Traffic Flow Analysis Using Uber Movement Data

Mackenzie Pearson
pearson3@stanford.edu

Javier Sagastuy
jvrsgsty@stanford.edu

Sofía Samaniego
sofiasf@stanford.edu

Abstract

Utilizing the Movement data set recently released by Uber, we propose a way to model traffic flow throughout the day by using centrality measures in a sequence of temporal directed graphs with dynamic weights. We analyze and discern traffic bottlenecks and rush hours in multiple cities by comparing these temporal graphs at different points in time to a spatial graph that models the underlying city structure. Further, we use our traffic flow models 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 identify communities in each city based on mean travel time and analyze how these communities change throughout the day as well as how they compare to communities based on distances in the spatial graph.

1 Introduction

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 as the New York City Open Data Project provided researchers with taxi trajectory data that enabled them to investigate different aspects of traffic flow.

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 was seldom included in traffic flow models. However, traffic demands exhibit a dynamic behaviour and, hence, urban flow cannot be modeled using solely static measures. Fortunately, with the boom of the digital age, human mobility and location data has dramatically increased in volume. We now have data of over two billion Uber trips at every hour of the day in seven different cities around the world starting in 2016, which is significantly more data than any other study in this topic that we've encountered. In this project, we provide a dynamic analysis of this brand new and very powerful data set and use our findings to model urban mobility.

In order to identify mobility patterns that exist in cities and compare them against each other, we construct a sequence of graphs with dynamic weights built from the mean travel times between sources and destinations of Uber trips across different times of day. We use centrality measures in these graphs throughout the day as a dynamic model of traffic flow and compare them to static centrality measures in a spatial graph that models the geographical structure of the city. We use this information to identify bottlenecks in traffic and rush hours. Finally, we analyze how communities in our temporal networks change throughout day as well as how they compare to clusters in the static spatial network. 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, so using our resulting models as an urban planning tool for coming up with solutions to traffic bottlenecks could have a great positive impact on millions of people!

2 Previous Work

In this section we summarize three papers that use GPS taxicab data as a tool to approach urban dynamics from different perspectives. These papers cover a broad range of topics such as different approaches for estimating urban flow, methods for uncovering travel communities and patterns in a city's structure, and techniques for identifying locations of traffic-flow co-behavior and potential flaws in city planning.

2.1 Understanding traffic flow characteristics

Historically, several authors have 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 [5] proposed betweenness centrality as a good predictor of traffic flow. We focus on the work of Gao, et al. [2], 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.

2.2 Revealing travel patterns and city structure

In 2015, Liu, et al. [4] 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.

2.3 Urban Computing with Taxicabs

Zheng, et al. [6] provide an interesting framework for analyzing taxicab data, which 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, and (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.

3 Preliminaries

In this section we describe the travel time data we worked with, the temporal and spatial graphs we constructed from it, and provide an extension of some key graph definitions to the weighted case that will prove useful in our analysis.

3.1 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].

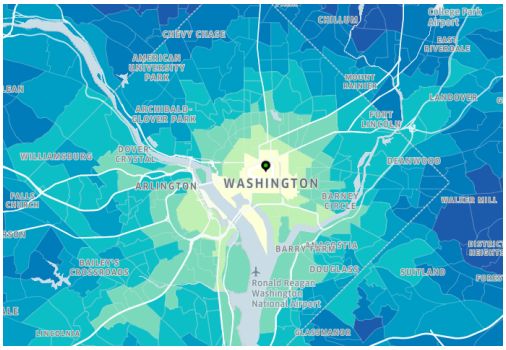


Figure 1: The Uber Web interface colors cells in the city grid based on the average travel time to them from the specified pin

3.2 Data processing

The data as provided by Uber cannot be directly used to build a graph on which we can detect traffic congestion. Instead, we need to build a graph that models the underlying city structure on which trips take place. To do so, we built a spatial graph to represent the adjacency of the zones on which Uber aggregates data. A GeoJSON file describing the polygons which delimit the zones in a city is provided with the data. Using the `igraph` and `rgeos` package for R we were able to load the geometry and compute an adjacency matrix for when any two given polygons were touching each other. The adjacency matrix could then easily be exported as an edge list to be imported into `snap`. Figure 2 shows how the adjacency graph is built from the set of Polygons.

