characterizing & representing New York City’s taxi transportation network

By

Nilesh Patil & Jiang Shang

Dr. Gourab Ghoshal

Progress Report

November 18, 2016

# Characaterizing & representing New york city’s taxi transportation network

**1. Introduction:**

Large scale transportation networks offer a fascinating opportunity to identify local population’s travel habits, aggregated daily routines and a way to augment city-planning decisions. In our analysis, we have focused on determining travel patterns of New York City residents from about 14 million taxi trips.

Our core assumption behind this analysis is that given the large number of rides being used, a generalization of travel patterns is plausible. For each element in the network, a trip to a specific hub should contribute to a specific type of travel pattern.

Since taxis represent an on-demand travel option, our analysis of taxi trips provides an insight into characterizing travel network congestion and regularities in the network.

**2. Prior work:**

GPS based transportation networks have been studied in detail for traffic flow analysis and determining social dynamics [1].

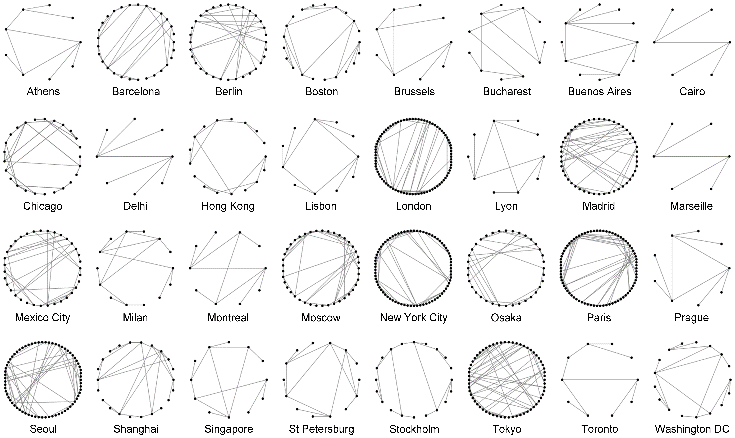
Bike sharing datasets have been used for clustering locations based on the usage profile [2] and predicting bike demand [3]

GPS based taxi datasets have been used to identify mobility patterns in Shanghai, China [4]

In [4], the trip distribution has been characterized as combination of 3 independent types and non-negative matrix factorization has been used to identify 3 patterns from 1.58 million trips in Shanghai, China. This is the core of our approach, as we are attempting to characterize taxi usage in New York City with a similar dataset.

Prior analysis has also been done to produce transportation network graphs from street geometries [5] and subway maps [6]. As our final submission, we also hope to submit a similar centrality based structural map, but created from current annual usage patterns instead of geometric features. Ideally, the network structure we generate from usage pattern would look as follows:

Fig 01: Circular layout of metro systems [6]



**3. Data Description:**

1. **Raw data:**

New York city taxi & Limousine Commission has made the taxi trips dataset available for public use since 2009 onwards[7] . We have used this dataset to perform our analysis.

The data contains ~150 million trips for each year & each row represents one trip, with features for starting and stopping point, distance travelled, taxi-charge, time taken etc. We are using 2015 dataset which contains about 146112990 trips in total.

1. **Data transformation:**

We use the following variable for each trip:

* 1. Trip starting timestamp
  2. Start point (Lat/Long)
  3. Trip stopping timestamp
  4. Stop point(Lat/Long)
  5. Charges

From these variables, we use the following subset of trips we select the subset of trips which originate and terminate in manhattan. Using these points, we build our directed graph with starting and stop points as nodes.

As an additional condition, we have started with most frequent trips (>500 in the year for a given pair of start & stop points). From prior work we summarized that rounding off location coordinates to 2 decimal points is also feasible and given our difficulties analyzing the dataset with 40,000+ nodes, we are now in the process of reducing our network by 2 separate approaches:

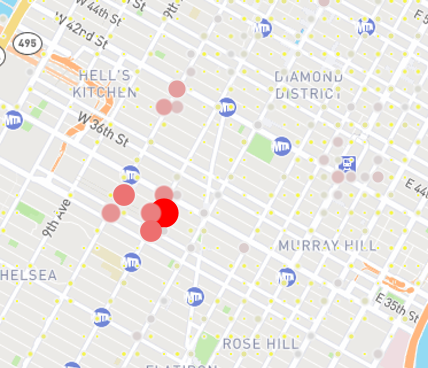
1. One node for 5 manhattan blocks
2. Using 6 million most frequent trips to get 1275 most frequently travelled edges.

