Improving Traffic Flow Using Uber Movement Data

Mackenzie Pearson pearson3@stanford.edu
Javier Sagastuy jvrsgsty@stanford.edu
Sofia Samaniego sofiasf@stanford.edu

Abstract

Utilizing the Movement data set recently released by Uber, we propose to analyze and discern traffic bottlenecks in multiple cities. Further, we plan to uncover general traffic movement patterns throughout these cities at various times in the day and identify the main differences in mobility behavior between them. Finally, we plan to come up with a solution to alleviate traffic on our identified bottlenecks and measure its impact through simulated data.

1 Reaction Papers on Traffic Flow

Studying travel patterns and the structure of cities has long been a research topic of great interest in urban planning. In the past decade, this interest has spiked due to an increase in availability of GPS and mobile phone data. Further, open-data initiatives such the New York City Open Data Project provided researchers with taxi trajectory data that enabled them to investigate different aspects of traffic flow. In this section we summarize and critique three papers that use GPS taxicab data as a tool to approach urban dynamics from different perspectives.

1.1 Understanding traffic flow characteristics

In this section, we investigate the problem of estimating and understanding traffic flow characteristics of cities. Historically, several authors claimed that the configuration of a city's street network plays an important role in vehicular flow and, hence, used centrality measures of a street graph to model and predict traffic. Specifically, authors such as Turner [4] proposed betweenness centrality as a good predictor of traffic flow. We focus on the work of Gao, et al. [1], who criticized this approach and proposed a new model of traffic flow based on the non-uniform distribution of human activity and the distance-decay law.

Gao, et at. argue that the betweenness centrality measure of a street network is static and thus can't be used to model the dynamic behavior of traffic demands. Further, the authors claim that this measure does not take into account the fact that travel demand (i.e. traffic flow) depends on the distance between origin and destination and, in particular, is decreasing as a function of trip length. To support their critique, the authors compute the weighted correlation (with weights given by street length) between "real" traffic flow, estimated through the line-density method using a one-week long GPS data set of 149 taxis in the core urban area of Jiaozhou Bay, and the betweenness centrality measure of the nodes of this city's street primal and dual networks. They find that this measure is not ideal by itself to predict urban traffic flow.

The alternative approach proposed by the authors is to construct a trip demand model that incorporates the heterogeneity in real human activities and the decay-distance law in trip demand. Specifically, they use the total call-traffic volume of base stations in Jiaozhou Bay in one hour, namely the Erlang values, to model the sample probabilities of origin and demand pairs (OD). Meanwhile, they model the probability of an edge existing between a sampled OD pair through a distribution that decreases exponentially as a function of the distance between the origin and distance nodes (power-law). Using this method they run Monte Carlo simulations to generate trip data and produce an estimate of traffic flow. Finally, they use weighted correlation to measure goodness of fit

between their simulated and observed taxi trajectory data. They conclude that the proposed model can interpret urban traffic flow well.

Critique: Despite the fact that this article was published in 2012, it seems very outdated. With the boom of the digital age, human mobility and location data has dramatically increased in volume. This information should be exploited when carrying out an analysis of this sort, and in this section we point out some ways in which this could be done.

First, we notice that the authors of this paper focus all of their attention in simulating trips between origins and destinations and then evaluate the validity of their model through comparison with "real" traffic flow estimated through a relatively small taxi trajectory data set. In contrast, with the release of "Uber Movement", we now have information of all Uber trips per hour in a collection of countries around the world since the beginning of 2016! Hence, we now know the traffic flow and can now focus on more useful tasks like: analyzing it, identifying bottlenecks, and suggesting ways of improving connectivity in cities.

Another flaw of the analysis presented in this paper is that it assumes that individuals rationally choose the geometrically shortest path; however, the taxi trips used in this data sets were performed by human drivers with way-finding behaviors that are not always rational. Meanwhile, our data consists of trips carried out by Uber drivers, who follow optimized routes given by a shortest path algorithm. This is more in line with the shortest path assumption.

Finally, the model presented in this paper does not take into account the fact that the probability of a person choosing a taxi as their means of travel varies depending on different places and over different time periods. Further, it assumes that the proportion of taxis is uniform across all streets of a city, which is clearly unrealistic.

One of the strengths of this work is that it provides a way to simulate new trips that follow relatively closely the real taxi trajectories. This could be useful if we ever needed to generate new trip data to evaluate the impact of a policy to improve traffic conditions.

1.2 Revealing travel patterns and city structure

In 2015, Liu, et al. [2] presented an analysis to infer travel patterns and city structure from data modeling traffic flow. By using taxi trip data from the city of Shanghai, they represented traffic flow as a directed graph and applied modern network analysis techniques to characterize it. Their approach revealed a two-level hierarchical structure of Shanghai based on the length of the taxi trips and contrasted the administrative boundaries of the city with the natural boundaries derived from the travel patterns. A modification of administrative and transportation planning boundaries is proposed to improve local mobility and current traffic analysis modeling to aid in urban planning.

To accomplish this, Shanghai was split into a 1×1 km cell grid; each of the resulting cells representing a node in the graph. Two nodes u and v were connected by a directed edge if a trip originating physically inside u and ending inside v existed. The edges were then weighted according to the number of existing trips between the same cells. Only data from Monday to Thursday was used, since this represents the most constant traffic flow due to an increased number in leisure and entertainment trips near the weekend.

The resulting network was then processed using community analysis to identify regions within which trips were common. Further, the detected communities were characterized by measuring graph density, node strength, closeness centrality and betweenness centrality for each of the nodes in a community. From this analysis, centers with a high degree of traffic flow (measured through node strength) were identified.

Critique: The way Liu, et al. construct the graph seems very appropriate for the problem they are trying to address. However, we might not be able to construct a similar model with the proposed data set due to the lack of information on the number of trips taken between two points. The authors might have, however, chosen to represent the distance between two nodes as a weight on an edge rather than by building different graphs for each distance "slice". The Uber Movement data can be represented as such a network without much preprocessing.

Node centrality, strength, and density seem like appropriate metrics with which to analyze the network in this paper. These metrics might not be suitable for the different type of graph we are proposing.

This paper clearly arrives at a model of the city structure in terms on how it is transited. The results might be useful to help with urban planning in the future. However, it fails to pinpoint specifically problematic areas or propose solutions to current traffic flow problems. We hope our approach will provide more immediate indicators of traffic flow problems which may lead to shorter term solutions.

1.3 Urban Computing with Taxicabs

Zheng, et al. [5] provide an interesting framework for analyzing taxicab data, which could be relevant for to our project. This framework consists of linking pairs of regions (i, j) to three key features:

1. The number of taxis going from region i to region j.
2. The average speed these taxi drives when commuting from region i to region j.
3. The ratio between the actual travel distance and the distance between the centroids of these two regions.

By mapping taxi trajectory data from 30,000 taxis driving in Beijing from March to May in 2009 and 2010 onto this framework Zheng et al. seek flaws in current urban planning.

Flaws are detected by finding obvious issues in these taxicab commutes. For example, if the flow from a region i to region j is high, but the average speed between these two regions is low and the actual distance traveled is high compared to the distance between the centroids of the two regions, then one could conclude that there is high traffic and the detours are slow. Zheng et al. compare and contrast these issues over two years to see if new roads or subways systems have had a clear impact on these problem areas.

Critique: Zheng et al. do well in removing outliers in their data set by mining consistent sub-graphs over a range of days. Specifically, they compare the traffic graphs between subsequent days for particular time intervals and find the common sub-graphs between these graphs. They also make the distinction between workdays and rest days, which include weekends and holidays, allowing the analysis to better reflect true traffic patterns.

We believe the framework Zheng et al. laid out could be useful and relevant to the objectives of our project. In particular, we could try to mimic their approach to finding obvious commute issues in our Uber Movement data. However, we may run into issues in doing something similar to the compare and contrast aspect of the paper, as new infrastructure is not the only possible reason for changes in traffic patters. In this sense the paper is limited in its approach to uncovering why the traffic patterns change over the two years.

Zheng et al. also fail to discuss the fact that taxi data is not representative of all vehicles. Additionally, speed may not be the best way to determine where traffic is, since speed limits were not incorporated into the analysis. Similarly, there may be issues with utilizing the "distance ratio", as there may be natural obstacles that make the ratio between distance traveled and distance between the centroids of two regions very high. Finally, an assumption is made that taxi drivers are optimizing their routes. It is unclear whether these taxi drivers were utilizing a route finding app to find their path. If not, these high density traffic areas may not be representative of poor urban planning, but of poor route finding.

1.4 Comparison

The three papers discussed above all use GPS taxi trajectories in the context of traffic flow; however, they cover very different aspects of this broad topic. Gao et al. compare two approaches for estimating urban flow: using centrality measures of a street graph and through Monte Carlo simulations. They evaluate the validity of these methods through their correlation with "true" traffic flow estimated from the taxicab data set; however, they never build a graph with the taxi trajectory data, only with street data. Further, they never use taxi data to uncover patterns in the city's structure or identify locations of traffic flow co-behavior. Meanwhile, Liu et al. build spatially-embedded networks on a massive taxi trip data set and try to use it to explore travel patterns and reveal structures

and travel communities in the city of Shanghai. Their network models traffic flow directly, whereas Gao's approach relies on simulation based on an underlying street graph. Similarly, Zheng et al. model traffic flow directly, but they do not take the next step of identifying potential flaws in city planning. By comparing two years of traffic data, they push the analysis, uncovering whether or not new infrastructure changed Beijing's traffic graph. They also leave the paper with the hope that their work will help advise city planners in the future.

1.5 Brainstorming

- What is particularly promising about our approach is that we have a brand new and very powerful data set that could help us learn a lot about the future of urban mobility.
- While there has been a lot of work in this area, GPS taxi trajectory data sets used in previous work usually only comprised small periods of time; hence, a temporal component has seldom been included in traffic flow models. We now have data of over two billion Uber trips in seven cities around the world starting on 2016, which is significantly more data than any other study in this topic that we've encountered.
- In particular, we want to identify mobility patterns that exist in cities and compare them against each other. Further, we aim to elaborate a sequence of graphs built with Uber trip durations across different times of day and use this information to identify bottlenecks in traffic.
- We have reasons to believe the introduction of affordable ride-share services such as Uber has significantly increased the scope and volume of people that chose cabs as a primary mode of transportation; hence, coming up with solutions to these bottlenecks could have a great positive impact on millions of people!

2 Project proposal

The project we propose is to uncover traffic bottlenecks within a given city at different hours of the day by using travel times between sources and destinations of over two billion Uber trips in the past two years. In the following sections we describe the data we will use to accomplish this, which techniques we plan to use, and what we expect to accomplish by the end of the quarter.

2.1 The data set

This January, Uber unveiled "Uber Movement", a tool intended for use by city planners and researchers looking into ways to improve urban mobility. The data set includes over two billion Uber trips in the cities of Bogotá, Boston, Johannesburg, Manila, Paris, Sydney, and Washington D.C., Specifically, it includes the arithmetic mean, geometric mean, and standard deviations for aggregated travel times over a selected date-range between every zone[1] pair in each of these cities. Uber Movement is open to the public and can be download in `.csv` format directly from [Uber Movement's Website].

2.2 Workplan

Initially we plan on choosing a specific city and mapping the data set from that city onto two graphs. Since the Uber Movement data aggregates source and destination of trips into a grid, we'll find the geographic centroids of each cell in the grid and create a complete graph in which edges are weighted according to the euclidean distance between centroids. We'll then create a second graph in which nodes u and v are linked if there exists a trip from u to v in a specific time frame. We'll do this for several different time frames. From these graphs we will be able to see how the network changes over time, permitting us to identify bottlenecks if the weights of a given edge (u, v) vary significantly. To facilitate the analysis of the graphs, we plan on applying a community detection algorithm and review key centrality measures of each of the graphs. After being able to pinpoint specific stress points on the global city network, we plan on simulating a solution designed to alleviate the stress on that region.

[1] A zone is a predefined region within the city, and each city consists of hundreds of zones.

2.3 Algorithms, techniques and models

Following previous literature on traffic flow networks, we will start by exploring some centrality measures in our graph. In particular, we will measure betweenness of our nodes to identify structurally important locations and closeness centrality to pinpoint which sources and destinations live in the core and periphery of our travel network. It is important to note that, even if our nodes represent geographical locations in a map, this doesn't mean that nodes that are together in space will be "close" in our network. Nonetheless, we hope that analyzing centrality in our network will reveal some structural characteristics of urban dynamic traffic flow and spatial human activity.

Additionally, we aim to study the temporal evolution of our traffic flow network by observing snapshots of it taken at regularly spaced points in time. The nodes of the graph, which represent the sources and destinations of our trips, will remain fixed across time periods; however, the weights of the edges will be dynamic, as average travel times change throughout the day. We hope that this analysis of progressing travel times will help us identify the traffic bottlenecks and rush hours in the different zones of our cities and shed some light into what is originating them.

One of our goals is to identify communities in our network in order to shed light on the structure of our traffic flow data. In order to accomplish this, we could try to cluster the nodes in our travel graph into dissimilar groups of similar items. In particular, we could use a spectral algorithm, since these methods produce results of demonstrably higher quality than competing methods in shorter running times [3]. This could allow us to pinpoint locations within or across our cities that exhibit co-behavior at a certain time of day or at a certain week of year.

2.4 Evaluation and deliverables

At first level, evaluation of our results will consist in how clearly we manage to identify points of conflict for traffic flow in cities. Me might further be able to evaluate the impact of our results by proposing a solution to alleviate the problem in the identified bottlenecks and measuring its impact in traffic flow through simulated data.

We expect to submit an overview and comparison of the identified bottlenecks for the different cities in the Uber Movement data set. Time-permitting, we would like to provide simulations of the effect a proposed fix would have on traffic flow in the network.

3 References

[1] Song Gao, Yaoli Wang, Yong Gao, and Yu Liu. Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality. *Environment and Planning B: Planning and Design*, 40(1):135–153, 2013.

[2] Xi Liu, Li Gong, Yongxi Gong, and Yu Liu. Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography*, 43:78–90, 2015.

[3] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[4] Alasdair Turner. From axial to road-centre lines: a new representation for space syntax and a new model of route choice for transport network analysis. *Environment and Planning B: Planning and Design*, 34(3):539–555, 2007.

[5] Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. Urban computing with taxicabs. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 89–98. ACM, 2011.