Mining and Visualizing Spatial Interaction Patterns for Pandemic Response *

Diansheng Guo†

Abstract

This paper views the movements of people among locations as a spatial interaction problem, where locations interact with each other via shared visitors. The connection between two specific locations can be weighted with several alternative weight definitions. The purpose of this research is to analyze the general characteristics of such spatial interactions and propose a strategy to explore and visualize such a large volume of data and reveal important patterns. The proposed strategy combines clustering, ordering, and visual techniques to efficiently present overall patterns and provide an interface for visual data mining and decision-support in response to possible disease outbreaks.

1 Introduction

The movements of individuals between specific locations and the contacts between different groups of people are essential in modeling disease spread [7]. The daily activity of people from place to place forms a complex and dynamic network of spatial interactions between locations.

The understanding of such spatial interactions can be difficult due to several reasons. *First*, the interaction network is extremely complex and difficult to model as many factors are involved and such factors also change over time. *Second*, such interaction data are often very difficult to acquire. However, researchers have begun to generate or discover such data sources, for example, generating simulation data of people's daily activities [3], [7] or using surrogate information (e.g., bank notes) to model human travel activities [5]. *Third*, such interaction data is often very large, unique, and complex (in terms of potential patterns), which demands special data mining algorihtms to process and effective visual approaches to reveal/present the patterns. Few existing methods can cope with such a large data volume and complexity.

The data used in this paper is a very large collection of simulated human activities in an urban setting for a normal day, which includes over 8 million records and each record shows a specific activity (e.g., the person ID, location ID, activity type, time, duration, etc.). There are altogether over 1.6 million people and 181,295 unique locations [7]. This data set is available from http://ndssl.vbi.vt.edu/opendata.

This activity data set can be analyzed from different perspectives. This paper views such activities as a spatial interaction problem [2], [6], where locations interact with each other via the visitors that they share. §2 introduces several basic definitions and parameters to characterize such a spatial interaction network. In §3 I present an overall analysis of the characteristics of the spatial interaction network formed with the simulation data. Based on the general analysis of the spatial interaction properties, §4 proposes a strategy to aggregate data, synthesize information, and visually present patterns to allow human interpretation and decision-support during pandemic outbreaks.

2 Problem Definition

The activity data that involve N locations and P people is tranformed to a $N \times N$ spatial interaction matrix. If the dynamics of interactions across time are also considered, then we have a space-time system defined by $N \times N \times T$, where T is time (in seconds). To quantify the "strength" of interaction between two specific locations, I define two measures in this section. The simple one is to use the "flow" of people between locations, i.e., the number of shared visitors between two locations. The other one uses the percentages of visitors they share instead of the raw count (see §2.3).

2.1 Bipartite graph. Bipartite graphs are often used to model the people-location relations [7]. As shown in Figure 1, for each day one person may visit several locations (places) and a location (place) may have many different visitors. However, existing analyses of a bipartite graph often focus on either people or locations and examine the degree distribution for each node [7], [5]. How to define the relationship between locations that do not have direct connections in the barpartite graph? In this paper, two locations are "connected" if they share at least one visitor for a normal day. This definition is similar to that presented

*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 U.S. Department of Homeland Security.

†Department of Geography, University of South Carolina. 709 Bull Street, Columbia, SC 29208. Email: guod@sc.edu

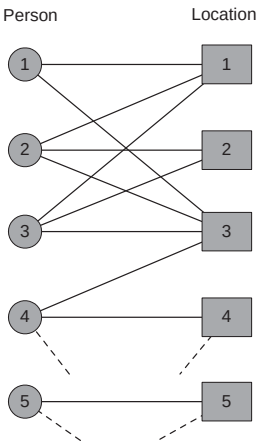

Figure 1: A bipartite of people's daily activities (ignoring time). Two locations are connected if they share at least one visitor. For example, locations 1 and 3 share three visitors (i.e., person 1, 2, and 3).

in [16] but is different from that defined in [7] where two locations connect only if there is at least one person whoe moves *directly* from one location to the other location.

