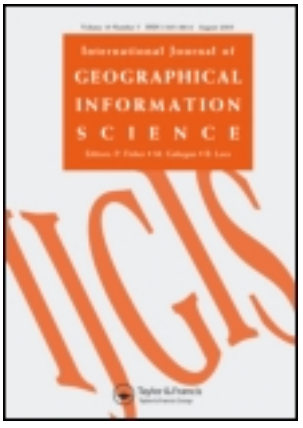
This article was downloaded by: [Arizona State University]
On: 12 January 2012, At: 13:28
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK




International Journal of Geographical Information Science

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

# Visual analytics of spatial interaction patterns for pandemic decision support

D. Guo [a]

[a] Department of Geography, University of South Carolina, 709 Bull Street, Columbia, SC 29208, USA

Available online: 17 Jul 2007

To cite this article: D. Guo (2007): Visual analytics of spatial interaction patterns for pandemic decision support, International Journal of Geographical Information Science, 21:8, 859-877

To link to this article: http://dx.doi.org/10.1080/13658810701349037

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

*International Journal of Geographical Information Science*
Vol. 21, No. 8, September 2007, 859–877

Taylor & Francis
Taylor & Francis Group

**Research Article**

# Visual analytics of spatial interaction patterns for pandemic decision support

D. GUO*
Department of Geography, University of South Carolina, 709 Bull Street, Columbia, SC 29208, USA

Downloaded by [Arizona State University] at 13:28 12 January 2012

Population mobility, i.e. the movement and contact of individuals across geographic space, is one of the essential factors that determine the course of a pandemic disease spread. This research views both individual-based daily activities and a pandemic spread as spatial interaction problems, where locations interact with each other via the visitors that they share or the virus that is transmitted from one place to another. The research proposes a general visual analytic approach to synthesize very large spatial interaction data and discover interesting (and unknown) patterns. The proposed approach involves a suite of visual and computational techniques, including (1) a new graph partitioning method to segment a very large interaction graph into a moderate number of spatially contiguous subgraphs (regions); (2) a reorderable matrix, with regions 'optimally' ordered on the diagonal, to effectively present a holistic view of major spatial interaction patterns; and (3) a modified flow map, interactively linked to the reorderable matrix, to enable pattern interpretation in a geographical context. The implemented system is able to visualize both people's daily movements and a disease spread over space in a similar way. The discovered spatial interaction patterns provide valuable insight for designing effective pandemic mitigation strategies and supporting decision-making in time-critical situations.

*Keywords*: Spatial data mining; Visual analytics; Spatial interaction; Graph partitioning; Pandemic; Decision support

## 1. Introduction

To prepare for possible outbreaks of pandemic diseases (e.g. influenza), it is critical to understand the time course and geographic spread of the outbreak and be able to design/plan effective containment strategies *before* it happens. Population mobility, i.e. the movement of individuals between specific locations and the contact between different groups of people, is one of the essential factors that determine the course of disease spread. Recently, researchers have begun to use complex simulation systems to generate near-realistic (and very large) data sets that depict individuals' daily activities and social contacts for an urban area or even the entire nation (Eubank *et al.* 2004, Ferguson *et al.* 2005, 2006, Germann *et al.* 2006). Such data hold great potential in providing unknown information regarding population mobility to assist decision-making in pandemic preparedness and response.

*Email: guod@sc.edu

*International Journal of Geographical Information Science*
ISSN 1365-8816 print/ISSN 1362-3087 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810701349037

Downloaded by [Arizona State University] at 13:28 12 January 2012

However, it is a challenging problem to analyse massive spatial interaction data, discover useful patterns, and facilitate the pandemic decision-making process. Individual-level movement data sets are often very large, containing millions of events that involve millions of people and locations. Few existing methods can deal with data of such a large volume and complexity. Traditionally, movement or interaction data are visualized with flow maps, as used in migration studies, epidemiology, and economic analysis of the flow of goods and services (Tobler 1976, Cliff and Ord 1981, Bailey and Gatrell 1995). However, flow maps are limited to relatively small data sets, e.g. migration among 50 US states (Tobler 1987).

The data analysed in this research include two parts: (1) a very large collection of human activities in the Portland metropolitan area for a normal day (simulated based on surveys and census data); and (2) a simulated pandemic outbreak for a 100-day period (Barrett *et al.* 2001, Eubank *et al.* 2004). The daily activity data set involves over 1.6 million people and more than 180 000 locations. Each record contains information for a specific activity event, including the person ID, location ID, activity type, time, duration, etc. The activities are simulated with agent-based modelling techniques, which are configured with social surveys, transportation simulations, and census data. The pandemic outbreak data set is generated via coupling epidemiological models with the above urban social contact information to simulate the time course and geographic spread of a pandemic disease over 100 days. Readers are referred to Barrett *et al.* (2001) and Eubank *et al.* (2004) for details about these datasets. (Synthetic Data Products for Societal Infrastructures and Proto-Populations: Data Set 1.0, NDSSL-TR-06-006, Network Dynamics and Simulation Science Laboratory, Virginia Polytechnic Institute and State University, 1880 Pratt Dr, Building XV, Blacksburg, VA, 24061, ndssl.vbi.vt.edu/Publications/ndssl-tr-06-006.pdf)

This research views both individual-level daily movements and the pandemic spread as spatial interaction problem, where locations interact with each other via the visitors that they share or the virus that is transmitted from one place to another. The goal of this research is to develop a general visual analytic approach to synthesize very large spatial interaction data and discover interesting (and unknown) patterns. The proposed visual analytical approach involves several visual/computational techniques: (1) a new graph partitioning method to segment a spatial interaction graph into a moderate number of spatially contiguous regions in such a way that locations in the same region interact more with each other than with locations in different regions; (2) a reorderable matrix, with regions 'optimally' ordered on the diagonal, to effectively present a holistic view of major interaction patterns/structures; and (3) a modified flow map, interactively linked to the reorderable matrix, to enable pattern interpretation in a geographic context. The proposed approach can help decision-makers understand how humans move or a pandemic spreads across space and time. The discovered patterns can then assist in designing effective mitigation strategies prior to a pandemic strike and responding to time-critical situations during an outbreak.

The remainder of the paper is organized as follows. Related work is reviewed in the next section. Section 3 provides an overview of the data and the proposed methodology. Section 4 presents the spatially contiguous graph partitioning method. The matrix-based visualization and its integration with a modified flow map are presented in section 5. Section 6 discusses the insights derived from spatial interaction visual analytics and their implications for designing effective pandemic

mitigation strategies. The paper is concluded with a summary and a discussion on future work.

## 2. Related work

Pandemic diseases (e.g. smallpox and influenza) are highly infectious and can spread across a large region or even worldwide (Potter 2001). Mitigation strategies for pandemic outbreaks can be classified into three categories (Ferguson *et al.* 2006): antiviral, vaccine, and non-pharmaceutical measures. Antiviral measures seek to promptly treat clinical cases and reduce their severity and infectiousness. Usually, cases are most infectious soon after symptoms develop, which requires very rapid reaction to new clinical cases. Therefore, early detection is critical for the containment of pandemic spread (Eubank *et al.* 2004). Vaccination protects people from being infected or from becoming a clinical case (and thus infectious). However, the effectiveness of vaccination depends on a stockpile of pre-prepared vaccine against the virus. The limited availability of vaccines and the time-consuming process to vaccinate the population require effective targeting and prioritizing to achieve maximum impact (Eubank *et al.* 2004, Ferguson *et al.* 2006). Non-pharmaceutical strategies include case isolation, household quarantine, school closure, border control, travel restriction, and other reactive responses (Ferguson *et al.* 2005, 2006).

