Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data

Diansheng Guo, Member, IEEE

Abstract—Spatial interactions (or flows), such as population migration and disease spread, naturally form a weighted location-to-location network (graph). Such geographically embedded networks (graphs) are usually very large. For example, the county-to-county migration data in the U.S. has thousands of counties and about a million migration paths. Moreover, many variables are associated with each flow, such as the number of migrants for different age groups, income levels, and occupations. It is a challenging task to visualize such data and discover network structures, multivariate relations, and their geographic patterns simultaneously. This paper addresses these challenges by developing an integrated interactive visualization framework that consists of three coupled components: (1) a spatially constrained graph partitioning method that can construct a hierarchy of geographical regions (communities), where there are more flows or connections within regions than across regions; (2) a multivariate clustering and visualization method to detect and present multivariate patterns in the aggregated region-to-region flows; and (3) a highly interactive flow mapping component to map both flow and multivariate patterns in the geographic space, at different hierarchical levels. The proposed approach can process relatively large datasets and effectively discover and visualize major flow structures and multivariate relations at the same time. User interactions are supported to facilitate the understanding of both an overview and detailed patterns.

Index Terms—Hierarchical clustering, graph partitioning, flow mapping, spatial interaction, contiguity constraints, multidimensional visualization, coordinated views, data mining.

1 INTRODUCTION

Both the physical environment and the human society are highly dynamic systems. Many elements in the system, such as humans, virus, commodities, funds, and the intangibles like information, constantly move from place to place on the Earth surface. Such location-to-location movements, referred to as spatial interactions, are among the essential forces that drive many physical and socioeconomic processes, and thus, are often a critical component in a wide range of research fields and decision-making, such as epidemiology, pandemics, demography, urban planning, transportation, economics, and emergency management.

However, it is a challenging problem to analyze and understand patterns in massive spatial interactions, which can easily have thousands of nodes (locations) and millions of connections. For example, as a moderate-sized dataset, the county-to-county migration data for the U.S. has about a million county-to-county migration paths among >3000 counties. Much larger datasets have also been emerging, for example, individual-based activities and pandemic spread data generated via survey and simulation models [8, 13].

In addition to its large volume, spatial interaction data is often characterized by having many variables. For example, a migration flow from one county to another can have detailed information such as the number of migrants for different age groups, income levels, and occupations. To discover the unknown information hidden in the data and address complex real-world problems, it is critical to be able to process and discover patterns across several spaces, i.e., the network space (e.g., to discover community structures), the multivariate space (e.g., to identify multivariate relations of migration flows), and the geographical space (e.g., to examine the spatial distribution of both flow structures and multivariate patterns).

Flow maps are commonly used to visualize spatial interaction data [29, 36, 39, 40]. However, existing flow mapping approaches have three major limitations. First, they are only effective in portraying small datasets and will quickly become cluttered as the dataset size increases. Second, it often uses the default geographic unit of the observational data (e.g., counties or states), which may not be the best unit to represent and uncover underlying patterns due to the dramatic differences (e.g., population and size) among the units. One can easily imagine that if locations are grouped into different regions, different flow patterns may emerge. Third, multivariate information cannot be visualized simultaneously with the flow patterns and thus it is difficult to perceive patterns that involve both flow structures and multivariate relations.

This paper presents a solution to the above problems, which is to:

- discover natural regions (communities) in spatial interactions, with spatially constrained graph partitioning and fine tuning;
- generalize flow patterns to higher abstraction levels using the discovered hierarchy of regions;
- perform multivariate clustering for the multivariate data of flows at the regional level; and
- visualize both the flow patterns and multivariate patterns in an interactive flow mapping component, which supports mapping at different hierarchical levels, visualization of different combination of variables, and dynamic linking and brushing across multiple views.

Throughout this paper, the 2000 U.S. county-to-county migration data is used to demonstrate the developed approach.