In our current network, each node represents 200m x 200m around it and each edge represents the total number of trips between two nodes in a given year.

We create features for month, day, wekday, period of the day etc from the timestamp. For this graph based on the annual dataset, we realized that we’ll have to manually tag multiple locations as a single node to filter out segmentation of trips into nearby locations.

As shown in the following plot, a big chunk of trips originate at or around Penn Station, due to our previous approach, they were getting divided into multiple nodes, thus making it diffult to cluster.

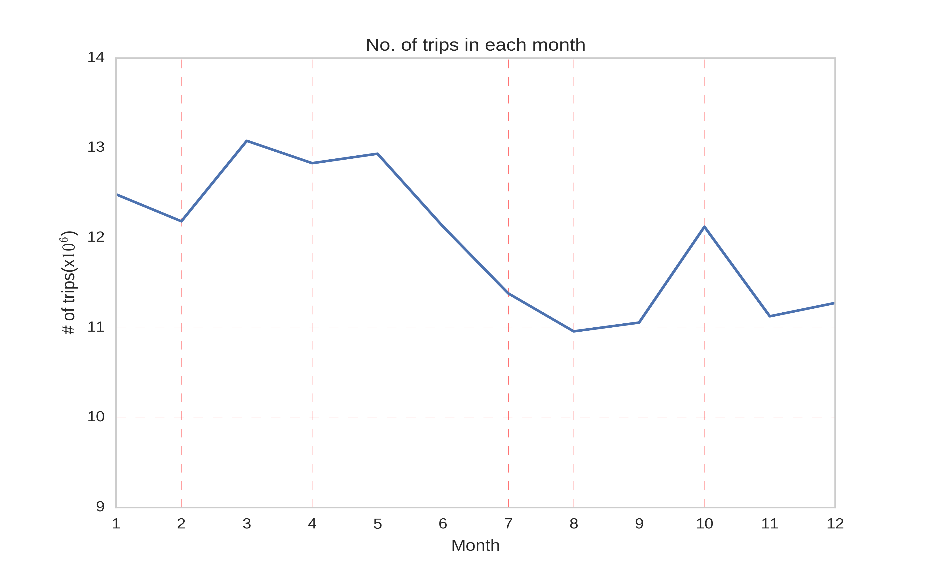
Same location : multiple nodes



We are now manually tagging such locations (going through the 1275 locations in our graph) and as an alternate, working on the automated decimal aggregation. Since the rest of our process remains same, comparison between these 2 approaches should take about 2-3 days and we’ll run our analysis for the final graph.

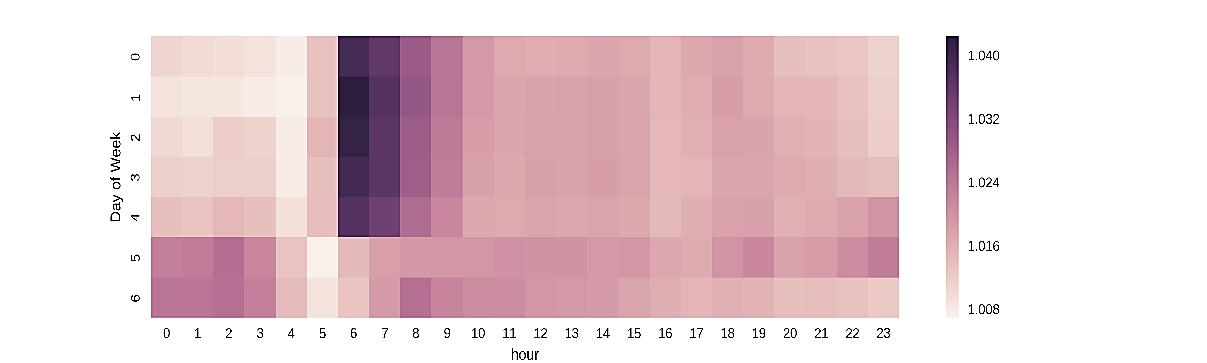
**4. Analysis:**

1. Trips taken in each month(fig-i) peaks between March-May and drops substantially during June onwards. This can be attributed directly to the weather pattern, as commuters are expected to avoid walking long distances during low temperatures or rainy weather.



