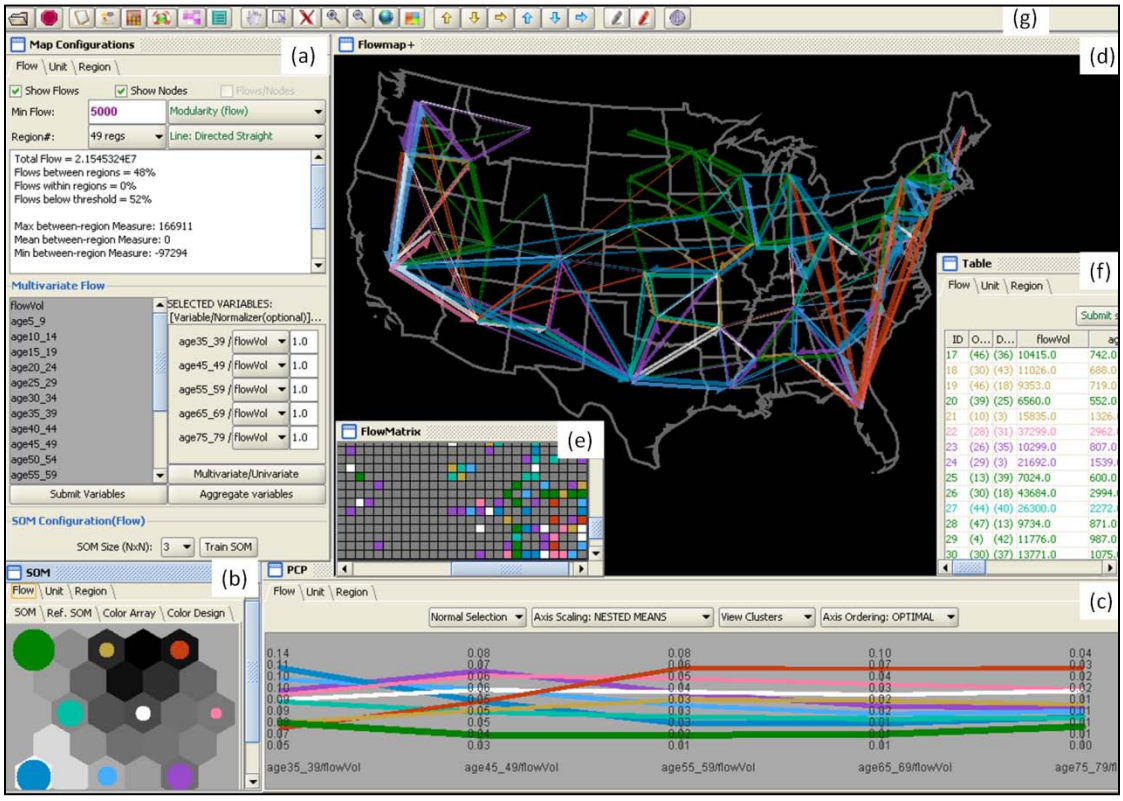


Figure 4.6 The components of the visual analytic system. (a) Configurations of the flow map. (b) Self-organizing map (SOM). (c) A parallel coordinate plot (PCP). (d) The flow map (FlowMap+). (e) A flow matrix. (f) A data table. (g) A menu bar. (a)-(f) are coordinated to represent community structures, spatial structures and multivariate structures.

The configuration panel (Figure 4.6a) allows the user to turn on/off the map, change flow measures, navigate up/down region levels, filter flows, access summary information of the mapped flows, and configure multivariate analyses. The self-

organizing map (SOM) (Figure 4.6b) includes an invisible computational module to conduct SOM clustering, a color assignment module to encode the clustering results, and a visual module to represent the results. The parallel coordinate plot (PCP) (Figure 4.6c) serves as a legend for colors derived by SOM. FlowMap+ (Figure 4.6d) is the center visualization of the system. The flow matrix (Figure 4.6e) offers a supplemental view of the flow structure. The rows and columns in the matrix can be reordered such that contiguous regions are adjacent to each other. The data table (Figure 4.6f) facilitates convenient examinations of numeric flow measures and multivariate information attached to the mapped flows. The menu bar (Figure 4.6g) provides quick access to a collection of functions, such as data loading, zoom in/out, and resizing flow symbols.

Table 4.1 The functions of the components in the visual system

| Component | Input | Function |
|---|---|---|
| configuration | user's selection | customizations of FlowMap+ |
| FlowMap+ | spatial information, region partition, region-level flows, and multivariate analysis results | show spatial structures and community structures of flows, multivariate patterns of flows/units/regions |
| flow matrix | region-level flows, and multivariate analysis results | show connection structure of region-level flows, and multivariate patterns of flows |
| data table | multivariate data of flows/units /SI regions | provide multivariate data of flows/units/regions in numeric forms |
| SOM | multivariate data of flows/units/regions | cluster flows/units/regions based on selected subset of variables |
| PCP | multivariate analysis results of flows/units/regions | explain the multivariate structure of flows/units/regions |
| menu bar | N/A | data loading, symbolic operations, and etc. |

### 4.2.2 FlowMap+

FlowMap+ is a comprehensive and dynamic visualization for SI data analyses developed in this dissertation research. Its efficiency and functions are significantly improved over traditional flow maps. Four types of spatial objects are involved in the visual system: the units/locations in SI data (e.g., counties), flows among original units/locations, SI regions, and the flows among the SI regions. SI regions are generated by the graph partitioning method, which groups the units in the SI data into regions based on flow connections (see Chapter 3). Multivariate information of SI regions is derived from the multivariate information of the units.

With SI regions, FlowMap+ can process and map large SI data. It is also able to visualize multiple data spaces of SI data, including the geographic space (for region-level flows, SI regions, and original units), network space (for unit-level flows), and multivariate space (for region-level flows, SI regions, and original units). Moreover, FlowMap+ is a scalable representation of SI data with a hierarchy of SI regions. Each hierarchical level corresponds to a unique set of SI regions. By changing the region level, one can view flow patterns at different resolutions.

Figure 4.7 shows the configuration options for region-level flows and location (unit) features. For flow configuration (Figure 4.7a), the top half of the control panel allows the user to configure visual symbols, choose flow measures, set a filtering threshold, and choose a SI region level. In additional to existing variables, FlowMap+ also allows dynamic extractions of new variables, e.g., the GINI index and net flow for locations (see Figure 4.7b).

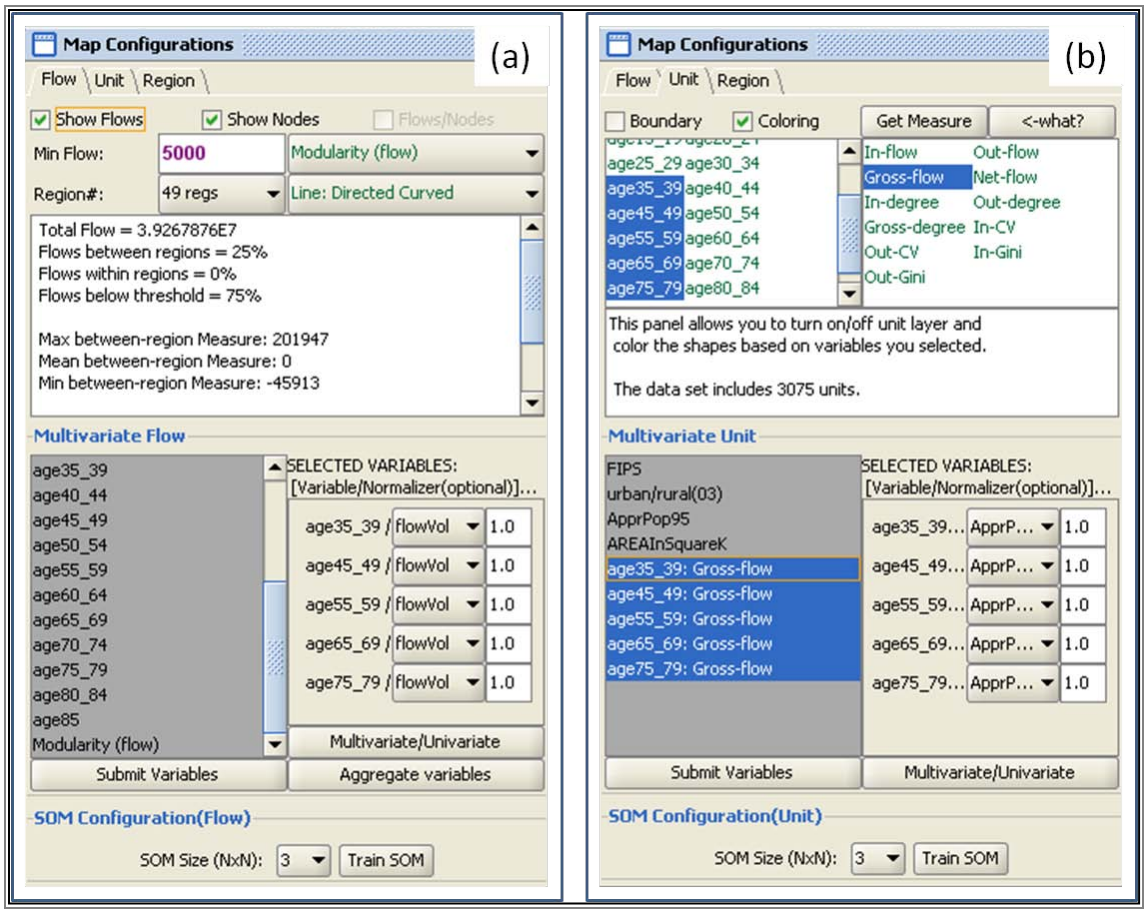


Figure 4.7 The FlowMap+ configuration interface. (a) For flows. (b) For units. The configuration interface for SI regions (not shown) is similar to that for units.

A set of popular flow measures is supported (for region-level flows), as listed in Table 4.2. $F_{AB}$ represents the total flow from region A to region B. $E_{AB}$ represents the expected flow from region A to region B.

Table 4.2 Alternative flow measures provided in the visual system

| Flow Metric | Calculation |
|---|---|
| Raw flow (A, B) | $F_{AB}$ |
| Net flow (A, B) | $F_{AB} - F_{BA}$ |
| Flow efficiency (A, B) | $(F_{AB} - F_{BA})/(F_{AB} + F_{BA})$ |
| Modularity (A, B) | $F_{AB} - E_{BA}$ |

"Raw flow" is the volume of region-level flows aggregated from flows between locations. "Net flow" is the difference of the inflow and the outflow. "Flow efficiency" is the normalized net flow by the gross flow. The "Modularity" is calculated as the difference of the actual flow and the expected flow. There are two different formulas for calculating the expected flow: one is flow based (see Chapter 3) and the other is population based, which replaces the flows of origins and destinations in the formulation of flow-based expectation with the population of origins and destinations. Population-based modularity is available when a population field is specified.

### 4.2.3 Multivariate Analyses

The multivariate analysis provided in this visual system represents a significant enhancement over traditional flow maps. If only one variable is selected, the analysis is univariate and the variable is classified. A set of classification methods are provided (e.g. natural break, equal internal). Otherwise, multivariate clustering is performed with SOM. Univariate classification and multivariate clustering are collectively referred to as multivariate analysis. The multivariate analysis is available for region-level flows, original units, and regions.

#### 4.2.3.1 Procedures of Multivariate Analysis

Figure 4.8 sketches the procedure of a multivariate analysis and shows how it interacts with user's input. It starts with a selection of variables, involves multivariate clustering and visualization, and ends with color assignments of the features (i.e. flows, locations, or regions) that represent their multivariate structures.

First, the selected variables can be transformed (or normalized). The transformation would be needed in many scenarios. For instance, a flow ratio can be obtained by normalizing (dividing) the number of male migrants by the total migrants, which are two variables of region-level flows. Next, transformed or untransformed variables are automatically standardized to Z-scores by subtracting the mean from the observed value and dividing it with the standard deviation. Since variables can be dramatically different in the scale and/or the range, this standardization makes them comparable, which is important to clustering analyses. Then weights can be specified for the standardized variables. If weights are not specified, the variables are treated equal.

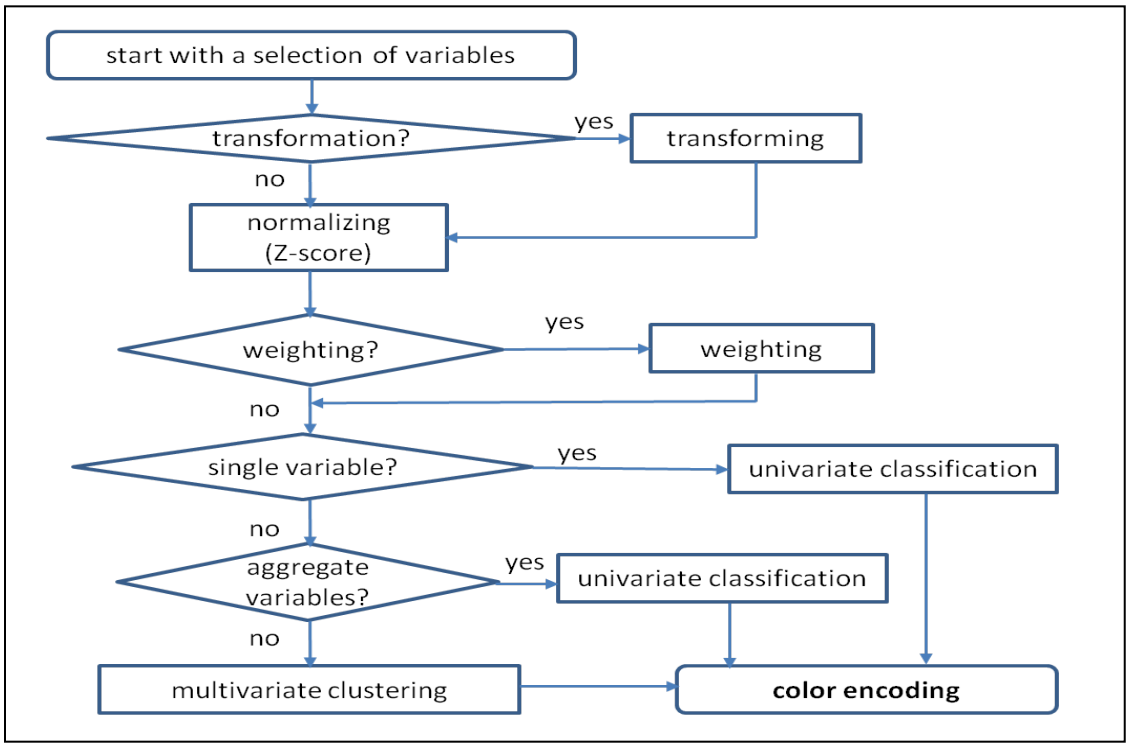


Figure 4.8 Procedures of multivariate analyses

If the user only selects one variable, a univariate classification is conducted. If two or more variables are chosen, the user can choose to aggregate them into a composite

variable. This option would be useful when the variables are addable and the resulting composite is meaningful. For instance, analysts may want to focus on elderly migrants while the migration flows are stratified by age groups of 5-year interval. In this case, analysts may sum several age groups into one to obtain a single variable of the elder migrants. If the user chooses to aggregate variables, the remaining analysis is univariate.

The procedures of multivariate analysis are the same for flows, units, and SI regions. The sole difference lies in the content of the variables. Different sets of variables are provided for them (Table 4.3). Network measures are only available for units and regions. Flow variables and flow meausres are supplied for (region-level) flows. Multivariate region-level flows are aggregated from the variables of original location-level flows. Flow measures are calculated on the fly. Four groups of area-oriented network measures can be derived for units and regions: (1) in-, out-, gross-, and net-flow; (2) in-, out-, and gross-degree; (3) in- and out-CV (Allison 1978); (4) in- and out-GINI (Plane and Mulligan 1997).

Table 4.3 Variable sets provided for flows/units/regions in the visual system

| | Flows | | Units | | Regions | |
|---|---|---|---|---|---|---|
| Variable type | provided | aggregated | provided | aggregated | provided | aggregated |
| Flow variables (input) | Yes | Yes | - | - | - | - |
| Flow measures (derived) | Yes | - | - | - | - | - |
| Network measures (derived) | - | - | Yes | - | Yes | - |

#### 4.2.3.2 Components of Multivariate Analysis

SOM (i.e. Self-Organizing Map) and PCP (i.e. Parallel Coordinate Plot) are coordiated to cluster multivariate flow data and visualize the discovered multivariate

patterns. Figure 4.9 presents an example of SOM and PCP. The SOM in this research bears two major improvements: (1) it is able to visualize a larger volume of data by using nodes to represent clusters with the radius indicating cluster sizes, i.e., the count of units assigned to it; (2) it offers flexible color design (Guo et al. 2005). Similar colors and the proximity of nodes in SOM suggest high similarity between clusters while the cluster sizes are represented by the radii of node.

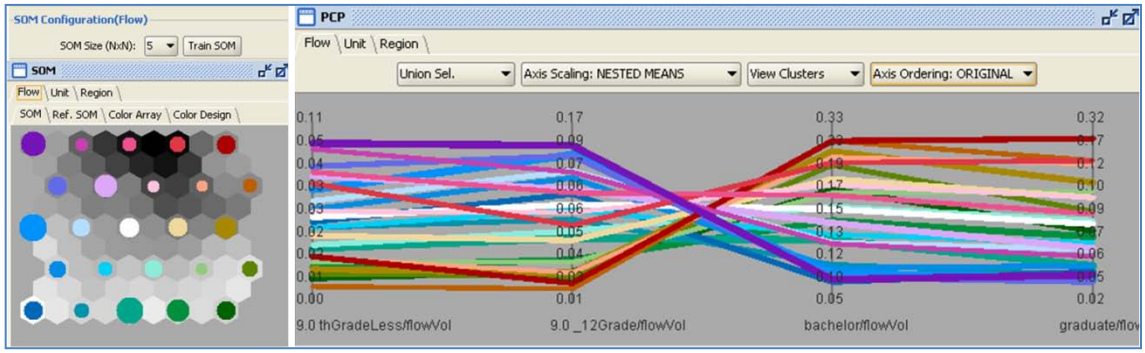


Figure 4.9 SOM and PCP for multivariate analyses

PCP can be viewed as the legend of the SOM and the flow map, from which we can interpret the meaning of clusters and colors derived by SOM. PCP was first introduced in (Inselberg 1985) and has been applied to multidimensional geographic data (Edsall 2003). It is constructed with a series of parallel axes, where data are shown as connected line segments. Each axis represents one variable and each connected set of line segments (i.e. a string) depicts one observation. A high point position where a line segment intersects an axis means a relatively high value of this observation on this variable. PCP suggests the relationship among variables: if line segments between two axes are parallel, that means these two variables are highly correlated. The advantage of PCP over traditional multivariate visual forms (e.g. scatter plot) is that it can visualize the relationship among several variables simultaneously.

The PCP in this visual system is enhanced over original PCPs. First, it provides the option to have the strings represent clusters, with the thickness proportional to the size of clusters or classes. That improves its capability to handle large data sets. Another enhancement is that it accommodates flexible interactions including selection modes, axis scaling, and the ordering of axes (Guo et al. 2006, Guo et al. 2005). The reorderable axes facilitate the discoveries of correlations between variables--"optimal" ordering arranges most correlated variables in adjacency.

### 4.2.4 Integration and Coordination

The reported visual system is implemented in JAVA programming language with component-oriented model. The idea of component-oriented model is to separate concerns in a wide range and place them onto individual components. Visual and computational components in this analytics are assembled into a system to help analysts understand the data and the discovered patterns. These components "talk" with each other in an "event-listener" mechanism which glues all components together. "Event-listeners" mechanism is utilized to manage interactions among the involved components.

Selection and coloring are the major interactions coordinated among the components of this system. Figure 4.10 describes how selection and coloring events are propagated among the five major visual components. SOM is the sole source of color events. Once the configuration of multivariate analysis is altered, it generates a color event (Figure 4.10a) containing the color assignments and sends it to PCP, FlowMap+, and flow matrix. These components respond by updating the features they represent. Selection events are freely broadcasted among the components (Figure 4.10b).

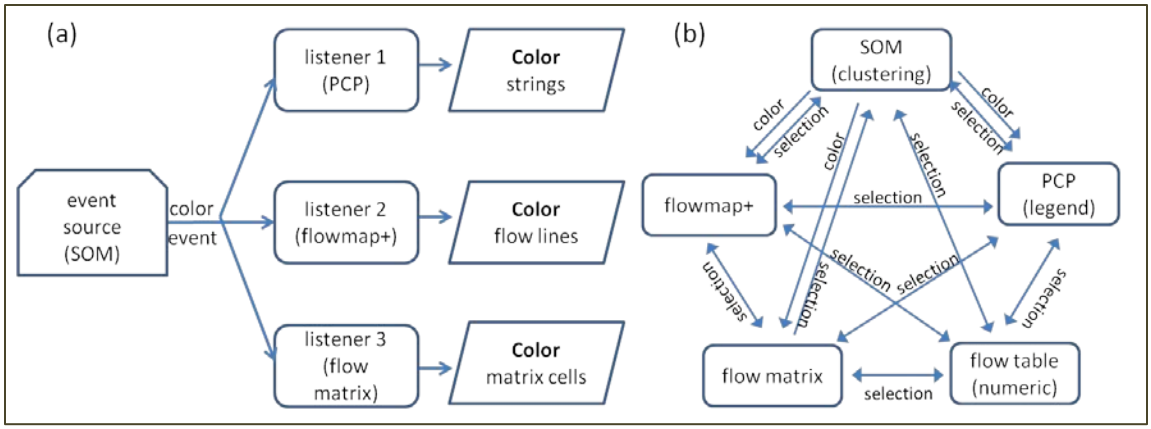


Figure 4.10 The propagation of selection and color events. (a) SOM sends color events to three other visual components which respond by updating the colors of the features they represent. (b) Selection event can be sent and responded by each visual component. Color events are solely sent by SOM.