2.2 Spatial interaction. Not all elements in the $N \times N$ spatial interaction matrix have values. Some locations do not share any visitor and thus have no connection to each other. The value for each connection, i.e., the weight (or connection strength) between two locations, can be defined by the number of visitors (P_s) they share, the location pair strength (see next subsection), or simply the geographic distance. Thus, we can view the spatial interaction matrix as a graph/network of locations. The density of a graph (see [17]) is defined as:

$$\Delta = 2L_{pc}/(L(L-1)), \tag{2.1}$$

where Δ is the graph density, L_{pc} is the number of edges (or connections) in the graph, and L is the total number of nodes (or vertices) involved. Here I define a similar but simpler measure, *connection ratio* (C_r):

$$C_r = L_{pc}/L, \tag{2.2}$$

which is more efficient to calculate for a very large data set. As we know that we need at least $L-1$ connections to connect L locations . Therefore, if C_r is close to 1.0, the network or graph is barely connected and easy to break.

2.3 Location pair strength. Instead of using the number of shared neighbors to define the connection weight, we also calculate a location pair strength (L_{ps}), as defined below:

$$L_{ps} = 10000P_s^2/(L_aL_b), \tag{2.3}$$

where P_s is the number of visitors between two specific locations a and b, and L_a and L_b are the total visitors to locations a and b, respectively. The constant 10000 is just to scale the measure so the it ranges from 0 to 10000. This measure takes into acount the location size (in terms of daily visitors) and the number of shared visitors. In other words, the strength is defined by the percentage of visitors that two locations share. Intuitively, the higher percentage they share, the more likely that those people contact each other.

3 General Properties

In this section I present a series of analysis of the location network to understand some general characteristics of spatial interactions. Such general characteristics are important for developing a successful strategy to explore and visualize the data and patterns.

3.1 Location pair count vs. shared visitors. Figure 2 shows the total number of location pairs (L_{pc}) that share exactly P_s visitors. The two variables exhibit a strong power-law relationship. This indicates that the spatial interaction network can be dramactically reduced and simplied if we remove location pairs that share less than a certain number of neighbors (e.g., 5).

Figure 3 shows the connection ratio C_r for all location pairs that share P_s visitors. It indicates that C_r can be dramactically decreased if weak connections (e.g., $P_s < 5$) are discarded.

3.2 Geographic distance vs. shared visitors. Figure 4 presents the relationship between the number of shared visitors (P_s) and the average distance of all location pairs that share exactly P_s visitors. Both the average and the standard deviation are shown. it shows that location pairs that share one or two visitors vary greatly in distances, with an average distance of about 8km. Location pairs that share more visitors tend to be closer in space. Again, it shows roughly a power-law trend, which means that strong connections (in terms of shared visitors) are often localized.

3.3 Location pair strength. Figures 5 and 6 are similar to Figures 2 and 3, respectively, except that the number of shared visitors (P_s) is replaced with the location pair strength (L_{ps}).

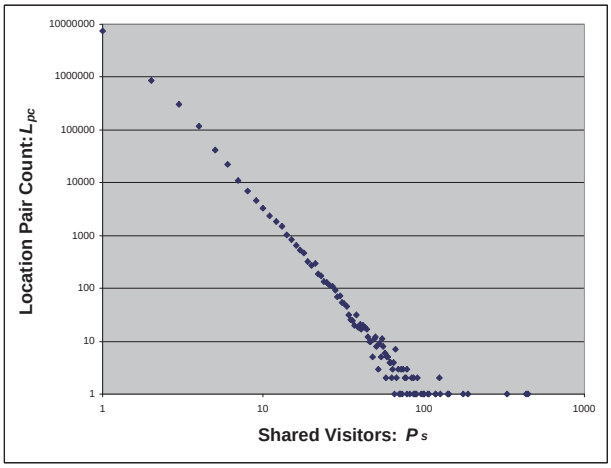

Figure 2: The total number of location pairs (L_{pc}) that share exactly P_s visitors.