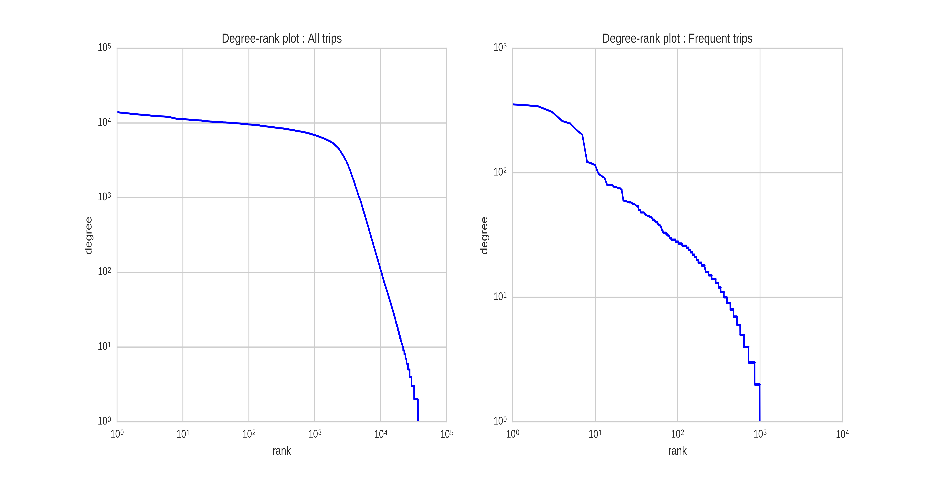
(fig-i)

2. The heat-map(ii) shows weekday-wise relative density of trips for each hour of the day. From this information, we summarize that the busiest hours are 6AM to 10AM. We attribute most of this traffic to business travel whereas there is a remarkable increase in density between 12AM to 4AM on weekends. We are looking for an approach to perform similar analysis on the network structure generated from our subset, to create temporal traffic density visualization.



(fig-ii)

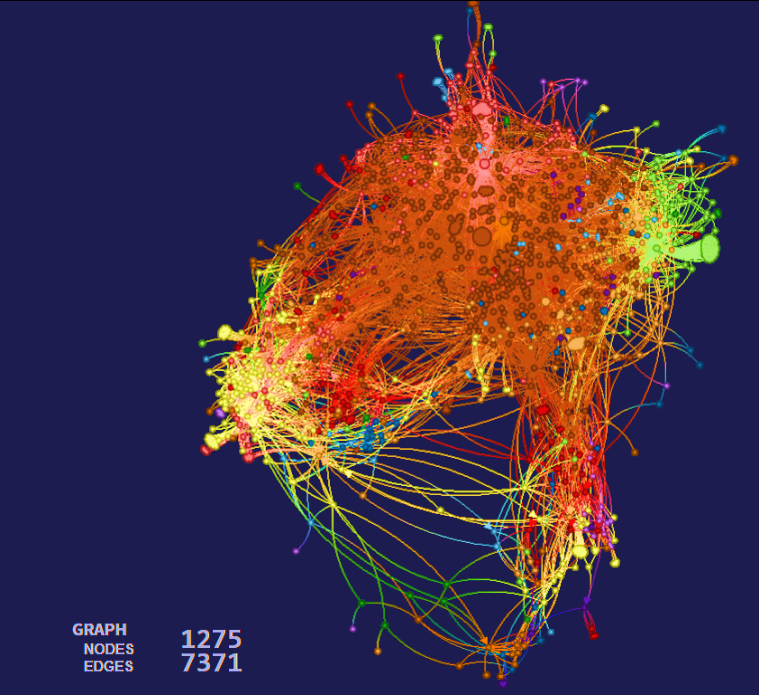
1. For our full dataset of Manhattan, the degree distribution is shown in first plot whereas the second plot shows degree distribution for graph generated using nodes with at least 500 trips in the year between them.



