A Complex Network Analysis of Macroscopic Structure of Taxi Trips

Homayoun Hamedmohaghadam-Rafati*, Ingrida Steponavice*, Mohsen Ramezani**, Meead Saberi*

* Monash University, Melbourne, Australia
(e-mail: {homayoun.hamedmohaghadamrafati, ingrida.steponavice, meead.saberi}@monash.edu).
** The University of Sydney, Australia
(e-mail: mohsen.ramezani@sydney.edu.au)

Abstract: Despite the growing availability of big mobility data in cities, methodologies to extract meaningful information from them are still scarce. In this paper, we investigate taxi trips in New York City, develop a large-scale weighted and directed mobility network, and apply a macroscopic methodology to extract the spatial-temporal structure of urban mobility. We also present a new approach to study weighted networks of mobility in which links in the network have journey speed or travel time attribute in addition to commonly used link weights representing number of trips between pairs of nodes. We show that the structure of mobility network in a city when temporal characteristics and variations are taken into account exhibit different properties than what was previously observed. Results provide a better understanding of mobility characteristics in cities.

Keywords: Parsimonious Modeling, Urban Mobility, Taxi, Big Data, Power Law

1. INTRODUCTION

Taxi trips constitute a significant portion of urban mobility in large cities. In New York City alone, more than 450,000 trips per day carrying about 1 million passengers, and over 170 million trips per year are served by taxis (Ferreira et al., 2013). The increasing availability and popularity of ride-sharing services such as Uber and Lyft has recently created a disruption in traditional taxi services and raised questions on the efficiency of the mobility services provided by taxi industry. Several studies (Schaller, 2005; Schaller, 2007; Çetin and Eryigit, 2011; Nourinejad and Ramezani, 2016) have explored passenger demand and taxi supply balancing problem. Other studies (Yang and Wong, 1998; Yang et al., 2000; Wong et al., 2001; Schaller, 2005; Chang et al., 2009; Yang and Gonzalez, 2014) examined the factors affecting taxi demand and developed models of taxi-passerger searching and prediction models to estimate taxi-passerger trips.

The growing availability of pervasive mobility data from non-traditional sources such as mobile phones, social media activities, and GPS traces has provided enormous opportunities to quantitatively and empirically study many urban and transportation problems. Recent studies by (Qian et al., 2015; Qian and Ukkusuri, 2015) used large-scale taxi data from New York to characterize urban dynamics and study spatial variation of taxi ridership. In another study, (Liu et al., 2015) extracted 860,905 taxi trips from GPS trajectories of more than 6,600 taxis in Shanghai to study travel patterns and structure of urban network communities.

Nevertheless, extracting meaningful and descriptive mobility information from travel "big data" is still a challenge. In this paper, we study taxi trips as a large-scale weighted and directed network and propose a macroscopic approach to extract and understand the spatial-temporal structure of taxi mobility. The methodology presented in this paper is built upon a recent study on spatial structure of mobility networks (Louail et al., 2015) in which large-scale origin-destination (OD) matrices of trips extracted from mobile phone data are converted into coarse-grained information represented by a $2 \times 2$ matrix. The paper demonstrates such macroscopic perspective of mobility demand structure reveals specific patterns on how people travel in cities. We apply the proposed coarse-graining method in (Louail et al., 2015) on a large-scale taxi data from New York consisting of about 13 million taxi trips over one month, and utilize the result in order to uncover the structure of taxi mobility network.

Furthermore, this paper presents a new approach to study weighted networks of urban mobility in which links in the network have journey speed or travel time attribute in addition to commonly used link weight representing number of trips between each origin-destination pair. Statistical properties of such networks are also presented and discussed.

The remainder of the paper is organized as follows. Next section includes a description on data, data cleaning, and pre-processing. In Section 3, the coarse-graining method for OD matrices is described and results of applying the method on taxi data are presented. Section 4 studies taxi trips data as network structured information to uncover the urban mobility network characteristics and discuss the results. Finally, conclusions are presented in Section 5.

2. TAXI TRIPS DATA

The data set used for our analysis comprises anonymous New York Taxi GPS data of 13,990,176 trips from February 2013. For each taxi trip, the data includes timestamps and location of pick-up and drop-off in the New York metropolitan area,
together with the distance recorded for each individual trip. Pick-up and drop-off locations are recorded with a latitude and longitude pair of values. Trip record vector also holds the journey speed attribute that is calculated from the raw data.

2.1 Data cleaning

We cleaned the original raw data based on the following rules to possibly eliminate outliers and trips with erroneous attributes:

- The calculated Euclidean distance between origin and destination cannot be longer than actual reported trip distance.
- The calculated speed using recorded trip duration and distance cannot be higher than 105 km per hour.
- Trip duration has to be longer than 60 seconds.
- Trip distance has to be longer than 0.3 km.
- Trip distance has to be shorter than 3 times Manhattan distance calculated using pick-up and drop-off coordinates.
- The latitude and longitude of both pick-up and drop-off locations have to be in the New York metropolitan range.

The cleaned data includes 12,617,928 trips consisting of 90.19% of original taxi trip records in the data.

2.2 Data transformation

To obtain a better understanding of taxi trip flows and systematically analyzing the data, we divide New York area into 13,161 zones of approximately one squared kilometer. Then for each trip record, the pick-up and drop-off locations are transformed to the zone number in which the location is enclosed. In total among all zones, we obtain only 3,845 active zones, i.e. zones with at least one recorded taxi trip starting from or ending to that zone. Through this study, taxi trip network is generated considering each active zone as a node. A directed link connects node i to node j if there exists at least one trip record from i to j and the link holds certain information on all trips from i to j as an attribute vector. Through the manuscript, we use zone and node interchangeably.

3. MACROSCOPIC STRUCTURE OF TAXI TRIPS

A recent study (Louail et al., 2015) proposed a method to characterize the structure of large, weighted, and directed networks. The purpose of the method is to efficiently extract the signature of a large OD matrix by aggregating the detailed OD information. Taking all origins and destinations, a number of nodes are selected as origin (or destination) hotspots according to the number of trips originated from (or attracted to) those nodes in the mobility network. An origin or destination node is a hotspot if according to a predefined criteria, it has a relatively large number of trips originated from or destined to, during a certain amount of time. A node which is not determined as a hotspot is defined as a non-hotspot. The core idea is to reduce the numerous OD demand flows, to only four types: (i) Integrated flow from hotspots to hotspots, (ii) Convergent flow from non-hotspots to hotspots, (iii) Divergent flow from hotspots to non-hotspots, and (iv) Random flow from non-hotspots to non-hotspots. The four flow types are denoted by I, C, D, and R, respectively.

In the proposed method, once the hotspots are determined, a reordering process is performed on the OD matrix. In the reordering process, the columns corresponding to destination hotspots should be all moved to the left side of the OD matrix and the rows corresponding to origin hotspots should be moved to the top of the OD matrix. Accordingly, an abstract matrix is extracted by normalizing the sum of all entries in each of the four quadrants determined by the border lines between hotspots and non-hotspots in the OD matrix. With the set of origin hotspots as M and the set of destination hotspots as P, the reduced 2 x 2 matrix would be

\[ \Lambda = \begin{pmatrix} I & D \\ C & R \end{pmatrix} \]

which is calculated as:

\[ I = \sum_{i \in M, j \in P} F_{ij} / \sum_{i,j} F_{ij} \]

\[ C = \sum_{i \in M, j \in P} F_{ij} / \sum_{i,j} F_{ij} \]

\[ D = \sum_{i \in M, j \in P} F_{ij} / \sum_{i,j} F_{ij} \]

\[ R = \sum_{i \in M, j \in P} F_{ij} / \sum_{i,j} F_{ij} \]

where \( F_{ij} \) is the flow from origin i to destination j. Note that \( I, C, D, R \in [0, 1] \) and \( I + C + D + R = 1 \). Each value in matrix \( \Lambda \) represents the proportion of trips associated with one demand (or flow) category, namely, I, C, D, and R flow.

For any selected time interval in the time period under study, an OD matrix can be built from the taxi trip record, started or ended within that particular time interval. We label nodes with a number of outgoing (or incoming) taxi trips being more than or equal to 1% of the total number of trips, as origin (or destination) hotspots. In total, we compose 28 daily OD matrices for each day in February 2013 and 28 x 24 = 672 hourly OD matrices. All hotspots found for each of the 672 hourly OD matrices happen to be positioned in a relatively small spatial area (namely Manhattan). Fig 1.a and 1.b respectively depict for every zone the probability of being an origin and destination hotspot, i.e. the number of hours that a zone is a hotspot divided by the total number of hours (i.e. 672). It is clear that the majority of the hotspots concentrate in the Manhattan area with a few others located at the airports.

Although, hotspots constitute a very small proportion of all active zones, i.e. about 10% for hourly and 3% for daily OD matrices, trips from hotspots to hotspots contribute the most to the total number of taxi trips during all hours with I value ranging from 0.46 to 0.74, see Fig. 2 for the temporal variations in I, C, D, and R values. For each particular hour a certain trip type value is calculated by averaging over values
of that hour interval in all days in February 2013. The second largest contributor is trips from hotspots to non-hotspots (divergent flow), with its maximum occurring in the fifth hour interval of the day. The contribution of trips from non-hotspots to hotspots (convergent flow) and from non-hotspots to non-hotspots (random flow) vary slightly over time at values near 0.1. I and D flows are found to have an inverse relationship. A positive correlation between D and R and a negative correlation between I and R is also observed.

![Figure 1](image1.png)

**Fig. 1. Probability of being a hotspot for all zones: (a) origin hotspots and (b) destination hotspots.**

To further evaluate the correlations observed in Fig. 2, we calculate the Pearson correlation coefficients for every pair of variables, see Table 1. The highest absolute correlations are found between I-D, I-R, and D-R pairs, with values very close to -1 for the first two pairs and close to +1 for the latter. Based on Fig. 2, one can expect correlation between C and other flows, but according to Table 1 there exist no significant correlation between C and any other flow. It appears that if we shift the C flow values one hour to the right, i.e. C_{i+1}, a relatively significant correlation about 0.9 with I positively and with both D and R negatively can be observed. Convergent flow includes mostly the commutes to work or other non-residential hotspots, as it increases drastically in the sixth hour interval of the day.

![Figure 2](image2.png)

**Fig. 2. Hourly average of I, C, D, and R demand values for all days in February 2013.**

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>D</th>
<th>C</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>-0.988</td>
<td>-0.991</td>
<td>0.682</td>
<td>1.0</td>
</tr>
<tr>
<td>C</td>
<td>-0.719</td>
<td>-0.768</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.979</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1. Pearson correlation coefficients between hourly values of I, C, D, and R flows.**

In addition to the temporal trends of flow types, we investigate the relationship of flow type values with hourly intensity of taxi trips. To test such relationship, number of trips is counted for each single hour interval corresponding to different flow types. The result shows that with increasing number of trips per hour, I has overall tendency to increase (see Fig. 3), but plateaus after reaching its peak around 22,000 trips per hour with a slight decrease with increase of total number of trips. Here again as showed earlier, overall trends of flow types are correlated; I is correlated negatively with both R and D but has positive correlation with C. All flow types change their behavior almost at the same point, around 22,000 trips per hour. A curve is fitted to each set of scattered data points in Fig. 3 to reveal the trends more clearly. Equal-frequency binning is used for data fitting, with every bin containing 56 data points. Bin center and value is calculated for each bin, by averaging over data points in the bin.

Figure 4 illustrates the trip distance characteristics of different flow types. For each flow type a distance histogram is created by rounding the trip distance values to the closest
that is significantly shorter than average trip distance of £, and
for $i \times \text{R}$, Toulouse, France, July 9-14, 2017
Proceedings of the 20th IFAC World Congress
considered as a node in the graph. To study the mobility in Section 2.2, each zone determined as an active zone is
drop-off points to zones based on zoning approach described
in Section 2.2, each zone determined as an active zone is
considered as a node in the graph. To study the mobility
network in any time interval of interest, a matrix with the size
of $n \times n$, with $n$ equals to the number of active zones in the
selected time interval, is generated with each entry containing
a vector of multiple values corresponding to different
attributes of that link. We calculate average speed, average
distance, and weight for every link. For the network of a time
interval, the link weight is defined as the number of trips
between the pair of nodes (zones) connected via the
Corresponding link, which start or end in that particular time
period. Link speed and link distance are also calculated by
averaging over trips which contribute to the weight. When
generating multiple networks for sequential time intervals,
one trip can be counted twice if its pick-up and drop-off time
take place in different time intervals; for example a trip
started before midnight and finished the next morning is
taken into account once in generating the network
corresponding to the first day, and once for the next daily
network.

4. CHARACTERIZING THE MOBILITY NETWORK

In this section, we view the mobility network of taxi trips as a
directed weighted graph with multiple link attributes (Saberi
et al., 2016). Accordingly, we scrutinize the hourly and daily
characteristics of such networks. After assigning pick-up and
drop-off points to zones based on zoning approach described
in Section 2.2, each zone determined as an active zone is
considered as a node in the graph. To study the mobility
network in any time interval of interest, a matrix with the size
of $n \times n$, with $n$ equals to the number of active zones in the
selected time interval, is generated with each entry containing
a vector of multiple values corresponding to different
attributes of that link. We calculate average speed, average
distance, and weight for every link. For the network of a time
interval, the link weight is defined as the number of trips
between the pair of nodes (zones) connected via the
Corresponding link, which start or end in that particular time
period. Link speed and link distance are also calculated by
averaging over trips which contribute to the weight. When
generating multiple networks for sequential time intervals,
one trip can be counted twice if its pick-up and drop-off time

4.1 Hourly Flux-Speed relationship

For each hour interval in the data, we generated a network
and calculated node speeds, i.e. aggregated speed for each
zone. For each node in the network, a speed attribute is
calculated as the average hourly speed of all trips to and from
the zone. An interesting yet intuitive finding is that node
speeds are higher during off-peak period. Flux-in and flux-
out are also calculated, i.e. sum of weights for incoming and
outgoing links, respectively. Most of the zones did not have
trips starting and ending in the same zone. Therefore, we
could not calculate the average speed inside these zones.

We observe a pattern in change of flux-in/out according to
increase in node speed. Both flux-in and flux-out follow the
similar pattern to a great extent as in Fig. 5. Average flux-
in/out is very low for small node speeds, but with a steep
(linear) increase starting at node speed value about 11 km/h,
reaches its maximum value at node speed around 15 km/h. For the latter, increase in node speed is a possible sign
of density decrease in the zone and thus the average flux-
in/out declines, while in the former range the opposite is the
case for flux-in/out value. Node speeds up to 10 km/h imply
severe congestion or gridlock phenomenon where the near
zero flux-in/out is probably due to jam density, thus for the
node speeds in the range [0, 15) km/h increase in node speed
relates to congestion or gridlock dissipation.

4.2 Exploring daily networks

Daily networks are constructed considering all the trips
within a particular day in the data. We find that the day-to-
day mobility network derived from data, exhibit certain
degree of similarities. This similarities among networks can
be verified with a number of their properties, such as size
(number of nodes), density (ratio of the number of edges to
the number of possible edges), etc. Table 2 summarizes a
selected set of network characteristics with mean, standard
deviation, minimum, and maximum values for 28 daily
networks of taxi trips.

![Fig. 4. Trip distance distribution for each category: (a) integrated trips, (b) convergent trips, (c) divergent trips, and (d) random trips.](image)
Fig. 5. Average (a) flux-in and (b) flux-out versus node speed. The average flux-in/out is calculated by averaging over the flux-in/out values corresponding to equal node speeds.

Table 2. Summary of selected characteristics for 28 daily networks.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std Dev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>size</td>
<td>1091.25</td>
<td>50.65</td>
<td>946</td>
<td>1187</td>
</tr>
<tr>
<td>density</td>
<td>0.0151</td>
<td>0.0018</td>
<td>0.0123</td>
<td>0.0187</td>
</tr>
<tr>
<td>Average in/out-degree</td>
<td>16.36</td>
<td>1.56</td>
<td>14.26</td>
<td>19.13</td>
</tr>
<tr>
<td>LWCC Size</td>
<td>1020.04</td>
<td>41.04</td>
<td>901</td>
<td>1080</td>
</tr>
<tr>
<td>LSCC Size</td>
<td>347.64</td>
<td>19.21</td>
<td>320</td>
<td>384</td>
</tr>
</tbody>
</table>

*LWCC and LSCC stands for the Largest Weakly and Strongly Connected Component.

Next, we study the distribution of node degree, node flux, and link weight of the mobility network of taxi trips. Many real-world phenomena exhibit heavy-tailed or even long-tailed distributions, such as power law:

\[ p(x) \propto x^{-\alpha} \]  

where \( \alpha \) is a constant scaling parameter.

A typical quantitative definition for a heavy-tailed distribution is that the tail of the distribution, i.e. the probability to observe very large values, is not exponentially bounded (Asmussen, 2008). Therefore, the minimum effort before asserting that a distribution is long-tailed, would be to compare a candidate long-tail fit and an exponential fit to the distribution of interest and see if the first fit performs better than the latter. Node flux is the sum of flux-in and flux-out, or equivalently the sum of weights for all the links connected to the node. Based on all daily networks, histograms are constructed for link weight (\( w \)), node degree (\( k \)), and node flux (\( f \)) values. We compare three candidate distributions, namely, exponential, log-normal, and power law as hypothesized models. Using log-likelihood ratio test proposed in (Alstott et al., 2014), we find out that power law provides the best fit for link weight, node degree, and also node flux with high level of significance in all comparisons (\( p \)-value < 0.1). Based on the widely accepted method of power law fitting by (Clauset et al., 2009), the best power law fit is found for each network measure. The best power law fit has two parameters, scaling parameter (\( \alpha \)) and the minimum value for which the quantity behaves similar to the best fit (\( x_{\text{min}} \)), see Fig. 6. As expected, more links have fewer weights, i.e. as the weight increases the number of links decreases. Weight values follow power law with \( x_{\text{min}} = 1 \) but with a cut-off at about \( w = 300 \). The difference between measures \( k \) and \( f \), is taking the link weight information into account for the latter but comparing Fig. 6.b with 6.c, the observed flux values show smaller discrepancy with the fitted model.

5. CONCLUSIONS

The paper has applied a coarse-graining method proposed by (Louail et al., 2015) to analyse large-scale OD trip matrices and studied the taxi mobility structure in New York. The method reduces all commuting patterns to four macroscopic flows, namely, \( I \), \( C \), \( D \), and \( R \) flows, which are asserted to be a signature of urban mobility structure.

In (Louail et al., 2015) \( I \), \( C \), \( D \), and \( R \) flows are compared in multiple cities. They observed a negative correlation between \( I \) and \( R \), as population changes. We explored the temporal trends and sensitivity to intensity of trips among the flow types. Beside the \( I-R \) relation, a number of significant correlations are uncovered and analyzed among flows in response to time and taxi transport intensity changes. Furthermore, we characterized commuting distances of different flow types, and our findings are consistent with the results in (Louail et al., 2015).

Taxi trip data is also viewed as a weighted and directed network, in this study. We generated network structures from large-scale temporal data, and showed that exploring networks of different time interval sizes can result different findings. We showed that our approach is effective in characterizing the mobility structure and leads to better understanding of the urban dynamics, which is of great importance to a number of fields including epidemics, cascade spreading, transportation planning, and urban policy making.
Fig. 6. The complementary distribution function \( \Pr(X \geq x) \) and maximum likelihood power-law fit with parameters \((\alpha, x_{\min})\) for (a) Link weight \((1.60, 1)\), (b) Node degree \((1.61, 8)\), and (c) Node flux \((1.38, 1)\).

REFERENCES