2 BACKGROUND AND RELATED WORK

Spatial interactions naturally form a network/graph, where each node is a location (or area) and each link is an interaction between two nodes (locations). Such spatial interaction networks (e.g., county-to-county migrations) normally consist of:

- \( S \): a set of locations (nodes), e.g., counties or states in the U.S.;
- \( F \): a set of flows (links) among locations, directed or undirected;
- \( V \): a set of variables for each flow, e.g., the number of migrants for different age groups, income levels, and occupations (Table 1).

The number of locations \( n \) in a spatial interaction data can range from dozens (e.g., U.S. states), through thousands (e.g., U.S. counties), to millions (e.g., mobile phones or individual household locations). The number of interaction links may vary from thousands (e.g., migration flows among 48 states), through millions (e.g., migration flows among >3000 counties), to billions (e.g., packages delivered by FedEx annually). The number of variables (for flows and locations) normally ranges from a dozen to several hundreds.
2.1 Flow Mapping

As an exploratory approach, flow maps are commonly used to visualize spatial interactions, such as migrations between 48 contiguous states in the U.S. [36, 41] and the air traffic over the U.S. and Canada in 24 hours [26]. In a flow map, the origin and the destination of a flow are connected with a straight or artificially curved line. The lines are usually symbolized by width or color to indicate the volume of the flow. However, as introduced above, traditional flow maps are only effective in visualizing small datasets. A flow map will become cluttered and difficult to read even for a relatively small dataset such as migrations among 48 U.S. states (see Fig. 1 left map). With much larger spatial interaction data sets becoming available at finer spatial resolution levels (such as counties or even household locations), there is a urgent need of new approaches to render flow maps and understand complex patterns hidden in the data.

To cope with the large number of flow lines and reduce the cluttering in a flow map, one strategy is to use sampling [7] or interactive queries to select and show only a small subset of data at a time [37] (see Fig. 1 right map). This of course cannot support an overview of general patterns in the data. Another type of strategy is to derive and visualize line densities [37], or group edges into bundles [6, 20, 36, 38], which can effective resolve the cluttering problem in the visual display. However, they have limitations in addressing the following two main concerns.

First, although data is available at a fine spatial resolution (e.g., counties), meaningful patterns and information to be discovered are often related to and expressed at a coarser resolution (e.g., regions). For example, we are more interested in migration patterns such as “senior migrants from the Northeast tend to move to the Southeast”. The key challenge is to discover the natural regions (communities) from the migration flows instead of using pre-defined political boundaries. Line density or edge bundling approaches use the given geographic units and group flow lines into bundles, which do not explicitly seek and provide a clear understanding of natural regions.

Second, in visualizing spatial interaction data such as migration, the flow lines in the map signify the connection between origins and destinations and optionally may also show information regarding the moving objects using visual variables such as colors and line widths. A line density or edge bundling approach, however, makes it difficult to perceive the correspondence between origins and destinations, e.g., [6] uses color and [21] allows interactive selection to alleviate this problem.

This research proposes an approach from a different perspective, which is to aggregate locations into regions based on the flow structure/topology (as oppose to grouping edges into bundles based on geometric adjacency).

2.2 Graph Partitioning

In non-spatial domains, graphs or networks have been widely used to represent interaction-based data such as social networks, citations, and webpage links. There are also extensive research on the extraction of structural patterns from such networks, using methods including hierarchical clustering [5, 30], graph partitioning [4, 10, 23], eigenvector-based approaches [31], and mixture models [34]. There are two main types of network structures that have been extensively studied: assortative mixing—members in each group are mostly connected to other members in the same group [1, 32], and disassortative mixing—members in the same group are more likely to connect to members of other groups [28, 42].

To perform hierarchical clustering of a network, a “connection strength” is first assigned to each pair of nodes, then the most strongly connected nodes are merged into a cluster, and this merging procedure is repeated until all nodes are in the same cluster. Graph partitioning methods, different from clustering methods, partition a network into two (or several) parts, each of which is then recursively partitioned. Different measures can be optimized during the partitioning, for example, minimizing the number of edges to be cut [24] or maximizing a modularity measure within subgraphs [31].

Note that the methods introduced above cannot handle spatial constraints (e.g., spatial contiguity) and thus are not directly applicable for discovering spatially contiguous regions (communities). This research adopts the notion of modularity [31] in partitioning spatial interaction graphs and discovering assortative mixing patterns (i.e., community structures) in migrations.