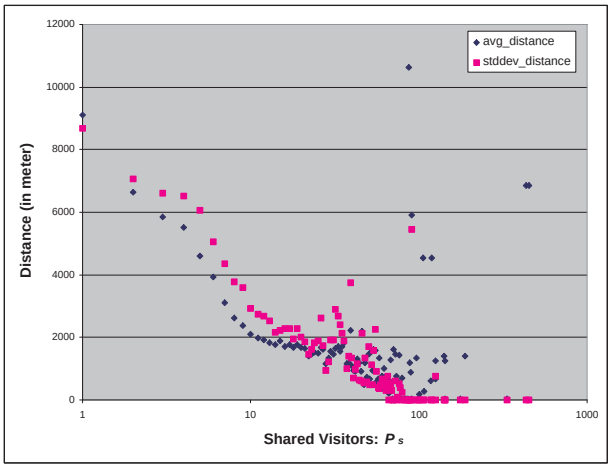

Figure 4: The relationship between the number of shared visitors (P_s) and the average distance of all location pairs that share exactly P_s visitors. Dark blue points are the average distances and pink points are the standard deviation values.

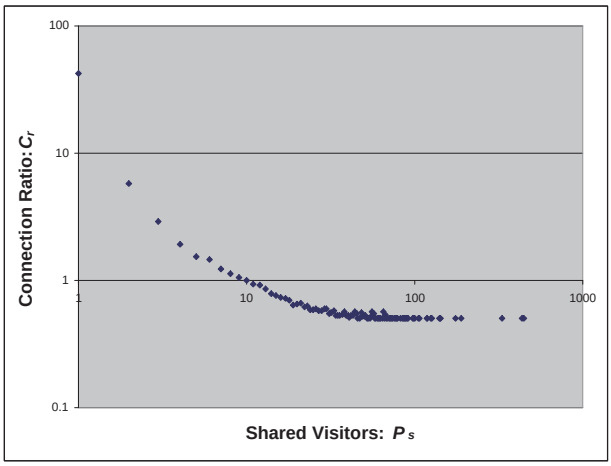

Figure 3: The connection ratio C_r for all location pairs that share P_s visitors.

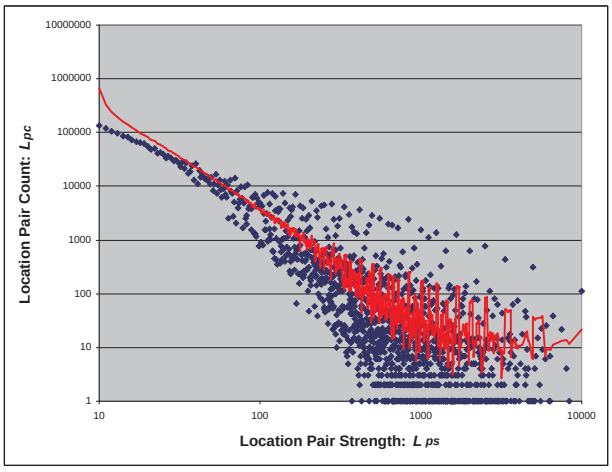

Figure 5: The relationship between the location pair strength (L_{ps}) and the number of all location pairs that have that strength. The red curve shows the moving average (with a window of size 10).