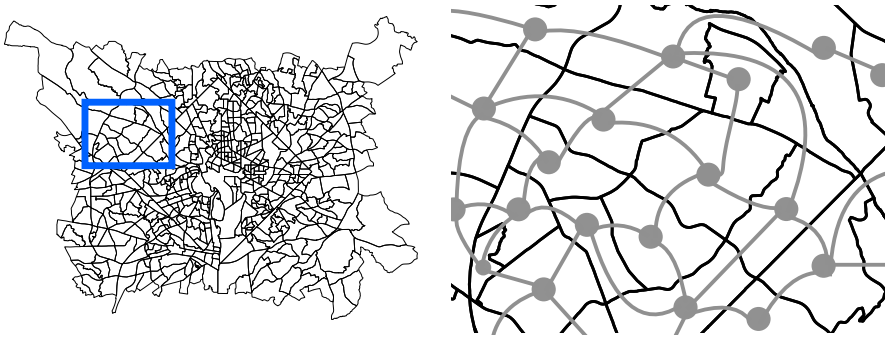


Figure 2: The GeoJSON file contains a set of polygons on which we'll zoom in (left). Looking closely at the region in the blue rectangle, we define a graph where every node represents a region and two regions are linked if they are adjacent (right)

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

Once we have a graph representing the structure of the city, we define two sets of weights which will result in two different interpretations of the resulting graph. First, we may think of the graph as being undirected and the weights being the distances between cells in the grid. To calculate these distances we rely once more on the geospatial packages available for R. We computed the geographic centroid of each cell and then measured pairwise distances between centroids using Haversine distance[2]. We'll further refer to this as the **spatial graph** $G_s = (V_s, E_s, w_s)$.

Second, we may wish to assign weights for each of the four statistics on trip length. In this interpretation of the graph, there exists a link between node u and v if there was at least one trip originating in cell u and ending in cell v. Since we may not have data for all possible node pairs in the graph for a given time-frame, the nodes in this graph will be a subset of V_s. Also note that the average time of travel from node u to node v may differ from the time of travel from node v to u. Thus this graph is directed and it may be weighted with any of the four aggregate measures on trip duration provided by Uber. We'll further refer to this graph as the **temporal graph** $G_t = (V_t, E_t, w_t)$.

We chose to work with data using hourly aggregates of the trip duration measures, since this was the most granular export provided by Uber. We restricted our analysis to the third quarter of 2017 since it proved to be the richest subset of the data across all cities. Further, we only processed data from weekdays following the intuition that movement during the weekend is related with leisure and entertainment and hence does not reflect the normal traffic flow in the city, as presented in Liu, et al. [4]. Thus, we generated 24 different temporal graphs for each city, each with its own set of weights. Only the arithmetic mean of travel time was considered as a weight for the edges on the graph for the scope of this project. The ability to compute the same metrics for a time series of graphs allowed us to analyze and visualize how the traffic patterns evolved over the course of one day across cities.

3.3 Mathematical background

As described in Section 3.2, we will be working with two weighted graphs: a spatial graph (undirected) and a temporal graph (directed). In this section we define some notions and matrices associated to weighted graphs that we will need for our analysis. We will denote a weighted graph by a triple $G = (V, E, w)$, where (V, E) is the associated unweighted graph, and w is a function from E to the real numbers.

In order to identify the structurally important nodes in our graph, we need to extend the centrality measures we learned in class to weighted directed graphs. Measures such as closeness and betweenness are defined in terms of node degrees and shortest paths, so we need to extend these definitions to the weighted case to be able to use these metrics in our analysis. In addition to these measures, we will use the Page Rank and HITS algorithms to determine the importance of our nodes. The HITS algorithm uses the adjacency matrix of a graph to identify hubs and authorities; meanwhile, the Page Rank algorithm uses the stochastic adjacency matrix of a graph, which is defined in terms of the out-degree of its nodes. In this section we provide the definition of these matrices in the weighted case.

3.3.1 Degree

In the case of undirected graphs, the weighted degree of a vertex is defined as the sum of the weights of its attached edges. Formally, the degree of a vertex v of a graph, d_v, is defined as

$$d_v = \sum_u w(u, v).$$

Similarly, in the case of directed graphs we define the weighted in-degree $d_v^{(\text{in})}$ and the weighted out-degree $d_v^{(\text{out})}$ of a vertex v as the sum of the weights of the edges with source v and the sum of the weights of the edges with destination v, respectively. These degrees can also be denoted by node in- and out-strength.

