New York Taxi Network: Community Structure and Predictive Analysis

Ben Weems, Emily Field, Tucker Ward

December 12, 2016

Abstract

This paper seeks to provide novel algorithms and analyses of taxi networks to understand and predict important features of human society. Recent work shows that transportation networks reflect social phenomena, and taxis are particularly responsive as they can rapidly adjust supply to meet demand [1] [2] [3]. We develop two novel analyses that show how taxi data can be used to find and predict changes in important social statistics. First, we develop a community structure algorithm around taxi networks that could enable city planners to better invest in infrastructure and transportation upgrades. Next, we use taxi data to predict housing price movement; a useful tool for everyone from investors to public housing administrators. Our results show that the power of taxis can unlock key insights into social phenomena before they are visible to the public.

1 Introduction

In New York City, people travel over one million miles a day by taxi [4]. Human transportation on this massive scale has potential to provide granular insights into activity patterns within a city. Transportation, and access to it, is intertwined with important metrics, from crime statistics to housing prices. Thus, understanding travel patterns has the potential to predict or explain social phenomena. We present that NY taxi data can help predict housing prices and discover connected communities, potentially helping focus infrastructure investment dollars. As of 2009, New York City releases a rich data set that includes pinpointed pickup and drop-off locations for all taxi rides [5]. This data set now includes over a billion taxi rides, and, despite being ripe for network analysis, previous work is shockingly sparse.

We show how taxi data can yield insights into social phenomena through careful analysis. In particular, we show how communities can be detected using taxi data, and how transportation networks can be used to predict housing prices. These serve as proofs of concept, providing a novel approach that paves the way for future research. In addition to showing utility for everyone from city planners to real estate investors, our fundamental contribution is demonstrating that transportation network data–specifically taxi data–is not only reflective of important social phenomena but predictive of it. Observe a map of New York taxi rides:

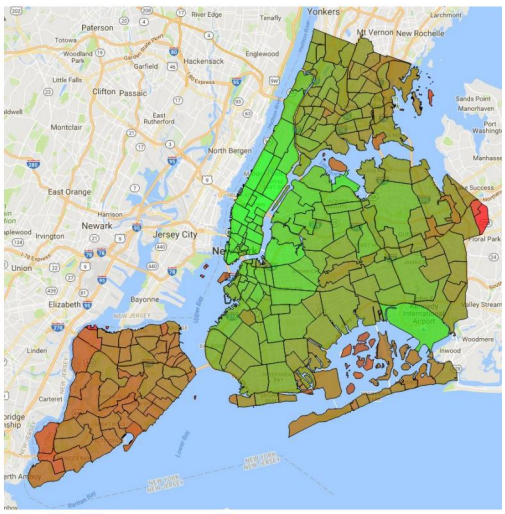

Figure 1: Heat map of trip volume by neighborhood, green represents high and red low volume.

2 Related Work

There is a robust body of literature focused on applying analyses of transportation networks to social metrics. Latora et al. explore the small-network characteristics of the Boston subway system [1], while Colizza et al. use the global airline network to predict the spread of global epidemics [3]. Guimera et al. also provide research on the global airline network, specifically attempting to identify the roles of individual cities [2]. Prominent statisticians have also shown how Uber disrupts traditional community structures and impacts public transportation networks, including taxis [21].

Moreover, previous work has used characteristics of relation graphs to reveal relationships between variables, such as collaboration tendencies in academic networks [10]. Significant climate change analysis also relies on graph theory [17], while other researchers tease out social dynamics from networks of humans and other biological networks [8]. All of that work convinced us that an exploration of social phenomenon using taxi transportation was likely to be fruitful.

That work served as inspiration for applications of transportation network structure, but for our algorithmic implementation we turned to West and Leskovec (WL)'s work, "Automatic Versus Human Navigation in Information Networks", which focuses on the localized knowledge a human or algorithm may have of some nodes during shortest path discovery [7]. However, the task of shortest paths loses significance in our NY Taxi data, as each edge is an independent trip; thus, we began to investigate community structure.

Girvan and Newman (GN) define the 'community structure' of a graph to be the concept that network nodes are joined together in 'tightly-knit groups between which there are only looser connections' [8]. GN's research presents two algorithms that provided a framework for how we would think about local structure in our graph. Namely, that there are two, equally valid, approaches: 1) starting with the original graph G, and peeling away nodes with high betweenness to reveal community structure 2) starting with an empty new graph, and joining those nodes in G that are very strongly connected into a community in the new graph. However, they analyzed only unweighted, undirected graphs, and thus, in addition to betweenness, they had to formulate an edge weight that would be used to measure connectedness (or lack of connectedness) between any two nodes.