The implementation and effectiveness of pandemic mitigation strategies depend heavily on the correct understanding of the time course and geographic spread of pandemic diseases. Population mobility in a highly fabricated urban (or even national and international), social, and spatial network is essential in determining how people contact each other and therefore how the disease spreads over time and space (Eubank *et al.* 2004, Ferguson *et al.* 2005, 2006). Various data sources, including simulation data, transportation (e.g. data from the Bureau of Transportation Statistics—BTS; Bureau of Transportation Statistics, http://www.bts.gov), and surrogate data sets (e.g. dollar-bill tracking) have been used to study the human travel patterns at different geographic scales (urban, national, and international) (Eubank *et al.* 2004, Brockmann *et al.* 2006). However, these studies have so far only focused on examining general characteristics (e.g. degree distribution) of human travels with summary statistics. Despite their vital role in shaping the pandemic spread process, spatial interaction patterns have not been examined and used adequately in designing mitigation strategies. For example, the travel restriction strategies (in combination with other strategies) evaluated by Ferguson *et al.* (2006) simply apply a certain distance threshold (e.g. 5 km) to prohibit long-range trips, attempting to contain or delay the pandemic spread.

The lack of consideration of specific spatial interaction patterns in current pandemic studies is, in part, because it is a challenging problem to synthesize, visualize, and discover patterns from *very large* spatial interaction data sets. Spatial interaction data are traditionally visualized with flow maps (Tobler 1976, 1981, 1987, Phan *et al.* 2005) or 3D views if the temporal dimension is included (Kwan 2000). (Tobler's Flow Mapper is freely available at http://www.csiss.org/clearinghouse/FlowMapper) However, a flow map (or its 3D version) has limited capability in visualizing very large data sets. Tobler (1987) discussed a variety of strategies to reduce the complexity of flow patterns and thus render legible flow maps. One approach is to show only the interaction or movement to a place or from a place. This, of course, can only present a small portion of the entire

Downloaded by [Arizona State University] at 13:28 12 January 2012

data. Another option is to remove the links that are below a certain threshold and only map the strong links or movements. A third approach is to aggregate adjacent places to reduce the data size. The fourth option is to allow flows take place only between geographical neighbours (and therefore introduce distortion in the data) to achieve cartographic neatness. In general, even with these various strategies, flow maps can only deal with relatively small data sets (e.g. 50 US states). Phan *et al.* (2005) uses multiple layers, each showing flows from (or to) a single place, which is still not capable of generating an overview of *all* flows in a large data set.

A matrix-based visualization and a graph-based view (with nodes and links) are two alternatives to visualize interaction data. Graphs and matrices are often used as parallel representations of the same data, and one can easily transform from one to another (Wong *et al.* 2006). However, studies show that when the number of vertices is larger than 20, a matrix-based visualization performs better than node-link graphs on most tasks (Ghoniem *et al.* 2004). To reduce visual clutter and simplify graphs, partitioning techniques have been designed to segment large graphs via aggregating nodes (vertices) that are strongly connected with each other (Karypis and Kumar 2000, Abou-rjeili and Karypis 2005). However, when used to partition a spatial interaction graph (with locations as nodes and interactions as links), existing graph-partitioning methods cannot guarantee that each sub-graph is spatially contiguous. There are parallel efforts in geographical research that seek to build regions from detailed geographic data (Haggett *et al.* 1977). A recent regionalization method is based on tree-partitioning (Assunção *et al.* 2006), which is similar to a graph partitioning approach but cannot handle very large data sets.

To render a matrix view, the ordering of rows (and columns) is critical (see figure 1 for an example). Originally introduced by Bertin in French in 1967 (translated in 1983), reorderable matrix is widely used in information visualization (Wilkinson 1979, Bertin 1983, Mäkinen and Siirtola 2000, Friendly 2002, Siirtola and Makinen 2005). It is an NP-hard problem to derive an optimal ordering of the rows (and/or columns) in a matrix, and so heuristic-based or approximate approaches are often used. Friendly and Kwan present a general framework for ordering information in visual displays (tables and graphs) according to the effects or trends (Friendly and Kwan 2003). An ordering can also be obtained from a hierarchical clustering result (Bar-Joseph *et al.* 2001, 2003, Guo *et al.* 2003,

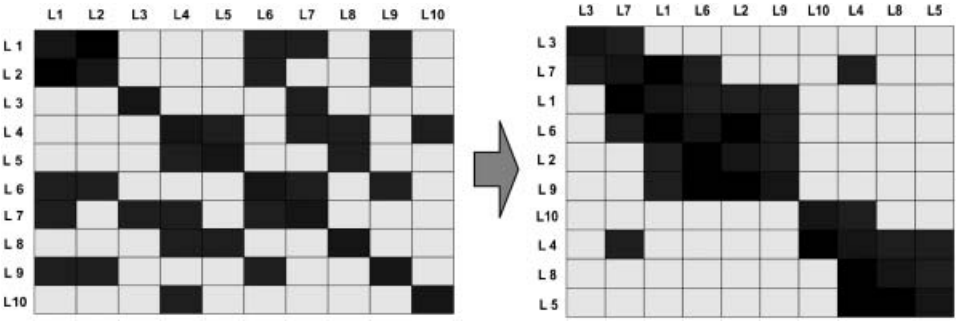



Figure 1. Illustration of the use of a reorderable matrix to visualize spatial interaction patterns among 10 locations. The two matrices show the same data, with dark colours representing strong interactions. The right matrix has columns and rows reordered according to interactions among locations.

Downloaded by [Arizona State University] at 13:28 12 January 2012

Guo and Gahegan 2006). Guo and Gahegan (2006) present an evaluation result for nine different ordering techniques, and this shows that the optimal ordering derived from a complete-linkage clustering result outperforms others in preserving patterns.

Visual analytics is an emerging research area that seeks to leverage the power of information visualization and computational analysis in the process of turning massive data into meaningful information and knowledge (Thomas and Cook 2005, Guo *et al.* 2006). Visual analytics tools and techniques are used to synthesize massive data, detect the expected and discover the unexpected, provide timely and understandable assessments, and communicate assessment effectively for action (Thomas and Cook 2006). This implies that a successful visual analytic approach should satisfy two primary requirements: (1) being able to synthesize a very large amount of data items, generate a holistic view of major structures in the data, and facilitate the discovery of meaningful (and oftentimes unknown) information; and (2) being able to transform the discovered information into forms that human users can easily recognize, interact, interpret, and ultimately integrate with domain knowledge in assisting real-world decision-making processes.

Downloaded by [Arizona State University] at 13:28 12 January 2012

## 3. Visual analytics of spatial interactions—an overview

### 3.1 *Data generation and transformation*

The data used in this paper consist of two main data files. One contains the simulated daily activities for a normal day in the Portland metropolitan area, and the other contains the outcome of a simulated pandemic spread over 100 days (Barrett *et al.* 2001, Eubank *et al.* 2004). (The data set is available at http://ndssl.vbi.vt.edu/opendata/)

The daily activity data involve over 1.6 million synthetic individual persons and 181 267 synthetic locations. The synthetic individuals carry with them a variety of demographic attributes collected from census data, including variables such as income level and age. Each of the synthetic locations, two per roadway link in the city, has an $(x, y)$ UTM coordinate. Each individual person is associated with a location as home. Although persons and locations are synthetic due to privacy concerns, they are consistent with census data at the block group level. For example, the total population or the average income for a block in the synthetic data is the same as in the census data. Each activity event includes the person ID, location (where the activity is performed), activity type, time, and duration. The simulation is based on agent-based modelling and configured with social surveys, transportation simulations, and census data.

The pandemic outbreak simulation integrates epidemiological models, urban social contact data, and agent-based techniques to simulate the time course and geographic spread of a pandemic disease over 100 days. The outcome keeps track of the path of a disease that is transmitted from person to person, including when, where, and from whom a person became infected. When a person is infected during the simulation at a specific location, the identification numbers of other infected people at the same location are recorded, and a generation number (which is one more than the minimum generation number among the infectious people present at that location) is assigned to the newly infected person.