3.3.2 Shortest path

The length of a path P in a weighted graph is the sum of the weights of the edges of P. That is, if P consists of edges $e_0, e_1, \ldots e_{k-1}$, then the length of P, denoted $w(P)$ is defined as:

$$w(P) = \sum_{i=0}^{k-1} w(e_i).$$

The distance from a vertex u to a vertex v in G, denoted by $d(u, v)$ is the length of the shortest path from u to v, if such path exists.

[2]The Haversine distance is the shortest distance between two points on the surface of a sphere and has proved to be a good enough approximation for distances between (latitude, longitude) pairs.

3.3.3 Adjacency matrix

The adjacency matrix of a graph with n nodes is a matrix with rows and columns labeled by graph vertices, with the weight of an edge or a zero in position according to whether the vertices are adjacent or not. Formally, the adjacency matrix is $W = (w_{uv}),\ \ u, v \in \{1, 2, ..., n\}$ where

$$W_{uv} = \begin{cases} w(u,v), & uv \in E \\ 0, & uv \notin E. \end{cases}$$

3.3.4 Stochastic Adjacency matrix

Recall that the column stochastic adjacency matrix M used in the page rank algorithm is defined in the following way. Let v have $d_v^{(\text{out})}$ out links. Then,

$$v \to u \implies M_{uv} = \frac{1}{d_v^{(\text{out})}}.$$

We can extend this matrix to the weighted case by simply considering the weighted out-degree as defined in Section 3.3.1.

4 Algorithms, Techniques and Models

In this section, we propose a way to model traffic flow based on the evoution of some key centrality measures on the temporal graph described in Section 3.2; namely, we claim that the dynamics of a set of importance measures of geographical regions throughout the day model the flow of different components of traffic. We provide the interpretation of the models resulting from each of these measures in terms of traffic flow and describe how we can extract insights from a comparison of these measures across different hours of the day and against these same measures in the spatial graph. Further, we present methods for identifying bottlenecks using betweenness centrality and hot-spots using PageRank scores. We also review a community detection algorithm, namely, Girvan Newmann's Stength of Weak Ties, to detect clusters in our spatial and temporal graphs. We interpret the evolution of the border of the communities found by the algorithm as an approximation of the flow of traffic. We will translate this information into specific examples of the resulting models in our cities and try to pinpoint specific stress points of each city network in Section 5.

4.1 Centrality

We begin by exploring what some key centrality measures mean in our temporal and spatial graphs. In particular, we analyze seven measures: weighted in- and out-degree, closeness and betweenness centrality, PageRank score, hubs, and authorities. Additionally, we suggest a way of using the resulting models to identify bottlenecks and rush hours and to pinpoint which sources and destinations live in the core and periphery of our travel network. We hope that analyzing centrality in our network will reveal some structural characteristics of urban dynamic traffic flow and spatial human activity.

4.1.1 Node degree

First, we consider an intuitive and conceptually simple measure of centrality, namely weighted in- and out- degree. The weighted in-degree of a node is calculated by summing the weights of the incident edges on that node. Similarly, out-degree is calculated by summing the weights of outgoing edges from that node. In terms of traffic flow, the sum of weights in our temporal graphs corresponds to total travel time, so we would expect regions with high weighted in- and out- degree to correspond to stress regions, either as zones where a high volume of trips finish or originate or as zones where long trips finish or originate. One drawback of using in- and -out degree as centrality measures is their inability to highlight nodes with low degree that are in fact structurally important for connecting regions. This drawback led to the consideration of the next centrality measure: betweenness centrality.

4.1.2 Betweenness Centrality

The betweenness centrality of a node v is the number of shortest paths in the graph that pass through it. Intuitively, this metric highlights the "gate keeper" nodes which are structurally important to the graph. In our context of urban dynamics, whenever a region is passed through on an Uber trip[3] from origin to destination, its betweenness value is

[3]Throughout this section, we will assume that uber trips correspond to the shortest temporal paths between origins and destinations, based on the fact that Uber drivers follow optimized routes given by a shortest path algorithm.