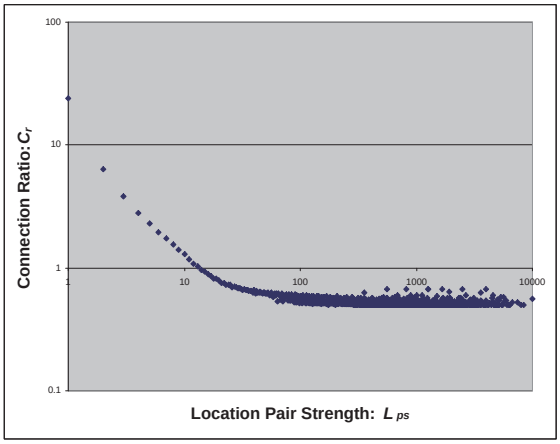

Figure 6: The connection ratio C_r for those location pairs that have a strength of L_{ps}

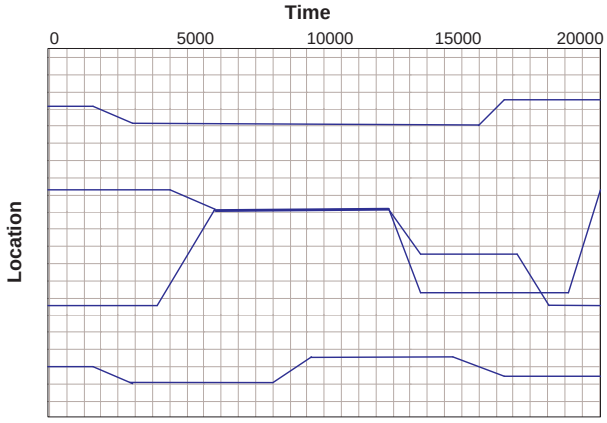

Figure 7: Space-time view of people's movements.

4 Mining and Visualizing Interaction Patterns

The above analysis result indicates taht it is possible to segment and partition the spatial interaction network to synthesize the data, explore patterns, and visulize patterns in an aggregated form. However, there are two major challenges to achieve this goal. First, the large data size (millions of records) requires that the data mining algorithm should be scalable [8]. Second, given the complexity of potential patterns in the $N \times N \times P \times T$ (i.e., location-location-people-time) space, effective visual approaches are needed help human users (or decision-makers) to interpret patterns [9], [10].

4.1 Hierarchical clustering of locations. Given the spatial interaction network/graph, locations are first grouped into a hierarchy of clusters. Graph-based clustering methods are potential candiated for this task [8], [11], [4]. However, most of these clustering methods are often of complexity $O(n^2 \log n)$ or $O(n^3)$, which is not efficient enough to process over 180,000 locations. Although the sinlge-link method is of complexity $O(n \log n)$, it is generally not as good as other clustering methods.

A two-step clustering may be adopted. The purpose of the first step is to aggregate locations, via sampling [14], simplifying (e.g., removing weak location connections), or partitioning [1]. Then the second step is to perform a hierarchical clustering with the much smaller set of aggregates (or samples).

4.2 1D Ordering of locations. An ordering can also be obtained from a hierarchical clustering result. A cluster hierarchy, represented by a dendrogram, is a binary tree with each data point as a leaf node. To derive a one-dimensional ordering from a cluster hierarchy, several different methods are available [9], [4]. The purpose of the 1D ordering is twofold.

First, to discover patterns accross several differents dimensions (e.g., the $N \times N \times P \times T$ space), we have to project some dimensions to a lower-dimensional space. In this case, we project the spatial locations to a 1D space so that we can add time to the view (see Figure 7).

Second, a one-dimensional ordering can preserve more information than a cluster hierarchy. An ordering tries to arrange locations so that locations with strong connections (more shared visitors, high pair strength, or short distance) are placed as close as possible to each other in the ordering. For example, if the weight for location connections is defined as the number of shared visitors, then locations that share many people will be close to each other in the ordering. If a person visits places that are far from each other in the ordering, then the activity of this person can be seen as an "outlier" [15].

4.3 Visualizing people movements across space and time. Now since locations are ordered in a 1D space, a space-time view can be constructed, with time as the horizontal axis and locations on the vertical axis. Then the daily movement of a person can be drawn as curve in the space-time view (see Figure 7). We can also aggregate nearby locations in the ordering to reduce the number of rows. Such an aggregation is possible because locations that are close in the ordering are also strongly connected in the interacion graph. Thus, we have a spatio-temporal view of all people.