When considering community structure alone (and thinking less about traffic flows), the edge weights are the most important, distinctive feature of our graph – the sheer magnitude of trips between regions. Thus, rather than try to construct custom weights for either of GN's approaches, or calculating betweenness, we searched the literature to find methods of using these edge weights to build meaningful communities. Lu et al. developed an algorithm for community structure on a weighted graph similar to the GN approach, and defined two key metrics to transform edge weights into a measure of connectedness:

1. Belonging Degree (denoted $B(n, C)$)
2. Conductance (denoted $\phi(C)$)

They use these measures of connectedness to greedily build up communities. We will use a modified Lu et al.'s community detection algorithm on our NY Taxi data.

3 Taxi Data, Analysis Tools, and Taxi Graph Summary

The City of New York publishes yearly taxi ride statistics of each individual taxi ride, including date, time, dropoff/pickup coordinates, fare paid, taxi number, and more. The data cover every year since 2009. The data naturally form a graph of taxi rides that connect all parts of New York City, with pickup/dropoff coordinates representing nodes and trips representing edges, weighted by total number of trips between those two vertices. There are approximately 165 million trips per year.

3.1 Data Pre-Processing

Each trip's origin and destination are measured in GPS coordinates with sub-meter accuracy, necessitating some pre-processing in order to generate a graph that is meaningful on a human scale. We divide trips by two demarcations: neighborhood and ZIP code, depending on the specific analysis requirements and data available. Mapping to neighborhood or ZIP code directly required nearly 30 processing hours per year of data, so we first mapped all trips to a 300x300 grid overlaying New York City. We mapped each trip to its nearest grid point and then converted those 90,000 grid points to neighborhoods and ZIP codes, resulting in a processing time of only 3 hours per year of

data. A 300x300 grid gives trip granularity of about 2 square blocks per grid coordinate, which is large enough to effectively reduce computation time but small enough to be precise on any desired human scale.

3.2 Graph Data Processing

Our neighborhood/ZIP framework allows for two arbitrary filtering techniques: filtering and grouping. Filtering removes trips from the final graph based on arbitrary user-defined criteria, eg, time of day. Grouping takes the filtered results and divides them into buckets, eg, each day of the week. The result is that trips can be filtered and grouped as desired, with processing time on the order of 3 hours per 1 year or 165 million trips. These final groups can then be output on a neighborhood, ZIP code, or raw 300x300 grid basis to form the final network used in analysis. As Snap.py does not support many needed features, such as edge weights, we implemented all graph representation, manipulation, and creation functionality from scratch, excepting coordinate-to-neighborhood functions which were drawn (with permission) from Austin Benson's work [11].

3.3 Graph Statistical Analysis

Our suite also supports automatic collection of key statistics. Statistics include betweenness centrality, closeness centrality, and more common metrics such as mean, median, and standard deviation of trip numbers for all edges. For example, our analysis shows that in December 2013 on weekdays there were on average 63 daily taxi trips from Penn Station to Rockefeller Plaza, with a standard deviation of 22.

3.4 Taxi Graph Summary Statistics

Given the extremely high variance of all statistics, individual metrics are rarely meaningful without context. Thus we focus on generating graphs and charts to express overall network topology.

Measurement	Median	Mean	Std Deviation
Edge Weights	18	9,766	147,190
Net Edge Flow	12	2,648	87,531
Node Volume	13,854	1,263,999	5,522,274

Figure 2: Summary Statistics of the Taxi Graph

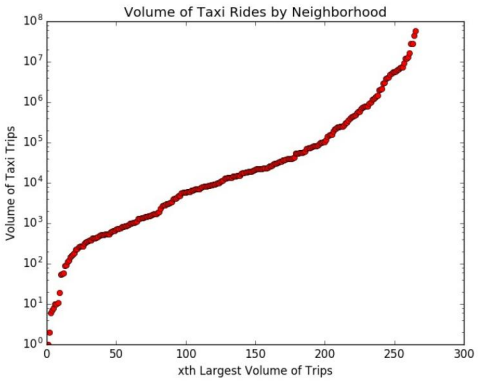

Figure 3: Neighborhood Volume Distribution

Figure 1 is a heat map with green representing high volume nodes and red representing low volume. The data show that taxi rides are heavily concentrated in the Manhattan borough, and volume rapidly decreases with distance from Manhattan – with a few notable exceptions such as JFK airport. We expect that there are three major reasons for this distribution: first, people who live or work in Manhattan can more easily afford taxi trips, second, people who live or work in Manhattan must use taxis because parking is expensive, and third, there is a higher population density near Manhattan. These three reasons together create a fourth reason, which is that taxis are readily available and people rely on them for everyday transportation.