incremented; hence, city zones that are usually transited through will take high values while seldom transited zones will take low values. Many authors have claimed that this measure seems to model movement intuitively and hence have proposed betweenness centrality of nodes in a street graph as a good predictor of traffic flow. However, this measure as used previously in literature is static, so it cannot be used directly to model the dynamic behavior of traffic flow. We introduce a temporal component to this model by analyzing the evolution of betweenness centrality in our temporal graphs throughout the day, rather than looking at this measure in the static spatial graph. Further, we suggest using the contrast between structurally important nodes in the spatial graph and structurally important nodes in our temporal graphs to identify traffic bottlenecks at different hours of the day. Intuitively, betweenness centrality in the spatial graph will highlight solely regions that appear in geographically shortest paths, while on the temporal graphs this measure will pick up on nodes that appear in shorter paths based on trip duration when geographically central nodes are congested. Hence, we can define a zone as a bottleneck if it's central in the spatial graph but isn't central at a given hour of the day in the temporal graph.

4.1.3 Closeness Centrality

The closeness centrality of a node v is defined as the reciprocal of the mean average shortest path from node v to all other nodes in the graph. Intuitively, nodes in the core of our network will have high closeness centrality and nodes in the periphery should have low closeness centrality. Closeness centrality highlights a node's ability to reach all the other nodes in the graph quickly; thus, in terms of our temporal graph, one would expect the nodes in the city's downtown area to have the highest closeness centrality. This would be in contrast to closeness centrality in the spatial graph, where the nodes with highest closeness would be geographically central in the given city map, despite wherever the true city center may lie.

4.1.4 PageRank

An alternative measure of centrality of a node is its PageRank score. This algorithm is based on the idea that links from important nodes count more; that is, that importance flows across the directed edges of a graph. In our context, this means that links to central city regions will count more than links to low transit zones and, hence, the importance of a region will depend on the importance of the regions adjacent to it. In our spatial graph, this measure will be uniform (except for the outer rim) if the city is geographically divided into a regular grid, as in the case of Manila, so it is not possible to extract any insights from it. Meanwhile, the PageRank of nodes in the temporal graph yield in hierarchy of nodes based the volume of traffic directed to them or to regions adjacent to them. We can use this centrality measure to identify hot-spots of our cities at different times of the day.

4.1.5 HITS

The authority scores of the vertices are defined as the principal eigenvectors of $W^\top W$, where W is the adjacency matrix of the graph. Meanwhile, the hub scores are the principal eigenvectors of $WW^\top$. Intuitively, hub scores rate nodes based on their quality as an expert and authorities rate nodes based on their quality as content providers. In our temporal graphs, the authority of city regions can be interpreted as a measure of how long it takes to get to a given region. Conversely, hub scores would be interpreted as a measure of the amount of time to drive out of a city region. In the temporal graphs, these identify city regions with a high degree of congestion, since at that point in time it takes a long time to either drive into that region (high authority score) or it takes a long time to drive out of that region (high hub score). Meanwhile, the analysis of these importance measures in the spatial graphs is not very interesting, since in this case regions are identified as structurally important if the distance between them and their neighbors is high. Thus, regions that are far away from everything would be identified as hubs or authorities.

4.2 Community Detection

One of our goals is to identify communities in our network in order to shed light into the structure of our traffic flow data. 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. In order to accomplish this, we will use a modification to the Girvan-Newmann Strength of Weak Ties algorithm for weighted directed graphs.

This community structure detection algorithm, invented by M. Girvan and M. Newman in 2002 [3], is based on the betweenness of the edges in the network. The idea is that the betweenness of the edges connecting two communities is typically high, as many of the shortest paths between nodes in separate communities go through them. So we gradually remove the edge with highest betweenness from the network, and recalculate edge betweenness after every removal. This way sooner or later the network falls off to two components, then after a while one of these components falls off to two smaller components, etc. until all edges are removed. This is a divisive hierarchical approach and the result is

a dendrogram. In the context of our temporal data, the identified communities can be interpreted following the same intuition we used for betweenness centrality. Since the algorithm splits the graph progressively based on computed betweenness measures, then the "borders" between communities progressively identify areas through which most of the trips flow. Therefore, the time series of the way the border evolves can be seen as the way in which traffic is flowing through the city. This can be contrasted with the spatial border which identifies the communities for which transit within a community implies covering a longer distance than transit to a different community.

5 Results and Findings

To uncover information about traffic flow within each city we began by constructing the measures described in Section 4. For the temporal networks, each measure was computed for each city and each hour of the day. Additionally, for each city, each measure was computed using spatial distances as edge weights. Armed with the 24 temporal graphs for each city with weights for each measure and the spatial reference graph for each city and measure, we deduce traffic flow. The visualizations presented in this section plot the nodes according to the geographical location of the centroid of the region each node represents. This allows for a clearer understanding of the underlying city structure on which trips take place.