However, given the huge number of people involved

(¿1.6 million), clustering is again needed to simplify the view yet perserve major patterns. In addition to derive clusters of people according to their interactions with locations, we can also resort to a 'visual' approach [13], which visualizes densities (i.e., how many person overlap at each location-time pixel) instead of each individual person.

This space-time view can facilitate the decision-making process in response to a pandemic outbreak. For example, suppose there are 10 person infected at the very beginning. We can show the activities of the 10 person for the past 24 hours in the space-time view. Then people that 'overlap' with one of those 10 person are selected and this view can potentially help identify the most severe areas to take immediate actions. Based on simulation data, it may also show the projected developement for the next 24 hours if no action was taken.

References

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

[2] T. C. Bailey and A. C. Gatrell, *Interactive spatial data analysis*, John Wiley & Sons, Inc., New York, NY, 1995.

[3] C. Barrett, J. Smith, and S. Eubank, *Modern Epidemiology Modeling*, Scientific American, March 2005.

[4] Z. Bar-Joseph, E. D. Demaine, D. K. Gifford, A. M. Hamel, T. S. Jaakkola, and N. Srebro, *K-ary Clustering with Optimal Leaf Ordering for Gene Expression Data*, Bioinformatics, 19 (2003), pp. 1070-1078.

[5] D. Brockmann, L. Hufnagel, and T. GEisel, *The scaling laws of human travel*, Nature, 439 (2006), pp. 462-465.

[6] A. D. Cliff and J. K. Ord, *Spatial Processes: Models and Applications*, Pion, London, UK, 1981.

[7] S. Eubank, H. Guclu, V. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, *Modeling Disease Outbreaks in Realistic Urban Social Networks*, Nature, 429 (2004), pp. 180-184.

[8] D. Guo, D. Peuquet, and M. Gahegan, *ICEAGE: Interactive Clustering and Exploration of Large and High-dimensional Geodata*, GeoInformatica 7 (2003), pp. 229-253.

[9] D. Guo, *Coordinating Computational and Visualization Approaches for Interactive Feature Selection and Multivariate Clustering*, Information Visualization, 2 (2003), pp. 232-246.

[10] D. Guo, M. Gahegan, A. M. MacEachren, and B. Zhou, *Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach*, Cartography and Geographic Information Science, 32 (2005), pp. 113-132.

[11] J. Han, M. Kamber, and A. K. H. Tung, *Spatial Clustering Methods in Data Mining: a survey*, Geographic Data Mining and Knowledge Discovery, edited by H. J. Miller and J. Han, Taylor & Francis, London and New York, pp. 33-50, 2001.

[12] N. M. Ferguson, D. A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke, *Strategies for Containing an Emerging Influenza Pandemic in Southeast Asia*, Nature, 437 (2005), pp. 209-214.

[13] J. Johansson, P. Ljung, M. Jern, and M. Cooper, *Revealing Structure within Clustered Parallel Coordinates Displays*, Proceedings of the 2005 IEEE Symposium on Information Visualization (INFOVIS'05), pp. 17-25, 2005.

[14] D. Rafiei and S. Curial, *Effectively Visualizing Large Networks Through Sampling*,Proceedings of the IEEE Visualization 2005 - (VIS'05), pp. 48-56, 2005.

[15] S. Shekhar, C.-T. Lu, and P. Zhang, *A Unified Approach to Detecting Spatial Outliers*, GeoInformatica 7 (2003), pp. 139-166.

[16] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos, *Neighborhood Formation and Anomaly Detection in Bipartite Graphs*, Fifth IEEE International Conference on Data Mining (ICDM'05), pp. 418-425, 2005.

[17] S. Wasserman, and K. Faust, *Social Network Analysis*, Cambridge University Press, Cambridge, UK, 1994.