The edge weight distribution (Figure 4) demonstrates that the vast majority of taxi trips occur between a few neighborhoods, most of which are located in Manhattan, while the outer boroughs have less volume. Neighborhood betweenness centrality scores (Figure 5) demonstrate that most neighborhoods are peripheral in the taxi network (neighborhoods with a 0 betweenness score are represented as a score of 0.1). As trips are independent, the significance of betweenness is diminished (as is any score based on shortest paths), but this graph shows that both edge weights and hub nodes are highly centralized based on taxi volume.

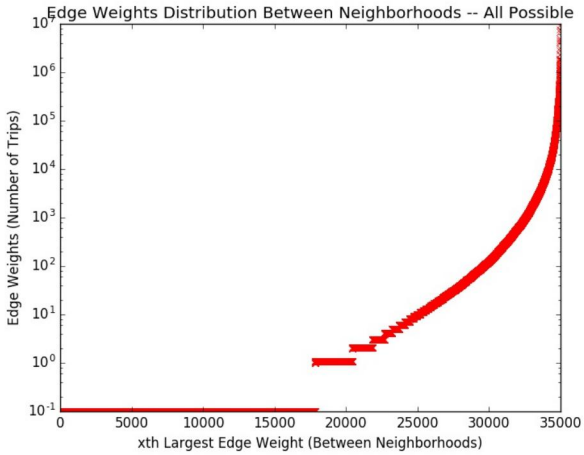

Figure 4: Edge Weight Distribution

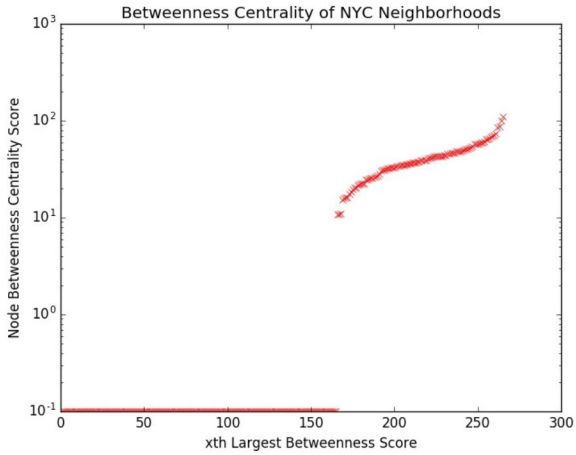

Figure 5: Neighborhood Betweenness Centrality

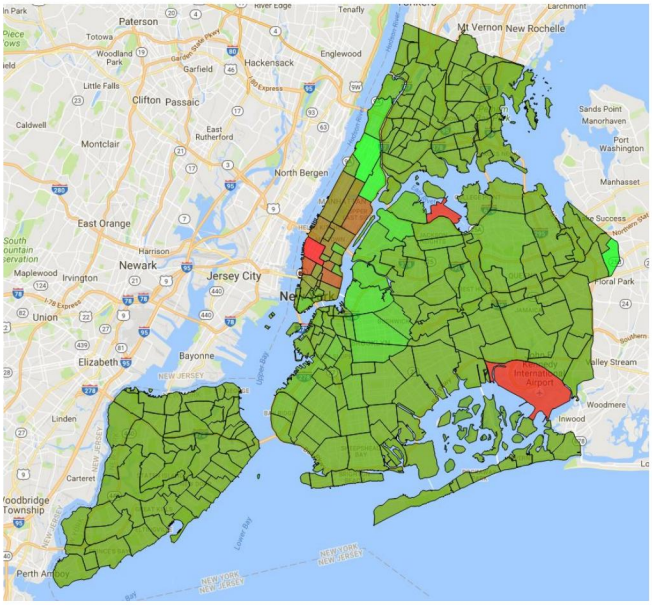

Figure 6: Net Node Flow

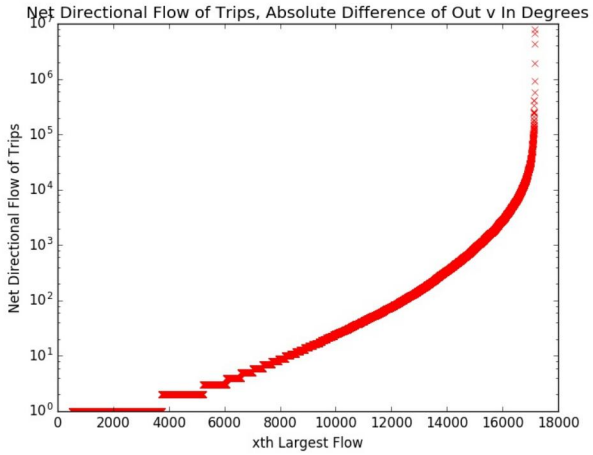

Figure 7: Net Edge Flow

Given the high concentration of edge weights, the net direction of taxi trip flows provides a further interesting perspective on the network's structure. Figure 6 shows the net flow of trips in to and out of each neighborhood, with red nodes a strong net sink and green nodes a strong net source. The red over JFK and LaGuardia airports are completely sensible as many people travel to the airport, as is the case with downtown Manhattan. The relative lack of outgoing taxi trips is likely explained by those nodes' hub status and use of public transportation. People who were in a hurry to make their flight might take a taxi there, but be unwilling to accept the cost – and airport surcharge – when time is not a factor. The corresponding graph for edge net flow in in figure 7, which shows that most edges have a net difference of less than 1,000 trips, but the top 10% of nodes have an extremely large net flow–as high as 1 million. This difference between in and out degrees is important in our later exploration of the relation of taxi data to real estate prices.

4 Community Detection and Analysis

New York City is famous for its distinct and diverse neighborhoods, and the NYC Taxi data provides a unique look into how people move within and between these neighborhoods. By analyzing the data with a community detection algorithm, we seek to reveal neighborhoods that have a large proportion of trips between each other that can be

considered a distinct community. This work is useful for city planning and understanding taxi data structure before applying predictive analysis.

4.1 Algorithm