## 4.3 An Illustration

This section demonstrates the main features of the reported visual analytic system with the 2000 Census county-to-county domestic migration data in the conterminous U.S.

### 4.3.1. SI Regions

This reported visual system takes the SI regions derived with the graph partitioning introduced in Chapter 3 as a data reduction strategy to convert the usually large SI data to a smaller region-level data, enabling legible mapping of large SI data. Compared to existing data aggregation schemes (i.e. state, urban-rural spectrum), SI regions have the advantage that they can at least partially cure data losses, which is an inherent limitation of data aggregation. Normally spatial variations are missing within the scale that data are aggregated at. In this visual system, as users scale along the hierarchy of SI regions, details at different resolutions automatically and gradually unfold.

The map in Figure 4.11 showing net migration rates of states comes from one of the "Census 2000 Special Reports" (Franklin 2003). The state-level net migration rates[6] are aggregated from Census county-to-county domestic migration data as described in Chapter 2. As the natural growth tends to be even and low across the U.S., migration has a significant influence on the redistribution of population and the demographic compositions. This map is created to capture the spatial variation of net migration rates and its impacts on the population redistribution in the U.S. The data aggregation strategy (i.e. state division), the visualization method (i.e. choropleth map), and the flow measure (i.e. net migration rate) used in this report are typical in current migration studies.

Figure 4.12 shows the migration rates of 49 SI regions. It uses the same data and the same flow measure (i.e. migration rates) as Figure 4.11, except that it utilizes SI regions to summarize the county-level migration data. This 49-level of SI regions is chosen to obtain a resolution similar to Figure 4.11. The same breaks for migration rates and similar color scheme are used as in Figure 4.11. Note Figure 4.12 only has 5 classes because of the different range of net migration rate but the 5 classes are defined with the breaks as the 5 lower breaks of Figure 4.11.

On one hand, it is shown that the map for SI regions (Figure 4.12), to a large degree, preserves major patterns identified in the map for states (Figure 4.11). Both maps indicate that the South, especially the Southeast, displays high level of gains. A general loss is found for the Northeast except for the New England states at the corner while New York lost most. The landscape of the West and the Midwest is mixed: Arizona, Oregon, Idaho, Colorado, and Washington gained while Nebraska and Illinois lost populations.

[6] Net migration rate is the (migration) flow difference of two directions divided by the population and multiplied by 1000. Negative and positive rates suggest losses and gains in population, respectively.

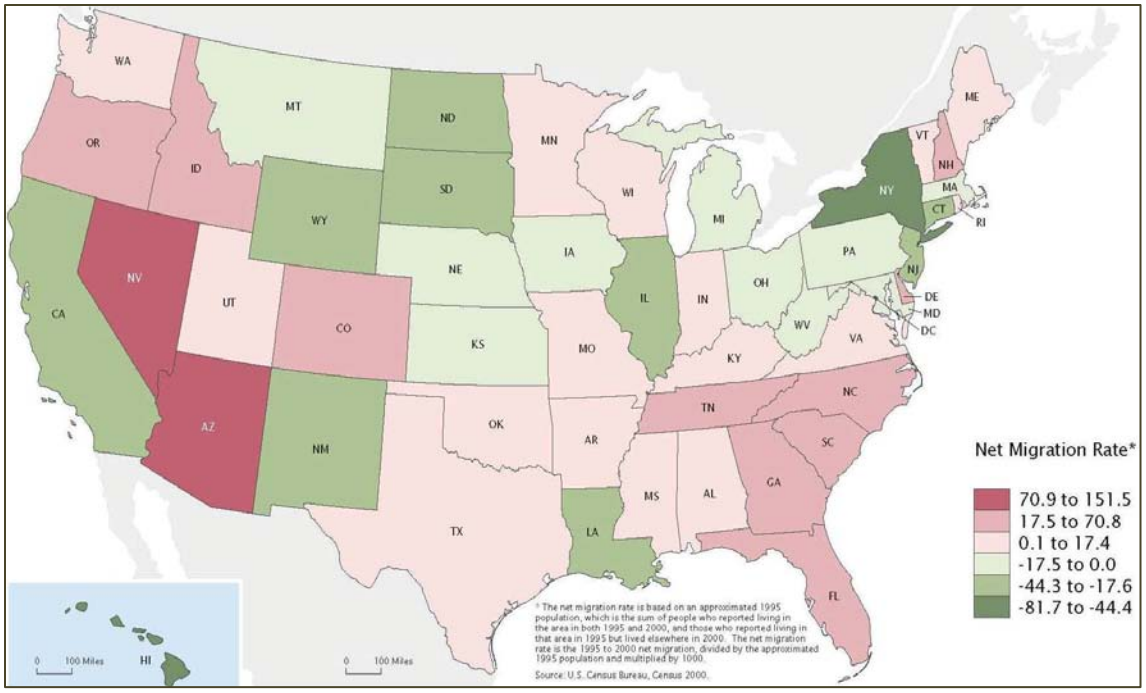


Figure 4.11 Net domestic migration rates of states (1995-2000) (source: (Franklin 2003))

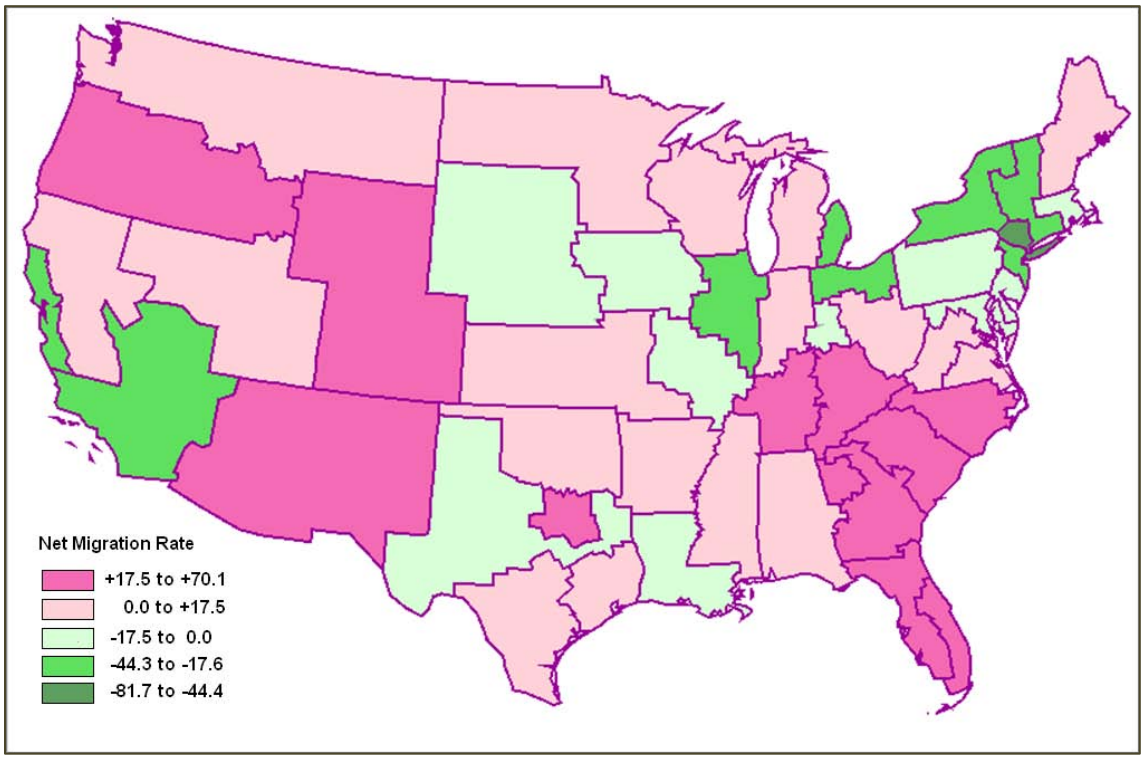


Figure 4.12 Net migration rates of SI regions at the 49-level (1995-2000)

On the other hand, variations between some metropolitan areas and rural areas within the same state are discovered in Figure 4.12. Note that an SI region that covers a large city or urban area is referred to by the city for convenient naming (e.g. Dallas or Detroit), although the SI region may not exactly match the urban area or city boundary. As an example, the Dallas SI region gained considerable population while the less urban area in Texas sent more population than it received. This shows that a map based on SI regions can show spatial variations, which are not evident in Figure 4.11. In particular, the urban-rural variations are meaningful and have been noted in many migration studies.

Moreover, opposite net migration rates are found in a few states in the two maps. This is understandable because these states are part of SI regions encapsulating multiple states. For instance, Montana has a small negative rate in Figure 4.11 but a positive rate Figure 4.12. This is because the SI region that Montana belongs to includes a portion of Washington and Oregon states, both of which gained population, as shown in Figure 4.11.

Figure 4.13 is generated to show the net migration rate of SI regions at the 70-level. Net migration rate is classified into 6 classes with the Jenks Natural Break method, which optimizes (minimizes) within-class variations. More metropolitan areas are identified as individual SI regions (e.g. Chicago and Denver) at this level. Some interesting details are presented. For example, the SI region encompassing Montana now shows a negative rate, as shown by Figure 4.11. At the same time, it becomes obvious that positive rate of this SI region is largely contributed by the positive rate in western Washington (see Figure 4.13).

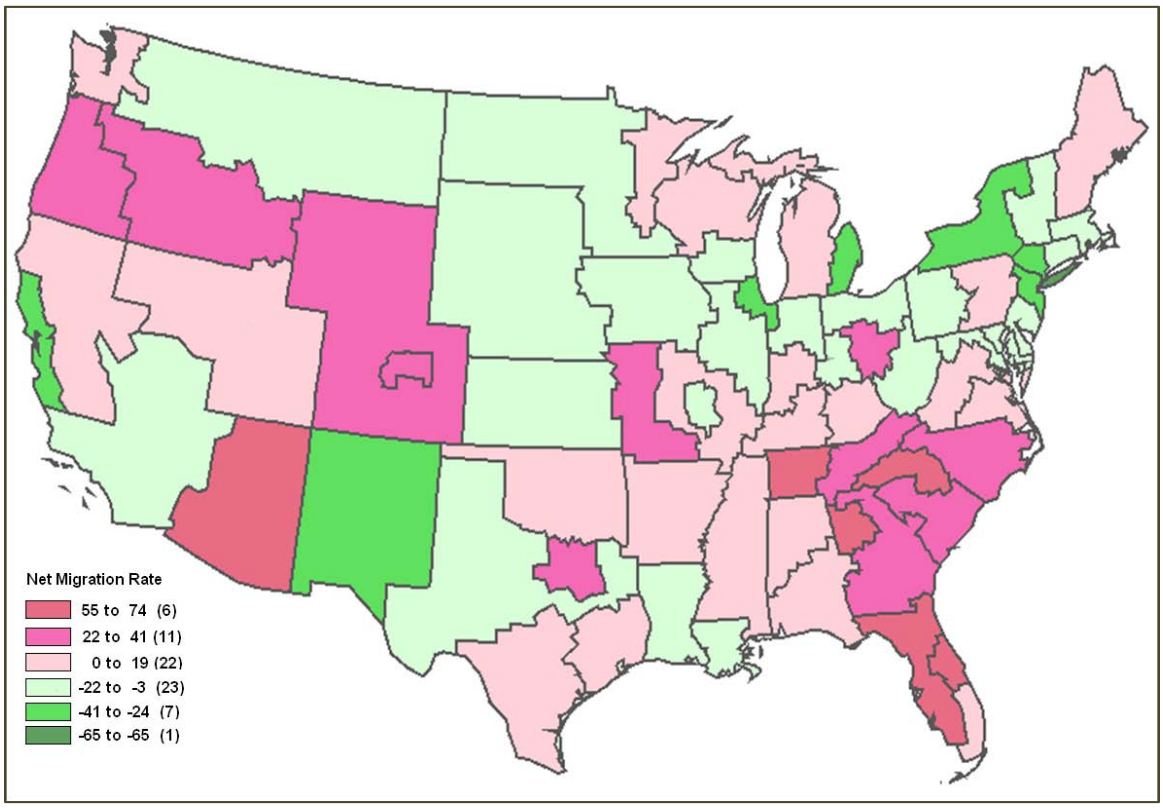


Figure 4.13 Net migration rates of SI regions at the 70-level (1995-2000)

### 4.3.2. FlowMap+

FlowMap+ is the primary visual form of the presented visual system. It integrates the patterns extracted from the three data spaces in SI data, facilitating comprehensive understanding of SI data and enhancing the communications of the understanding.

Figure 4.14 presents a FlowMap+ showing two flow clusters extracted from all above-expectation flows relevant to 49 SI regions. The first flow cluster is dominated by poorly educated ("12$^{th}$ grade or less") migrants and the other one is dominated by well-educated migrants ("bachelor" or "graduate"). In this map, curved flow lines are chosen for better visual clarity. That is, the flow lines are curved at the origin location and become straighter on the destination location. The widths of flow lines are proportional to

the magnitude of the flow measurement. End points of flow lines are the centroids or population weighted centroids. Regions are shaded based upon the population density.

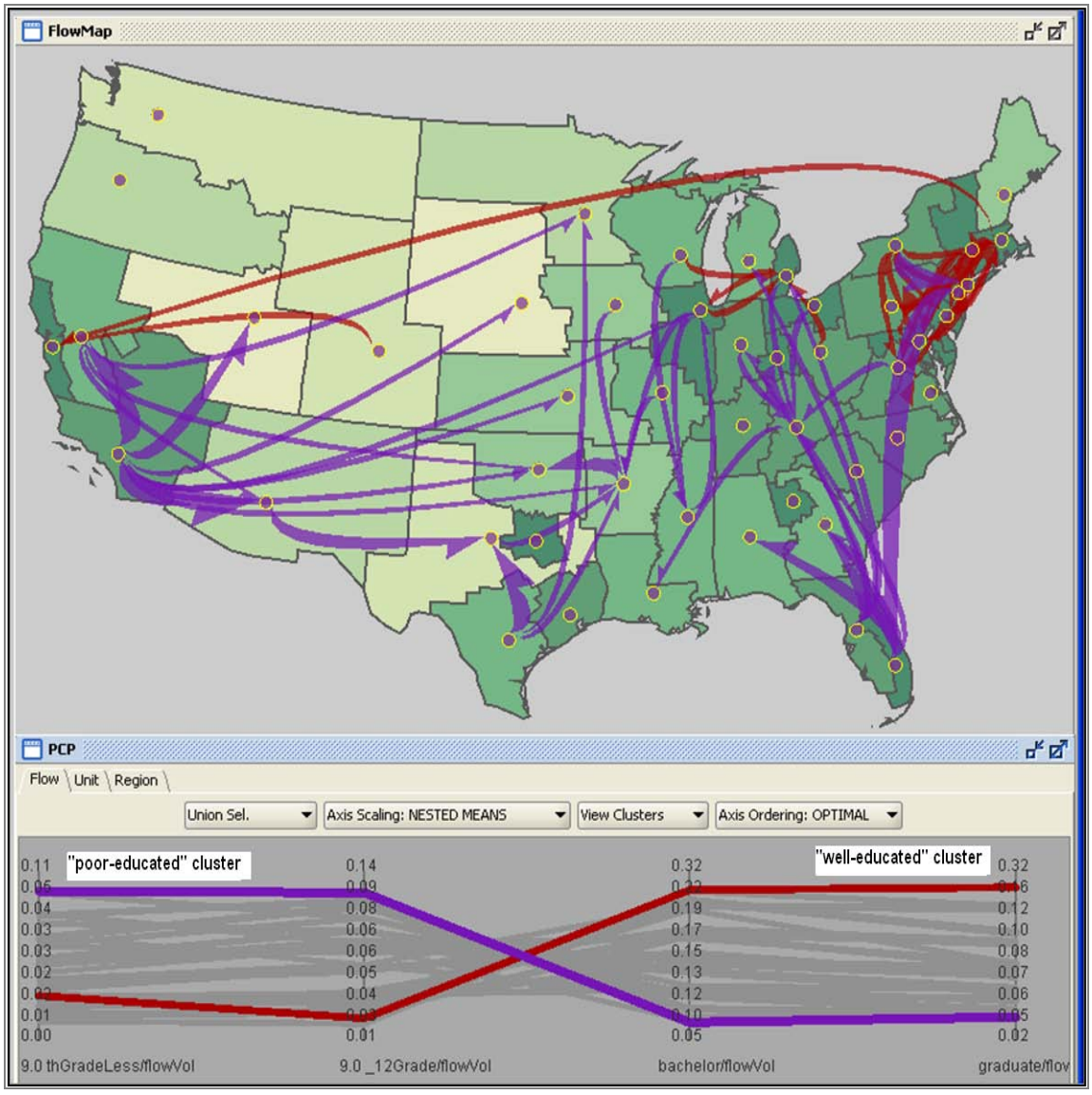


Figure 4.14 Flow visualization with FlowMap+. (1) Line width: magnitude of flow measure. (2) Flow color: univariate or multivariate information of flows. (3) Region boundary: community strctures. (4) Region color: univariate or multivariate information of regions.The filled circles reprsent the (county) population-weigthed centriods of regions. The flow lines are curved at the origin location and become straighter on the destination location.

Figure 4.14 shows that FlowMap+ is able to simultaneously present the multiple data spaces of SI data: (1) the graph space by presenting the boundaries of SI regions; (2) the geographic space by preserving the spatial information of SI regions, which are spatially contiguous clusters of the original units (i.e. counties); (3) the multivariate space through color-encoding of flows and regions. Although not shown here, the multivariate patterns of units can be displayed as well. Figure 4.14 only represents one configuration of FlowMap+. Integrated and coordinated with a set of visual and computational components, FlowMap+ is able to accommodate more creative and in-depth examinations.

### 4.3.3. Multivariate Analyses

The visual system in this research provides a range of options to customize multivariate analyses. There are a number of migration studies focusing on the moving behavior of elderly population, such as (Conway and Houtenville 2001, Walters 2002, Kim 2010). The census data sets used in this research provides the age information of migrants by decomposing each migration flows into seventeen 5-year age groups. These age groups cannot be directly used to study the migration of the elderly since "elderly" population covers a larger age range than 5-year and needs to be considered as a single group. The analysis below provides an example to show that the visual system can solve the conflicts between the data and the research topic by customizable multivariate analyses.

Figure 4.15a presents the age composition of 360 flows with the flow efficiency larger than 20%. To analyze the age structures of flows, the counts of migrants of the five

5-year age groups (age 65-69, 70—74, 75-79, 80-85, and above 85) are first normalized by the volume of flows containing all age groups. The resulting ratios are then summed to obtain a composite ratio, which indicates the portion of elderly people in the migration flows. Thus, a low percentage of this composite indicator means a low percentage of elder migrants and a high percentage of younger migrants and vice versa. Figure 4.15b show the red flow cluster with a high ratio in elderly migrants (65-years or older). Figure 4.15c shows the opposite: the flow cluster high in younger migrants (64-yeasr or younger).

## 4.4 Summary and Discussions

The overall objective of the presented visual system is to map and visualize large and complex SI data to facilitate the exploration, understanding, sharing, and communicating of spatial interaction information. This visual system takes the SI regions derived with the graph partitioning introduced in Chapter 3 as a data reduction strategy to dynamically transform large SI data into smaller region-level data upon the user's choice of scale (or hierarchical level), enabling legible mapping and comprehensive analyses of data with multiple variables and multiple scales.

The flexibility of the visual system includes: (1) freely changing the scale (i.e., number of regions) in the hierarchy of SI regions to interactively obtain flow patterns at different resolutions; (2) highly customizable configurations of multivariate analyses; (3) a myriad of interaction techniques (e.g. selection, filtering) to select and configure the data to map; and (4) multiple flow measures and area-based network measures automatically derived from original data for further analyses.

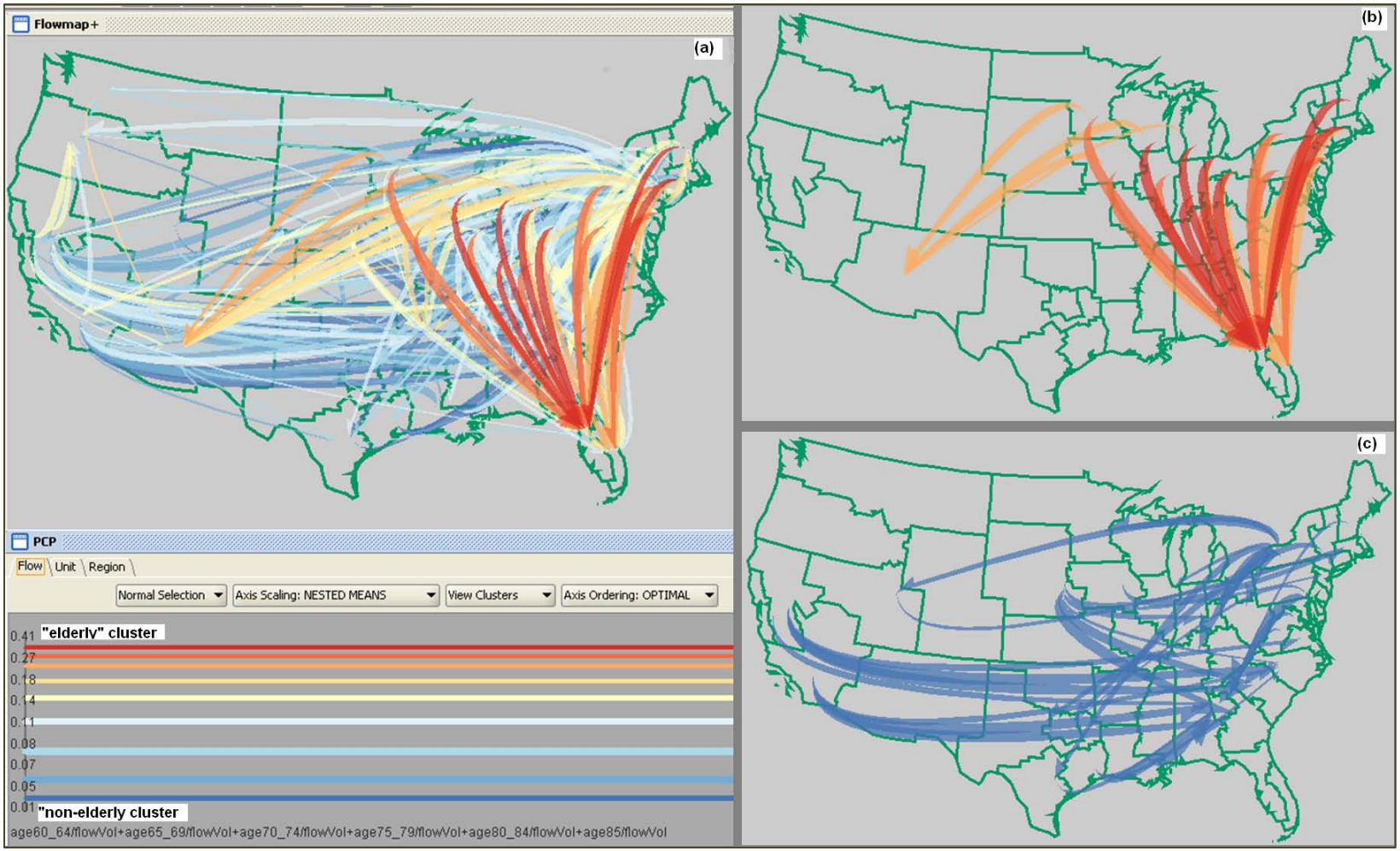


Figure 4.15 Flows with different age compositions for 49 regions. (a) The overall pattern. (b) The flow cluster high in old migrants. (c) The flow cluster high in young migrants. 360 flows with the flow efficiency (i.e. a ratio of net flow to the gross flow) equal to or larger than 20% are chosen for the analysis. Red color is for the flows biased on the elderly while blue color is for flows with a high portion in the younger migrants. The flow lines are curved at the origin location and become straighter on the destination location.

The capability of this visual system primarily comes from its integration of the multiple data spaces in SI data in the data analysis process. Traditional flow maps emphasize the spatial variation of flows can only handle small data sets. On the other hand, current research on graph patterns in SI data (Masser and Brown 1975, Slater 1975) is isolated from the multivariate information and their spatial variations. There is few method that is able to combine the three data spaces in the visualization and analyses of spatial interaction data and information. Using SI regions to aggregate SI data and exploring the aggregated data with a much enhanced flow map (i.e. FlowMap+) can enable a meaningful combination of the three data spaces. FlowMap+ is coordinated and integrated with a set of computational and visual components (i.e. SOM, PCP, data table, flow matrix) to simultaneously represent the patterns extracted from the three spaces of SI data in an easy-to-understand manner.

Future research will further and systematically evaluate the usability of the visual system through a series of designed experiments. The assessment will be task-based and concentrate on users satisfaction and system's effectiveness in accomplishing certain tasks, as suggested by Koua et al. (2006). Currently, this system only supports flow filtering based on a single flow measures. In certain scenarios, researchers may want to focus on flows based on two or more characteristics of flows. Moreover, query functionalities will be a useful in addition to the interaction strategies. Another important expansion of this visual system would incorporate the time dimension and support analyses of time-varying nature of SI information.

# CHAPTER 5

## CASE STUDY: 1995-2000 DOMESTIC MIGRATIONS IN THE U.S.

This chapter examines the presented approach and system from an application perspective. The first component of the approach, graph partitioning, has been evaluated with synthetic data sets from a methodological perspective in Chapter 3. The visual system has been applied to the Census migration data in Chapter 4 to demonstrate its usage and main features. The case study in this chapter focuses on the meaning of SI regions derived with the partitioning methods and the application and interpretation of the visual system.

This chapter starts with a brief review of migration analyses that involve spatial aspects. Then an analysis is conducted to compare the derived SI regions and two widely used aggregation approaches in migration studies: state divisions and county classifications of urban/rural sectors. Next, the visual analytic system is applied to investigate "income migration" (Plane 1999a), which focuses on the flows of incomes associated with migrants or "the capacity to receive income (or lack thereof)" of migrants (Manson and Groop 2000). Income flows have been found influential to the transformation of local economy (Shumway and Otterstrom 2001).

## 5.1 Background: Migration Studies

Geographers, demographers, sociologists, and economist have made voluminous contribution in migration studies. Below is a brief review of selected work in this area.

Spatial distribution and variation of migration have been inspected at various geographical scales. For instance, Bell et al. (2002) compared the migration in Australia and Great Britain using a number of migration measures. Ambinakudige and Parisi (2011) examined the effects of migrations among four county categories and large cities in the U.S. Rogers and Raymer (1998) analyzed the spatial concentration of migration flows at the state level in the U. S. with four indicators. In Frey's studies (1996), distinct patterns have been found on the immigration and domestic internal migration between states and major metropolitan areas. The net domestic migration losses in high-immigration gateway cities have been linked to the net domestic migration gains in nonmetropolitan areas, based on the 1990-1996 data (Frey and Liaw 1998). In another study, Frey (2002b) suggests a classification of states into three regions: "New Sunbelt", "Melting Pot", and "Heartland" based upon the immigration and domestic migration flows.

Among other demographic variables, age is a critical factor in determining the moving behavior of migrants. Research results have shown that migrants' decisions to move and where to move are heavily conditioned on their positions in the life course and the age of household members (Plane and Heins 2003, Plane et al. 2005). Using various visual forms (i.e. arrow graph, tables, and maps), Plane and Jurjevich (2009) found that senior and young migrants exhibit very different preferences in making migration decisions and choosing destinations (i.e., up or down the urban hierarchy). There is also

research focusing on groups of migrants at the same stage of life course and sharing the same motivation to move, such as college freshmen (Alm and Winters 2009) and graduates to enter labor force (Gottlieb and Joseph 2006, Kodrzycki 2001). Some other studies focus on age-specific migration flows are seen in (Rogers et al. 2002, Morrill 1994, Mueser et al. 1988, Longino et al. 1984, Johnson et al. 2005).

In addition to the age, other characteristics of migrants have also been examined in migration studies. As an example, the study on the migration selectivity of various occupations leads to the conclusion that highly educated and skilled workers tend to be attracted more by economic opportunities than less-educated and less-skilled workers (Reisinger 2003). A number of researches are concerned with the moving behavior of migrants by race (Krieg 1993, Frey and Farley 1996), the birth place (i.e. foreign-born and native born) (Frey 2002a, Newbold 1999, Belanger and Rogers 1992), or gender (He and Gober 2003, Faggian et al. 2007, Weber and Munst 2009, He and Pooler 2002).

From the above brief review, we can see that spatial locations and migrant characteristics are two of the most important factors that migration studies have long been investigating. To perform spatial analyses of migrations, it is necessary to choose a suitable spatial unit (such as state, county, or metropolitan area). However, most existing studies use existing high-level administrative units (e.g., states), which do not necessarily reflect that inherent structure in migration flows. One of the contributions in this research is to discover regions based on the graph structure in flows, instead of using existing arbitrary units. Therefore, in the following section, SI regions derived from the data are compared with existing boundaries (such as state) to examine their meaning and effects in migration analyses.

## 5.2 Network-derived SI Regions

Aggregation is typically used in migration analyses to reduce data size and enable visual examination of flows. State divisions and the urban-rural classification are two popular aggregation approaches in U.S. migration studies. These two divisions represent two different strategies: states are spatially contiguous while urban-rural classes are not; states represent an administrative division while urban-rural classes are defined from the socioeconomic perspective.

This research utilizes contiguous SI regions derived from the graph partitioning method to aggregate flow data. In order to bridge SI regions and current practices in migration analyses, the difference and correspondence between the data-driven SI regions and the two popular divisions are examined. If not specified otherwise, SI regions in this chapter refer to the partitioning hierarchy containing 100 continuous levels, derived with the partitioning method developed in this research (ALK initialization) (see Chapter 2) from the county-to-county migration data (see Chapter 3).

SI regions are distinguished from existing divisions by two major advantages. First they capture the community structures in migration connections and thus represent interesting patterns in the graph space, which maximize internal flow connections and minimize external connections among SI regions. Spatial units/locations within a SI region are spatially contiguous and tend to have stronger interactions with each other than with units from the outside. Second, SI regions are automatically derived from the data and form a hierarchy that represents natural regions embedded in the data at different scales. Therefore one can flexibly derive unique regions for different data sets (thus do

not need to use the same predefined boundaries such as states) and can examine patterns at different scales (instead of using the same predefined partition).

### 5.2.1 SI Regions vs. State Divisions in Migration Analyses

States are often used to aggregate large migration data, such as county-to-county migration, to reduce the data volume and complexity (Perry 2003, Ashby 2007, Slater 1976b, Rebhun and Raveh 2006). The wide use of state divisions is partially due to the fact that state is the primary administrative subdivision and the number of states is relatively small. However, predefined boundaries (such as states) often vary dramatically in size. More importantly, they do not necessarily reflect the true partitions manifested by the flows. Therefore, using administrative boundaries to aggregate migration flows may hide true patterns.

To investigate how the SI regions are different from and correlated with the U.S. states divisions, SI region boundaries are overlayed on top of the state boundaries in Figure 5.1. The mapped SI regions are obtained with the partitioning method with ALK initialization (see Chapter 3) from the county-level migration data (see Chapter 2). For better compatibility, 49 regions are used, which is the same as the number of states (including Washington DC) in the study area (i.e. conterminous U.S.). In Figure 5.1, blue lines represent state boundaries while red lines represent SI region boundary. The lines become purple when the region boundaries match (overlap with) the state boundaries. County boundaries are drawn in light gray.

SI regions in general are more balanced in size (in terms of population), where large states (such as California and Texas) are divided into several SI regions while in the

less populated Midwest an SI region may contain multiple states. It is also interesting that SI regions match state boundaries to a considerable extent. In other words, state boundaries are discovered from the migration data by the graph partitioning method.

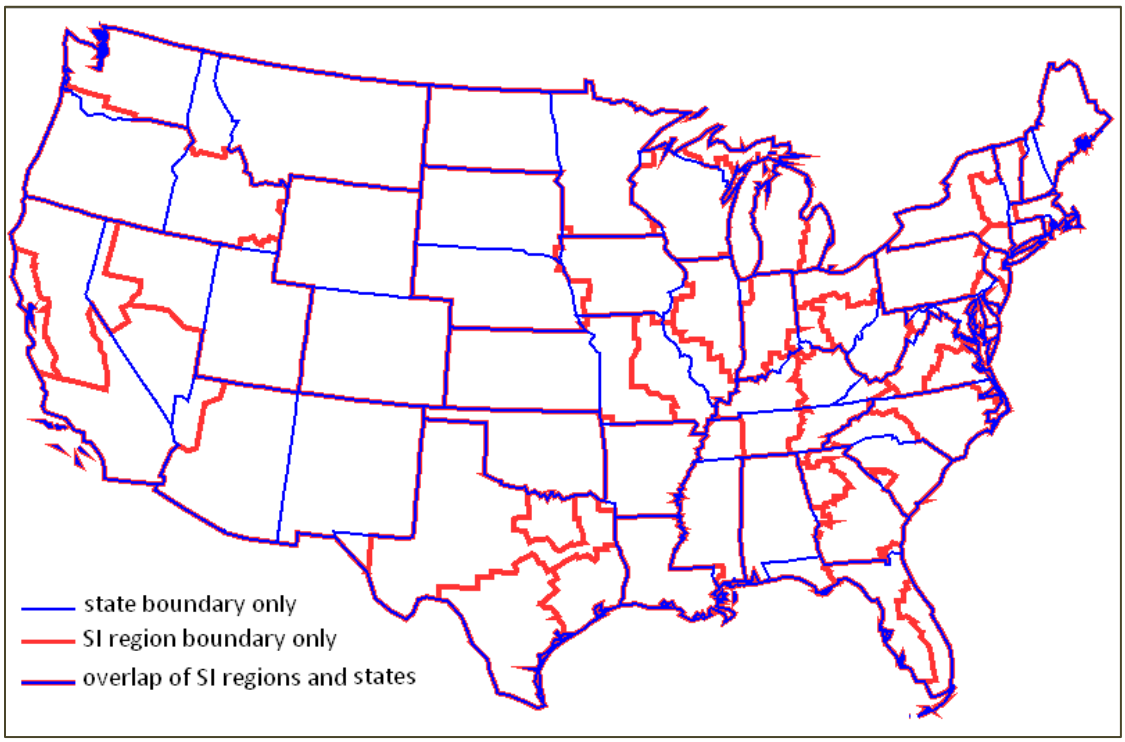


Figure 5.1 Comparison of 49 SI regions and state boundaries of the conterminous U.S.

To better understand the agreement and difference between SI regions and state boundaries, states were roughly grouped into four categories according to their matching degrees with SI regions (see Table 5.1). Below each category is examined and interpreted to understand why the SI regions indeed capture the inherent graph structure in migration connections. In the "Excellent" category, there are 9 states that almost perfectly match their corresponding SI region (i.e., each state is recognized as a community in terms of migration connections). This is a very interesting result since the graph partitioning method "found" these states is solely based on county-to-county migration connections.

On one hand, it shows that the partitioning method works well since it discovers real-world structures without knowing them. On the other hand, it is interesting to learn that administrative boundaries such as states do have significant impact on migration choices.

Table 5.1 Matching degrees of state divisions and SI regions at the 49-level

| Matching code | States | Count |
|---|---|---|
| 1: excellent match (single state-single region) | Wisconsin, Oklahoma, Louisiana, Mississippi, Alabama, Iowa, Arkansas, Indiana, Pennsylvania | 9 states (9 regions) |
| 2: good match (single state-multiple regions) | Texas (4 regions), Georgia (2 regions), Florida (2 regions), Michigan (2 regions), Virginia (2 regions) | 5 states (12 regions) |
| 3: good match (multiple states-single region) | North Dakota and Minnesota, South Dakota and Nebraska, Wyoming and Colorado, Washington and Montana , Idaho and Oregon, Arizona and New Mexico, Rhode Island and Massachusetts, Maine and New Hampshire, District of Colombia (DC) and Maryland | 18 states (9 regions) |
| 4: mixed (multiple states-multiple regions) | Delaware and New Jersey (2 regions), Kansas, Missouri and Illinois (3 regions), New York , Connecticut, and Vermont (3 regions), South Carolina and North Carolina (2 regions), California, Nevada and Utah (4 regions), Tennessee and Kentucky (2 regions), West Virginia and Ohio (3 regions) | 17 states (19 regions) |

The second category has 5 states, which are divided into 12 SI regions, meaning that a single state is divided into 2 or more SI regions. For example, Texas is detected as a single SI region at a higher level but at the 49-region level it is divided into 4 regions, including three metropolitan areas (i.e. Dallas, Houston, and San Antonio) and the rest of Texas. This shows that the graph partitioning method can detect the inherent hierarchical structure of spatial interactions at different scales.

The third category has 18 states, which are grouped into 9 SI regions. These states are primarily small states. Together with the second category introduced above, this

shows that optimizing the modularity measure tends to produce regions of balanced size (population), with larger states divided into more regions while small states merged into larger regions. This is a desirable feature in mapping migration flows as it can alleviate the size impact and help discover true patterns.

The states and SI regions in the last category do not match well: state boundaries intersect with region boundaries. These states are mostly in areas where major cities and their extended urban areas are located on or near state boundaries, such as Delaware and New Jersey (with Philadelphia and New York City on their borders), Kansas, Missouri and Illinois (with Kansas City and Saint Louis cross the state boundaries), New York, Connecticut, and Vermont (with New York city on state borders), South Carolina and North Carolina (with Charlotte on the border), California, Nevada and Utah (with Las Vegas on the borders), and so on. It is understandable why the migration patterns in these areas do not respect state boundaries, because of the migration flows between these large cities tend to be intensive.

The above observations are for the 49-SI region level (derived from the county-to-county migration data). Later in this chapter, Figure 5.2 shows 70 regions, where more metropolitan areas emerge and some small states become regions. For example, Arizona and New Mexico are in the same SI region at the 49-level and split into two regions at the 70-level following their state boundary.

.

### 5.2.2 SI Regions vs. Urban-Rural Classes in Migration Analyses

Urban-rural and urban-urban migrations have been a prominent dimension in domestic migration analyses. There are extensive research efforts examining the

consequences of and the relationship between migrations across urban and rural areas (Morrill 2006, Fuguitt 1985, Fulton et al. 1997, Long and Deare 1988, Rayer and Brown 2001, Renkow and Hoover 2000, Gottlieb 2006, Slifkin et al. 2004, Wilson 1987, Kephart 1988). For example, between 1995 and 2000, migration into nonmetropolitan areas (6.17 million) is greater than migration to metropolitan areas (5.66 million), resulting in a net gain of 0.51 million in nonmetropolitan areas (Schachter et al. 2003).

Two commonly used urban-rural classification schemes are provided by the Office of Management and Budget (OMB) and Public Use Microdata Series (PUMS). As a county classification scheme, the OMB standard assigns an urban-rural code to each county based upon a series of economic and social criteria. Specifically, the urban-rural scheme released by OMB in 2003 classifies all counties and county equivalents into two categories: metro and non-metro, based on the population and the percentage of commuting workforce according to the 2000 Census data. The metro category is further divided into three classes based upon the population magnitude of the Metropolitan Statistical Area (MSA) that the county belongs to. Each MSA must contain counties covering a core urban area with 500,000 or more population and neighboring counties with a high degree of social economic attachment to the core urban area. The non-metro category is further divided into six classes based upon the aggregated urban population, adjacency and functional connections to the metro area. Thus, 9 urban-rural classes are defined. In Figure 5.2, counties of the three metro classes are in pink shades. Spatially contiguous metro counties form a densely populated area centered on a core city.

To compare the SI regions with the metropolitan areas, the boundaries of 70 SI regions are shown in Figure 5.2 in blue. Note that some SI regions are conveniently

named by the major city that it covers. It is obvious that the SI regions successfully recognize major metropolitan areas, such as San Francisco, Phoenix, Denver, Chicago, Dallas, Houston, New Orleans, Saint Louis, Minneapolis, Detroit, Atlanta, and New York City. The match of major metro areas with SI regions (within states) is meaningful. OMB uses the "percentage of the work-force commuting to the central city" to enforce the "core city-suburban" association in the designation of metro counties. The agreement of SI regions with OMB standard indicates that the SI regions can capture the strong "core-suburban" integration from a network perspective.

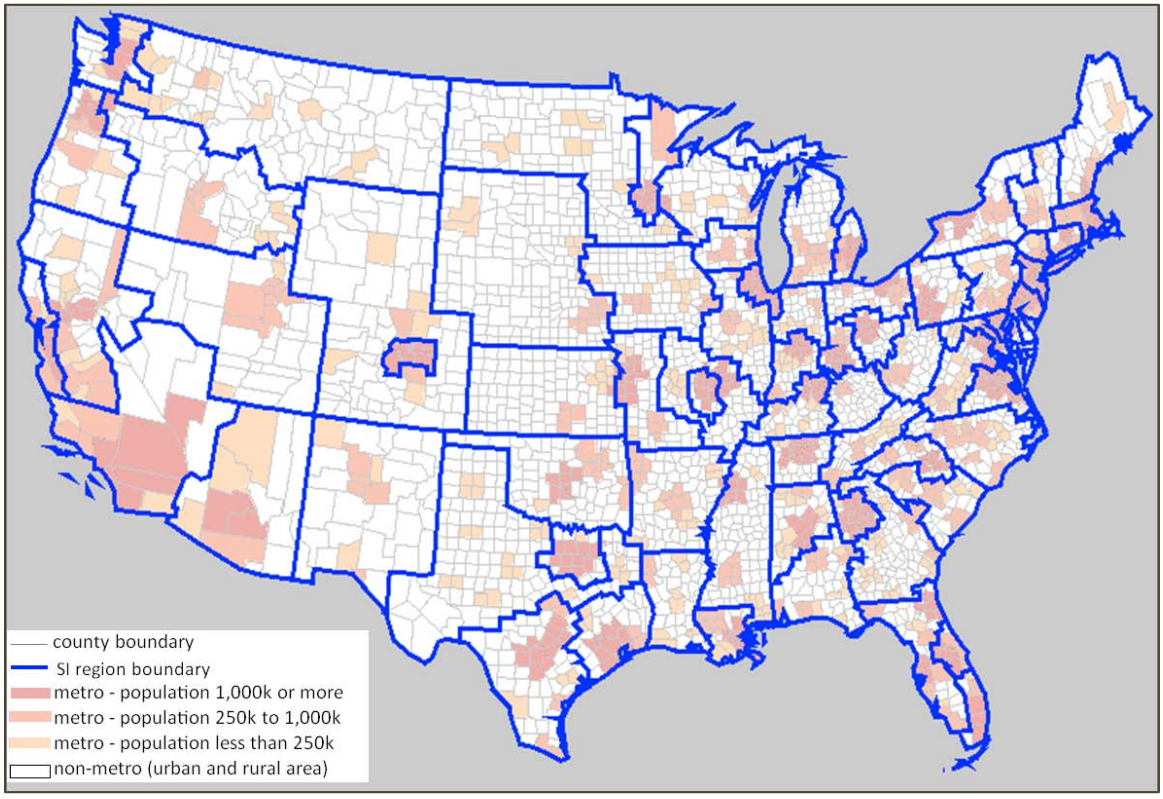


Figure 5.2 70-level SI regions and the metro counties designated by OMB

Based on the comparison and interpretation of SI regions, it is clear that the graph partitioning method can capture the underlying spatial structures embedded in county-to-county migration data at different scales such as states and metropolitan areas. This is a

significant improvement over existing methods and provides valuable insights on the spatial aspect of network structures in migration connections. Next I will examine how the SI regions can better facilitate the mapping and understanding of migration patterns than with states as used in existing research.

