Final Report

Data Science and Traffic Patterns

Fall 2017

Faculty Name: Dr. Richard Sowers

Project Leaders: Vipul Satone

IGL Scholars: Ziying Wang, Wenting Xu, Anthony Fontana, Pranav Bhardwaj, Jiahao Zhu, Kexin Fang, Eden Brewer

Persistent Homology:

The idea of Persistent Homology comes from computing topological features of shapes by counting the holes in them. If they share the same number of holes, by twisting or some other deformation methods, we consider them as same shape. In this research, Persistent Homology is applied in the following manner:

1. By changing the speed in one fixed time slice, we analyze the topological properties of the congestion regions of New York City.
2. By releasing the time restriction, we try to analyze those properties dynamically.

To consider the congestion region of New York City, the connected components and strongly connected components (SCC) with the average speed of specific link as criteria were visualized. Whole New York City was considered as a graph and each link of New-york city was defined as edge and each node as a vertex.

A barcode represents each persistent generator with a horizontal line beginning at the first filtration level where it appears, and ending at the filtration level where it disappears. An undirected graph is connected if there is a path between any two vertices, and a disconnected graph consists of several components. Two vertices are in the same component if and only if there exists a path between them. However, considering traffic flow has direction, and many roads in New York city are one-way, we then used Strongly Connected Component(SCC). A directed graph is strongly connected if there is a directed path from any vertex to any other.

By setting specific speed criterion, we select those links whose average speed is lower than the criterion and count the number of strongly connected components. Intuitively, for two arbitrary nodes i, j in one strongly connected component, we can always find a path from i to j or a path from j to i. Figure 1 is an example of barcode on 2011-01-01 at 6:00 AM with a speed criterion of 6 mph and Figure 2 is the corresponding map and SCC. In Figure 1, The x-axis represents the average speed (mph), and the y-axis represents the components. Smaller y value represents that the component emerged earlier. The length of each bar represents the life of each bar, while the left end of each bar represents the emergence of that components and the right end of each bar represents the disappearance of that components. When two components merge under a certain criterion, we define the younger components merges to the older one.

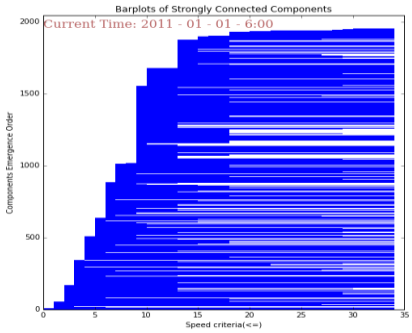

Figure 1: Example of Barcode

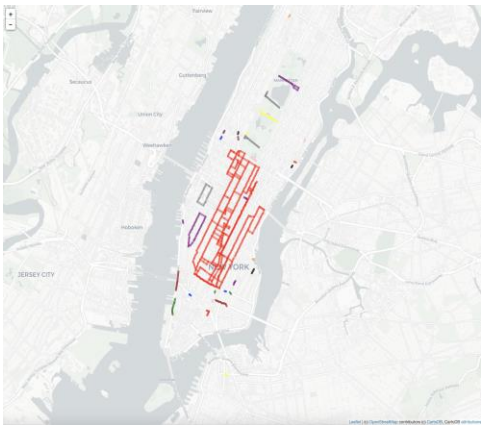

Figure 2: Visualization of strongly connected components

Intuitively, the merger represents that the expansion of the congestion region. More traffic is expected during major events and therefore mergers should be more intense if special events happen. The previous algorithm has been applied to data on the day of NYC Marathon and a Concert in 2011.

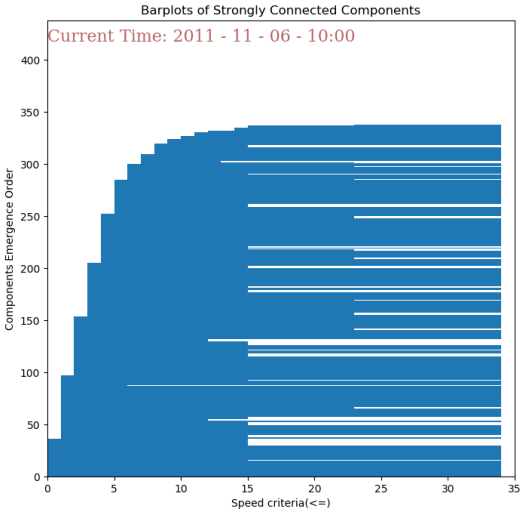

Figure 3: Barcode on day of event

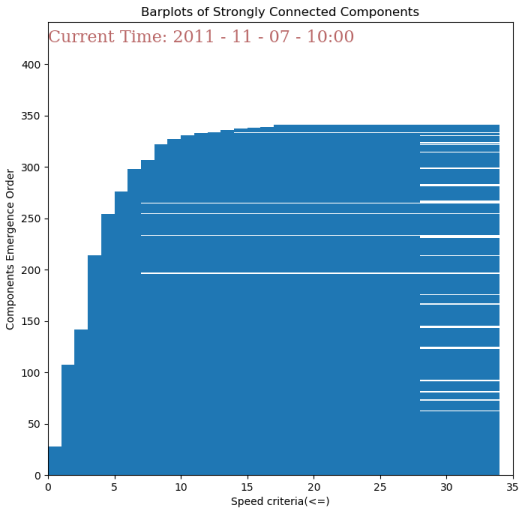

Figure 4: Barcode on regular day

From the above two barcodes, it can be seen that the performance of the congestion components are different. The merger of components is more intense in Figure 1 (on the day of event), compared to Figure 2 the day after event (regular day).

Future Directions:

We can visualize the congestion through barcodes. In the future we should be able to visualize strongly connected components before and after major events in congested regions. From visualization of congestion near the event area using past data we should be able to predict future congestion patterns allowing us to take precautions to avoid it in advance.

References:

Erickson, J. http://jeffe.cs.illinois.edu/teaching/algorithms/

Parking:

The goal of the parking team is to identify patterns in parking occupancy throughout several different urban settings in hopes of predicting parking demand and better understanding urban planning in the future.

Freedom of Information requests were submitted to several cities asking for parking data from 2010 onward. About 157,000 lines of parking transactions were received from Seattle from January 1-7, 2015. Over 1,048,000 lines of parking transactions were received from San Francisco from January 2012 to May 2013.

Non-negative matrix factorization (NMF or NNMF), is a group of algorithms in multivariate analysis and linear algebra where a matrix D (the data matrix) is factorized into two matrices W and H, with the property that all three matrices have no negative elements. For our problem we use the Sparse Non Negative Matrix Factorization. SNMF can be used to enforce sparseness on the basis/mixture matrices, i.e, on the W and H matrices. We use the Nimfa python library for implementing SNMF which uses a fast non negativity constrained least squares algorithm. Consider data matrix , D , of shape T x L. The rows of D is indexed by time and the columns are indexed by location. The entries of D represent the parking lot attributes (occupancy of parking lot / number of vehicles entering a parking lot). SNMF factorizes this into matrix W of shape T x N and H of shape N x L , where N is the rank of the matrix factorization. W represents the elementary behavioral signatures and H represents the coefficients.

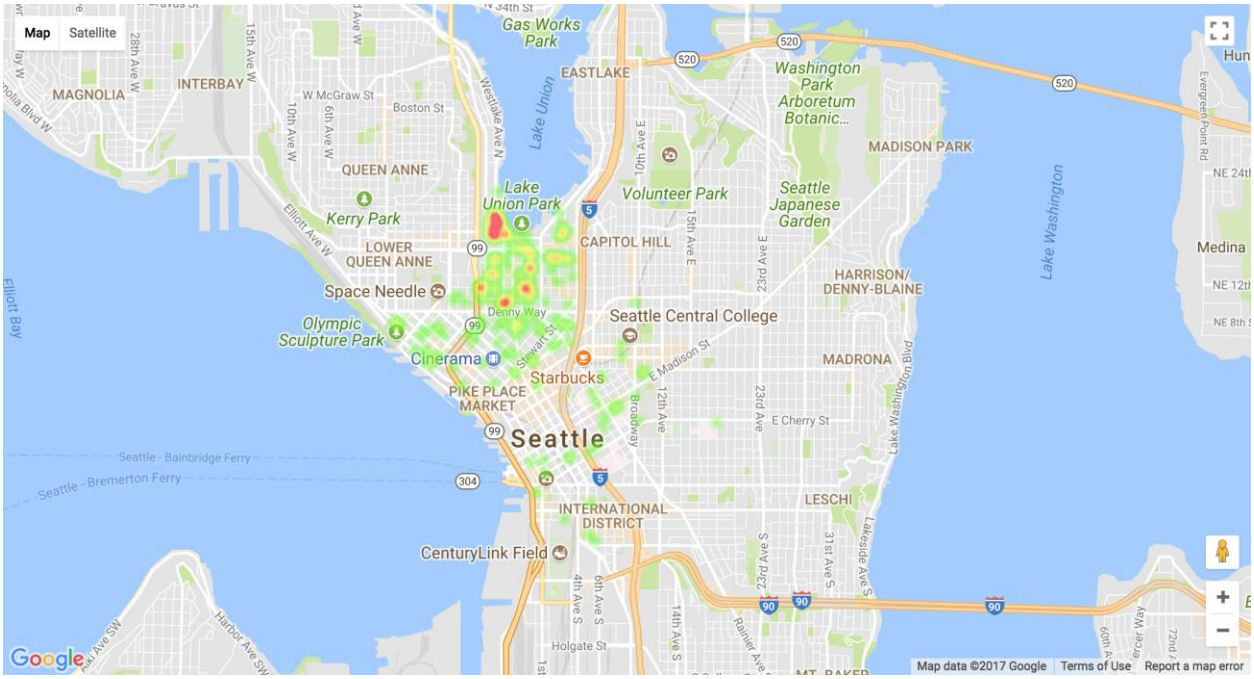

Figure 5: This is a parking density heatmap showing Seattle at 6:30 a.m. on Jan 2, 2015. A high parking density can be observed around Lake Union, which may be the result of people gathering for exercises in the morning.

Future Directions:

The parking team is currently in the final stages of optimizing the parameters (such as rank, beta-controls sparsity, etc.) to get the best possible factorization. Following this, we will be able to visualize and hence identify various parking trends in San Francisco and Seattle. These trends vary from different times of the day, weeks of the month, and months of the year. We may observe peculiar behavior in the parking trend which may be attributed to unique events such as holidays, sporting events, and parades.