To describe the core of the Lu et al. community detection algorithm, we define a community C as a set of two or more neighborhood nodes, and belonging degree as

$$B(u, C) = \frac{\sum_{v \in C} w_{uv}}{\sum_{k \in N_u} w_{ku}}$$

where N_u is the set of all nodes that have a weighted edge with u in G. A high belonging degree means a node should likely be in the given community. We define conductance as

$$\phi(C) = \frac{cut(C, G/C)}{w_c}.$$

where $cut(C, G/C)$ is the sum of the edge weights going outside of the community from inside the community, and w_C is the total edge weights of all nodes in the community (to nodes both in and out of the community) [14]. Intuitively, it measures how strongly this group of nodes is connected to the rest of the graph – low conductance communities are isolated and thus good communities [14].

The algorithm works as follows:

1. Find the two nodes with the highest edge weight between them, our initial community C.
2. Calculate the conductance of this community C, $\phi(C)$.
3. Out of all the nodes that have an edge to at least one node in the community C, but are not in the community C, find the node with the highest belonging degree, $B(n, C)$.
4. Add that node n to the community, and calculate its new conductance $\phi(C_n)$.
5. If $\phi(C_n)$ is less than $\phi(C)$, add the node n to C and repeat the process. If it is not, mark C as a complete community, remove all of the edges within the community C from the global network, and repeat the process. Continue until no edges are left.

This algorithm solves the majority of the shortcomings of GN's analysis, but we still had to address the challenges unique to our data set, while simultaneously ensuring that we don't over fit the algorithm to our data. Our modifications and potential improvements are described below:

1. Lu et al. do not address directed edges. To tackle this problem, we considered the relevance of the direction in our edges. From *Upper East Side* to *Midtown*, there might be 5 million edges, and in the reverse direction, there may only be 4 million edges. Thus, to just consider the connectedness of the community, we sum the two edge weights to serve as a single edge weight on an undirected graph between the two nodes.

2. Lu et al. did not add the constraint that communities are distinct. Nodes could appear in several communities, but if any two communities overlapped by greater than 50 percent, they were merged. We made a slight modification, forcing a uniqueness constraint that each neighborhood could only belong to one community.