Both individual-based daily activities and the pandemic spread can be modelled with a spatial interaction graph, where locations are linked with each other if they

share visitors or a virus is transmitted from one location to another. Two different but similar spatial interaction graphs can be extracted from the above two data files, respectively. Hereafter, they are referred to as the *activity graph* and the *spread graph*. In the activity graph, two locations are linked if they share at least one visitor. The activity graph has 181 267 locations and about 9 million links. The links in the activity graph have no direction. In the spread graph, location $A$ is connected to location $B$ if a person is first infected at location $A$ and later infects another person at location $B$. Thus, the links in the spread graph have a direction and a weight, which is the total number of people that are infected at $A$ and then spread the disease at $B$.

### 3.2 *Proposed visual analytic approach*

This research proposes a visual analytical approach for synthesizing spatial interaction data and discovering useful patterns for pandemic decision support. The approach consists of three phases: (1) graph partitioning and data reduction; (2) matrix rendering and map linking; and (3) pattern interpretation and decision-making support. Sections 4, 5, and 6 elaborate on these three phases, respectively. Below is a brief overview.

Given the large data size (i.e. 181 267 locations) and complexity (i.e. 9 million links) of the activity graph, it is not practical (or useful) to visualize the entire graph with either a node-link graph or row–column matrix view. For example, a normal screen space is about 1600 by 1200 pixels, which are several magnitudes smaller than the matrix size if each individual node (location) occupies a row (and column). Moreover, the discovery of major patterns in such a large data set often requires that the data be synthesized first to allow salient structures to emerge and be easily recognized. To reduce the graph (or matrix) while preserving major interaction patterns, this research develops a new graph partitioning method to segment the activity graph into a moderate number of subgraphs in such a way that strongly connected locations will be ideally in the same subgraph. The partitioning method guarantees that each of the subgraphs is *spatially contiguous* and that subgraphs are of similar sizes.

A normalized interaction strength value is calculated for each pair of subgraphs (i.e. regions) based on the total links connecting the two regions. Pairwise interaction strength values among all regions are visualized with a reorderable matrix, where regions are placed on the diagonal and ordered so that strongly connected regions are next to each other. This matrix view effectively presents a holistic image of the major patterns/structures in the activity graph. The matrix is interactively linked to a modified flow map. Through selection and brushing, users can understand and interpret patterns in a geographic context. The spread graph can also be visualized in a similar way except that it will use the regions derived with the activity graph and keep the same ordering of those regions. Therefore, the matrix (and map) view for the spread graph is directly comparable to that for the activity graph.

The discovered spatial interaction patterns and pandemic spread patterns provide valuable insight for designing effective pandemic mitigation strategies. Specifically, the discovered patterns can support informed decision-making on: (1) identifying critical locations and regions for a future pandemic outbreak, which are important factors for early detection and efficient targeting; (2) designing more effective travel restriction policies to delay or contain the spread while minimizing the consequent

Downloaded by [Arizona State University] at 13:28 12 January 2012

inconvenience or economic cost; (3) allocating limited resources (e.g. vaccines, personnel, facilities, etc.) geographically to effectively prepare for an imminent outbreak; and (4) facilitating decision-making in time critical situations during an outbreak, e.g. taking actions upon a new clinical case, or adjusting strategies according to updated spread information.

## 4. Partitioning very large spatial interaction graph

This section introduces the new graph partitioning method that segments the activity graph into subgraphs while preserving major interaction patterns. Specifically, the partitioning method satisfies the following four requirements: (1) clusters in the graph (i.e. sets of strongly connected locations) are preserved as much as possible; (2) each partitioned subgraph is spatially contiguous; (3) subgraphs are of similar sizes; and (4) the method is efficient to deal with the large data size. The proposed method is based on a spatially constrained minimum spanning tree clustering approach, using a unique similarity measure to quantify the inherent relationship between two locations in the activity graph. The tree is then partitioned from the top down under a size constraint to produce spatially contiguous subgraphs of similar sizes.

### 4.1 *General properties of the activity graph*

It is important to understand the general properties of the activity graph before trying to partition and visualize it. Figure 2 shows two important characteristic distributions of the activity graph. One is the degree distribution for all locations in the graph. The degree of a location is the number of links incident to that location in the graph (i.e. the number of neighbours of that location in the graph). The degree distribution follows a power-law curve (figure 2(*a*)), which means that most locations only have a few links while a small number of high-degree locations are linked to many other locations. This type of graph is also called a scale-free network (Barabási and Albert 1999), which is usually made of hub nodes (i.e. high-degree nodes) and many small nodes (i.e. low-degree nodes).

The distance distribution for all the links in the activity graph is shown in figure 2(*b*). Since each node is a geographical location, this distribution summarizes how the activity graph spans across the geographical space. It shows that most links

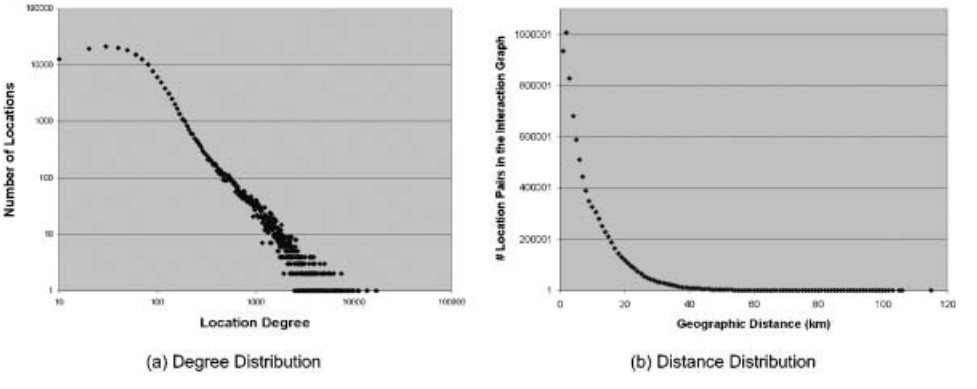



Figure 2. (*a*) Degree distribution of the activity graph. Note that both axes use a logarithmic scale. (*b*) Distance distribution of all links in the activity graph.

Downloaded by [Arizona State University] at 13:28 12 January 2012

are connecting locations that are geographically close (in terms of the straight-line Euclidean distance). The degree and distance distributions indicate that it is possible to partition the activity graph into spatially contiguous subgraphs while preserving the major structure of the original graph.

### 4.2 *Shared neighbours as similarity measure*

The similarity between two locations in the activity graph is defined as the number of neighbours that they share. This is similar to the shared nearest-neighbour (SNN) clustering approaches (Jarvis and Patrick 1973, Ertoz *et al.* 2003). The neighbours of a location in the graph are those locations that have direct links to it. A location is also a neighbour of itself. Figure 3 shows a simple scenario to demonstrate the shared-neighbour (SN) similarity measure. Location D has eight neighbours (including itself): B, C, D, E, F, G, H, and I. Location H has five neighbours: C, D, G, H, and I. The intersection of the two sets contains five locations, which are the shared neighbours of D and H. Therefore, the weight for the link between D and H is set as five.

### 4.3 *Graph partitioning via spatially constrained clustering*

The proposed graph partitioning method consists of three steps. First, a Delaunay triangulation (DT) is constructed for all locations (figure 4(*b*)), using the Guibas–Stolfi algorithm (Guibas and Stolfi 1985), which is of $O(n\log n)$ complexity. The triangulation captures the spatial neighbourhood for each location since every DT link connects two spatial neighbours. The weight for each DT link (labelled in figure 4(*b*)) is the number of shared neighbours in the interaction graph for the two locations (as defined above; see figure 3).