2.3 Regionalization

Regionalization is to group spatial units into spatially contiguous regions while optimizing an objective function [17, 18]. This is a combinatorial problem and it is impossible to enumerate all possible groupings to find the best one. Existing regionalization methods may be classified into four groups: (1) non-spatial clustering followed by spatial processing [19]; (2) non-spatial clustering with a spatially weighted dissimilarity measure [43]; (3) local search (e.g., hill climbing) and optimization from a starting partition [35]; and (4) spatially constrained hierarchical clustering and partitioning [3, 14].

To help analyze and map spatial interaction data, regionalization can reduce spurious data variations caused by uneven unit sizes or small base populations, and generalize large spatial interaction data to discover general flow patterns. The key requirement is that the regionalization should preserve major patterns in the network while suppressing details.

Guo [14] proposes a family of regionalization methods based on spatially constrained hierarchical clustering, including the single-linkage (SLK), complete-linkage (CLK), and average-linkage (ALK) methods. Evaluation results show that the average-linkage and complete linkage methods, with spatial contiguity constraints, can derive regions of significantly better quality (in terms of the objective function value) than other existing methods. However, it remains an open research problem to design even better methods to achieve more “optimal” solutions.

The research presented in this paper develops a fine-tuning procedure, which can be combined with an existing regionalization method and significantly improve the quality of regions (in terms of optimizing an objective function).

2.4 Multivariate Visualization

Another limitation of existing flow mapping methods is that they cannot display multivariate information, such as age compositions or income levels of migrants in each flow (see Table 1).

Table 1: A demonstrative example of flow variables.

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>Age 0-5</th>
<th>Age 6-10</th>
<th>Income &lt;10k</th>
<th>Income 10-20k</th>
<th>White</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>5</td>
<td>8</td>
<td>55</td>
<td>68</td>
<td>101</td>
<td>39</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>12</td>
<td>10</td>
<td>22</td>
<td>97</td>
<td>163</td>
<td>25</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>3</td>
<td>9</td>
<td>21</td>
<td>33</td>
<td>60</td>
<td>33</td>
</tr>
<tr>
<td>B</td>
<td>E</td>
<td>86</td>
<td>99</td>
<td>67</td>
<td>278</td>
<td>500</td>
<td>120</td>
</tr>
<tr>
<td>C</td>
<td>F</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>18</td>
<td>115</td>
<td>23</td>
</tr>
</tbody>
</table>
A variety of methods have been developed in the literature for visualizing multivariate data, such as scatterplot matrices [2], pixel-oriented approaches [25], and parallel coordinate plots (PCP) [22]. However, in the presented research, we not only want to visualize the multivariate data of flows, but also need to visualize both the flow structure and multivariate patterns simultaneously. Guo, et al [16] develops an integrated approach to visualize multivariate spatial patterns by encoding multivariate clusters (derived with a self-organizing map) with a 2-D color scheme and then use the colors (which signify multivariate information) to render a multivariate map. This research extends the above multivariate mapping approach to visualize multivariate flow patterns.

3 Overview of Methodology

This paper presents a new methodological framework for the exploratory analysis and visualization of large spatial interaction data. Given a set of locations and a set of interactions (links) among those locations, the proposed analysis framework consists of the following procedures to summarize the data, discover general patterns and produce an interactive flow map for user exploration.

First, a newly developed regionalization method is used to extract natural regions (community structures) by “optimally” partitioning the spatial interaction network so that there are much more flows within regions than across regions (see Fig. 2 for a demonstrative example). Section 4 presents the new regionalization method (or specifically, a fine-tuning procedure) that derives high-quality regions by maximizing an objective function (i.e., a modularity measure).

Second, the original spatial interactions and their associated multivariate information are aggregated based on the discovered regions so that a flow map can be rendered at a higher abstraction level to reveal major flow patterns. Since the regions form a hierarchy, the flow map naturally supports pattern exploration at different abstraction levels.