3. Similar to GN, we found that running the distinct community algorithm on a small test set will sometimes classify outlier nodes (nodes that connect communities) into one of the communities, rather than identify each one as a distinct community. This error is because of the simplicity of the belonging degree calculation – if edge weights are small overall, then there can be a high belonging degree as it is a fraction of total edge weight. To address this issue, we slightly modified the last step of the algorithm, to read: if $\phi(C_n)$ is less than α $\phi(C)$, where $\alpha < 1$, add the node n to C and repeat the process with an alpha of .95. This constraint only allows the node to be added to the community if it reduces conductance by at least 5 percent – therefore, outliers will be less likely to end up in a particularly strong community, and will instead form their own, weaker communities. Since we are looking at how key communities change over different splits of the data, we can ignore these outlier, low volume communities.

4.2 Results and Insights

Rather than simply run the community detection algorithm at a point in time, we ran the algorithm on variations of the network as described in the below three scenarios:

1. Weekday vs. Weekend: We divided all taxi trips for 2013 into two networks, one network has all trips that occurred from Monday to Friday, and the other has all trips that occurred on Saturday and Sunday.

2. Uber's Entrance Into NYC: Uber officially launched in May of 2011, but only began to represent a meaningful amount of trips in 2012 and 2013 – reaching 1 million trips in May 2013 [20]. They grew substantially in 2014 and 2015, increasing their market share from 4 percent to 15 percent, and they hit 100 million trips in August of 2016 [21]. Thus, we created two networks, the first with trips from 2012-2013, and the second from 2014-2015.

3. Midday vs. Late Night: We divided all taxi trips for 2013 into two networks, one network has all trips that occurred between 12:00 PM and 4:00 PM, and the other has all trips that occurred between 8:00 PM and 12:00 AM.

Weekday vs. Weekend

Weekday	% Of Trips Within Community	Total Trip Volume
Chelsea, Hell's Kitchen, Theater District, Upper West Side, Central Park	15.6	46,242,019
Upper East Side, Midtown	15.2	48,507,237
West Village, East Village, Greenwich Village, SoHo, Tribeca, Battery Park City	13.7	27,035,508
Williamsburg, Lower East Side, Greenpoint, Bushwick	7.7	4,829,633
Gramercy, Murray Hill, Kips Bay, Flatiron District, Stuyvesant Town, NoHo	7.6	20,817,711
Astoria, Long Island City, Ditmars Steinway	6.3	1,771,118
Harlem, East Harlem, Morningside Heights	5.8	5,377,791

Weekend	% Of Trips Within Community	Total Trip Volume
Upper East Side, Midtown	11.6	16,622,546
Hell's Kitchen, Chelsea, Upper West Side	11.3	15,533,908
Williamsburg, Lower East Side, Greenpoint, Bushwick	8.8	3,453,916
Financial District, SoHo, Tribeca, Battery Park City	7.5	5,423,548
West Village, East Village, Greenwich Village	7.2	9,781,015
Astoria, Long Island City, Ditmars Steinway	7.0	1,014,820
Harlem, East Harlem, Morningside Heights, Washington Heights	6.6	2,582,744

The total trip volume is the total number of edges to and from every neighborhood in the community, from every node in the network. The percentage of trips within the community is the total edge volume exclusively between the neighborhoods in the community, divided by the total trip volume. We notice a few key differences between the weekday and weekend communities. The third community listed for Weekday trips, beginning with *West Village*, splits into two distinct communities on the weekend *West Village, East Village, Greenwich Village* and *SoHo, Tribeca, Battery Park City* with the addition of the *Financial District*; the neighborhoods in this latter community are largely business districts that include the New York Stock Exchange, One World Trade Center, and the headquarters of Goldman Sachs, Deutsche Bank, and many more large corporations [18]. *West Village, East Village, Greenwich Village* are largely residential neighborhoods that boarder one another geographically. Within these neighborhoods, there are many small businesses, and a major university, NYU [19]. Thus, it seems that the influx between these two groups of neighborhoods is driven by the weekday commute to work, and that people tend to travel within the distinct communities during the weekends.

While the community of *Williamsburg, Lower East Side, Greenpoint, and Bushwick* does not change from weekdays to weekends, it is one of the few meaningful communities that spans across a bridge (to cross the East River from *Lower East Side* to the three Brooklyn neighborhoods). Also, the community of *Upper East Side and Midtown* has high volume, and remains a distinct community both during the week and on weekends. It is surprising that we don't see strong communities during the week between *Upper East Side* and downtown communities like *Battery Park City* – this suggests that long, downtown commutes may be taken by subway, Uber/Lyft, or independent driving.