Second, a minimum spanning tree (MST) is constructed from the DT (Guo *et al.* 2003), which is also of $O(n\log n)$ complexity. To derive the MST, the length for each DT link is set as 1/(weight + 1). If two links have the same weight (and length), their geographic distances are used to break the tie. Note that removing any link in the

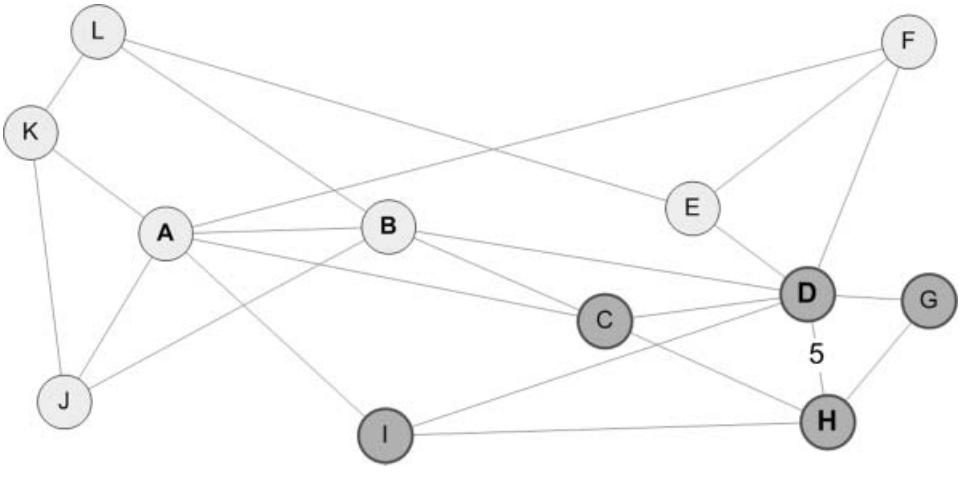



Figure 3. Illustration of the five neighbours {C, D, G, H, I} shared by two selected locations (D and H) in the activity graph.

Downloaded by [Arizona State University] at 13:28 12 January 2012

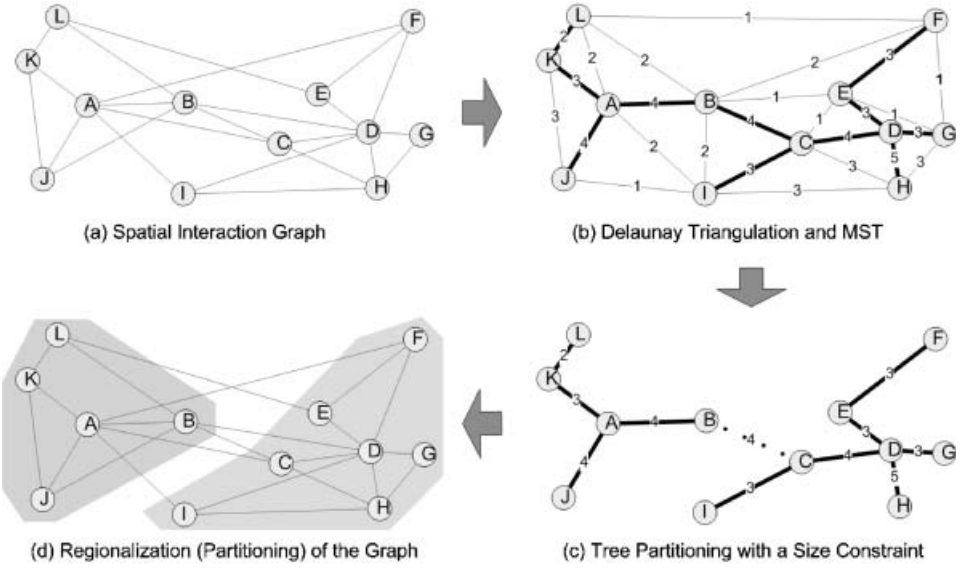



Figure 4. (*a*) Example of a spatial interaction graph. (*b*) Each link in the Delaunay triangulation (DT) is labelled with its weight, which is the number of shared neighbours (in the interaction graph) of the two incident locations. An MST is derived from the DT. (*c*) The MST is partitioned under a size constraint (figure 5). (*d*) Regionalized graph, with subgraphs (i.e. spatially contiguous regions) highlighted (see text for details).

MST will generate two subtrees, each of which is spatially contiguous (Assunção *et al.* 2006). The DT and MST are shown in figure 4(*b*).

Third, the MST is partitioned from the top down, starting from the least weighted link (figure 5). During this partition, a size constraint is enforced, which requires

| Sorted MST Edges | Edge Cut | Resulted two sub-trees (regions) | Meet Size constraint (i.e., >= 5)? |
|---|---|---|---|
| KL | Cut KL → | {L} and {A, B, C, D, E, F, G, H, I, J, K} | No |
| EF | Cut EF → | {F} and {A, B, C, D, E, G, H, I, J, K, L} | No |
| CI | Cut CI → | {I} and {A, B, C, D, E, F, G, H, J, K, L} | No |
| AK | Cut AK → | {K, L} and {A, B, C, D, E, F, G, H, I, J} | No |
| ED | Cut ED → | {E, F} and {A, B, C, D, G, H, I, J, K, L} | No |
| DG | Cut DG → | {G} and {A, B, C, D, E, F, H, I, J, K, L} | No |
| BC | **Cut BC →** | **{A, B, J, K, L} and {C, D, E, F, G, H, I}** | **Yes** |
| AJ | Remaining edges are then separated into two sets (one for each of the resulted sub-trees). Each set is processed as above until no trees can be partitioned further under the size constraint. | | |
| AB | | | |
| CD | | | |
| DH | | | |

Less weight ↓ Processing Order

Figure 5. Top-down partitioning of the MST (see figure 4(*c*)) under a size constraint of 5, i.e. each subgraph should have at least five locations. See text for details.

Downloaded by [Arizona State University] at 13:28 12 January 2012

that each subgraph cannot be smaller than the size limit. For the example shown in figure 5, the size limit is 5, and so it is not permissible to remove either KL, EF, CI, AK, ED, or DG. The first successful partition is to remove edge BC and generate two subtrees {A, B, J, K, L} and {C, D, E, F, G, H, I} (figures 4(*c*) and 5). After a successful partition, the remaining (i.e. unvisited) MST links are separated into two sets, one for each subtree. Each subtree is then recursively processed until no subtree can be further partitioned under the size constraint. According to the tree partitioning result, the input graph is partitioned into a set of spatially contiguous subgraphs (figure 4(*d*)).

The complexity analysis of the proposed graph partitioning method is as follows. Let $n$ be the total number of locations. The total number of links in the MST is $n-1$. The partitioning method essentially visits each MST link once and obtains the sizes for the two subtrees if that link is cut (which takes $O(|T|)$ time, with $|T|$ being the subtree size). The worst-case complexity of the partitioning method is $O(n^2)$. IsIdeally, $|T|$ is reduced by a half after each partitioning, in which case the method is much more efficient. The actual running time for partitioning the activity graph (which has 181 267 locations) is about 10 min on a desktop computer with 2 GB of RAM memory and a 3.60-GHz Pentium 4 CPU.