Third, a self-organizing map is used to perform clustering analysis of the flow variables and encode clusters with a 2-D color scheme. The flow map shows flows in colors (which represent clusters) and therefore allows the understanding of multivariate information and flow structure at the same time. As the user navigates up or down the region hierarchy, the set of flows and their associated multivariate values will change, and therefore the clustering will be performed again automatically.

Fourth, a variety of user interactions are supported to efficiently facilitate the exploration and accurate interpretation of spatial interaction patterns.

The main contribution of this research is the development of the fine-tuning procedure (i.e., a more effective regionalization method) for deriving natural regions according to flows and the integration of a suite of methods for the analysis and visualization of massive spatial interaction data. This paper focuses on the overall analysis framework and does not provide detailed algorithms and evaluations for the regionalization method due to limited space.

4 Regionalization of Spatial Interactions

4.1 Modularity Measure

To derive regions from spatial interactions, a similarity measure or connection strength has to be defined for each pair of locations (or regions). This research adopts the concept of modularity measure, which is defined as in equation (1) [33].

\[
\text{Modularity} = \text{Actual Flows} - \text{Expected Flows}
\]  

(1)

The above definition is a general framework that can be implemented or extended in different ways according to application contexts. Take the migration data as an example. The population of each state or county may vary dramatically, which will significantly influence the migration to and from it. For example, in the state-level flow map shown in Fig. 1 (left), California, Florida, and Texas have large volumes of migration, which is primarily due to the fact that these states all have large population. Therefore, it is necessary to take into account of the background population to calculate an expected migration value for each pair of states and then examine the modularity (i.e., the difference between the actual migration and the expected migration) instead of the raw migration counts.

Different statistical models can be used to calculate the expectation. In this paper, the simplest model is used, which assumes interactions among locations are random and proportional to the origin and destination population. For example, in the migration data, we assume each individual has the same probability to move and the choice of destination is proportional to the population of the destination (see equation (2)).

\[
\text{Expected Flows}(A,B) = P_A P_B F / P_S^2
\]  

(2)

where \( P_s \) is the total population for all locations \( S \), \( P_A \) is the population of region \( A \subseteq S \), \( P_B \) is the population of \( B \subseteq S, A \cap B = \emptyset \), and \( F \) is the total flows among all locations (including flows within the same location, i.e., internal flows). Equation 2 ensures that the total of expected flows is the same as total actual flows.

If there is no internal flow within the same location, Equation (3) should be used instead to calculate the expectation between two different regions and Equation (4) for within-region expectations, where \( P_i \) is the population for a specific location (e.g., a county). In this paper, the migration data does not include within-county migrations and thus equations (3) and (4) are used.

\[
\text{Expected Flows}(A,B) = P_A P_B F / (P_S^2 - \sum_{i \in S} P_i^2)
\]  

(3)

\[
\text{Expected Flows}(A,A) = (P_A^2 - \sum_{i \in A} P_i^2) F / (P_S^2 - \sum_{i \in S} P_i^2)
\]  

(4)

4.2 Spatially Constrained Hierarchical Clustering and Partitioning to Construct Regions

After calculating the modularity measure for each pair of locations, we have a weighted graph, where the weight for each link is the modularity. The higher the modularity, the stronger the connection between the two locations is. Given such a graph (or similarity matrix), an average-linkage clustering based regionalization method, i.e., the full-order-ALK (hereafter ALK) [14], is used to construct spatially contiguous regions. Algorithmic details of this method are available in [14]. Below is a brief and conceptual instruction. The ALK method takes two steps to derive regions.

First, it constructs a hierarchy of clusters from the bottom up by iteratively merging the most connected clusters (just as a normal average-linkage clustering does except that it ensures every cluster at each hierarchical level is spatially contiguous). Therefore, the method needs a contiguity matrix as input. The output is a spatially contiguous tree, where each edge connects two geographic neighbors and the entire tree is consistent with the cluster hierarchy.