Midday vs. Late Night

Midday (12 PM - 4 PM)	% Of Trips Within Community	Total Trip Volume
Hell's Kitchen, Chelsea, Theater District, Upper West Side, Central Park, Morningside Heights	17.9	13,245,926
Upper East Side, Midtown	15.7	13,689,303
West Village, SoHo, Tribeca	7.8	3,926,127
Gramercy, East Village, Kips Bay	6.1	3,657,313
Murray Hill, Flatiron District, Greenwich Village, Stuyvesant Town, NoHo	4.7	4,605,918
Williamsburg, Lower East Side, Nolita, Greenpoint, Chinatown	4.3	1,223,391
Financial District, LaGuardia Airport, Battery Park City	3.4	3,176,008

Late Night (8 PM - 12 AM)	% Of Trips Within Community	Total Trip Volume
Upper East Side, Midtown	11.8	13,972,228
Upper West Side, Theater District, Morningside Heights, Harlem, Washington Heights	8.4	7,717,175
Williamsburg, Lower East Side, Greenpoint	7.2	2,619,535
Hell's Kitchen, Chelsea	6.1	8,404,921
West Village, East Village	4.0	6,744,118
Flatiron District, Kips Bay, Stuyvesant Town, NoHo, Nolita, Battery Park City	3.5	4,770,299
Financial District, Tribeca	2.8	2,561,534

By filtering in these 4 hour blocks, we eliminate the vast majority of commuter traffic from the graph. Interestingly, we again see the neighborhood crossing the bridge from *Lower East Side* into Brooklyn both during the morning and the evening, although with a greater percentage of total trips within the community during the evening. We also notice two distinct communities that form at night *Chelsea, Hell's Kitchen* and *West Village, East Village* with very high volume. Both of these communities contain two neighborhoods that are geographically connected – perhaps in the evening people are less willing to walk short distances or take the subway due to safety concerns.

Uber's Entrance Into NYC

2012-2013	% Of Trips Within Community	Total Trip Volume
Upper East Side, Midtown	14.2	129,663,201
Hell's Kitchen, Chelsea, Theater District, Upper West Side, Central Park	15.8	128,859,595
West Village, East Village, Greenwich Village, SoHo, Tribeca, Battery Park City, Financial District	16.0	87,954,750
Gramercy, Murray Hill, Kips Bay, Flatiron District, Stuyvesant Town, NoHo	7.4	57,422,694
Williamsburg, Lower East Side, Greenpoint, Bushwick	7.7	15,814,285
Harlem, East Harlem, Morningside Heights, Washington Heights	6.6	16,886,045
Astoria, Long Island City, Ditmars Steinway, Sunnyside, Woodside	8.9	7,030,301

2014-2015	% Of Trips Within Community	Total Trip Volume
Upper East Side, Midtown	14.2	120,367,946
Chelsea, Hell's Kitchen, Theater District, Upper West Side	14.2	114,117,427
West Village, East Village, Greenwich Village, SoHo, Tribeca, Battery Park City	13.9	71,617,249
Sunnyside, Woodside, Maspeth, Elmhurst, Jackson Heights, Corona, East Elmhurst	8.4	3,328,674
Gramercy, Murray Hill, Kips Bay, Flatiron District, Stuyvesant Town, NoHo	7.8	53,644,257
Williamsburg, Lower East Side, Greenpoint, Bushwick	7.5	15,066,194
Astoria, Long Island City, Ditmars Steinway	7.1	5,367,099

Surprisingly, we did not see much change in community structure as Uber grew in New York. We do notice a decrease in volume in the top communities in 2014-2015, which coincides with Uber's growth to 15 percent market share in New York City in 2015 [21]. While we expected Uber's growth to be concentrated in quickly gentrifying neighborhoods, the results suggest that Uber's growth is likely evenly distributed throughout the city. This even distribution is also reflected in FiveThirtyEight's analysis of Uber's year over year growth in each neighborhood [21].