The activity graph is partitioned into 258 subgraphs (regions) with a size constraint of 500. Region sizes range from 500 to 1252, while most regions are of a size between 500 and 800. The 258 regions are shown in figure 6, which are used for subsequent analysis and visualization. One can also try a different size constraint

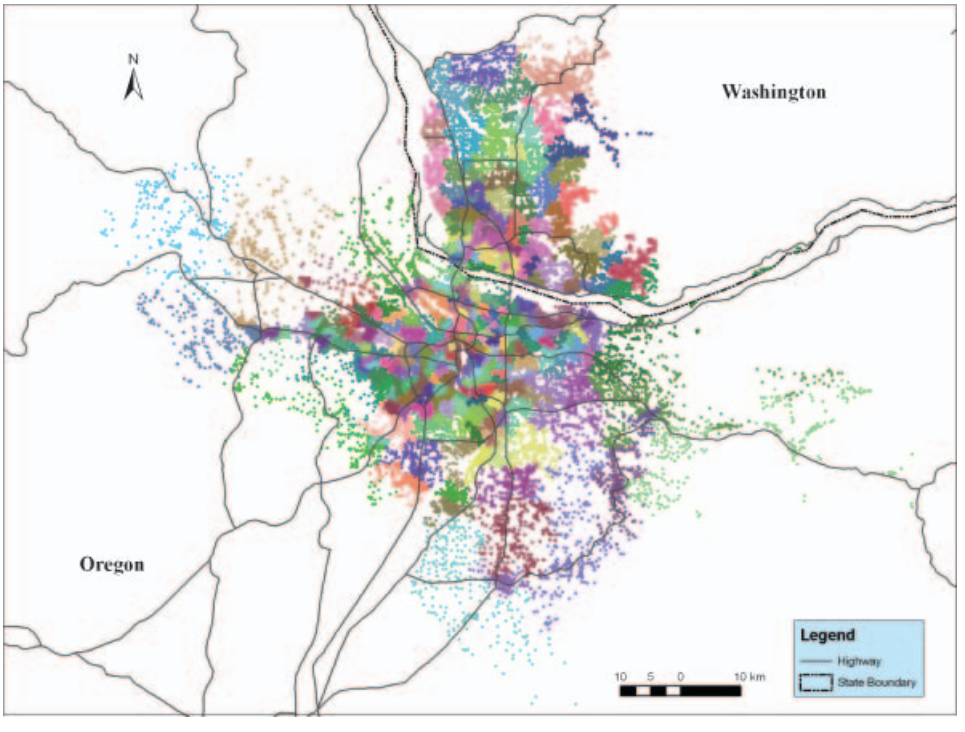



Figure 6. The partitioning result of the activity graph. 181267 locations in Portland are partitioned into 258 regions based on the activity graph. Locations of the same region are in the same colour. Major highways are also shown. Due to the limited number of different colours, it is possible that two nearby regions are in the same colour.

Downloaded by [Arizona State University] at 13:28 12 January 2012

and produce more or less regions to examine patterns at different granularity levels. However, since a minimum spanning tree has a well-known 'chain effect' problem (Hastie *et al.* 2001, Guo *et al.* 2003), the graph partitioning method cannot guarantee that all locations within a region are strongly connected *if the region is large*. The primary advantage of this partitioning method is that it is fast and works reasonably well to derive *small* regions. Therefore, this research synthesizes very large spatial interaction datasets with two steps. The first step is to segment the interaction graph into a moderate number of small regions with the proposed graph partitioning method. The second step further synthesizes and visualizes the interactions among regions with a better (but more time-consuming) clustering and ordering method, which is used to construct a reorderable matrix and present a holistic view of interaction patterns at various hierarchical levels (see section 5).

Downloaded by [Arizona State University] at 13:28 12 January 2012

## 5. Visualizing spatial interactions

### 5.1 *A measure of interaction strength*

Now, the original interaction data (i.e. the activity graph) is reduced to a 258 by 258 matrix. A normalized interaction strength measure for a pair of regions is defined as:

$$I_{ab} = C_{ab}\left(1000^2/(S_a S_b)\right) \tag{1}$$

where $I_{ab}$ is the interaction strength between region $a$ and $b$, $C_{ab}$ is the number of links (in the original activity graph) that connect the two regions, and $S_a$ and $S_b$ are the region size for $a$ and $b$, respectively. This measure is a normalized value that takes into account region sizes. For example, a value of 500 for $I_{ab}$ indicates that 500 links are expected between the two regions if each of them contains 1000 locations. Note that $a$ and $b$ can be the same region, in which case the interaction strength measures the internal interactions within that region.

### 5.2 *Linear ordering of regions*

As introduced in section 2 and demonstrated in figure 1, a linear ordering of the rows (and columns) is critical for a matrix-based visualization to accentuate patterns. Given the matrix of pairwise interaction strength values for all regions, a one-dimensional ordering of the 258 regions is derived using an ordering method that is based on the complete-linkage hierarchical clustering (Guo and Gahegan 2006).

The clustering-based ordering method first groups regions into a hierarchy of clusters according to the interaction strength matrix (which can be viewed as a similarity matrix in the clustering context). However, a cluster hierarchy cannot determine a unique ordering—there are as many as $2^{n-1}$ ($n$ is the number of regions) different orderings that are consistent with the same cluster hierarchy (i.e. regions in the same cluster are also contiguous in the ordering) (Bar-Joseph *et al.* 2001). The ordering method seeks to find an 'optimal' ordering by placing strongly connected regions as close as possible to each other while maintaining the cluster hierarchy (Guo and Gahegan 2006). Therefore, the ordering encodes more information and preserves more patterns than a cluster hierarchy does.

Figure 7 shows the linear ordering of the 258 regions based on their mutual interaction strengths. Note that geographic distances are not involved in the derivation of the ordering. However, the ordering largely follows a geographic

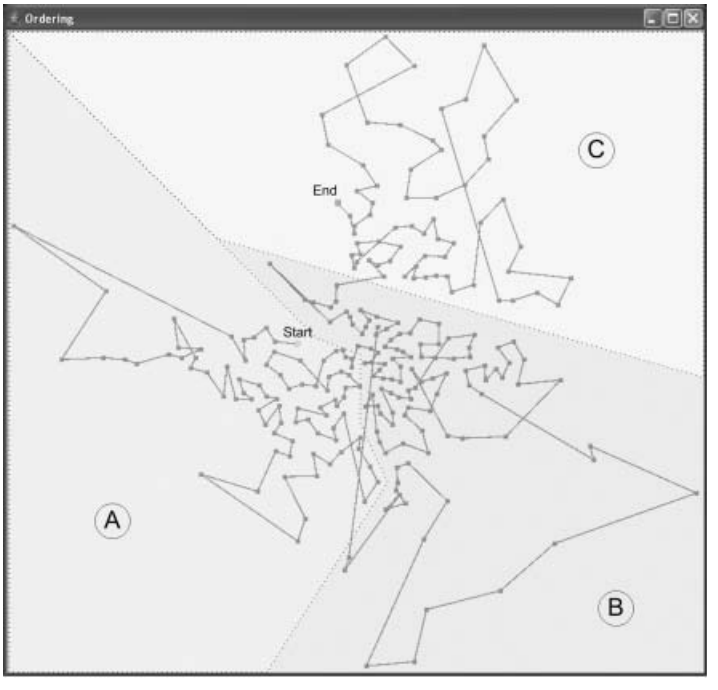



Figure 7. Linear ordering of the 258 regions (represented by their centroids). According to the patterns shown in the matrix view (figure 8), three top-level clusters of regions are highlighted and labelled. Note that these three clusters (A, B, and C) can be further decomposed into smaller (and more strongly connected) regions, which are obvious in the matrix (figure 8) but not recognizable in the maps (figures 6–7 and 10).

order, i.e. the ordering tends to connect spatial neighbours. This indicates (or verifies) that strong interactions are mostly local. This spatially dominated ordering helps the user understand and interpret patterns in the matrix view (see section 5.3).

### 5.3 *Visualization and user interactions*

Given the linear ordering of regions derived above, it is rather straightforward to visualize the matrix of pairwise interaction strengths for all regions. The matrix view shown in figure 8 is constructed by placing the 258 location clusters on the diagonal, following the one-dimensional ordering, with the start region (figure 7) at the top-left corner of the matrix and the end region in the bottom-right corner. Each grid cell, representing the interaction strength between the row region and the column region, is assigned a colour according to a five-class classification. Higher interaction strengths are represented with a darker colour.