(Fig-iii)

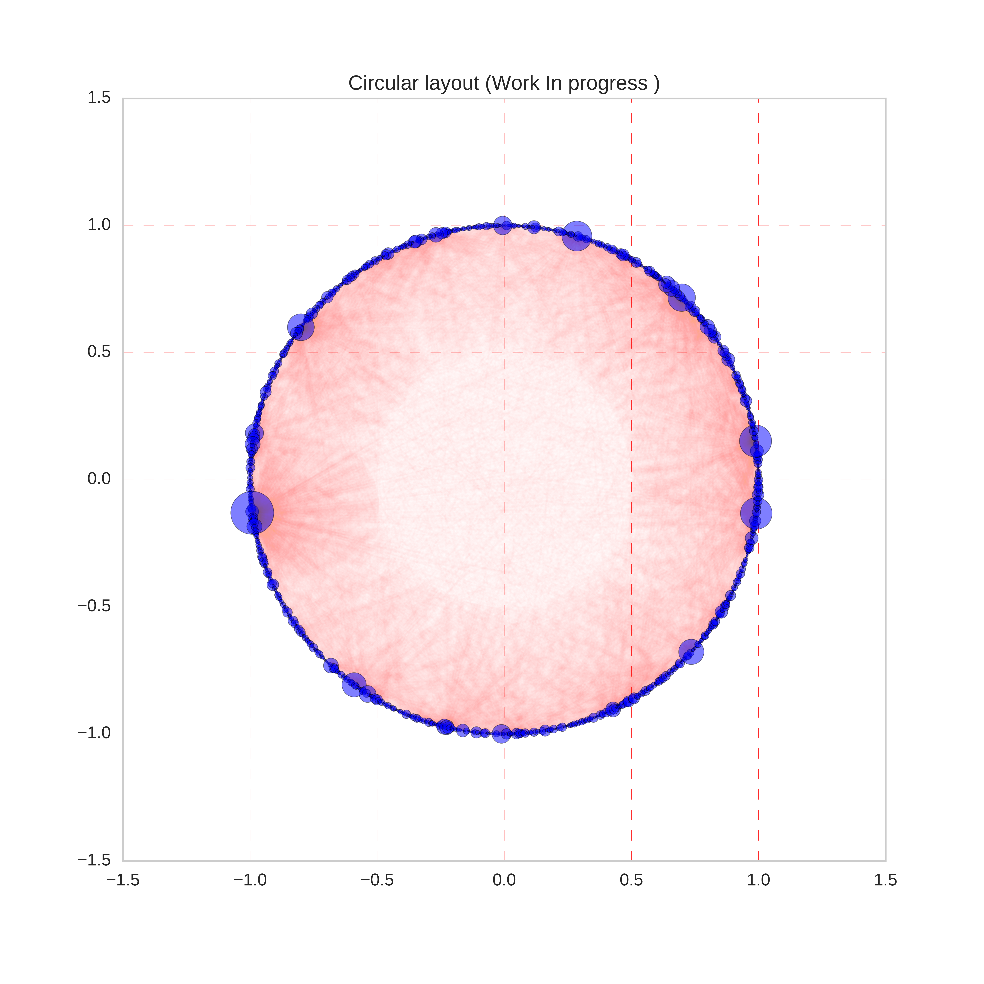
1. Based purely on connectivity between nodes, we use PageRank centrality as basis for clustering which generated the following rough cluster structure. It was at this point that we realized the problem caused by our choice of nodes using 3 decimal points. From this structure we inferred 2 key points:
2. The network seems to split between at-least 4 clusters (fig-iv) – Centered around Penn station, Airports, mid-town & upper-east Manhattan. Right now, from preliminary analysis, we observe that out-degree is highest for the nodes representing airports while in-degree is maximum for points representing Penn Station.
3. We can reduce the complexity of these clusters by converting physically close points (those which essentially represent frequent trips to the same location) to a single point.

(Fig-iv) : **Approximate community structure**



**5. Challenges & further analysis**

**Challenges in data handling:**

1. We were shortsighted in our plans to manage this dataset. Since early on, we were hampered by the data size and since we couldn’t access university’s data services, every aggregation and data manipulation step was run on a personal cluster, which was pretty limited.
2. A large number of data issues became apparent only after we received access to CIRC services a couple of weeks ago, the most important being possible normalization of data by aggregating points representing a common physical location, for which we had assumed, incorrectly, that rounding off to 3 decimal points was enough.
3. For our graph of 1275 nodes, an automated rounding off to 2 decimals makes the area under consideration much more general in terms of land use while our current 3 granularity makes it highly segmented to analyze efficiently, so we now have to manually tag these nodes.

**Challenges in analyzing:**

1. We are attempting to define travel pattern based on clusters formed from aggregated trips and similar usecases in previous research have focused more on complete path data [1] or physical network data [5][6]. We are in process of reviewing usecases which align more closely to our data structure.
2. Charactering trips based on nodes depends on land use datasets and given our inexperience in understanding these datasets, it is expected to take much longer time if we use government surveys of land use [8].
3. Key outputs from our analysis are expected to be graphical representations of the taxi trip network. Given the number of nodes & corresponding edges in our network, we are facing difficulties to analyze and represent it suitable using a single structure. Our Current graphical outputs are:
   1. Approximate clusters as shown in fig-iv. This can be simplified by a geometry-like feature based aggregation [9] and we are attempting to follow this for final viz. We would like to represent edge colors as congestion in network for our final submission.
   2. A circular graphical representation similar to metro-analysis [6]. Our current work-in-progress layout is based on purely degree centrality and we are attempting to figure out force-directed

(**fig-v: Circular Layout – NYC taxi network**)

edge bundling in our context to make a better representation of the graph.

References:

1. P. S. Castro, D. Zhang, C. Chen, S. Li, and G. Pan. From taxi gps traces to social and community

dynamics: A survey.ACM Comput. Surv, December 2013.

2. C. Etienne and O. Latifa. Model-based count series

clustering for bike sharing system usage mining.

ACM Trans. Intell. Syst. Technol., July 2014

3. D. Singhvi, S. Singhvi, P. Frazier, S. Henderson,

E. Mahony, D. Shmoys, and D. Woodard. Predicting

bike usage for new york city's bike sharing system. In

AAAI Workshops, 2015.

4. Peng C, Jin X, Wong K-C, Shi M, Lio P (2012) Collective Human Mobility Pattern from Taxi Trips in Urban Area. PLoS ONE 7(4): e34487. doi:10.1371/

journal.pone.0034487

5. P. Crucitti, V. Latora, and S. Porta. Centrality measures in spatial net-works of urban streets. PHYSICAL REVIEW E, 73(3):036125, 2006.

6. Derrible S (2012) Network Centrality of Metro Systems. PLoS ONE 7(7): e40575. doi:10.1371/journal.pone.0040575

7. NYC taxi data:

[http://www.nyc.gov/html/tlc/html/about/trip\_record\_data.shtml](http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml 8)

[8](http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml 8). NYC land use: <https://www1.nyc.gov/site/planning/index.page>

9. W. Cui, H. Zhou, H. Qu, P. C. Wong and X. Li, "Geometry-Based Edge Clustering for Graph Visualization," in IEEE Transactions on visualization and Computer Graphics, vol. 14, no. 6, pp. 1277-1284, Nov.-Dec. 2008.  
doi: 10.1109/TVCG.2008.135

10. Holten, D, & Wijk, J 2009, 'Force-Directed Edge Bundling for Graph Visualization', *Computer Graphics Forum*, 28, 3, pp. 983-990, Business Source Premier, EBSCO*host*, viewed 12 November 2016.