Second, the spatially contiguous tree is partitioned from the top down, by finding the best edge to remove. In other words, the removal of this edge will obtain two regions and achieve a maximum
gain of within-region modularity, which is the total modularity within each region. By repeating this step for each new region, a hierarchy of regions is constructed. During this partitioning process, additional constraints can be enforced, for example, we may want to impose a minimum population size for each region, i.e., a region cannot be further partitioned if one of the two resultant regions does not meet the minimum population requirement.

To summarize: the first step builds a hierarchy of clusters from the bottom up, using location-to-location modularity values; the second step partitions the spatially contiguous tree (which is obtained through the above hierarchy) to obtain regions while maximizing within-region modularity and enforcing other constraints.

As evaluated in [14], this regionalization method significantly outperforms other regionalization methods. However, in this research, a significant improvement is made over this method by developing a fine-tuning procedure to improve each partitioning.

4.3 A Fine-Tuning Procedure

During the second step (tree partitioning) introduced above, each iteration partitions a region into two, by removing the best edge. Given these two newly generated regions, a fine-tuning procedure is developed to modify the boundaries between them (by moving locations from one region to the other) to further optimize (i.e., maximize) the total within-region modularity.

Suppose the tree-partitioning step suggests dividing a region into two regions A and B (each of which is spatially contiguous). The fine-tuning algorithm will find the best location (among all the locations in A and B) that, when moved to the other region, will increase the overall modularity the most. If no location can be moved to increase the overall modularity, then the one that will cause the least decrease in modularity will be moved. While moving a location from one region to the other, the spatial contiguity of both regions and other optional constraints must be enforced. In other words, moving a location to or from a region should not break the contiguity of that region. Above moves are made repeatedly but each location can only be moved once. When all the possible locations have been moved once, the entire sequence of moves will be analyzed and the sub-sequence (e.g., the first 10 moves) that gives the maximum increase in modularity will be accepted (i.e., a new partition is generated using this sub-sequence of moves). Then, take this new partition as the starting point, the above procedure is repeated again until there is no further improvement in modularity.

Although in spirit similar to the procedure introduced in [33], the fine-tuning procedure developed in this research has two contributions. First, it starts with the ALK regionalization result (note: the fine-tuning procedure needs a reasonably good regionalization to start with). Second, it ensures spatial contiguity and other constraints (such as minimum region population), which makes the algorithm much more complex.

Please note that, the fine-tuning procedure allows locations to move even if they cause temporary decrease in modularity, as long as subsequent moves can make up the loss and eventually achieve a better modularity. As such, the fine-tuning procedure can escape local optima and has a chance to reach (or get close to) the global optima. Fig. 3 shows the results of partitioning the county-to-county migration data in the U.S. using the ALK method alone and the ALK method coupled with the fine-tuning procedure. It is evident that the fine-tuning procedure can significantly improve the regionalization quality (in terms of maximizing within-region modularity).

The high quality of this new regionalization method (specifically the fine-tuning procedure) is achieved at the cost of computing time. The worst-case time complexity for the fine-tuning procedure is $O(n^4)$, where $n$ is the number of locations. For the migration data used in this research, which has about 3100 counties, the regionalization procedure takes about 15-20 minutes (while it takes only 30 seconds without the fine-tuning procedure). However, this is only a one-time cost—once the regions are constructed, they can be saved and loaded later. For much larger datasets, an approximate but more efficient version of this fine-tuning procedure is needed.

4.4 Interpretation of Migration Regions

Migration is an intriguing topic to regional scientists, demographers, sociologists, and geographers [11, 12]. This research uses the county-to-county migration data (1995-2000) in the U.S. to demonstrate the developed analysis and visualization framework. The migration data contains >3000 counties and about 800,000 origin-destination flows. To derive regions, the directed flow matrix is converted to an undirected symmetric matrix by adding in both directions for each pair of counties. The expectation matrix is by definition symmetric. Modularity is calculated using equations 1, 3 and 4. The ALK method, coupled with the fine-tuning procedure, is applied to discover regions. Fig. 4 shows maps of: (a) 4 regions, (b) 7 regions, (c) 12 regions, and (d) 45 regions.