This matrix is an overview of the major spatial interaction patterns. One can see that there are distinct clusters of regions that have strong interactions internally but far fewer connections to the outside. For example, each of the top three clusters (marked in both figures 7 and 8) forms a dark rectangular area along the diagonal in the matrix. Specifically, such an ordered matrix view can help human users recognize two important types of interaction patterns that may be difficult to see otherwise (e.g. with a flow map alone). First, one can recognize the hierarchical structure of interactions among regions—smaller clusters of regions are often

Downloaded by [Arizona State University] at 13:28 12 January 2012

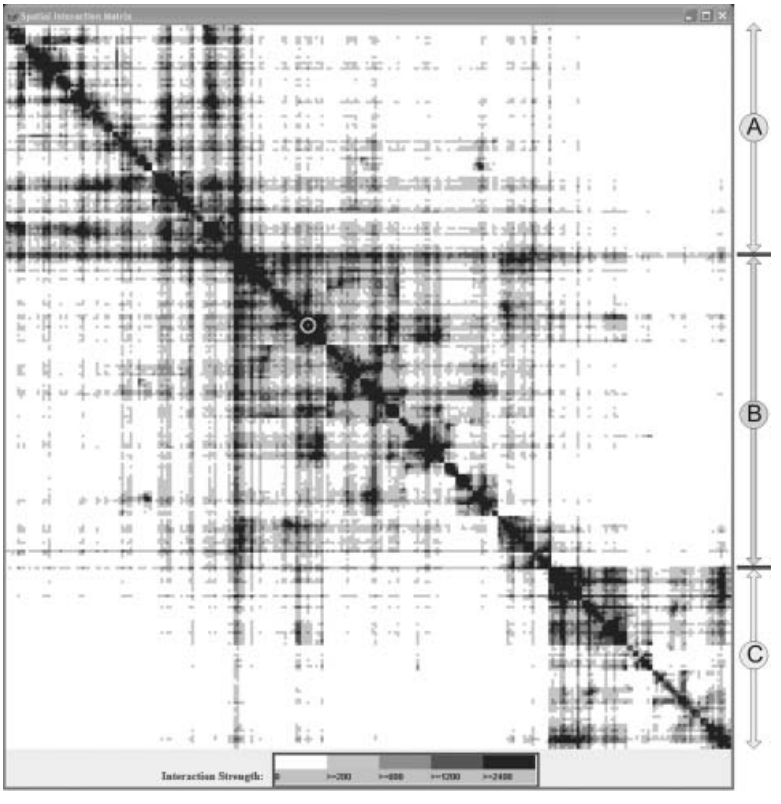



Figure 8. Matrix view of the activity graph, with locations aggregated into 258 regions. The circle in the matrix marks a selected region, which is shown in a flow map in figure 10(*a*). Three obvious top-level clusters of regions are labelled with A, B, and C, each of which contains a hierarchy of internal clusters. Note that the matrix ordering is consistent with the hierarchical clustering structure, and strongly connected regions are close to each other on the diagonal.

Downloaded by [Arizona State University] at 13:28 12 January 2012

embedded inside a larger cluster of regions. Second, one can also recognize hub regions that have strong connections to many other regions across different clusters. The importance and usefulness of these patterns will be explained in section 6 in relation to pandemic response.

To reveal patterns regarding the geographic spread of pandemic disease, the spread graph can be visualized in a similar way. Note that a link in the spread graph has a direction and weight (section 3.1). The regions derived with the activity graph and the linear ordering of the regions are reused here. According to the spread graph, the interaction strength from region $A$ to region $B$ is the total number of people that are infected at locations in region A and later spread the disease at locations in region B. This measure quantifies how severely region B is affected by Region A during the pandemic spread. The interaction strength measure is then normalized in the same way as shown in equation (1) (section 5.1). Note that the spread interaction strength is not symmetrical, i.e. the interaction strength from $A$ to $B$ may not the same as that from $B$ to $A$.

Figure 9 shows the spread graph for the entire 100-day period (aggregated to a $258 \times 258$ matrix). Although the matrix is not symmetrical, it is very close to a symmetrical matrix. Despite the difference in the interaction strength measure and

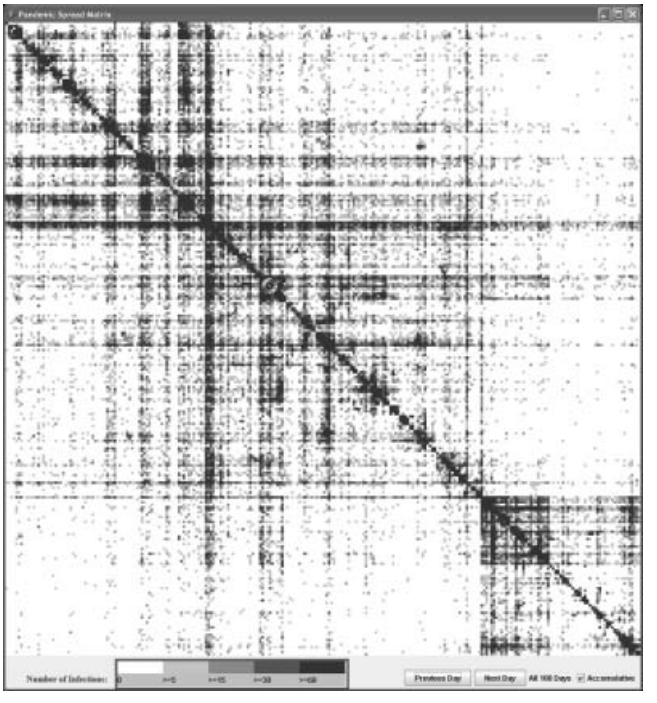


Figure 9. Pandemic spread patterns for the entire 100-day period. Each pixel represents the normalized spread interaction strength, i.e. the normalized number of infections (see text for explanation). The circle marks a selected location, which is shown in a flow map (figure 10(*b*)).

the classification scheme, the spread matrix (figure 9) and the activity matrix (figures 8) exhibit a great similarity. For example, the distinct clusters of regions that have strong internal interactions (figure 8) match very well with the regions observed in the pandemic spread matrix (figure 9). The hub locations observed in figure 8 also stand out clearly in figure 9. Given the unlikely availability of timely and realistic pandemic spread information, daily population movement patterns are among the most important information that can be used to predict the pandemic spread process and thus help design effective containment policies (especially for geographically targeted prophylaxis) before the outbreak and make response decisions during the break.

However, a matrix view (as shown in figures 8 and 9) alone cannot effectively present interaction patterns within a geographic context. On the other hand, as reviewed in section 2, flow maps are generally effective for visualizing one-to-many flows but not many-to-many flows. Therefore, this research links a matrix view to a flow map and allows user interactions to select part of the matrix to visualize in maps. Specifically, user interactions can be supported at three levels: cell selection, row/column selection, and window selection.

At the cell level, the user can mouse-click an individual cell in the matrix, which represents the interaction between an origin region (*O*) and a destination region (*D*), to highlight the two in the flow map (with a large cross representing *O* and a small cross for *D*). The colour of *O* in the map indicates the internal interaction strength within *O*, while the colour of *D* indicates the flow (or interaction) from *O* to *D*.

Downloaded by [Arizona State University] at 13:28 12 January 2012

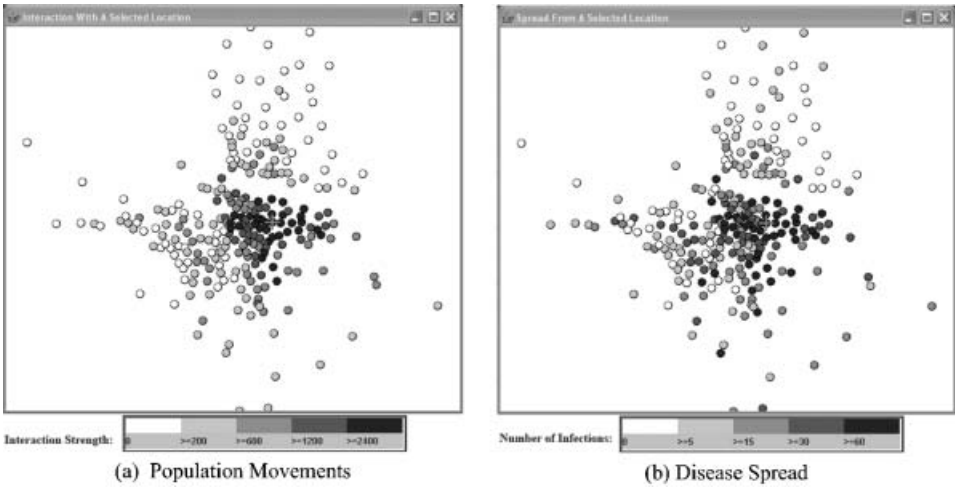



Figure 10. (*a*) Flow map of daily population movement from a selected location (marked with a cross) to other locations. (*b*) Disease spread from the same location to all other locations.