Note: For brevity in this report we chose to only include a small subset of this data, focusing only on one hour of the day and on specific city measurement pairs. There is a clear drawback of analyzing static images in a limited amount of space compared to viewing these images together in a timeline as seen in the animations we put together here: [`https://goo.gl/NDKtfZ`].

5.1 Centrality

In this section we analyze various centrality measures for multiple cities considering both the temporal and spacial graphs. In-degree, Out-degree, Betweenness and PageRank scores were constructed by employing various functions in `snap`. To compute closeness scores, hubs and authorities we utilized the `igraph` [1] package in R.

5.1.1 Node Degree

We begin by exploring our most central nodes in the temporal graph by weighted in- and out-degree, as defined in Section 4.1.1. The top nodes by degree in Johannesburg at 9:00 a.m. are shown in Figure 3 and 4. As expected, we see the city center darkening in color for both the in- and out- degree weighted graphs as rush hour continues through the morning.

Figure 3: Johannesburg: In-Degree

Figure 4: Johannesburg: Out-Degree

5.1.2 Betweenness Centrality

Figures 5 and 6 show the nodes (regions) of Manila colored by betweenness centrality in the temporal graph at 9:00 a.m. and in the static spatial graph, respectively. As in the rest of the figures in this section, darker tones of blue denote higher betweenness and the coordinates of the nodes in the graph correspond to their true latitude and longitude.

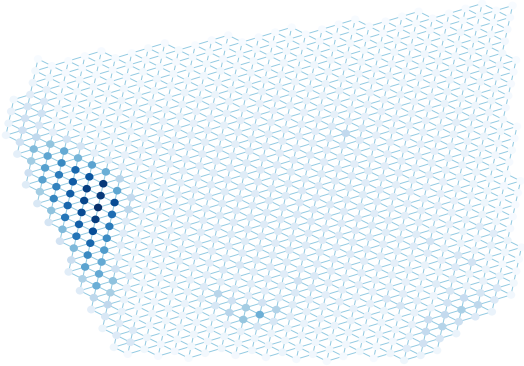


Figure 5: Manila Betweenness Centrality: Temporal

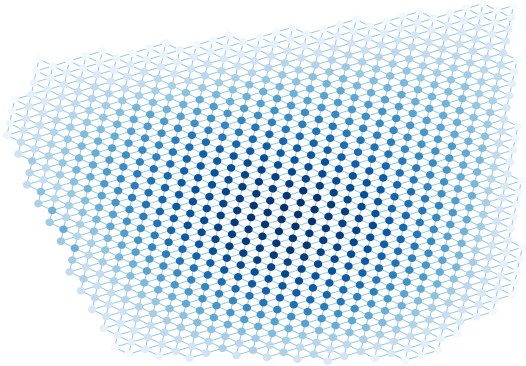


Figure 6: Manila Betweenness Centrality: Spatial

Since the regions in Manila are constructed using a regular grid, betweenness centrality in the spatial network decreases uniformly as a function of distance to the (geographical) center of the network, as shown in Figure 6. However, we notice that the top central nodes in terms of temporal betweeness are located in only a few sub-regions of the network. We found that the biggest "hot" spot, located on the East coast of Manila, corresponds to the city center, where most of the urban movement is concentrated. The nodes in these sub-region, along with other city zones highlighted in darker tones on blue, correspond to regions where traffic flow is high during weekday rush hours. This is a good example of how we lose valuable information by looking only at static centrality measures, particularly when the regions are modeled through an equidistant grid.

5.1.3 Closeness

We see in Figure 7 and Figure 8 that the closeness scores for the temporal graph versus spatial graph in Washington D.C. at 9:00 a.m. are in fact quite similar. Both graphs seem to decay uniformly from the center towards the outer rim. This result makes intuitive sense as the Washington D.C. graph produced by Uber was centered around Washington's city center. An interesting difference that appears is the lighter nodes found at the top middle of temporal graph versus the very dark nodes in the distance graph. These nodes could represent a less developed part of downtown Washington D.C. which could be developed further to help alleviate stress on the nodes seen south of this area. Another interesting takeaway from these graphs is that most of the urban mobility is concentrated around the river which corresponds to the white space in the center of the graph.