The circle symbol for each region indicates its internal modularity (i.e., above-expectation migration within the region) and is positioned at the population centroid of the region. The maximum total modularity of all regions is achieved at 12 regions (see Fig 4-c), which means that none of the 12 regions can be partitioned without decreasing the overall modularity. If the purpose is to maximize modularity and discover community structures, then the regionalization will stop at 12 regions. It is interesting to see that migration data exhibit strong community structures, in a spatial hierarchy. For example, with four regions (Fig. 4-a), it is clear that people tend to move within the Northeast, East, Midwest, and the West regions. With 12 regions, we can also see how the migration naturally defines local regions such as the Southeast and the West.

To continue the partition, a minimum region population threshold of 3.5 million is imposed, which means that the regionalization method will try to maximize (or minimize the decrease of) modularity while satisfying the region size constraint. Following the convention in migration analysis, the estimated population for year 1995 is used, excluding those died before 2000. With the 3.5 million threshold, 49 regions are obtained, i.e., no region can be further partitioned to have two regions that both have 3.5 million or more population. Note: the region size constraint only takes effect after the maximum modularity has been achieved.

It is striking to see that, although state information is not used in the process, many of the discovered regions follow state boundaries closely (see Fig. 4-d). This probably indicates that people tend to move within administrative boundaries.
4.5 Hierarchical Flow Mapping at Region Levels

With locations aggregated into meaningful and natural regions, it becomes possible to visualize the original large spatial interaction data at a higher abstraction level. Fig. 5 shows the flows with 45 regions. The thickness of each flow indicates its modularity value. In other words, the flow map only shows above-expectation flows. The map to the left is an undirected flow map and the one on the right is directed. In the directed flow map, we can see some one-way flows, i.e., the flow is only significant (above expectation) for one direction.

As the interface snapshot in Fig. 6 shows, the user can change the number of regions to view the flow patterns at different levels. The user may also change the flow (modularity) threshold to view more or less flows at the same level. If the threshold is set as zero, only above-expectation flows are shown.

There are three visual clues for flow directions: arrows, right-hand traffic rule, and (optionally) quadratic Bezier curves [9]. Flow lines between two locations follow the right-hand traffic rule: pointing to its destination, a flow line is on the right side of the "road" (see Fig. 5 and Fig. 6). When arrows are too small and flows are mostly one-way (so that it is difficult to tell which side of the "road" they are on), Bezier curves can be used, where a flow line is curvy at the origin and straight on the destination end (see Fig. 7).
5 Multivariate Flow Mapping

Spatial interaction data often contains multivariate information. For example, each migration flow at the county level contains the origin county, destination county, the number of migrants, and dozens of other flow variables such as counts of migrants for different age groups, income levels, etc. To fully understand spatial interactions and its driving processes, it is important to examine information across different perspectives and obtain a holistic understanding of the overall patterns. For example, while looking at the flow maps shown in Fig. 5, many follow-up questions may arise, such as: What type of migrants are moving to the South (e.g., Florida or Arizona)? Where are the young population moving? Do people of different occupation tend to move differently? To answer above questions, we need to integrate multivariate information into the analysis and visualization.

This research extends an existing multivariate clustering and mapping approach [15, 16] to visualize multivariate flow information. The unique challenge for this extension is that, flows and their attributes change when the user navigates up and down the region hierarchy. Therefore, the clustering is performed automatically whenever flows and their attributes change.

5.1 Multivariate Clustering and Pattern Encoding

A self-organizing map is used to identify clusters in the multivariate flow data and order clusters in a two-dimensional layout so that nearby clusters are similar [27]. Each SOM node (cluster) is represented with a circle, whose size (area) represents the number of flows that it contains (see Fig. 6, bottom left). Behind the nodes (circles) there is the U-matrix layer, where hexagons are shaded to show the multivariate dissimilarity between neighboring nodes, with darker tones representing greater dissimilarity.

Traditionally each SOM node is labelled to indicate the data items that are covered by the node. However, such a labelling strategy is not useful here since it is difficult to name flows. The approach introduced in [16] uses a 2D color scheme to assign a unique color to each SOM cluster so that nearby (and therefore similar) clusters have similar colors (see Fig. 6: bottom left). Colors are then passed on to other visualization components where their multivariate information can be revealed and understood.