### 5.2.3 Region-based Migration Visualization and Analysis

In order to determine the different effects that state divisions and SI regions may have on flow visualization, county-level flows are aggregated by states and 49 SI regions respectively, which are then mapped in Figure 5.3 and Figure 5.4. Flow-based modularity is chosen as the flow measure, which is the difference between the actual flow and the expected flow. The threshold of modularity is set at zero so that the maps show all above-expectation flows, where the expectation is flow-based (see Figure 3.3). The county population-weighted centroids are indicated by filled small circles in each region (the same for the remaining figures in this chapter).

There are two notable differences between the flow patterns shown in the two maps. First, some strong flows in the state map are "missing" in the SI region map, such as the flow between California and Texas. This is because the state map tends to highlight flows of large states due to the dramatic difference in population among the states, although the modularity flow measure has partially alleviated the size effect. As observed earlier, SI regions are more balanced in population. Thus the flow patterns presented in Figure 5.4 can better avoid the size bias. Second, SI regions reveal more details about flows within populated states (such California and Texas) while suppressing the details among small states in the Mid-West.

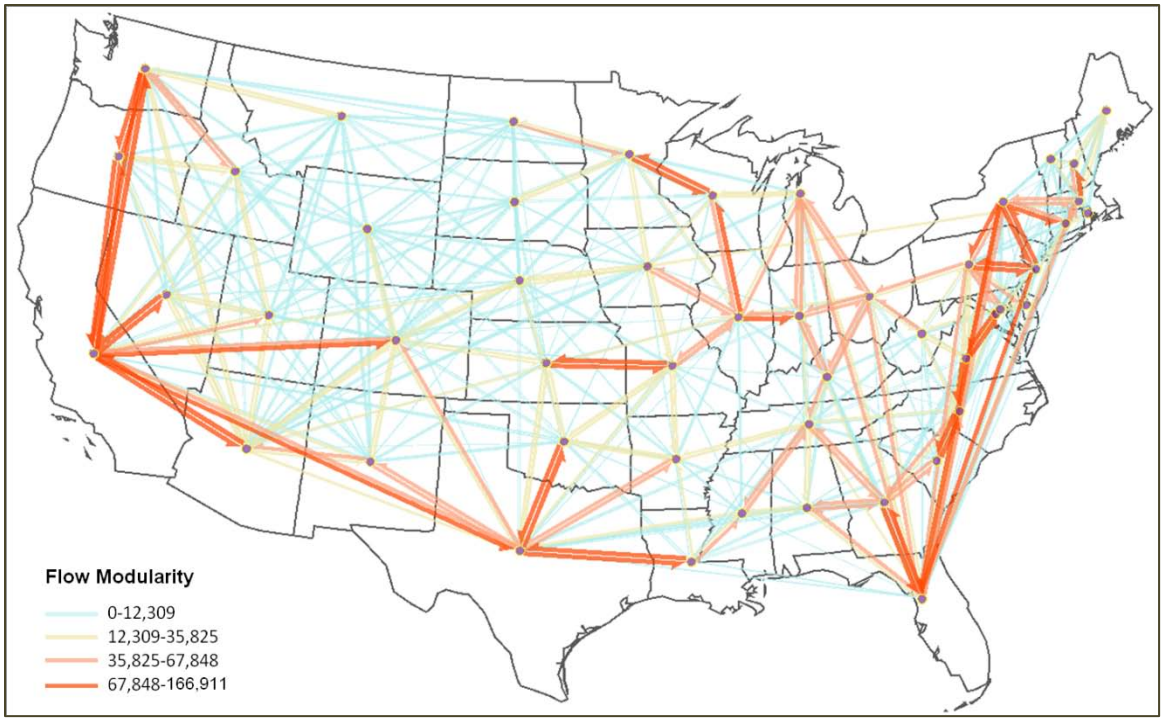


Figure 5.4 Flow modularity among states. Above-expectation flows are shown. Circles represent (county) population weighted centroids.

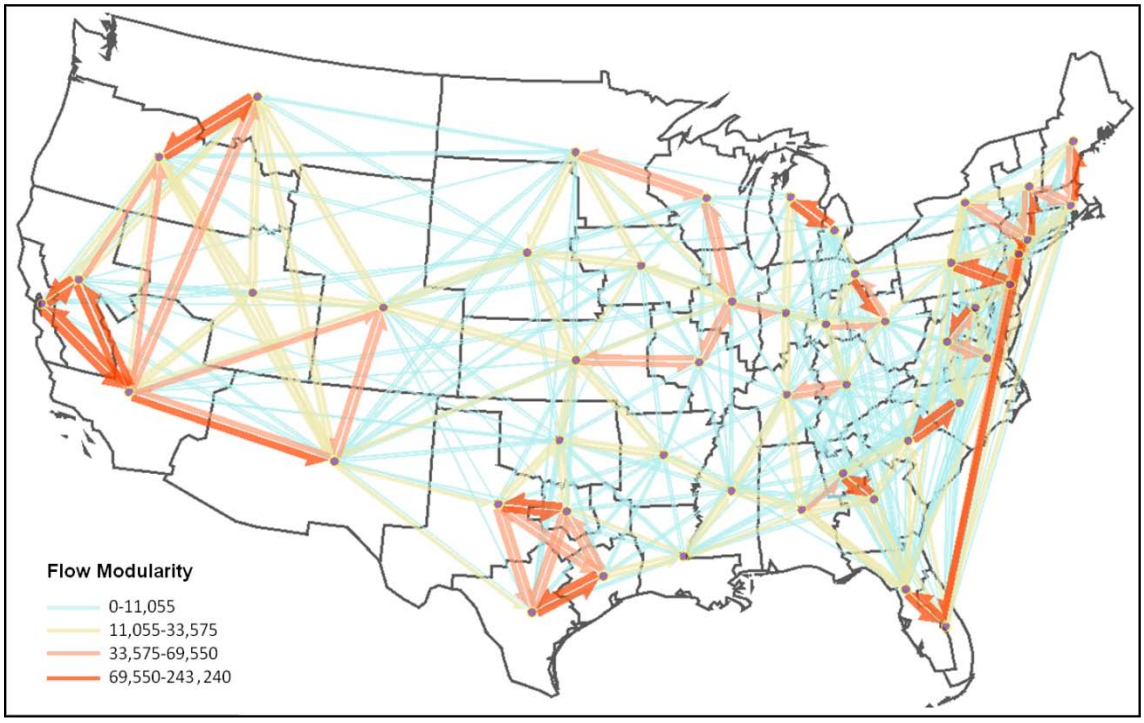


Figure 5.5 Flow modularity among SI regions (49-level). Above-expectation flows are shown. Circles represent (county) population weighted centroids.

One of the most important advantages of using SI regions is that we can change the resolution of the flow map on the map by changing the number of regions (i.e., choosing a hierarchical level). For example, one can choose fewer regions to see more general migration patterns such as migration from the West to the East and from the Mid-West to the South-East. One may also use more regions (e.g., 70 regions) to examine flows among metropolitan areas.

As the state map shows (Figure 5.5a), Florida receives a significant amount of migration from the Northeast and the North (e.g., Michigan and Ohio). The flow map based on SI regions (Figure 5.5b-c) confirms this pattern but also reveals the difference between the east and the west of Florida. The East Florida region mainly connected to the Northeast (Figure 5.5b) while the West Florida region is more connected to the Midwest (Figure 5.5c). Furthermore, the migration from New York to East Florida is more significant than to West Florida (Figure 5.5b), in terms of the flow modularity.

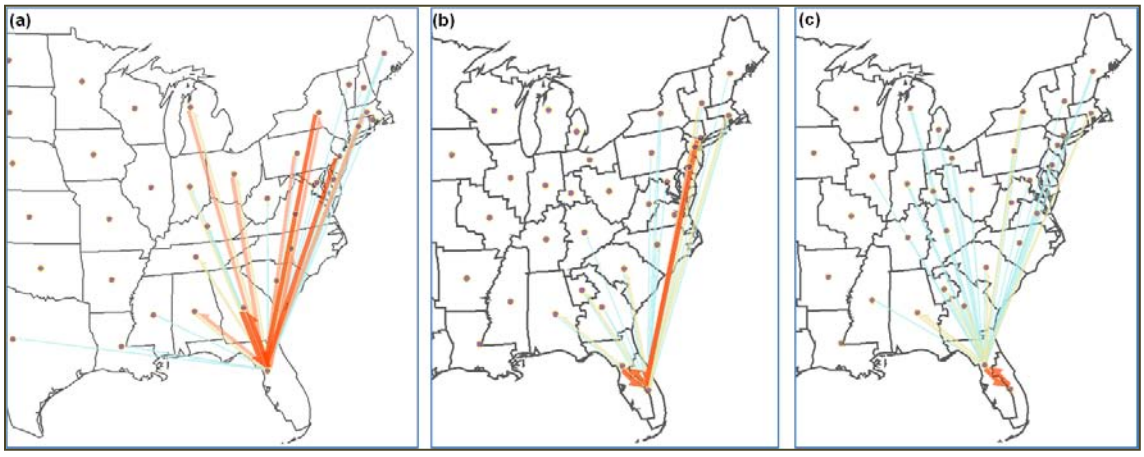


Figure 5.6 Migration flows (modularity) to Florida. (a) Flows to Florida from other states. (b) Flows to the east of Florida from other SI regions. (c) Flows to the west of Florida. Colore represent the modularity (see the legend of Figure 5.4 for the precise meaning).

## 5.3 Income Migration in the U.S.

In this section, the developed visual analytics system is applied to analyze the geography of "income migration" (Plane 1999a). There are two reasons for choosing this topic. First, income migration represents a new direction and current concern of migration analyses. Second, analyses of income migration are based on the income information of migrants and thus require an examination of the multivariate patterns (i.e. income stratifications) of location-to-location flows, which match the capability of the developed system. This provides an opportunity to assess the functionality of multivariate analysis supported in the system.

### 5.3.1 The Context: Income Migration

For a long time, little attention has been paid to the consequences of migration (Greenwood 1975, Nelson 2005). Since the 1980's, there has been a growing interest on the flows of income associated with migrants. To distinguish from population migration, they are referred to as "income flow" or "income migration" (Plane 1999a). There exists a variety of approaches to studying income migration, in terms of the aggregation approaches and the mapping methods. Table 5.2 presents the enumeration unit, the aggregation scheme, the study area, and the main findings of some income migration studies between 2000 and 2010.

Shumway and Otterstrom (2001) aggregate income migration by "county groups", which are identified with a clustering method based upon a set of criteria (e.g. economic types, natural amenity index, and recreation measures. The income gains received by the

"New West" cluster helped its transformation from rural economy to "*one based on preservation of environmental amenities*" (Shumway and Otterstrom 2001: 492).

Table 5.2 A comparison of selected income studies

| Study | Enumeration unit | Aggregation scheme | Study area | Main findings |
|---|---|---|---|---|
| Shumway and Otterstrom (2001) | county | clustering | rural area in the West | The "New West" cluster gained most income via migration. |
| Nelson (2005) | Public Use Microdata Areas unit | urban-rural | nationwide | 1. Sunbelt and Rocky states gained nonearning income while the Plains, Great Lakes, Mideast and New England lost. 2. Nonearning incomes shifted from metropolitan to nonmetropolitan areas. |
| Manson and Groop (2000) | County and some large cities | urban-rural | nationwide | Migrants and incomes tend to flow from central cities to suburbs and from suburbs to rural counties. |
| Ambinakudige and Parisi (2011) | county | urban-rural | nationwide | Large metropolitan counties and rural counties export migrants and incomes to small metropolitan counties and counties adjacent to metropolitan counties. |
| Shumway and Otterstrom (2010) | city | urban-rural | nationwide | Income flows tend to be from the expensive housing areas to the cheaper areas (e.g. the rural areas). |

The nonearning-income migrations of the metropolitan sectors in the 9 Bureau of Economic Analysis (BEA) regions are compared with the non-metropolitan counterparts in Nelson (2005). It is found that over the three five-year periods (i.e. 1975-1980, 1985-1990, and 1995-2000), Sunbelt (i.e. Far West, Southwest, Southeast) and Rocky states gained nonearning income while the Plains, Great Lakes, Mideast and New England lost. Further, it is revealed that nonearning incomes shifted from metropolitan to nonmetropolitan areas over two periods of time (i.e. 1985-1990 and 1995-2000). Manson

and Groop (2000) use a 6-category urban-rural classification scheme similar to the one proposed in (Butler et al. 1994) and study the income migrations of counties and some large cities in 1994-1995. They find that migrants and incomes tend to flow from central cities to suburbs and from suburbs to rural counties. A similar result is obtained by Ambinakudige and Parisi (2011) from the 2006-2007 migration data. Shumway and Otterstrom (2010) find that the most effective income flows are from the expensive housing areas to the cheaper areas, especially between three high-price areas (i.e. Bay area in California, suburbs in California and New York City, coastal mega metropolitan) and the rural cluster.

A general finding of these studies is that the central cities or populated urban areas tend to lose income to suburban or rural areas. This finding counts on an urban-rural classification scheme or a cluster approach to aggregate data. Such aggregation strategies do not consider the spatial information (such as contiguity or distance) and therefore the aggregated area is usually not spatially contiguous. It is then difficult to visualize and comprehend specific origin-destination flows between these areas. For instance, Nelson's inferences (2005) on flows between metro and non-metro areas rely on the metro-nonmetro dichotomy. Manson and Groop (2000) map the in-migration and out-migration respectively for a few big cities. Similar strategy is also seen in (Henrie and Plane 2008). Therefore, existing methods are limited to flows attached or aggregated to one place and are unable to analyze the overall flow connections (i.e., connections among places).

An area-based measurement of income flows, "income effectiveness" (Plane 1999a), is often adopted to measure income flows to/from places (Manson and Groop

2000). "Income effectiveness" is a variant of migration effectiveness (or migration efficiency) as used in (Podolák 1995). It is calculated as the ratio of the net aggregate income to the gross aggregate income: $E = \frac{Y_i - Y_o}{Y_i + Y_o}$. The net income is the difference between the aggregate income of in-migrants ($Y_i$) and that of out-migrants ($Y_o$), while the gross income is the sum of the aggregate income in both directions. Income effectiveness can be considered as a normalized index of "net income migration".

Using the relative term (i.e. income effectiveness) along with the absolute term (i.e. net aggregate income), Plane (1999a) studies the state-level income migration between 1993-1994. The top five gaining states are identified including Nevada (37%), Arizona (31%), Idaho (27%), Florida (26%), North Carolina (22%), and Colorado (21%). The top five losing states include California (-34%), New York (-32%), and District of Columbia (-21%), Rhode Island (-16%), and Illinois (-16%).

### 5.3.2 Income Migration Analysis with the Visual Analytics System

This analysis starts with a primary inquiry of an existing research: the spatial differentials of area-based income effectiveness (Plane 1999a). Different from existing research, the research reported here has two contributions. First, it goes beyond area-based analysis and analyzes origin-destination income flows. Second, it incorporates multivariate income composition of flows to examine the income redistribution through migration.

The Census migration data provides income information for each origin-destination pair, which is summarized from personal income information collected for migrants over 16-years old in year 1999. An 11-level income stratification is provided for

each non-zero migration flow: 0K, 0-5K, 5-10K, 10-15K, 15-20K, 20-25K, 25-35K, 35-50K, 50-75K, 75-100K, and 100K+. In other words, for each origin-destination pair, we have the number of migrants falling in each income range.

In order to obtain the income flow for each origin-destination pair, the migration volume of each income level is first weighted by the income level and then summed to obtain its income flow volume, based on the definition of income effectiveness (i.e. the ratio of the net aggregate income to the gross aggregate income). The income weight is the average of the two ends of an income range, with two exceptions: 0 for level "0k" and 100K for level "100k+". For instance, suppose there are 30 migrants moving from place *A* to place *B*, out of which 20 migrants earn an income of "5K to 10K" and the remaining 10 earn an income of "25K to 35K". The income flow from place *A* to place *B* would be: 20*(5K+10K) /2+10*(25K+35K)/2. The income flow from *B* to *A* can be calculated in the same manner. Dividing the difference (i.e., net income flow) by the sum of the two direction (i.e., gross income flow) leads to a flow-oriented "income effectiveness" measurement (denoted by "*e*"). An area-oriented "income effectiveness" (denoted by "*E*") can be obtained by first aggregating the net income flows and the gross income flows respectively for a place and then dividing the total net flow with the total gross income flow for the place.

Nine income sources are included in the migration data: (1) wages, salary, commissions, bonuses, or tips from all jobs; (2) self-employment income from own non-farm businesses or farm businesses, including proprietorships and partnership (net income after business expenses); (3) Interest, dividends, net rental income, royalty income, or income from estates and trusts; (4) Social Security or Railroad Retirement ;

(5) Supplemental Security Income (SSI); (6) any public assistance or welfare payments from the state or local welfare office; (7) retirement, survivor, or disability pensions (excluding Social Security); (8) any other sources of income received regularly and (9) unemployment compensation, child support, or alimony.

In the analyses below, both "income flow effectiveness" (including both area-based and flow-based) and the absolute net income flow are used. The analyses are performed with the set of 70-regions derived with the county-to-county migration data (see Chapter 2).

### 5.3.3 Spatial Variations of Area-based Income Effectiveness

To compare with existing research and findings, states (used in existing research) and 70 SI regions are used to aggregate the same income migration data (i.e. 2000 Census county-to-county migration data). Income effectiveness is calculated accordingly for each state and SI region. In both maps (Figure 5.6 and Figure 5.7), income effectiveness is classified into 9 groups with the Jenks natural break method. Blue color represents large income losses, red represents high income gains, and yellows represents moderate gains or losses. Despite that different data sources are used, the overall pattern with states in Figure 5.6 is similar to existing studies (Plane 1999a, Nelson 2005). That is: states with high income effectiveness clusters in the Southeast and the West; states experiencing income losses are mostly found in the Northeast, the Great Lakes, and California.

The top tier receiving the highest income inflows ($E \geq 20\%$) includes Nevada, Arizona, and three Southeastern states (i.e. North Carolina, Georgia, and Florida). While Nevada shows the highest positive effectiveness (37%), its absolute gain ($5.3 billion) is

less than Florida ($15.0 billion), Arizona ($7.8 billion), Georgia ($7.7 billion), and North Carolina ($6.7 billion).

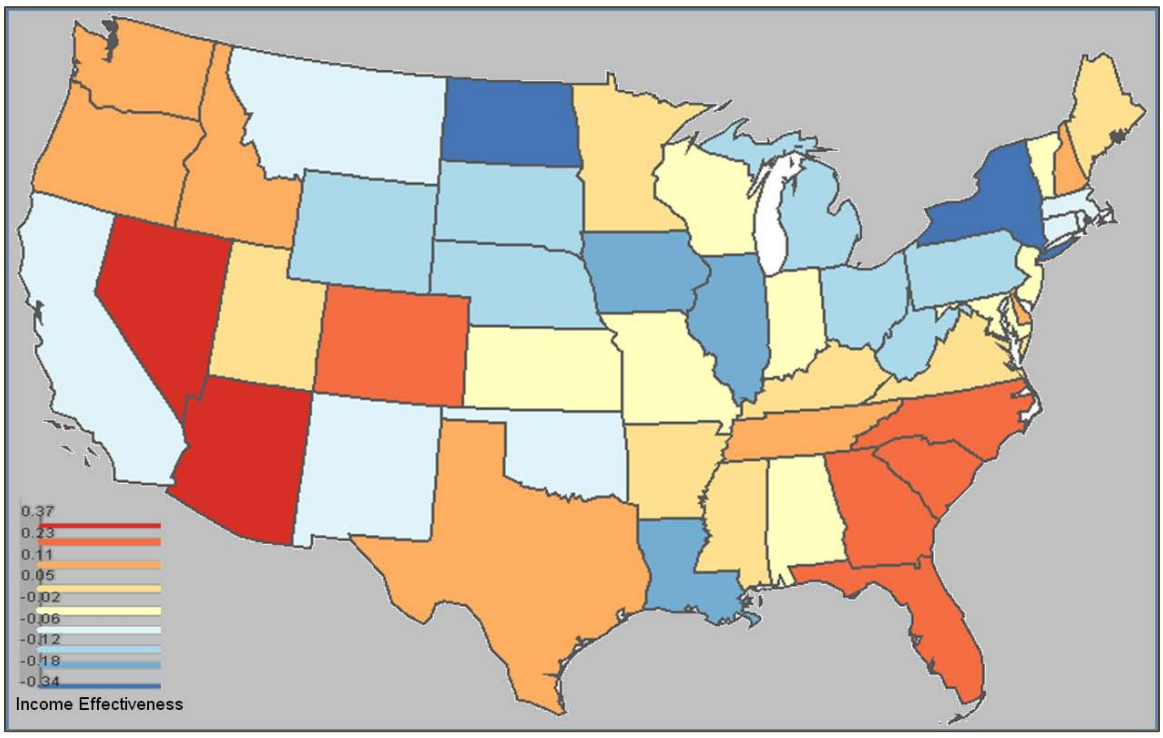


Figure 5.7 Income effectiveness of states based on the 2000 Census county-to-county migration data (Income effectiveness = net income flow/gross income flow)

Among the states of the lowest negative effectiveness ($E \leq$ -20%), the worst is New York (-33.5%). Next to New York are North Dakota (-30.2%) and District of Columbia (-20.8%). New York also lost the most income ($-19.2 billion). Other states that lost more than $5 billion include Illinois ($-6.8 billion), California (-$6.6 billion), and Pennsylvania ($-5.2 billion). Moderate losses are found for the remaining states in this group ($E \leq$ -20%), with District of Columbia losing $1.6 billion and North Dakota $0.8 billion.

Figure 5.7 presents the income effectiveness for 70 SI regions to be compared with the findings at the state-level (as used in existing research). The most noteworthy

difference between Figure 5.6 and Figure 5.7 is the discovery of variance between metro SI regions and the encompassing SI region.

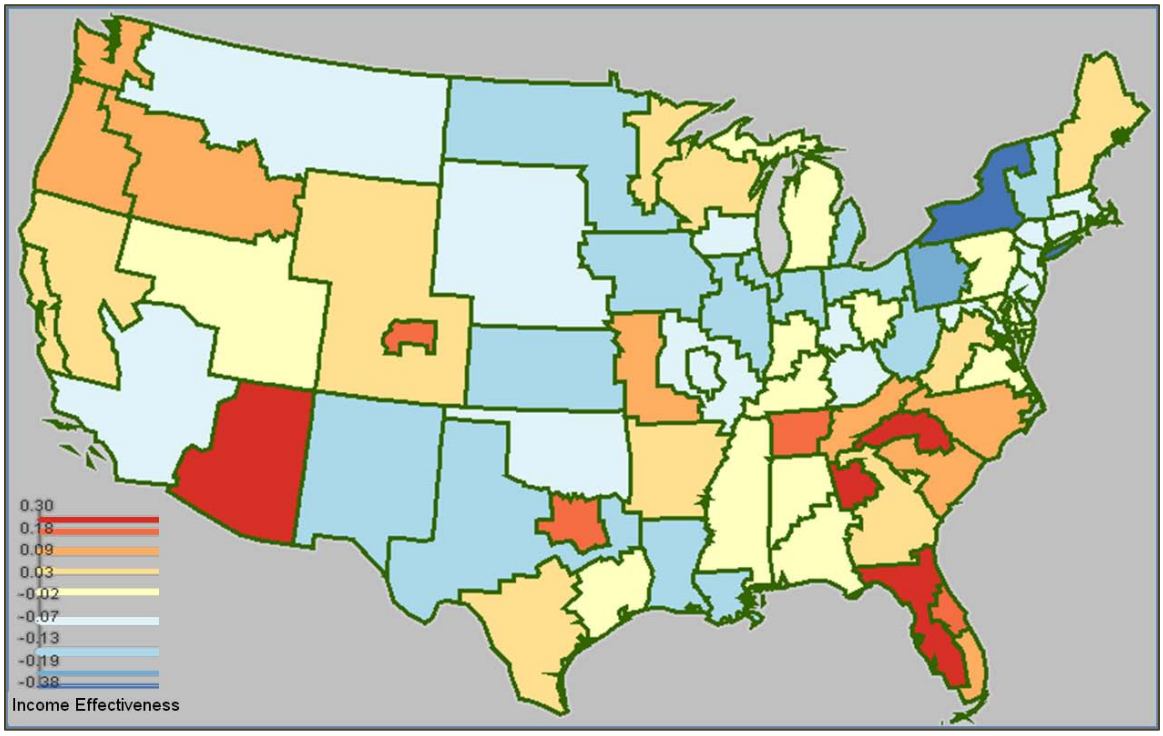


Figure 5.8 Income effectiveness of SI regions at the 70-level. Variances among metro SI regions and the encompassing SI region are discovered (e.g. Atlanta, Denver).

As an example, the three metro SI regions within Texas all received net income gains between 1995 and 2000, while the remaining less urban areas lost considerable income. Detroit lost more income, while the other portion of Michigan gained. Also, Atlanta ($E$=0.28) apparently is more attractive to income flows than the remaining area in Georgia ($E$=0.04). Similarly, Denver has a higher effectiveness index ($E$=0.20) than the encompassing SI region ($E$=0.07). Notable variation is found in Florida as well. On the opposite, St. Louis in Missouri ($E$=-0.08) does not demonstrate a dramatically different level from the encompassing SI region ($E$=-0.06).

Another new finding in Figure 5.7 is related to Charlotte. The state map (Figure 5.6) shows positive gains for both North and South Carolina. The 70-level region partition identifies Charlotte as an independent SI region, which experienced the most effective income gains ($E$=0.30) among all 70 SI regions. Its net income gain is $4.3 billion with a remarkable ranking ($5^{th}$). These indicate that Charlotte is an important income receiver in Carolinas. The significance of Charlotte may be related to the rapid growth of employment opportunities. According to the Census, Charlotte experienced a population increase of 36% or approximately 145,000 from year 1999 to 2000. It became one of the major financial centers in the U.S.

As far as the losing areas are concerned, the large portion of New York state ($E$=-0.37) is found to have the lowest negative income effectiveness. It lost 6.1 billion (ranking $2^{nd}$). New York City ($E$=-0.36) experienced unprecedented income loss ($-10.0 billion). The $3^{rd}$ largest losses occurred in Chicago ($-4.9 billion).

### 5.3.4 Spatial Variations of Flow-based Income Effectiveness

Most studies about income migration focus on the spatial variability of area-based effectiveness. Little attention has been paid to origin-destination income flows. Where did the "income" go from the areas losing most? Which places contribute most to the gaining areas? Answers to these questions can help us better understand the processes that led to the spatial variation of area-based income patterns (Mueser 1989). Existing methods for origin-destination flows (Manson and Groop 2000, Henrie and Plane 2008) are limited to flows associated with a single place. This section utilizes the flow-based income effectiveness to examine origin-destination income flows. A flow-based

effectiveness is denoted by "*e*" to discern it from the area-based effectiveness "*E*". It is the ratio of the net income flow to the gross income flow for a pair of places. Note that, income flow effectiveness (i.e., net income flow divided by the gross income flow between two locations) has a direction, pointing to the gaining area.

Between these 70 SI regions, there are 2,414 net income flows (i.e., only one pair of regions has no flow between them). One third of these income flows has a low effectiveness value ($e$ <0.1) and 60% of these flows are less than 0.2. Figure 5.8a presents the most effective flows ($e$ >0.6), including 22 flows primarily from the Northeast to the Southeast. The total income flow associated with this group is $11 billion, about 3% of the total income flows in the data ($347 billion). The most effective flow is from the east part of New York to the Charlotte region ($e$ = 0.86). Note that, the Charlotte SI region (see Figure 5.7) is larger than the actual metropolitan area of Charlotte.

The second map (Figure 5.8b) contains 59 flows, with $e$ ranging from 0.5 to 0.6. While the northeast-to-southeast pattern is still evident, flows from the Midwest are remarkable. There are also many flows from the Northeast to Arizona. 11 out of the 59 (20%) flows in this group went to the Charlotte SI region, including one from California. The long-distance flow connecting New York and the San Francisco region is also notable. The total income flow associated with this group is 15 billion, which is about 4.3% of all flows.

Figure 5.9 shows the net income flows pointing to six significant (most attractive) SI regions, including West Florida, Arizona, Atlanta, Charlotte, San Francisco Bay Area, and Dallas. Note again that these are SI regions that are conveniently named by the major city that it covers. Please refer to the map in Figure 5.7 for their precise coverage.

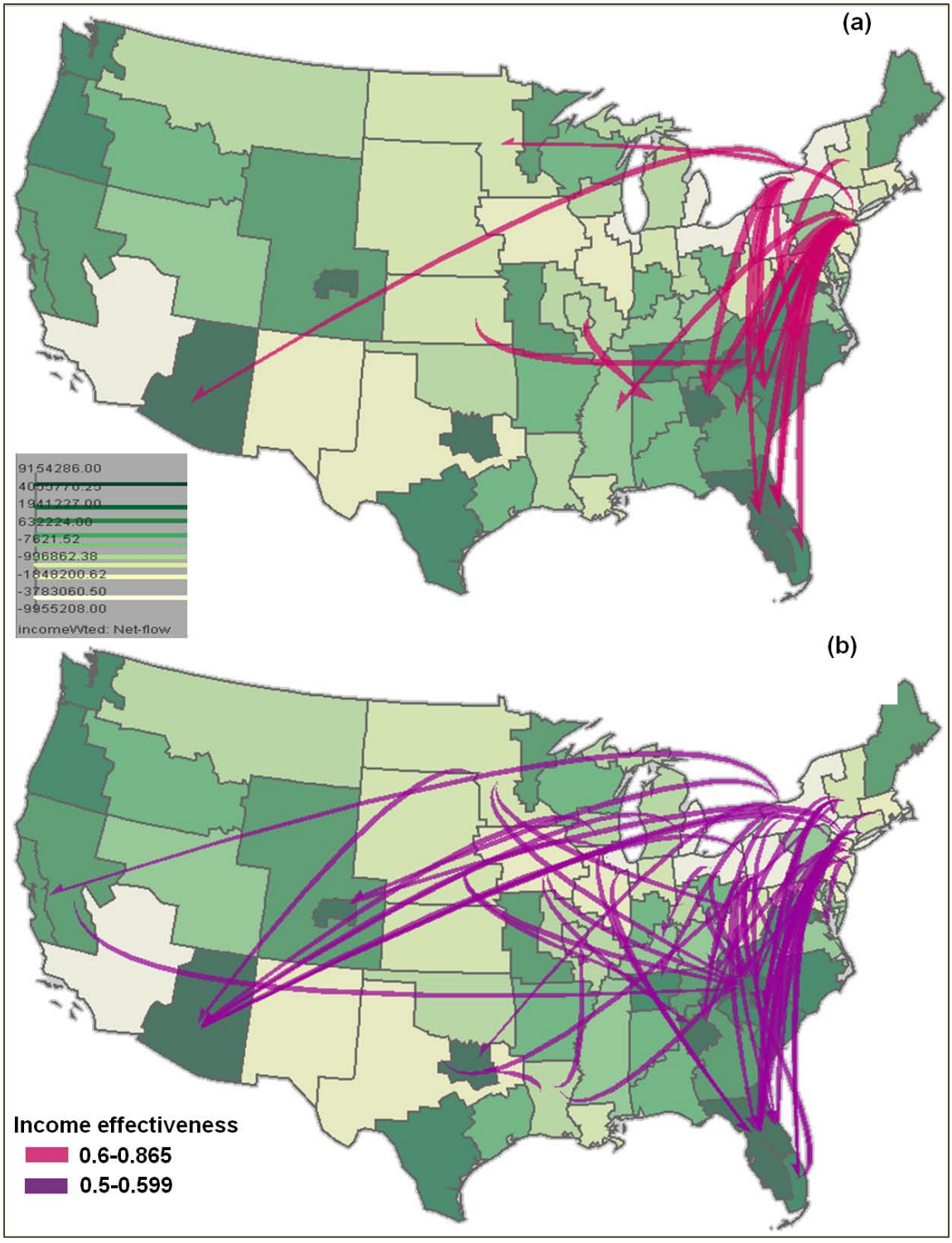


Figure 5.9 Distribution of significantly effective income flows. (a) The most effective class of flows. (b) The 2nd most effective class of flows. SI regions are shaded by the net income. The flow lines are curved at the origin location and become straighter on the destination location.

The regions in Figure 5.9 already stand out in the previous analysis with area-based flow measures. In Figure 5.9, net flow is selected as the flow measure and the threshold is set at zero, i.e., all net flows related to the six regions are shown. In other words, each region has 69 net flows, either coming in or going out.

The SI region of the west Florida (Figure 5.9a) received a total income of $24.1 billion from 56 incoming flows and 992,010 in-migrants, with an average income of 27.3K. This region ranks the 1st on the net income ($9.2 billion). Arizona (Figure 5.9b) received a net income of $7.5 billion (rank 2nd). Flows connected to Arizona are dominantly inflows (67 out of 69), with the only exceptions to New Mexico and the Austin/San Antonia area in Texas. These two maps show the national significance of Arizona and Florida in the domestic migration. While both areas received income flows nationwide, the west Florida received most net incomes from the Northeast, while Arizona received considerable net income from the Midwest and the West, in addition to the Northwest.

The next four maps in Figure 5.9 present the spatial distribution of net income flows related to four large metropolitan SI regions (i.e. Atlanta, Charlotte, San Francisco, and Dallas). Atlanta received $7.1 billion net income (ranked 3rd), close to the total gain of the entire North Carolina. The largest income flows to Atlanta are from the Northeast, Midwest, California, Alabama, Florida, and the rest of Georgia. Charlotte (Figure 5.9d) received a total income gain of $4.3 billion (ranked 5th), with prominent net income gains from New York and East Florida.

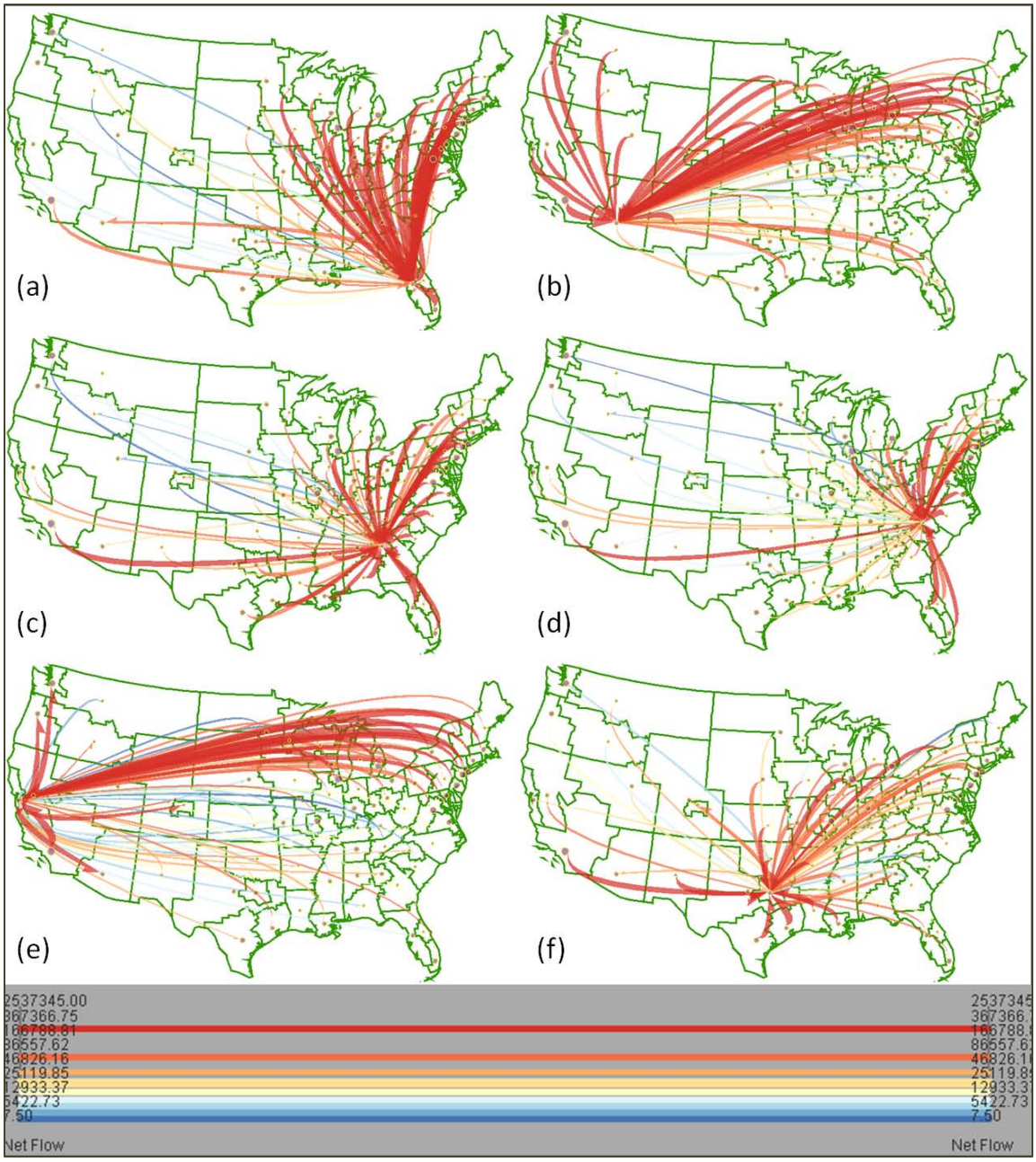


Figure 5.10 Net income flows related to the areas gaining most incomes. (a) West Florida. (b) Arizona. (c) Atlanta. (d) Charlotte. (e) San Francisco. (f) Dallas. The flow lines are curved at the origin location and become straighter on the destination location. The dots reprsent the (county) population-weigthed centriods of regions.

What is most interesting about the net flows related to San Francisco Bay Area (Figure 5.9e) is the spatial distribution of flows. The San Francisco area attracted

significant income flows from the Northeast and the North but lost large incomes to the rest of western regions such as Arizona, Oregon, Washington, and Denver. The Dallas region (Figure 5.9f) is another magnet for income flows, with a net gain of $5.3 billion (ranked 4$^{th}$). Its main income providers include Texas, California, the Northeast and the Midwest.

Figure 5.10 focuses on 4 regions experiencing the worst losses: New York City ($-10.0 billion), west of New York state ($-6.1 billion), Chicago ($-4.9 billion), and southern California/Las Vegas ($-3.5 billion).

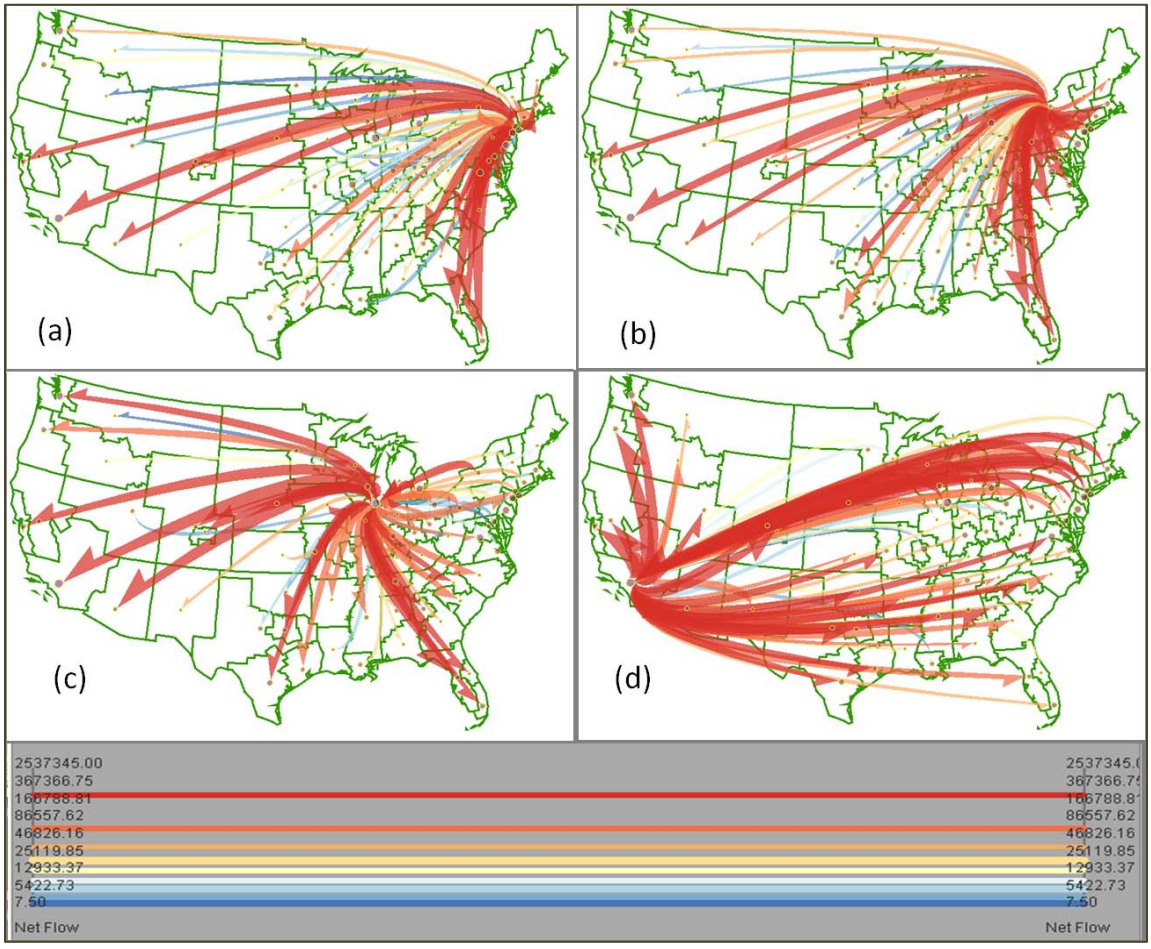


Figure 5.11 Net income flows of the areas losing most. (a) New York City. (b) West of New York state. (c) Chicago; (d) Southern California and Las Vegas. The flow lines are curved at the origin location and become straighter on the destination location The dots reprsent the (county) population-weigthed centriods of regions.

Unsurprisingly, outflows dominate these areas. Florida, California, Texas, and Arizona are the most preferred destinations. The impact of New York City (Figure 5.10a) is nationwide. The net loss of New York City is close to 10% of the total net income flows. Despite the loss, it gained income from several Midwest areas, including Illinois (excluding Chicago), Iowa, Indiana, Milwaukee, New Orleans, west Pennsylvania and west of New York state). West New York (Figure 5.10b) sent net income to all other 69 regions. Despite its considerable loss, Chicago (Figure 5.10c) received net incomes from 20 regions, including some areas in the Northeast or those areas in proximity, e.g., Ohio, Detroit, Iowa, Missouri (excluding St. Louis). Similar to San Francisco, Southern California and Las Vegas (Figure 5.10d) demonstrates an interesting diverging flow pattern: in flows from the Northeast and outflows to the South and the West.

### 5.3.5 Understanding through Demographic Lens: Income, Age, and Education

The above two sections examined the spatial variation of area-based income effectiveness (which leads to a collection of areas of interest) and analyzed origin-destination income flows (which unveils spatial patterns of income flows). This section will further exploit the multivariate analysis capability of the visual system to analyze multivariate factors of income flows. Specifically, income, age, and education information are analyzed for income flows to understand income migration.

In addition to the income stratification as used in previous sections, each non-zero migration flow has the number of migrants for seventeen age groups and seven education groups. The age stratification starts with “age 5-9” and ending with “age 85 and above”. Since the income information is collected for migrants of 16-year old and older in 1999

and the age data is collected for migrants of 5-years and older in year 2000, the two youngest age groups (i.e. age "5-9", age "10-14") are excluded in this research. Therefore, 15 age groups are included in the analysis. The seven education categories are: "$9^{th}$ grade or Less", "9-$12^{th}$ grade", "high school", "college", "associate degree", "bachelor ", and "graduate".

The following six figures (from Figure 5.11 to Figure 5.16) present the analysis result for the three demographic stratifications (i.e., income, age, and education). The six maps are configured in the same way as following: (1) income effectiveness is the chosen flow measure; (2) the threshold of income effectiveness is set at 0.2 (i.e., weaker flows are not shown in the maps); (3) each variable is normalized by the number of migrants of aged 16 and over (i.e., each variable is a ratio); (4) income flows are grouped into 49 clusters with SOM based on selected variables (note: each map may use different variables). The colors of flow lines represent multivariate patterns/structures, i.e., similar colors represent similar multivariate characteristics.

The map in Figure 5.11 shows two flow clusters based on the percentage of total migrants (in the flow) that falls in each income range. The red cluster has the highest percentage values in the four highest income levels (i.e. 35-50k, 50-75k, 75-100k, and 100k+). In other words, the flows in the red cluster are the wealthiest in terms of migrants' income. The other cluster (in blue) represents the opposite—flows of lower income groups (i.e. 5-10k, 10-15k, 15-20k and 20-25k). The spatial patterns of these two clusters, as shown in the map, suggest that the wealthiest flows mainly pointing to urban areas such as San Francisco Bay Area, New York City region, Atlanta, and Dallas. On the other hand, the relatively poor flows mostly go to less urban (more rural) areas.

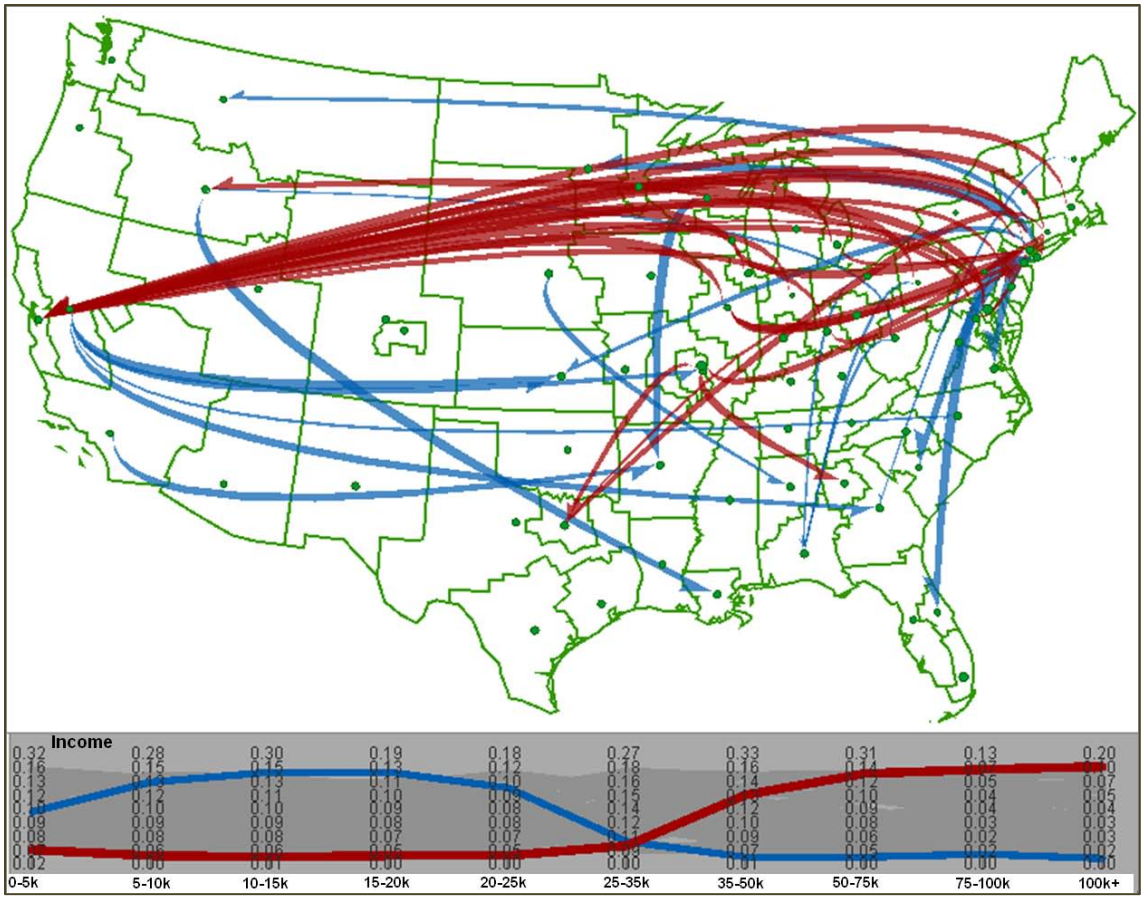


Figure 5.12 Spatial patterns of two flow clusters. The two clusters are the "wealthiest" flows (red) and the relatively "poor" flows (blue). The flow lines are curved at the origin location and become straighter on the destination location The dots reprsent the (county) population-weigthed centriods of regions.

For the SI regions receiving the most income gains, as we found in previous sections, it is interesting to examine their income structure. Figure 5.12 shows two maps of the flows associated with West Florida and Denver, with colors indicating the income composition of the migrants in each flow. As evident in the maps, although both West Florida and Denver attracted considerable incoming flows, their income compositions are quite different.

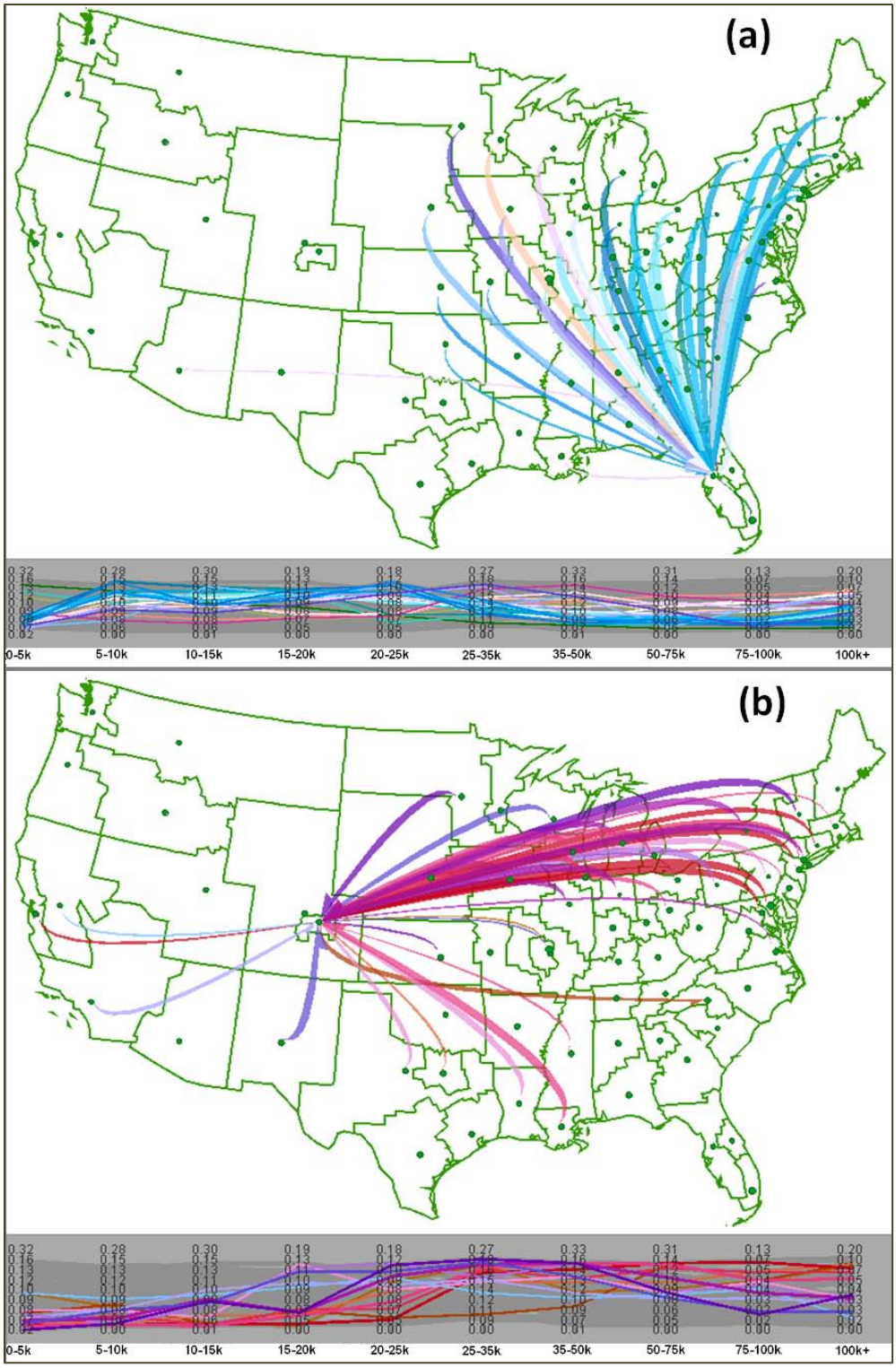


Figure 5.13 The income structure of flows relevant to West Florida (a) and Denver (b) (both received high income gains). The flow lines are curved at the origin location and become straighter on the destination location The dots reprsent the (county) population-weigthed centriods of regions.

The incoming migrants to West Florida are mostly of low-medium income (thus a majority of flows in light blue, see Figure 5.12a). Similar patterns are found for Arizona ($2^{nd}$ income gains) and Orlando ($6^{th}$ income gains) (not shown). Denver (Figure 5.12b) also received a huge net income migration of $3.1 billion ($7^{th}$) but its incoming migrants are wealthier. This pattern is common for several other metropolitan areas (e.g. Atlanta, Dallas, and Charlotte). Further, from the flow table (not shown) we can see considerable difference in the average income of migrants: 25k for migrants to West Florida and 32k for those to Denver.

Figure 5.13 shows four clusters identified based on age variables. Each curve in the PCP represents a cluster of flows. As the PCP shows, the two clusters in blue consist of mostly elder migrants (i.e. dominated by migrants of age 55 and above). On the opposite, the two red clusters are characterized by a high proportion of young migrants (of age 25 to 39). Their spatial patterns are strikingly different: flows dominated by senior migrants tend to move from the north to the south and highly prefer Florida and Arizona. Moreover, Florida is the favorite of flows from the Northeast and the North while Arizona is more attractive to the Northwest and Midwest. On the other hand, flows dominated by young migrants were mainly in east-west directions and predominantly targeting urban areas (e.g. San Francisco, Seattle, Denver, New York City, and Washington DC).

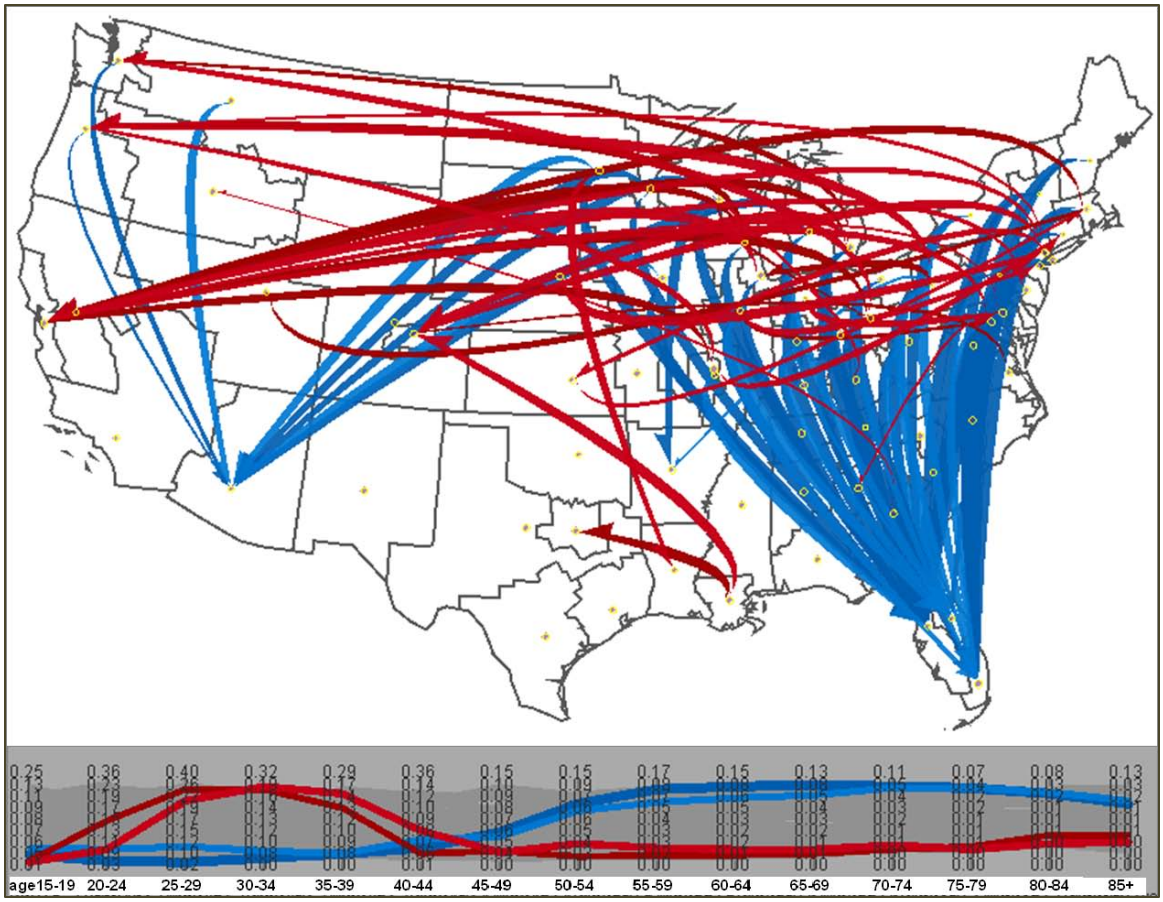


Figure 5.14 Four flow clusters with different age compositions. The two blue clusters are flows of mainly elderly migrants. The two red clusters consist of flows of young migrants. The flow lines are curved at the origin location and become straighter on the destination location The dots reprsent the (county) population-weigthed centriods of regions.

Figure 5.14 compares the age structures of flows related to Denver and Charlotte. The two regions are noteworthy because of the large income gains they received. The flow maps and the PCP unveil different patterns: migrants to Denver are relatively young (with more migrants in age groups of 25-35yrs) while migrants to Charlotte are relatively older (with larger portions in the middle-age groups (35-55yrs).

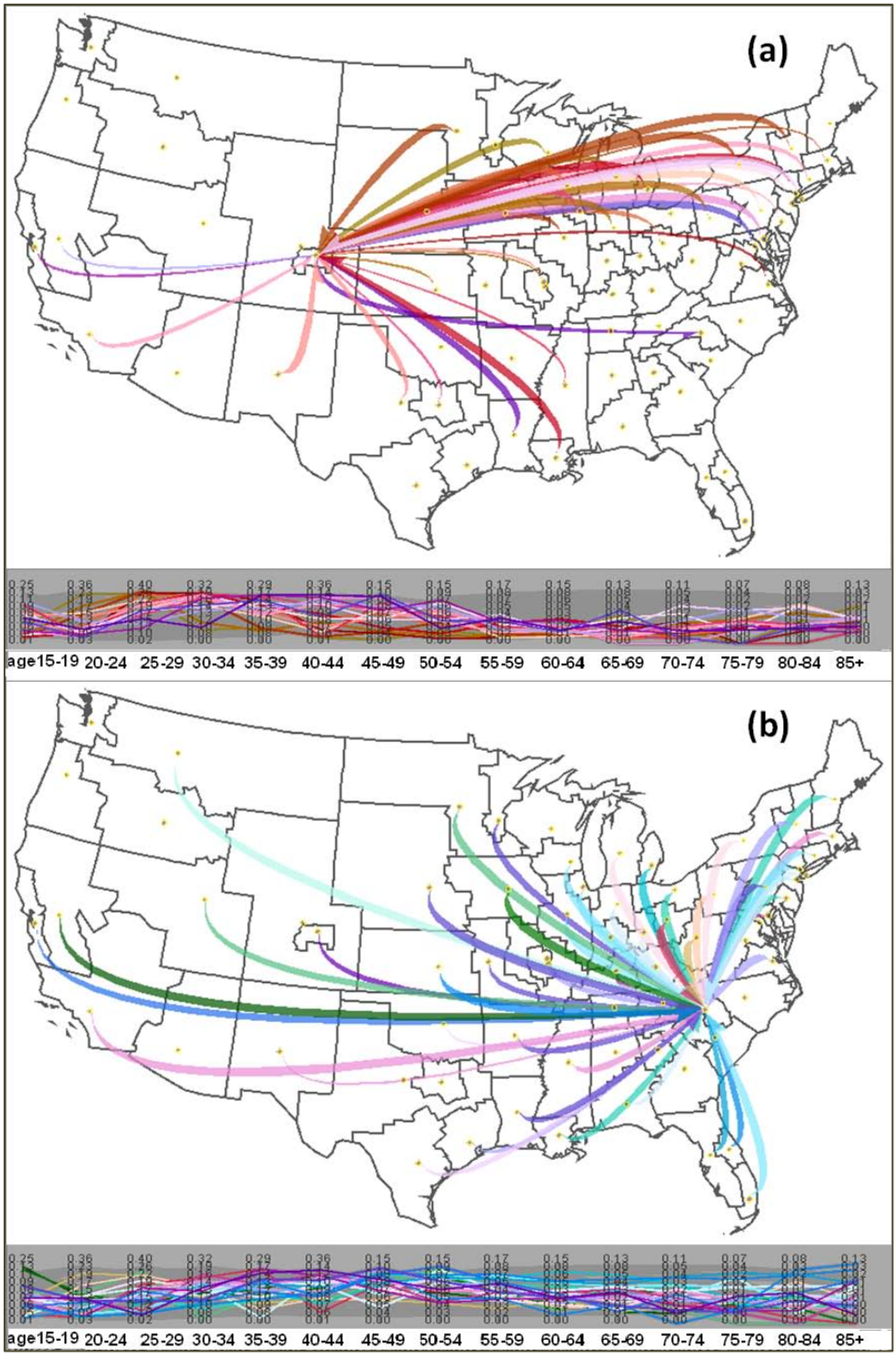


Figure 5.15 Age compositions of flows relevant to Denver and Charlotte: (a) Denver: high percentage of migrants of young ages (25-35yrs); (b) Charlotte: high percentage of migrants in middle age groups (35-55yrs). The flow lines are curved at the origin location and become straighter on the destination location The dots reprsent the (county) population-weigthed centriods of regions.

Last, the focus is turned to the education composition. Similar to the procedures used to examine the income and the age compositions, the education variables are normalized by the number of migrants of aged 16 and over. The SOM groups flows into 49 clusters based on their education structures. Figure 5.15 shows the characteristic of these clusters. Among others, the flows with a high percentage of migrants holding bachelor and graduate degrees are grouped in the blue clusters. The opposite extreme is the purple clusters, which have the highest proportion of least educated migrants (receiving "12$^{th}$ grade or less" education). Three other notable clusters are notably high in "high school" (in red), "college" (in yellow/tan), and "associate" (in green). These clusters will be mapped below to see their spatial patterns.