Instead of using colours, an alternative approach is to draw a directional edge from $O$ to $D$, with the edge width proportionally configured using the interaction strength.

At the row/column level, the user may select a row (i.e. an origin) or a column (i.e. a destination) to visualize interactions between the selected region and all other regions. For example, a region is selected in the activity matrix (figure 8), and its corresponding colour-based flow map (figure 10(*a*)) shows interactions from that location to all other locations, using the same classification and colour scheme as used in the matrix. The same location is also selected in the spread matrix (figure 9), and its corresponding flow map is shown in figure 10(*b*), which shows how the disease spreads from the selected location to other regions. Similarly, directional edges (instead of colours) may be used to create a traditional flow map to show flows from one location to all other locations.

The user can also select a rectangular area in the matrix to show flows from one group of regions to another group of regions. However, to map flows for such a window selection, only the traditional flow map (with edges and arrows) can be used. Note that, if the selection window is too large, the resulting flow map can be too cluttered to convey any useful information. To summarize, user interactions with both the matrix and the flow map can help the user understand the spatial context of the information shown in the matrix and thus facilitate a comprehensive interpretation of the discovered patterns.

The spread graph can be broken down to each day for the 100-day period. One can animate the spread matrix day by day or view the accumulated patterns up to a certain day. Figure 11 shows the cumulative spread patterns for the first 2 weeks and for the first month. It is clear that similar patterns emerge at a very early stage during the spread. Note that the classification scheme in figure 11 is proportionally reduced from that used in figure 9 (which is for the entire 100-day period) due to the small number of infections at the early stage. One may also notice that most of the infections early in the spread occurred locally (i.e. within the same region).

Downloaded by [Arizona State University] at 13:28 12 January 2012

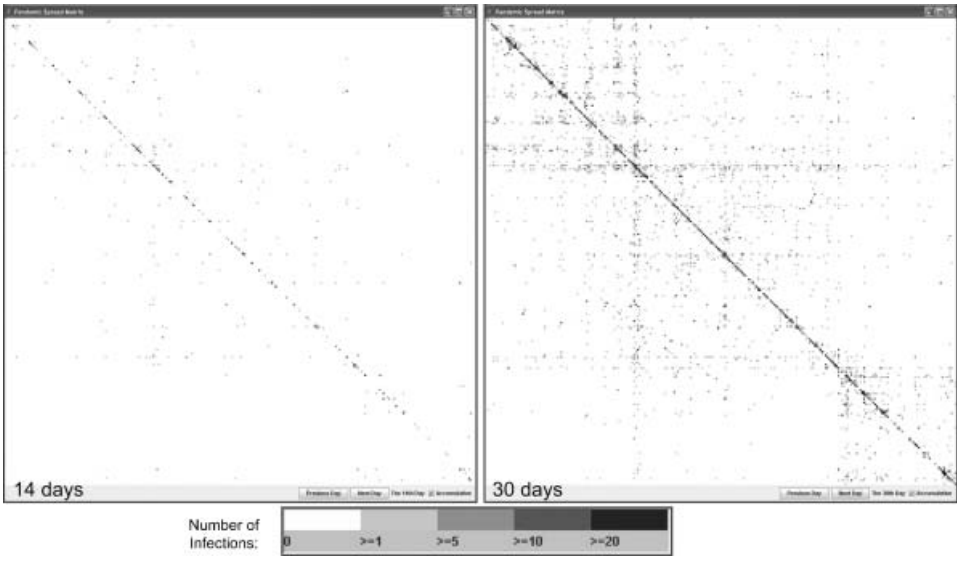



Figure 11. Cumulative spread patterns for the first two weeks (left panel) and for the first month (right panel). Note that the classification scheme used here is proportionally smaller than that used in figure 9 owing to a small number of infections at the early stage.

## 6. Implications for pandemic preparedness and decision support

As reviewed in section 2, mitigation strategies for pandemic outbreaks can be classified into three categories: antiviral, vaccine, and non-pharmaceutical measures. Due to limited resources and time constraints, the success of any of these measures depends on efficient targeting, early detection, and proper response in a timely manner. The discovered spatial interaction patterns and their relationship to pandemic spread patterns provide a valuable insight for designing effective pandemic mitigation strategies and helping meet the above challenges. Specifically, the discovered patterns can support an informed decision-making process in four major areas (in addition to other possible uses).

First, spatial interaction patterns can help identify critical locations, regions, and/or clusters of regions, for controlling a pandemic outbreak. For example, the hub locations/regions identified in the matrix view (figure 8) are important places to install sensors for early detection of infectious people (Eubank *et al.* 2004) and to reduce infections that can initiate infection in previously unaffected regions. Strongly connected clusters shown in the matrix view suggest that each of them should be treated as a unit, since it is very difficult to separate the locations inside a cluster.

Second, the discovered patterns can help decision-makers design more effective travel restriction policies to delay or contain the spread while minimizing the consequent inconvenience or economic cost. Existing travel restriction policies, e.g. those suggested by Ferguson *et al.* (2006), often use a simple distance threshold (e.g. 5 km) to prohibit long-range trips. However, the spatial interaction patterns (figure 8) and the spread patterns (figure 9) indicate that the most dangerous trips are those involving different regions (or clusters of regions). For example, a trip from cluster B to cluster C (figure 8) is more dangerous than any trip within B or C, regardless of the trip distance. It is also practically easier to implement a policy that

Downloaded by [Arizona State University] at 13:28 12 January 2012

prohibits trips between well-defined geographic regions than to prohibit trips longer than a certain distance. Moreover, since there are significantly fewer population movements between clusters than within each cluster, prohibiting or reducing trips across clusters can cause relatively less inconvenience than other restrictions do.

Third, the interaction patterns can help geographically optimize the allocation of limited resources to prepare for an imminent pandemic outbreak. For example, the coverage of each emergence centre should be assigned according to the discovered hierarchical structure of regions. In other words, a strongly connected region should be covered by the same centre. Vaccines and human resources can also be allocated according to the size and priority of each region (or cluster of regions) to reduce response time and achieve maximum impact.

Fourth, they can help achieve effective targeting in time critical situations during an outbreak. For example, when an infectious individual is reported at a certain place, the discovered cluster structure of spatial interactions can help identify the region(s) of the highest priority for immediate action, e.g. distribution of vaccines, vaccinating the population of that area, and/or travel restrictions to that area.

Downloaded by [Arizona State University] at 13:28 12 January 2012

## 7. Conclusion and future work

This research proposes a visual analytical approach to explore spatial interaction patterns in very large datasets of individual-based population movements and simulated pandemic spreads. The approach combines graph partitioning, linear ordering, matrix-based visualization, and flow maps to synthesize, visualize, and interpret spatial interaction patterns. The discovered spatial interaction patterns provide valuable insights and can potentially improve pandemic mitigation strategies and support decision-making in time-critical situations. The proposed approach is efficient in processing very large data sets and effective in presenting a holistic view of all major patterns, which can then guide more focused and detailed exploration.

Results are promising and yet reveal a need for future work. Further research is needed to develop an effective approach for the exploration and presentation of dynamic interaction patterns that evolve over time. The research reported here relies on a user-controlled animation to visualize the pandemic spread day by day but is not able to present an overview of patterns over time (and space). It will also be useful to examine interaction and spread patterns in relation to other attribute information (e.g. age, income, activity type, etc.)