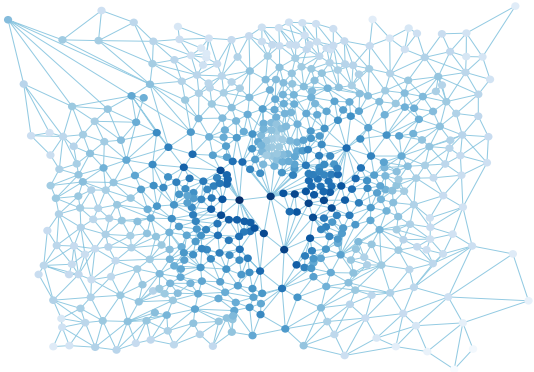


Figure 7: Washington Closeness: Temporal

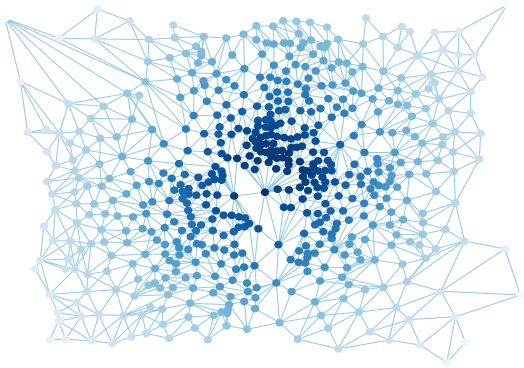


Figure 8: Washington Closeness: Spatial

5.1.4 PageRank

We compute the PageRank scores for all nodes in Paris using the `GetWeightedPageRank()` of snap. The resulting graphs for the 9:00 a.m. temporal graph and the spatial graphs are shown in Figures 9 and 10. Darker tones of blue denote higher PageRank score; meanwhile, the coordinates of the nodes in the graph correspond to their true latitude and longitude.

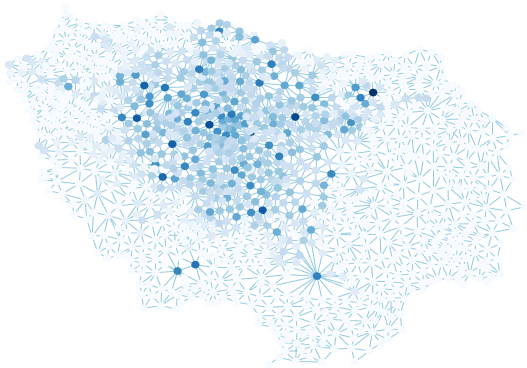


Figure 9: Paris PageRank: Temporal

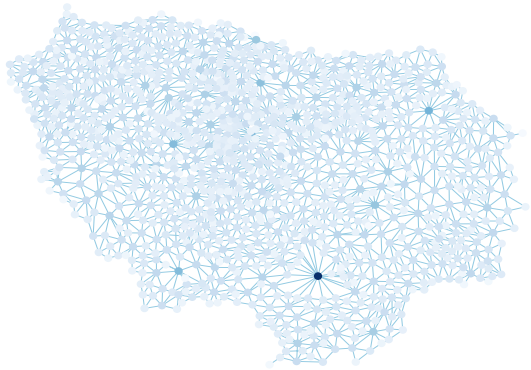


Figure 10: Paris PageRank: Spatial

PageRank centrality in the temporal graph captures the ring structure of the Paris: in the core lives the city of Paris, where most of the urban dynamics are conglomerated; meanwhile, the concentric rings around the "small ring" exhibit decreasing importance as measured by temporal PageRank score. As noted in Section 4, nodes with high page rank in the spatial graph are not very interesting, as this measure is relatively uniform across nodes, except for regions that are further away from their neighbors, which obtain a high PageRank score.

5.1.5 HITS

To compute the hubs and authorities in our graphs we used the `hub_score()` and `authority_score()` functions of the `igraph` [1] package in R. In Figures 11 and 12 we plot nodes according to their hub and authority score, respectively, for Johannesburg at 8:00 a.m. Nodes with a higher score are plotted in a darker shade. We note that the nodes identified as hubs and the nodes identified as authorities are very similar. We can also see that central nodes based on these measures are mostly located in the periphery of the city, which is expected to exhibit high travel times to and from it.