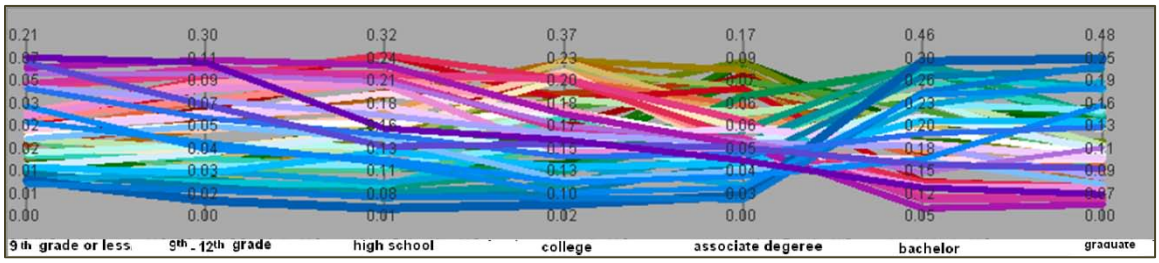


Figure 5.16 Education compositions of flows.

Figure 5.16 maps three clusters of distinctive education structures. The blue cluster is dominated by well-educated migrants (high in both "bachelor" and "graduate"), the other two clusters (purple and red) contain flows dominated by less educated migrants ( "12$^{th}$ grade or less" or "high school"). For the "well-educated" cluster, approximately 56% of its migrants have "bachelor" or "graduate" degrees. This cluster demonstrates two prominent spatial tendencies: (1) from the Northeast and the North Central (east portion of Midwest) to San Francisco, Seattle, and Oregon; and (2) from the Midwest to

Massachusetts, Rhode Island, New York, New Jersey, and Connecticut. No matter eastwards or westwards, these "well-educated" flows mostly target major urban areas.

Flows of less educated migrants display strikingly different patterns. Flows with the highest ration in "high school" (in red) tend to move from the Midwest and the Northeast to Florida or from the Midwest to Arkansas. On the other hand, many flows highest in "12$^{th}$ grade or less" (in purple) left California, the Midwest for the less urbanized portion of the South.

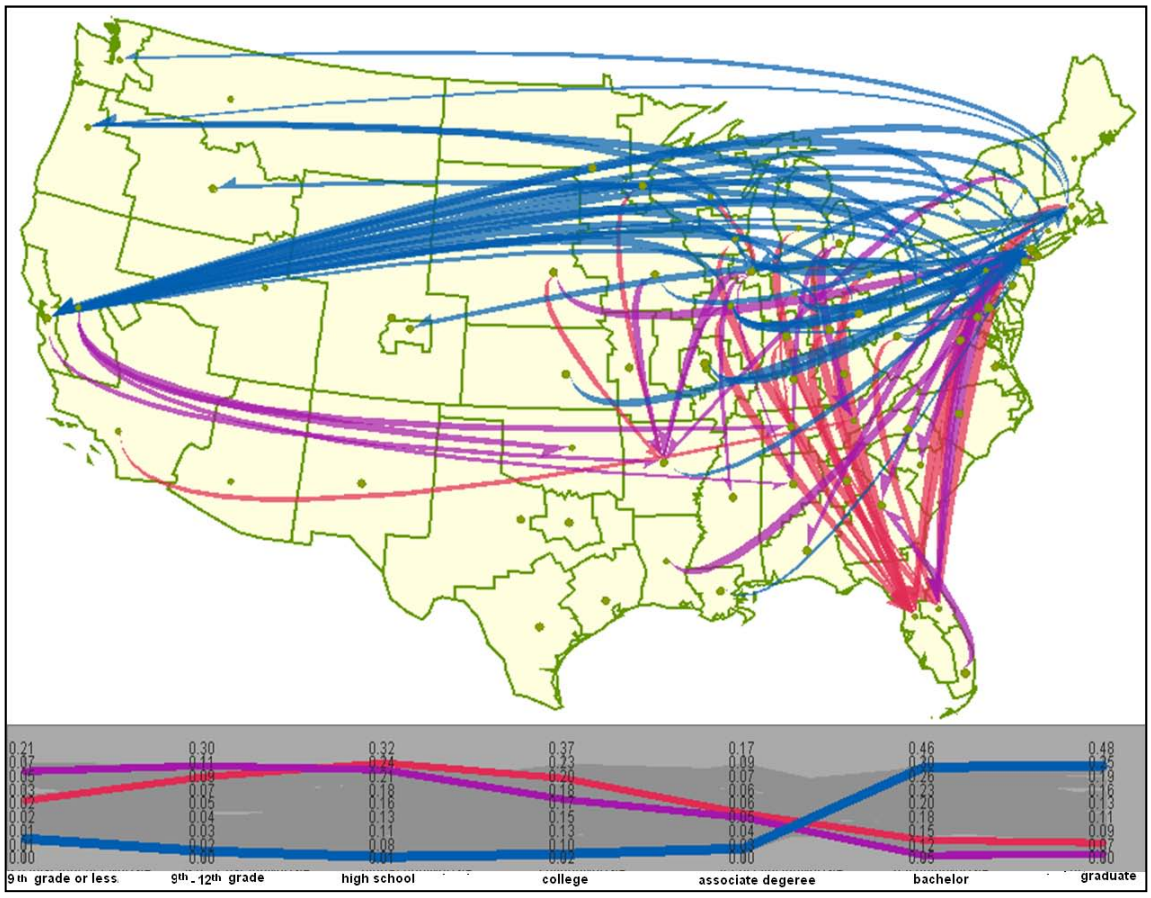


Figure 5.17 Three flow clusters with different education compositions. The blue cluster contains flows dominated by well-educated migrants. The red cluster consists of flows with a high ration in migrants finishing "high school". The purple cluster are mainly flows receiving "12$^{th}$ grade or less" education. The flow lines are curved at the origin location and become straighter on the destination location The dots reprsent the (county) population-weigthed centriods of regions.

The above analyses on the income, age, and education compositions of income flows are only three example analyses. Given the rich information in the data and the flexible capability of the developed system, one can perform many related analyses upon the initial discoveries. With that being said, the above analyses already provided several interesting findings, including: (1) migrants leaving the Northeast and the Midwest for the San Francisco area tend to have higher income, young (25-39 yrs), and well-educated, (2) migrants from the Northeast and the Midwest to Florida and Arizona are primarily elder people; (3) out-migrants from California tend to be less-educated. They prefer less urban areas in the South. Given the income effectiveness of the analyzed flows (20% the minimum) and the geographic coverage of the investigations, these patterns would generate significant impacts on the income redistribution and also the demographic landscape across the nation.

## 5.4 Summary and Discussions

Given a high-resolution and large SI data set, flow aggregation is often needed in migration studies. Existing investigations often rely on states or other predefined divisions/categorization of places. This chapter demonstrates, through a series of comparisons and analyses, that SI regions provide a meaningful and better aggregation scheme for SI data. Specifically, (1) the state division is automatically discovered to a considerable degree in the SI regions derived from the migration network; (2) the SI regions can also identify major metropolitan areas (e.g., San Francisco, Detroit, Denver, Dallas, and Atlanta) at a more detailed scale; and (3) SI regions can naturally support the analysis at different scales while existing boundaries cannot support this.

Moreover, using SI regions in the analyses can better detect patterns, which may be missed if inappropriate aggregation is performed with predefined boundaries. Please note that, although state names or city names are often used to "name" SI regions, SI regions are more meaningful in capturing true "migration regions", which often go beyond state or city boundaries. SI regions can unveil real-world structures such as the strong "core-suburban relationship" from a network perspective.

The visual analytics system developed in this research integrates SI regions, multivariate clustering and visualization, and a much enhanced flow map to create a comprehensive and efficient exploratory system for SI data. A focused investigation is carried out to study the income flows induced by migration, a newly emerged research topic. A majority of current research on income migrations is area-oriented. The investigation carried out in this research expands the analyses by examining origin-destination flows and demographic factors of the flow data (e.g., the income, the age, and the education composition of migrants). Thus, the visual system facilitates an integration of the graph space, the geographic space and the multivariate space in SI data, enhancing our understanding of the spatial and multivariate patterns of flows.

This focused investigation on income flows between 1995 and 2000 leads to some interesting findings. First, diverging income effectiveness is detected between urban and rural areas (e.g. Dallas in Texas gained notable income while the less urban area in Texas experienced income loss). Second, flows from the Northwest and the east portion of the Midwest to San Francisco tend to be high-income, young (25-39 yrs), and well-educated (bachelor degree or above). On the other hand, out-migrants from California are relatively poor and less- educated. They prefer less urban areas in the South. The urban-

rural theme is prominent as revealed throughout the analysis suggesting the usefulness of the reported approach to studies on migration, which is characterized by the interactions of urban and rural areas. Given the evaluating objective, this case study remains preliminary. Abundant domain knowledge and more inclusive explorations may lead to more inspiring and profound discoveries by using the integrated method in this research.

In this case study, flows are mapped at the 50 and the 70-level of SI regions. It would be interesting to see know how the 20 new regions at the 70-level are connected with the SI regions at the 50-level. Future work would offer more details on the relationship of different hierarchical levels of SI regions. Such an effort may also create an opportunity to use domain knowledge to justify a region level based upon the relationship among various hierarchical SI region levels.

As shown in the focused study of income migration, feature selections (i.e. flows, regions, and locations/places) are important for interactive data explorations. With the current implementations, feature selections are limited to drawing rectangle on the map, clicking on map, clicking rows in data tables and etc. In the future, users shall be allowed to choose features based on queries of multiple variables. For instance, a user shall be able to select flows with the gross volume higher than 5,000 and the ratio of foreign-born migrants more than 20%. In this example both the raw flow and the ratio of foreign-born migrants are needed to decide whether a flow shall be included in the analyses or not. Moreover, it would be a convenient addition to allow users to simultaneously view several flow maps that are produced with various configurations. That would make it easier and more efficient to compare flows relevant to different areas or different time stamps, which are often seen in migration analyses.

# CHAPTER 6

# CONCLUSIONS

Spatial interactions (SI) represent an essential force that drives many physical and socioeconomic processes. Spatial interactions are very complex in nature. They often involve: (1) distinctive data spaces (i.e., geographic space, network space, and multivariate space), (2) various spatial constraints (e.g., travel distances, geographic contiguity, and physical barriers), (3) many variables for locations and interactions (flows), i.e. high-dimensionality, and (4) very large data volume--even a moderate-sized dataset may involve thousands of locations and millions of connections.

The complexity of spatial interactions poses great challenges for processing, analyzing, understanding, and communicating spatial interaction data and information. There is a lack of powerful and comprehensive approaches to analyze and extract the rich information lurking in large and complex SI data. This dissertation develops an integrated computational-visual approach to examining SI data from different perspectives and synthesizing them into a holistic understanding. This approach has three main strengths.

*First*, the graph partitioning method of this approach is able to discover spatial structures in spatial interactions (SI regions). The SI regions can then be used to aggregate the original large spatial interaction data while preserving spatial patterns, which is better than using predefined political boundaries (such as states) and can reveal

structures that existing methods cannot. This new partitioning method is more effective and computationally efficient than traditional methods, as demonstrated in the evaluations with synthetic benchmark data.

*Second*, this research combines the three SI data spaces in data exploration and representation. In addition to representing graph patterns, SI regions are used to summarize massive spatial flows and reduce the data size. Thus, the use of SI regions facilitates the mapping and visualization of large SI data. More importantly, the dual role of SI regions bridges the graph space and spatial patterns of flows and the multivariate structures represented by the data.

*Third*, a novel and interactive visual analytic system is developed and implemented to analyze and visualize SI regions, multivariate patterns, and geographic patterns of SI flows. It creates a flexible and comprehensive environment to explore SI data from different perspectives and obtain holistic understandings. In addition to integrating different methods and data spaces, this system also allows the user: (1) to move up or down the SI region hierarchy to examine flow and multivariate patterns at different scales (or detail levels); (2) to configure multivariate analyses via subsetting, normalizing, and deriving new variables on the fly; (3) to choose different flow measures and area-based network measures derived from the original data.

A large inter-county migration data set of the U.S. is used to assess the developed approach and implemented visual analytic system from an application perspective. The data contains over 700,000 county-to-county migration flows (i.e., origin–destination pairs). Given such a high-resolution and large SI data set, traditional methods often aggregate the data by using states or other predefined divisions/categorization of places.

As an alternative aggregation strategy, this research detects SI regions from the data, which captures the underlying community patterns in the data. A series of comparisons and analyses demonstrate that SI regions provide a meaningful and better aggregation scheme for SI data. Specifically, the SI regions can unveil real-world structures such as the strong "core-suburban relationship" from a network perspective. A focused study on income migration shows that the developed visual analytic system can facilitate new and comprehensive analyses that existing research methodologies cannot support, better facilitate the understanding of the spatial and multivariate patterns of income migrations.

In its current form, the developed approach in this research has several limitations. *First*, the partitioning method cannot automatically determine the number of communities to detect. Future research may investigate how to suggest the best hierarchical level based on certain objective measures, such as the modularity gain at each partition and/or the trend of modularity values at each level. *Second*, feature selections in the visual system are now limited to drawing rectangle on the map, clicking on map, clicking rows in data tables and etc. In the future, users shall be allowed to choose features based on queries of multiple variables. Moreover, users shall be allowed to simultaneously view several flow maps that are produced with different configurations. That would make it possible to compare flows relevant to different areas or different time stamps. *Third,* in a larger research scope, future research will also address the temporal dimension of SI data with specifically designed methodologies to help understand shifting trends and structures in spatial interaction across both time and space. Additionally, future research will evaluate the usability of the visual system through designed user tests and usability evaluations.

## REFERENCES


Abello, J., Koutsoflos, E., Gansner, E. R., and North, S. C. 1999. Large Networks Present Visualization Challenges. *ACM SIGGRAPH Computer Graphics,* 33(3): 13-15.

Allison, P. D. 1978. Measures of Inequality. *American Sociological Review,* 43(6): 865-80.

Alm, J. and Winters, J. V. 2009. Distance and Intrastate College Student Migration. *Economics of Education Review,* 28(6): 728-38.

Ambinakudige, S. and Parisi, D. 2011. Internal Migration Effectiveness and Income Effectiveness in the Most Populous Cities in the United States. *Population Review,* 49(2).

Anderson, T. R. 1955. Intermetropolitan Migration: A Comparison of the Hypotheses of Zipf and Stouffer. *American Sociological Review,* 20(3): 287-91.

Andresen, M. A. 2009a. Regionalizing Global Trade Patterns, 1981-2001: Application of a New Method. *Canadian Geographer-Geographe Canadien,* 53(1): 24-44.

--- 2009b. Trade Specialization and Reciprocal Trading Relationships in Canada and the United States, 1989 and 2001. *Annals of the Association of American Geographers,* 99(1): 163-83.

Andrienko, G. and Andrienko, N. 2008. 'Spatio-Temporal Aggregation for Visual Analysis of Movements.' in *Spatio-Temporal Aggregation for Visual Analysis of Movements*, 51-58. Columbus, OH: IEEE.

Andrienko, G., Andrienko, N., Dykes, J., Fabrikant, S. I., and Wachowicz, M. 2008. Geovisualization of Dynamics, Movement and Change: Key Issues and Developing Approaches in Visualization Research Introduction. *Information Visualization,* 7(3-4): 173-80.

Arenas, A., Duch, J., Fernandez, A., and Gomez, S. 2007. Size Reduction of Complex Networks Preserving Modularity. *New Journal of Physics,* 9: 15.

Ashby, N. J. 2007. Economic Freedom and Migration Flows between Us States. *Southern Economic Journal,* 73(3): 677-97.

Belanger, A. and Rogers, A. 1992. The Internal Migration and Spatial Redistribution of the Foreign-Born Population in the United States: 1965-70 and 1975-80. *International Migration Review,* 26(4): 1342-69.

Bell, M., Blake, M., Boyle, P., Duke-Williams, O., Rees, P., Stillwell, J., and Hugo, G. 2002. Cross-National Comparison of Internal Migration: Issues and Measures. *Journal of the Royal Statistical Society Series a-Statistics in Society,* 165: 435-64.

Berry, B. J. 1966. 'Essays on Commodity Flows and the Spatial Structure of the Indian Economy.' in *Essays on Commodity Flows and the Spatial Structure of the Indian Economy*, 348. CHICAGO: UNIV ILL DEPT OF GEOGRAPHY.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. 2008. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics-Theory and Experiment*: 12.

Boettcher, S. and Percus, A. G. 2001. Optimization with Extremal Dynamics. *Physical Review Letters,* 86(23): 5211-14.

Boulding, K. E. 1970. *Economics as a Science.* New York: McGraw-Hill Education.

Butler, M. A., Beale, C. L., Agriculture, U. S. D. o. A. E. R. S., and Division, R. E. 1994. *Rural-Urban Continuum Codes for Metro and Nonmetro Counties, 1993*. US Dept. of Agriculture, Economic Research Service, Agriculture and Rural Economy Division.

Chun, Y. and Griffith, D. A. 2011. Modeling Network Autocorrelation in Space-Time Migration Flow Data: An Eigenvector Spatial Filtering Approach. *Annals of the Association of American Geographers,* 101(3): 523-36.

Clark, D. E. and Hunter, W. J. 1992. The Impact of Economic Opportunity, Amenities and fiscal Factors on Age-Specific Migration Rates. *Journal of Regional Science,* 32: 349–65.

Clark, G. L. 1982. Volatility in the Geographical Structure of Short-Run United-States Interstate Migration. *Environment and Planning A,* 14(2): 145-67.

Clauset, A., Newman, M. E. J., and Moore, C. 2004. Finding Community Structure in Very Large Networks. *Physical Review E,* 70(6): 6.

Condon, A. and Karp, R. M. 2001. Algorithms for Graph Partitioning on the Planted Partition Model. *Random Structures & Algorithms,* 18(2): 116-40.

Conway, K. S. and Houtenville, A. J. 2001. Elderly Migration and State Fiscal Policy: Evidence from the 1990 Census Migration Flaws. *National Tax Journal,* 54(1): 103-23.

Cui, W., Zhou, H., Qu, H., Wong, P. C., and Li, X. 2008. Geometry-Based Edge Clustering for Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG: Proc. of InfoVis'08),* 14(6): 1277-84.

Curry, L., Griffith, D. A., and Sheppard, E. S. 1975. Those Gravity Parameters Again. *Regional Studies,* 9(3): 289-96.

Cushing, B. and Poot, J. 2004. Crossing Boundaries and Borders: Regional Science Advances in Migration Modelling. *Papers in Regional Science,* 83: 317-38.

Danon, L., Diaz-Guilera, A., and Arenas, A. 2006. The Effect of Size Heterogeneity on Community Identification in Complex Networks. *Journal of Statistical Mechanics-Theory and Experiment*: 12.

Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A. 2005. Comparing Community Structure Identification. *Journal of Statistical Mechanics-Theory and Experiment*: 10.

Du, H., Feldman, M. W., Li, S., and Jin, X. 2007. An Algorithm for Detecting Community Structure of Social Networks Based on Prior Knowledge and Modularity: Research Articles. *Complex.,* 12(3): 53-60.

Duch, J. and Arenas, A. 2005. Community Detection in Complex Networks Using Extremal Optimization. *Physical Review E,* 72(2): 4.

Duncan, O. D. 1957. The Measurement of Population Distribution. *Population Studies,* 11(1): 27-45.

Eades, P., Feng, Q.-W., and Lin, X. 1996. 'Straight-Line Drawing Algorithms for Hierarchical Grphs and Clustered Graphs.' Paper presented at Proceedings of the 4th Int. Symposium on Graph Drawing.

Edsall, R. M. 2003. Design and Usability of an Enhanced Geographic Information System for Exploration of Multivariate Health Statistics. *The Professional Geographer,* 55(2): 146-60.

Faggian, A., McCann, P., and Sheppard, S. 2007. Some Evidence That Women Are More Mobile Than Men: Gender Differences in U.K. Graduate Migration Behavior. *Journal of Regional Science,* 47(3): 517-39.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. 1999. 'On Power-Law Relationships of the Internet Topology.'

Fan, Y., Li, M. H., Zhang, P., Wu, J. S., and Di, Z. R. 2007. Accuracy and Precision of Methods for Community Identification in Weighted Networks. *Physica a-Statistical Mechanics and Its Applications,* 377(1): 363-72.

Fienberg, S. E. 1970. An Iterative Procedure for Estimation in Contingency Tables. *Annals of Mathematical Statistics,* 41(3): 907-&.

Fischer, M. M., Essletzbichler, J., Gassler, H., and Trichtl, G. 1993. Telephone Communication Patterns in Austria - a Comparison of the Ipfp-Based Graph-Theoretic and the Intramax Approaches. *Geographical Analysis,* 25(3): 224-33.

Fisher, M. and Gopal, S. 1994. Artificial Neural Networks-a New Approach to Modeling Interregional Telecommunication Flows. *Journal of Regional Science,* 34(4): 503.

Fortunato, S. 2010. Community Detection in Graphs. *Physics Reports-Review Section of Physics Letters,* 486(3-5): 75-174.

Fotheringham, A. S., Rees, P., Champion, T., Kalogirou, S., and Tremayne, A. R. 2004. The Development of a Migration Model for English and Wales: Overview and Modeling out-Migration *Environment and Planning A,* 36(9): 1633-72.

Fowlkes, E. B. and Mallows, C. L. 1983. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association,* 78(383): 553-69.

Franklin, R. S. 2003. 'Domestic Migration across Regions, Divisions, and States: 1995 to 2000.' in *Domestic Migration across Regions, Divisions, and States: 1995 to 2000*. U.S. CENSUS BUREAU.

Frey, W. H. 1996. Immigration, Domestic Migration, and Demographic Balkanization in America: New Evidence for the 1990s. *Population and Development Review,* 22(4): 741-63.

--- 2002a. 'Census 2000 Reveals New Native-Born and Foreign-Born Shifts across Us.' in *Census 2000 Reveals New Native-Born and Foreign-Born Shifts across Us*. POPULATION STUDIES CENTER, UNIVERSITY OF MICHIGAN.