It remains an interesting and challenging research problem to design a comprehensive and systematic user interaction strategy to support the discovery and understanding of various spatial interaction patterns. There are two major challenges. One is that spatial interaction patterns are diverse, complex, difficult to visualize, and oftentimes non-intuitive to recognize. Another challenge is that related computational and visual approaches are still in their forming stage. The proposed visual analytic approach needs improvements to facilitate more intuitive linking and interaction between the matrix view and the flow map. How can the matrix incorporate spatial information so that users can easily recognize spatial patterns from the matrix? What is the most intuitive way to link a flow map and a matrix to facilitate the understanding of the overall spatial interaction patterns?

Another important area that calls for future attention and research efforts is to develop a visual analytic platform for designing, testing, and evaluating mitigation or containment strategies in a highly interactive and iterative manner, through

dynamic integration of simulation models, newly discovered patterns, and domain expert knowledge.

**Acknowledgement**

This research was partially supported by the United States Department of Homeland Security through the National Consortium for the Study of Terrorism and Responses to Terrorism (START), grant number N00140510629. However, any opinions, findings, and conclusions or recommendations in this document are those of the authors and do not necessarily reflect views of the US Department of Homeland Security.

Downloaded by [Arizona State University] at 13:28 12 January 2012

**References**

Abou-rjeili A. and Karypis, G., 2005, Multilevel algorithms for partitioning power-law graphs. Technical Report (TR 05-034). Minneapolis, MN, Department of Computer Science and Engineering, University of Minnesota.

Assunção, R.M., Neves, M.C., Câmara, G. and Freitas, C.D.C., 2006, Efficient regionalization techniques for socio-economic geogrpahical units using minimum spanning trees. *International Journal of Geographical Information Science*, **20**, pp. 797–811.

Bailey, T.C. and Gatrell, A.C., 1995, *Interactive Spatial Data Analysis* (New York: Wiley).

Barabási, A.-L. and Albert, R., 1999, Emergence of scaling in random networks. *Science*, **286**, pp. 509–512.

Bar-Joseph, Z., Demaine, E.D., Gifford, D.K., Hamel, A.M., Jaakkola, T.S. and Srebro, N., 2003, K-ary clustering with optimal leaf ordering for gene expression data. *Bioinformatics*, **19**, pp. 1070–8.

Bar-Joseph, Z., Gifford, D.K. and Jaakkola, T.S., 2001, Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, **17**(Supplement 1), pp. S22–S29.

Barrett, C., Beckman, R., Berkbigler, K., Bisset, K., Bush, B., Campbell, K., Eubank, S., Henson, K., Hurford, J., Kubicek, D., Marathe, M., Romero, P., Smith, J., Smith, L., Speckman, P., Stretz, P., Thayer, G., Eeckhout, E. and Williams, M.D., 2001, TRANSIMS: Transportation Analysis Simulation System, Technical Report LA-UR-00-1725. Los Alamos National Laboratory Unclassified Report.

Bertin, J., 1983, *Semiology of Graphics: Diagrams, Networks, Maps* (Madison, WI: The University of Wisconsin Press).

Brockmann, D., Hufnagel, L. and Geisel, T., 2006, The scaling laws of human travel. *Nature*, **439**, pp. 462–465.

Cliff, A.D. and Ord, J.K., 1981, *Spatial Processes: Models and Applications* (London: Pion).

Ertoz, L., Steinbach, M. and Kumar, V., 2003, Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Third SIAM International Conference on Data Mining (SDM '03)*, San Francisco.

Eubank, S., Guclu, H., Kumar, V.A., Marathe, M., Srinivasan, A., Toroczkai, Z. and Wang, N., 2004, Modeling disease outbreaks in realistic urban social networks. *Nature*, **429**, pp. 180–184.

Ferguson, N.M., Cummings, D.A., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S. and Burke, D.S., 2005, Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, **437**, pp. 209–214.

Ferguson, N.M., Cummings, D.A.T., Fraser, C., Cajka, J.C., Cooley, P.C. and Burke, D.S., 2006, Strategies for mitigating an influenza pandemic. *Nature*, **442**, pp. 329–484.

Friendly, M., 2002, Corrgrams: exploratory displays for correlation matrices. *The American Statistician*, **19**, pp. 316–324.

FRIENDLY, M. and KWAN, E., 2003, Effect ordering for data displays. *Computational Statistics & Data Analysis*, **43**, pp. 509–539.

GERMANN, T.C., KADAU, K., LONGINI, I.M. Jr. and MACKEN C.A., 2006, Mitigation strategies for pandemic influenza in the United States. *Proceedings of the National Academy of Sciences*, **103**, pp. 5935–5940.

GHONIEM, M., FEKETE, J.-D. and CASTAGLIOLA, P., 2004, A comparison of the readability of graphs using node-link and matrix-based representations. In *Proceedings of the IEEE Symposium on Information Visualization*, pp. 17–24.

GUIBAS, L. and STOLFI, J., 1985, Primitives for the manipulation of general subdivisions and the computation of Voronoi diagrams. *ACM Transactions on Graphics*, **4**, pp. 74–123.

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**, pp. 1461–1474.

GUO, D. and GAHEGAN, M., 2006, Spatial ordering and encoding for geographic data mining and visualization. *Journal of Intelligent Information Systems*, **27**, pp. 243–266.

GUO, D., PEUQUET, D. and GAHEGAN, M., 2003, ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata. *GeoInformatica*, **7**, pp. 229–253.

HAGGETT, P., CLIFF, A.D. and FREY, A., 1977, *Locational Analysis in Human Geography* (London: Arnold).

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J., 2001, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (New York: Springer).

JARVIS, R.A. and PATRICK, E.A., 1973, Clustering using a similarity measure based on shared near neighbours. *IEEE Transactions on Computers*, **22**, pp. 1025–1034.

KARYPIS, G. and KUMAR, V., 2000, Multilevel k-way hypergraph partitioning. *VLSI Design*, **11**, pp. 285–300.

KWAN, M.P., 2000, Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. *Transportation Research Part C-Emerging Technologies*, **8**, pp. 185–203.

MÄKINEN, E. and SIIRTOLA, H., 2000, Reordering the reorderable matrix as an algorithmic problem. In *Theory and Application of Diagrams, Diagrams 2000, Lecture Notes in Artificial Intelligence 1889, Edinburgh, Scotland, September 2000*, pp. 453–467 (Berlin: Springer).

PHAN, D., XIAO, L., YEH, R. and HANRAHAN, P., 2005, Flow map layout. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium*, pp. 219–224.

POTTER, C.W., 2001, A history of influenza. *Journal of Applied Microbiology*, **91**, pp. 572–579.

SIIRTOLA, H. and MAKINEN, E., 2005, Constructing and reconstructing the reorderable matrix. *Information Visualization*, **4**, pp. 32–48.

THOMAS, J. and COOK, K.A., 2006, A visual analytics agenda. *IEEE Computer Graphics and Applications*, **26**, pp. 10–13.

THOMAS, J.J., and COOK, K.A. (Eds), 2005, *Illuminating the Path: The Research and Development Agenda for Visual Analytics* (Los Alametos, CA: IEEE Computer Society).

TOBLER, W.R., 1976, Spatial interaction patterns. *Journal of Environmental Systems*, **6**, pp. 271–301.

TOBLER, W.R., 1981, A model of geographical movement. *Geographical Analysis*, **13**, pp. 1–20.

TOBLER, W.R., 1987, Experiments in migration mapping by computer. *American Cartographer*, **14**, pp. 155–163.

WILKINSON, L., 1979, Permuting a matrix to a simple pattern. In *Proceedings of the Statistical and Computing Section of the American Statistical Association*, pp. 409–412.

WONG, P.C., FOOTE, H., MACKEY P., PERRINE, K. and CHIN, G. Jr., 2006, Generating graphs for visual analytics through interactive sketching. *IEEE Transactions on Visualization and Computer Graphics*, **12**, pp. 1386–1398.

Downloaded by [Arizona State University] at 13:28 12 January 2012