5.2 Multivariate Visualization and Flow Mapping

To reveal the multivariate meaning of flow clusters (which are now represented by colors), a parallel coordinate plot (PCP) is used to visualize the multivariate data [15]. The PCP can visualize the data at two different levels: the cluster level (see Fig. 6) or the data item level (see Fig. 7). At the cluster level, the PCP shows each cluster as a single string, which has the same color as the cluster does in the SOM. The thickness of each string is proportional to the cluster size. At the data item level, each string in the PCP will represent an individual region-to-region flow, in the same color of its cluster.

With the clusters (and colors) derived by the SOM, and their meaning revealed by the PCP (as the legend), it is straightforward to pass on the colors to the flows in the flow map (see Fig. 6: top-right), where it is obvious that there are multivariate spatial patterns of flows. For example, red colors (clusters) represent migration flows that have a relatively high percentage of senior migrants (with age between 65 and 69). In the flow map, we notice these red flows are pointing to either the Florida region or the Arizona region.

Fig. 6: Multivariate flow mapping at the regional level. A self-organizing map (bottom left), parallel coordinate plot (bottom right), and a flow map (top right) are coordinated to present flow structure, multivariate information, and spatial patterns at the same time.
5.3 User Interactions
The three visual components (i.e., SOM, PCP, and Flow Map) allow a variety of user interactions such as selection-based brushing and linking (see Fig. 7). A selection can be progressively refined by, for example, adding or subtracting new selection(s). The user may select at either the cluster level or the data item level. To respond to item-level selection, the SOM will highlight the cluster that contains the selected data item and change the circle to a wedge accordingly if some items in that cluster are not in the selection.

It is worth emphasizing that the user can easily change the selection and combination of variables to examine different multivariate flow patterns or choose different number of regions to look at patterns at different levels. Such flexibility and interaction can efficiently support the exploration and understanding of large and complex datasets that have many variables. One may also assign different weights to variables. If only one variable is selected, then a natural-breaks classification method will be used (instead of SOM) to derive classes and assign colors.

6 Conclusion and Future Direction
This paper presents a methodological framework for the exploratory analysis and visualization of large spatial interaction data. The framework consists of methods for hierarchical regionalization, flow mapping, multivariate clustering and visualization. It supports a variety of user interaction features to facilitate exploration of spatial interactions patterns across different spaces. Using migration data analysis as a demonstrative example, the research shows that the analysis and visualization framework can process rather large data sets and discover complex and interesting patterns. The construction of natural regions (communities) is the core of the developed approach, which summarizes large spatial interaction data while preserving major patterns. The regionalization result depends on three factors: the algorithm, the modularity measure, and other constraints (such as minimum region size/population).

A reasonable concern is that how sensitive or robust the derive regions are in relation to those factors. Experiments (not included in this paper) have shown that the fine-tuning procedure creates similar regions when coupled with regionalization methods other than the ALK. However, future research is needed to better understand how close the fine-tuning result is to the global optima. The modularity measure is flexible to accommodate different definitions and models, which require domain knowledge and depend on the research questions. Different measures will lead to different regions. For example, the modularity measure used in this paper is designed to detect assortative patterns—members in each region are mostly connected to other members in the same region. It will be interesting to add other measures to detect different types of patterns (such as disassortative patterns). The region size constraint only takes effect after the maximum modularity has been achieved. A smaller size constraint will create more regions. Given the uniqueness and complexity of the patterns, which involve hierarchical structures, multivariate information, and the geographic space, future research needs to develop new strategies for user interactions for the exploration of spatial interactions.

The software will be available at www.spatialdatamining.org.

Fig. 7: User interactions. A section is made in the PCP on migration flows that have a relative high percentage of senior population (age between 65 and 69). It is strikingly clear in the flow map that senior (or retired) population tends to move to Florida (if they were from the north and east) or to Arizona if they were in the west part of the country.
ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation under Grant No. 0748813. I would also like to thank the anonymous reviewers for their insightful comments and suggestions.

REFERENCES: