Dimensionality Reduction, Clustering and PlaceRank Applied to Spatiotemporal Flow Data

Blake Shaw
Tony Jebara
Columbia University
Sense Networks

Introduction

Motivation

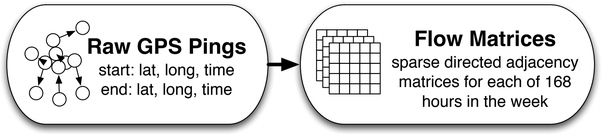
On average over 250,000 yellow taxi cab trips are recorded every day in New York City, capturing the collective movement patterns of millions of individuals. This massive high-dimensional dataset necessitates tools such as dimensionality reduction and clustering algorithms in order to better understand the flow patterns of an urban area.

New York City Taxi Cab Data

The dataset consists of 22.5 million trips spanning 6 months between January and June 2009. Each trip record contains the latitude, longitude, and timestamps for the start and endpoints. For our analysis we consider the 2000 busiest city blocks, each of which has a minimum of 20 pickups or dropoffs per day.

Preprocessing with Hadoop

We use an 8-node Hadoop cluster to process over 50GB of raw GPS logs, producing sparse directed adjacency matrices which capture the flow between the 2000 busiest city blocks for each of the 168 hours of the week.



A_{ij}^{t} - the number of trips between place i and place j at weekhour t

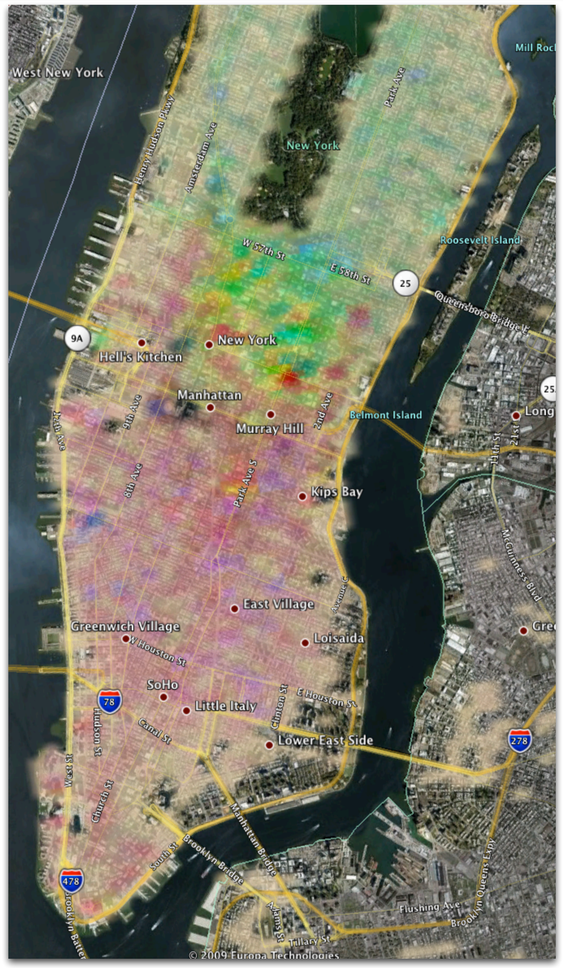
Similarity Between Places

Two places can be considered similar if they have similar amounts of flow to/from other places at similar times of the week, thus yielding the linear kernel:

$$W = \sum_{t=1}^{168} \frac{1}{2}\left(A^t A^{t\top} + A^{t\top} A^t\right)$$

Cab Density Map for Madison Square Garden / Penn Station

MVE



Dimensionality Reduction with MVE

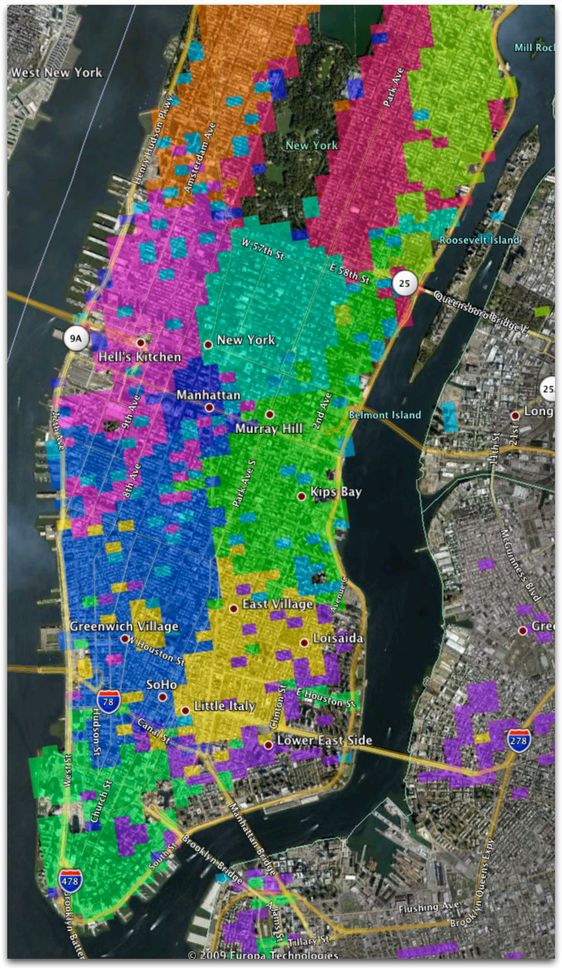
Given W we can apply Minimum Volume Embedding (MVE) to reduce the original 672,000 dimensional flow vectors to 3 dimensional vectors, and use these as RGB values to visualize the smooth modes of variation in the flow patterns of places in the city. MVE preserves over 99% of the variance of the original data using only 3 dimensions.

MVE alternates between an SDP and an SVD optimization to maximize the amount of variance of the data that is captured by the low-dimensional embedding while preserving local distances measured along the natural manifold of the data.

$$\max_{K\in\mathcal{K}} F(K) \quad = \quad \max_{K\in\mathcal{K}} \sum_{i=1}^{d} \lambda_i - \sum_{i=d+1}^{N} \lambda_i$$

$$\mathcal{K} = \begin{cases} K \in \mathbb{R}^{N\times N} \\ K \succeq 0 \\ \sum_{ij} K_{ij} = 0 \\ K_{ii} + K_{jj} - K_{ij} - K_{ji} = W_{ii} + W_{jj} - W_{ij} - W_{ji} \\ \forall_{i,j} \text{ s.t. } C_{ij} = 1 \end{cases}$$

Spectral Clustering



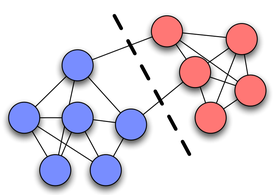
Spectral Clustering

Similarly, we can apply spectral clustering to the kernel W to assign one of k=12 labels to each place, thus visualizing the natural neighborhoods that emerge out of the data.

To perform spectral clustering we first form the Laplacian:

$$L = D - W \quad D = \text{diag}(W\vec{1})$$

and then apply k-means on the k eigenvectors of L with the smallest non-zero eigenvalues. Spectral clustering approximates finding the smallest normalized cut on the graph.



PlaceRank

Hubs **Authorities**

Hubs and Authorities

Similar to how PageRank finds the most authoritative places on the web, we can compute PlaceRank to find the most authoritative places in the physical world. The figures above show heat maps representing the hubs and authorities in the cab flow network which were computed by finding the stationary vector of the total flow adjacency matrix and its transpose using the power method. Places with high authority values have inbound traffic from many other high authority places; similarly places with high hub values have outbound traffic to many other hubs.

References

[1] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Neural Information Processing Systems 14*, 2001.

[2] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web, 1999.

[3] Blake Shaw and Tony Jebara. Minimum volume embedding. In Marina Meila and Xiaotong Shen, editors, *Proc. of the 11th International Conference on Artificial Intelligence and Statistics*, volume 2 of JMLR: W&CP, pages 460–467, March 2007.

[4] Blake Shaw and Tony Jebara. Structure preserving embedding. In Léon Bottou and Michael Littman, editors, *Proceedings of the 26th International Conference on Machine Learning*, pages 937–944, Montreal, June 2009. Omnipress.