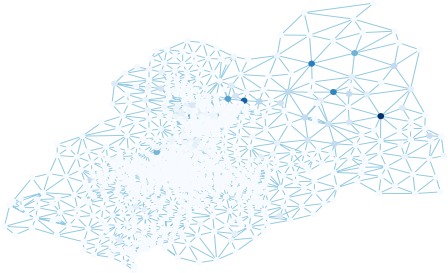


Figure 11: Authorities

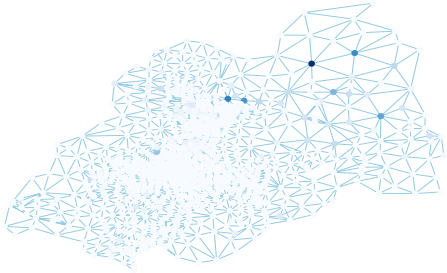
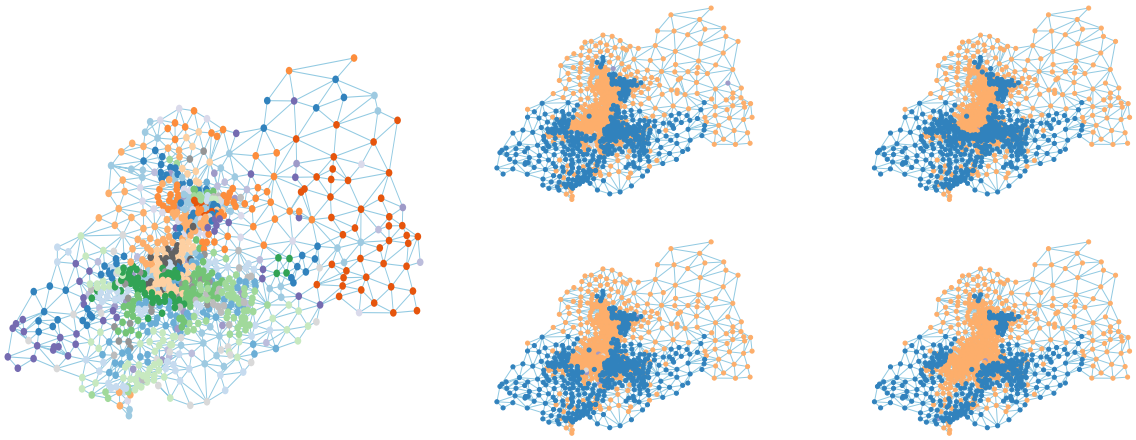


Figure 12: Hubs

5.2 Community Detection

We use the `igraph` package in R to find communities in our temporal graphs using the Girvan Newman algorithm. Figure 13a shows the detected communities for Johannesburg when using the spatial graph. In contrast, Figure 13b which shows the detected communities at four different times of day: 12:00 a.m., 1:00 a.m., 5:00 a.m. and 10:00 a.m. We identify two communities in the temporal graph: the blue and the orange regions. We can clearly see how the orange community grows in size as traffic starts flowing towards the city center during morning rush hour. If we had continued this analysis later into the day, we would see a contracting behaviour as it gets later in the night.



(a) Distance based communities

(b) Travel time based communities

Figure 13: Johannesburg Communities

6 Conclusions and Future Work

The findings of this study reveal the usefulness of centrality measures and community detection for modeling traffic flow. By analyzing how these measures change throughout the day in our temporal networks as well as how they compare to measures in the static spatial network, we were able to identify mobility patters and pinpoint traffic bottlenecks rush hours across cities. The clear next step would be to propose a solution to alleviate the problem in the identified bottlenecks and measure the impact of the solution on traffic flow through simulated data.

In the future, we would like to perform a similar analysis on weekend traffic data to reveal insights into how people's travel patterns change throughout the week and how the key stress points of cities change from weekdays to weekends. Similarly, we would like to extend our model to incorporate data from the rest of the year or even previous years in the hopes of uncovering seasonality patters in urban mobility.

Finally, we found Manila's lattice graph to be very useful for obtaining insights versus the seemingly arbitrarily constructed regions in the other cities to which we were constrained by the way the Uber data was already aggregated. In further work we would seek to ensure more regular graphs are constructed prior to starting the analysis. This would allow modeling the natural traffic flow in a city, rather than being constrained by a predefined set of boundaries as commented on Section 2.2.

7 References

[1] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[2] 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.

[3] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[4] 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.

[5] 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.

[6] 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.