--- 2002b. Three Americas: The Rising Significance of Regions. *Journal of the American Planning Association,* 68(4): 349-55.

Frey, W. H. and Farley, R. 1996. Latino, Asian, and Black Segregation in Us Metropolitan Areas: Are Multiethnic Metros Different*. *Demography,* 33(1): 35-50.

Frey, W. H. and Liaw, K. L. 1998. Immigrant Concentration and Domestic Migrant Dispersal: Is Movement to Nonometropolitan Areas "White Flight"? *The Professional Geographer,* 50(2): 215-32.

Friendly, M. and Kwan, E. 2003. Effect Ordering for Data Displays. *Computational Statistics & Data Analysis,* 43: 509-39.

Fuguitt, G. V. 1985. The Nonmetropolitan Population Turnaround. *Annual Review of Sociology,* 11: 259-80.

Fuguitt, G. V. and Beale, C. L. 1993. The Changing Concentration of the Older Nonmetropolitan Population, 1960-90. *Journals of Gerontology,* 48(6): S278-S88.

Fulton, J. A., Fuguitt, G. V., and Gibson, R. M. 1997. Recent Changes in Metropolitan-Nonmetropolitan Migration Streams. *Rural sociology,* 62(3): 363-84.

Ghoniem, M., Fekete, J.-D., and Castagliola, P. 2004. 'A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations.' Paper presented at IEEE Symposium Information Visualization.

--- 2005. On the Readability of Graphs Using Node-Link and Matrix-Based Representations: A Controlled Experiment and Statistical Analysis. *Information Visualization,* 4(2): 114-35.

Gober, P., Jeffery, C. F., and McHugh, K. E. 1996. Using Moving-Industry Data to Depict Us Migration Patterns. *Growth and Change,* 27(2): 231-51.

Gottlieb, P. D. 2006. "Running Down the up Escalator": A Revisionist Perspective on Decentralization and Deconcentration in the United States, 1970-2000. *International Regional Science Review,* 29(2): 135.

Gottlieb, P. D. and Joseph, G. 2006. College-to-Work Migration of Technology Graduates and Holders of Doctorates within the United States. *Journal of Regional Science,* 46(4): 627-59.

Greenwood, M. J. 1975. Research on Internal Migration in United-States - Survey. *Journal of Economic Literature,* 13(2): 397-433.

--- 1997. 'Internal Migration in Developed Countries.' in *Internal Migration in Developed Countries* eds. M. R. Rosenzweig & O. Stark. Amsterdam: Elsevier.

Greenwood, M. J. and Hunt, G. L. 2003. The Early History of Migration Research. *International Regional Science Review,* 26(1): 3-37.

Griffith, D. A. and Jones, K. G. 1980. Explorations into the Relationship between Spatial Structure and Spatial Interaction. *Environment and Planning A,* 12(2): 187-201.

Guimer, agrave, Roger, Sales-Pardo, M., Amaral, L., iacute, and s, A. N. 2004. Modularity from Fluctuations in Random Graphs and Complex Networks. *Physical Review E,* 70(2): 025101.

Guo, D. 2007. Visual Analytics of Spatial Interaction Patterns for Pandemic Decision Support. *International Journal of Geographical Information Science,* 21(8): 859-77.

Guo, D., Chen, J., MacEachren, A. M., and Liao, K. 2006. A Visualization System for Space-Time and Multivariate Patterns (Vis-Stamp). *IEEE Transactions on Visualization and Computer Graphics,* 12(6): 1461-74.

Guo, D. and Gahegan, M. 2006. Spatial Ordering and Encoding for Geographic Data Mining and Visualization. *Journal of Intelligent Information Systems,* 27: 243-66.

Guo, D., Gahegan, M., MacEachren, A. M., and Zhou, B. 2005. Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach. *Cartography and Geographic Information Science,* 32(2): 113-32.

Haenszel, W. 1967. Concept, Measurement, and Data in Migration Analysis. *Demography,* 4(1): 253-61.

He, C. and Gober, P. 2003. Gendering Interprovincial Migration in China. *International Migration Review,* 37(4): 1220-51.

He, J. S. and Pooler, J. 2002. The Regional Concentration of China's Interprovincial Migration Flows, 1982-90. *Population and Environment,* 24(2): 149-82.

Henrie, C. J. and Plane, D. A. 2008. Exodus from the California Core: Using Demographic Effectiveness and Migration Impact Measures to Examine Population Redistribution within the Western United States. *Population Research and Policy Review,* 27(1): 43-64.

Hirst, M. A. 1977. Hierarchical Aggregation Procedures for Interaction Data - Comment. *Environment and Planning A,* 9(1): 99-103.

Hollingsworth, T. 1971. 'Gross Migration Flows as a Basis for Regional Definition: An Experiment with Scottish Data.'

Holmes, J. 1978. Transformation of Flow Matrices to Eliminate the Effects of Differing Sizes of Origin-Destination Units: A Further Comment. *Systems, Man and Cybernetics, IEEE Transactions on,* 8(4): 325-32.

Holten, D. and Wijk, J. J. v. 2009. Force-Directed Edge Bundling for Graph Visualization. *Computer Graphics Forum (Proceedings of Eurographics/IEEE-VGTC Symposium on Visualization),* 28(3): 983-90.

Inselberg, A. 1985. The Plane with Parallel Coordinates. *The Visual Computer,* 1: 69-97.

Isserman, A. M., Plane, D. A., and McMillen, D. B. 1982. Internal Migration in the United-States - an Evaluation of Federal Data. *Review of Public Data Use,* 10(4): 285-311.

Johnson, K. M., Voss, P. R., Hammer, R. B., Fuguitt, G. V., and McNiven, S. 2005. Temporal and Spatial Variation in Age-Specific Net Migration in the United States. *Demography,* 42(4): 791-812.

Keane, M. J. 1978. Functional Distance Approach to Regionalization. *Regional Studies,* 12(3): 379-86.

Kephart, G. 1988. Heterogeneity and the Implied Dynamics of Regional Growth-Rates - Was the Nonmetropolitan Turnaround an Artifact of Aggregation. *Demography,* 25(1): 99-113.

Kim, S. 2010. Intra-Regional Residential Movement of the Elderly: Testing a Suburban-to-Urban Migration Hypothesis. *Annals of Regional Science,* 46(1): 1-17.

Kodrzycki, Y. K. 2001. Migration of Recent College Graduates: Evidence from the National Longitudinal Survey of Youth. *New England Economic Review*: 13-34.

Koua, E. L., MacEachren, A., and Kraak, M. J. 2006. Evaluating the Usability of Visualization Methods in an Exploratory Geovisualization Environment. *International Journal of Geographical Information Science,* 20(4): 425-48.

Krieg, R. G. 1993. Black-White Regional Migration and the Impact of Education: A Multinomial Logit Analysis. *The Annals of Regional Science,* 27(3): 211-22.

Lancichinetti, A. 2008. 'Benchmark Graphs to Test Community Detection Algorithms.' in *Benchmark Graphs to Test Community Detection Algorithms*.

Lancichinetti, A. and Fortunato, S. 2009. Benchmarks for Testing Community Detection Algorithms on Directed and Weighted Graphs with Overlapping Communities. *Physical Review E,* 80(1): 8.

Lancichinetti, A., Fortunato, S., and Radicchi, F. 2008. Benchmark Graphs for Testing Community Detection Algorithms. *Physical review E,* 78(4): 5.

Leicht, E. A. and Newman, M. E. J. 2008. Community Structure in Directed Networks. *Physical Review Letters,* 100(11): 4.

Long, L. and Deare, D. 1988. United-States Population Redistribution - a Perspective on the Nonmetropolitan Turnaround. *Population and Development Review,* 14(3): 433-50.

Longino, C. F., Wiseman, R. F., Biggar, J. C., and Flynn, C. B. 1984. Aged Metropolitan-Nonmetropolitan Migration Streams over 3 Census Decades. *Journals of Gerontology,* 39(6): 721-29.

Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., and Dawson, S. M. 2003. The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations. *Behavioral Ecology and Sociobiology,* 54(4): 396-405.

Makinen, E. and Siirtola, H. 2000. Reordering the Reorderable Matrix as an Algorithmic Problem. *Theory and Application of Diagrams, Proceedings,* 1889: 453-67.

Manson, G. A. and Groop, R. E. 2000. Us Intercounty Migration in the 1990s: People and Income Move Down the Urban Hierarchy. *Professional Geographer,* 52(3): 493-504.

Marble, D. F., Gou, Z., Liu, L., and Saunders, J. 1997. 'Recent Advances in the Exploratory Analysis of Interregional Flows in Space and Time.'

Masser, I. and Brown, P. J. B. 1975. Hierarchical Aggregation Procedures for Interaction Data. *Environment and Planning A,* 7(5): 509-23.

Masser, I. and Scheurwater, J. 1980. Functional Regionalization of Spatial Interaction Data - an Evaluation of Some Suggested Strategies. *Environment and Planning A,* 12(12): 1357-82.

Mitchell, W. and Watts, M. 2010. Identifying Functional Regions in Australia Using Hierarchical Aggregation Techniques. *Geographical Research,* 48(1): 24-41.

Morrill, R. 1994. Age-Specific Migration and Regional Diversity. *Environment and Planning A,* 26(11): 1699-710.

--- 2006. Classic Map Revisited: The Growth of Megalopolis. *Professional Geographer,* 58(2): 155-60.

Morrill, R. L. 1988. Migration Regions and Population Redistribution. *Growth and Change,* 19(1): 43-60.

Mueser, P. 1989. The Spatial Structure of Migration: An Analysis of Flows between States in the USA over Three Decades. *Regional Studies,* 23(3): 185 - 200.

Mueser, P. R., White, M. J., and Tierney, J. P. 1988. Patterns of Net Migration by Age for United-States Counties 1950-1980 - the Impact of Increasing Spatial Differentiation by Life-Cycle. *Canadian Journal of Regional Science-Revue Canadienne Des Sciences Regionales,* 11(1): 57-75.

Murtagh, F. 1985. A Survey of Algorithms for Contiguity-Constrained Clustering and Related Problems. *Computer Journal,* 28(1): 82-88.

Nelson, P. B. 2005. Migration and the Regional Redistribution of Nonearnings Income in the United States: Metropolitan and Nonmetropolitan Perspectives from 1975 to 2000. *Environment and Planning A,* 37(9): 1613–36

Newbold, K. B. 1999. Internal Migration of the Foreign-Born: Population Concentration or Dispersion? *Population & Environment,* 20(3): 259-76.

Newbold, K. B. and Peterson, D. A. 2001. Distance Weighted Migration Measures. *Papers in Regional Science,* 80(3): 371-80.

Newman, M. and Girvan, M. 2004. Finding and Evaluating Community Structure in Networks. *Physical Review E,* 69(2): 26113.

Newman, M. E. J. 2001. The Structure of Scientific Collaboration Networks. *Proceedings of the National Academy of Sciences of the United States of America,* 98(2): 404.

--- 2004a. Analysis of Weighted Networks. *Physical Review E,* 70(5): 9.

--- 2004b. Fast Algorithm for Detecting Community Structure in Networks. *Physical Review E,* 69(6): 5.

--- 2006a. Finding Community Structure in Networks Using the Eigenvectors of Matrices. *Physical Review E,* 74(3): 19.

--- 2006b. Modularity and Community Structure in Networks. *Proceedings of the National Academy of Sciences of the United States of America,* 103(23): 8577-82.

Ng, R. C. Y. 1969. Recent Internal Population Movement in Thailand. *Annals of the Association of American Geographers,* 59(4): 710-30.

Pandit, K. 1994. Differentiating between Subsystems and Typologies in the Analysis of Migration Regions - a United-States Example. *Professional Geographer,* 46(3): 331-45.

Pellegrini, P. A. and Fotheringham, A. S. 2002. Modelling Spatial Choice: A Review and Synthesis in a Migration Context *Progress in Human Geography,* 26(4): 487-510.

Perry, M. J. 2003. 'State-to-State Migration Flows: 1995 to 2000.' in *State-to-State Migration Flows: 1995 to 2000*. US Census Bureau.

Phan, D., Xiao, L., Yeh, R., Hanrahan, P., and Terry, W. 2005. 'Flow Map Layout.' Paper presented at IEEE Symposium on information visualization.

Plane, D. A. 1984a. Migration Space - Doubly Constrained Gravity Model Mapping of Relative Interstate Separation. *Annals of the Association of American Geographers,* 74(2): 244-56.

--- 1984b. A Systemic Demographic Efficiency Analysis of Us Interstate Population Exchange, 1935-1980. *Economic Geography,* 60(4): 294-312.

--- 1999a. Geographical Pattern Analysis of Income Migration in the United States. *International Journal of Population Geography,* 5(3): 195-212.

--- 1999b. Migration Drift. *Professional Geographer,* 51(1): 1-11.

Plane, D. A. and Bitter, C. 1997. The Role of Migration Research in Regional Science. *Papers in Regional Science,* 76: 133-53.

Plane, D. A. and Heins, F. 2003. Age Articulation of U.S. Inter-Metropolitan Migration Flows. *The Annals of Regional Science,* 37(1): 107-30.

Plane, D. A., Henrie, C. J., and Perry, M. J. 2005. Migration up and Down the Urban Hierarchy and across the Life Course. *Proceedings of the National Academy of Sciences of the United States of America,* 102(43): 15313-18.

Plane, D. A. and Jurjevich, J. R. 2009. Ties That No Longer Bind? The Patterns and Repercussions of Age-Articulated Migration. *Professional Geographer,* 61(1): 4-20.

Plane, D. A. and Mulligan, G. F. 1997. Measuring Spatial Focusing in a Migration System. *Demography,* 34(2): 251-62.

Podolák, P. 1995. Interregional Migration Pattern in Slovakia: Efficiency Analysis and Demographic Consequences. *Geoforum,* 26(1): 65-74.

Poon, J. P. 1997. The Cosmopolitanization of Trade Regions: Global Trends and Implications, 1965-1990. *Economic Geography,* 73(4): 390-404.

Pujol, J. M., Bejar, J., and Delgado, J. 2006. Clustering Algorithm for Determining Community Structure in Large Networks. *Physical Review E,* 74(1): 9.

Rae, A. 2009. From Spatial Interaction Data to Spatial Interaction Information? Geovisualisation and Spatial Structures of Migration from the 2001 Uk Census. *Computers Environment and Urban Systems,* 33(3): 161-78.

Rand, W. M. 1971. Objective Criteria for Evaluation of Clustering Methods. *Journal of the American Statistical Association,* 66(336): 846-&.

Ravenstein, E. G. 1885. The Laws of Migration *Journal of the Royal Statistical Society,* 48: 167-235.

Rayer, S. and Brown, D. L. 2001. Geographic Diversity of Inter-County Migration in the United States, 1980–1995. *Population Research and Policy Review,* 20: 229–52.

Rebhun, U. and Raveh, A. 2006. The Spatial Distribution of Quality of Life in the United States and Interastate Migration, 1965–1970 and 1985–1990. *Social Indicators Research,* 78: 137–78.

Reisinger, M. E. 2003. Sectoral Shifts and Occupational Migration in the United States. *The Professional Geographer,* 55(3): 383-95.

Renkow, M. and Hoover, D. 2000. Commuting, Migration, and Rural-Urban Population Dynamics. *Journal of Regional Science,* 40(2): 261-87.

Rogers, A. and Raymer, J. 1998. The Spatial Focus of Us Interstate Migration Flows. *International Journal of Population Geography,* 4(1): 63-80.

Rogers, A., Raymer, J., and Willekens, F. 2002. Capturing the Age and Spatial Structures of Migration. *Environment and Planning A,* 34(2): 341-59.

Rogers, A. and Sweeney, S. 1998. Measuring the Spatial Focus of Migration Patterns. *The Professional Geographer,* 50(2): 232-42.

Roy, J. R. and Thill, J. C. 2004. Spatial Interaction Modelling. *Papers in Regional Science,* 83(1): 339-61.

Schachter, J. P., Franklin, R. S., and Perry, M. J. 2003. 'Migration and Geographic Mobility in Metropolitan and Nonmetropolitan America: 1995 to 2000.' in *Migration and Geographic Mobility in Metropolitan and Nonmetropolitan America: 1995 to 2000*. U. S. Census Bureau.

Schaffer, D., Zuo, Z., Greenberg, S., Bartram, L., Dill, J., Dubs, S., and Roseman, M. 1996. Navigating Hierarchically Clustered Networks through Fisheye and Full-Zoom Methods. *ACM Transactions on Computer-Human Interaction (TOCHI),* 3(2): 162-88.

Schuetz, P. and Caflisch, A. 2008. Multistep Greedy Algorithm Identifies Community Structure in Real-World and Computer-Generated Networks. *Physical Review E,* 78(2): 7.

Schwartz, A. 1973. Interpreting the Effect of Distance on Migration. *The Journal of Political Economy,* 81(5): 1153-69.

Sheppard, E. S., Griffith, D. A., and Curry, L. 1976. Final Comment on Mis-Specification and Autocorrelation in Those Gravity Parameters. *Regional Studies,* 10(3): 337-39.

Shumway, J. M. and Otterstrom, S. 2010. Us Regional Income Change and Migration: 1995-2004. *Population Space and Place,* 16(6): 483-97.

Shumway, J. M. and Otterstrom, S. M. 2001. Spatial Patterns of Migration and Income Change in the Mountain West: The Dominance of Service Based, Amenity Rich Counties. *The Professional Geographer,* 53(4): 492-502.

Siirtola, H. and Makinen, E. 2005. Constructing and Reconstructing the Reorderable Matrix. *Information Visualization,* 4: 32-48.

Slater, P. B. 1975. Hierarchical Regionalization of Rsfsr Administrative Units Using 1966-69 Migration Data. *Soviet Geography Review and Translation,* 16(7): 453-65.

--- 1976a. Hierarchical Internal Migration Regions of France. *Ieee Transactions on Systems Man and Cybernetics,* 6(4): 321-24.

--- 1976b. The Use of State-to-State College Migration Data in Developing a Hierarchy of Higher Educational Regions. *Research in Higher Education,* 4(4): 305-15.

--- 1980. State Boundary Length as a Determinant of Migration Regions. *Ieee Transactions on Systems Man and Cybernetics,* 10(10): 678-83.

--- 1984. A Partial Hierarchical Regionalization of 3140 United-States Counties on the Basis of 1965-1970 Intercounty Migration. *Environment and Planning A,* 16(4): 545-50.

--- 2009. A Two-Stage Algorithm for Extracting the Multiscale Backbone of Complex Weighted Networks. *Proceedings of the National Academy of Sciences of the United States of America,* 106(26): E66-E66.

Slifkin, R. T., Randolph, R., and Ricketts, T. C. 2004. The Changing Metropolitan Designation Process and Rural America. *Journal of Rural Health,* 20(1): 1-6.

Stouffer, S. A. 1940. Intervening Opportunities: A Theory Relating Mobility and Distance. *American Sociological Review,* 5(6): 845-67.

Sun, Y., Danila, B., Josic, K., and Bassler, K. E. 2009. Improved Community Structure Detection Using a Modified Fine-Tuning Strategy. *Epl,* 86(2): 6.

Thiemann, C., Theis, F., Grady, D., Brune, R., and Brockmann, D. Y. N. 2010. The Structure of Borders in a Small World. *PLoS ONE,* 5(11): e15422.

Tobler, W. R. 1981. A Model of Geographical Movement. *Geographical Analysis,* 13(1): 1-20.

--- 1987. Experiments in Migration Mapping by Computer. *American Cartographer,* 14: 155-63.

Tufte, E. R. 1986. *The Visual Display of Quantitative Information.* Graphics Press.

Tyree, A. 1973. Mobility Ratios and Association in Mobility Tables. *Population Studies-a Journal of Demography,* 27(3): 577-88.

Walters, W. H. 2002. Later-Life Migration in the United States: A Review of Recent Research. *Journal of Planning Literature,* 17(1): 37-66.

Weber, S. and Munst, A. S. 2009. Migration and Mobility in an Enlarged Europe. A Gender Perspective. *Population,* 64(2): 417-19.

White, M. J. 1986. Segregation and Diversity Measures in Population Distribution. *Population index,* 52(2): 198-221.

Wilson, F. D. 1987. Metropolitan and Nonmetropolitan Migration Streams - 1935-1980. *Demography,* 24(2): 211-28.

Wong, D. W. S. 1992. The Reliability of Using the Iterative Proportional Fitting Procedure. *Professional Geographer,* 44(3): 340-48.

Wood, J., Dykes, J., and Slingsby, A. 2010. Visualisation of Origins, Destinations and Flows with Od Maps. *Cartographic Journal,* 47(2): 117-29.

Xiang, B., Chen, E.-H., and Zhou, T. 2009. 'Finding Community Structure Based on Subgraph Similarity.' in *Finding Community Structure Based on Subgraph Similarity*, 73-81. Springer Berlin / Heidelberg.

Yan, J. and Thill, J. C. 2009. Visual Data Mining in Spatial Interaction Analysis with Self-Organizing Maps. *Environment and Planning B-Planning & Design,* 36(3): 466-86.

Ye, Z., Hu, S., and Yu, J. 2008. Adaptive Clustering Algorithm for Community Detection in Complex Networks. *Physical Review E,* 78(4): 046115.

Young, D. A. 2002. A New Space-Time Computer Simulation Method for Human Migration. *American Anthropologist,* 104(1): 138-58.

Zachary, W. W. 1977. Information-Flow Model for Conflict and Fission in Small-Groups. *Journal of Anthropological Research,* 33(4): 452-73.

Zipf, G. K. 1946. The P 1 P 2/D Hypothesis: On the Intercity Movement of Persons. *American Sociological Review,* 11(6): 677-86.