Conclusion and Challenges It is inherently difficult to evaluate community detection on real networks, as the underlying community structure is unknown. However, Lu et al. do run their conductance based algorithm on synthetic weighted networks with power-law distributions of node degrees against two other weighted network community detection algorithms; they found that their conductance based algorithm outperforms the other two algorithms with regards to the communities, their conductance, and their average modularity [14]. Also, the Girvan and Newman approach would not provide meaningful results on our graph, as it is designed for unweighted networks, and our

weighted graph is nearly complete. Furthermore, it is important to note that unlike some of the graphs that have very strong, well-defined community structure, like Girvan and Newman's college football conferences, the taxi trips within New York are relatively evenly distributed throughout the city, as seen by the scale of the percentage of trips within the community.

5 Housing Data

Now that we have explored and understood the structure of the network, we will apply patterns in its structure to investigate correlations between the NY taxi data network and economic and sociological metrics. We focus on housing data, specifically the median cost per square foot (CPSF) in US Dollars of housing in New York separated by zip code. We separate by zip codes because Zillow provides monthly average CPSF for every zip code in the country [15]. Zip code to neighborhood conversion is not straightforward and could sacrifice the integrity of the data. When we layer this data on the network, we find many correlations between the NY taxi data and the cost of housing in NYC. We are even able to generate predictive correlations about housing price based on the role a zip code plays in the network. We break the analysis into three sections: node degree vs. housing price, expected CPSF of neighboring zip codes, and predictive price analytics.

5.1 Node Degree vs. CPSF

In this section, we discuss the relationship between a node's degree in the network and the median CPSF of that node. We examine in degree, out degree, and total degree for comparison. If the in or out degree is 0, the point is not plotted due to the logarithmic scale.

5.1.1 Results

We see a strong positive correlation between degree and CPSF. The functional relationship between these properties is a power law, with negative α. The relationship is stronger for in degree, shown in Figure 8, than it is for out degree, shown in Figure 9. We see that $\alpha_{in} = -0.185$ compared to $\alpha_{out} = -0.150$. The correlation of the logarithms is $r = .824$ when using in degree, compared to $r = .678$ for out degree. The relationship for total degree, shown in Figure 10 is slightly stronger than that seen for in degree, with a correlation coefficient of $r = 0.836$. All of these correlations have $p < 10^{-9}$.

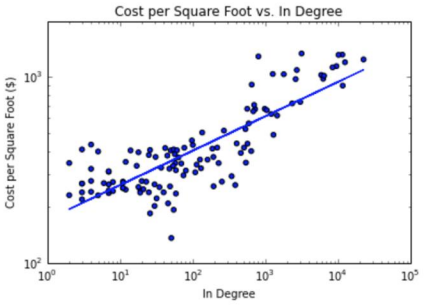

Figure 8: Log-log scatter plot of CPSF vs. In Degree of zip codes in NY taxi network.

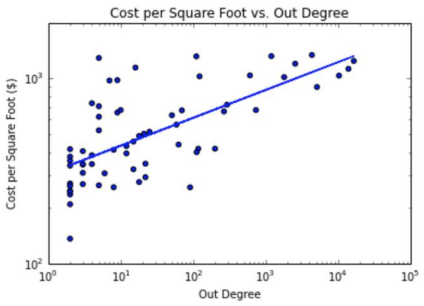

Figure 9: Log-log scatter plot of CPSF vs. out degree of zip codes in NY taxi network.

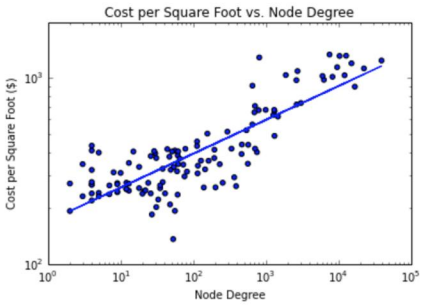

Figure 10: Log-log scatter plot of CPSF vs. total degree of zip codes in NY taxi network.

5.1.2 Analysis

The strong correlation between high income areas and high degree shows that high CPSF nodes bookend most trips. This makes sense based on observation, as Manhattan districts have higher CPSF and more taxis. However, the out degree plot, Figure 8, provides a surprising element: nodes with low out degree but high CPSF and high in degree. One hypothesis for these outliers is that these are high income areas outside the heart of the city without easily available cabs. Residents may use public transportation or ride-sharing options from home, but then when taxis are available inside the city will take advantage of them to return home.

5.2 Expected CPSF of Neighbors

This section explores the expected cost of housing if a random edge is followed into or out of a target zip code. We analyze both "in" and "out" neighbors, where an "in" neighbor means that a trip was taken from that zip code into the target zip code, and an "out" zip code means a trip that originated in the target zip code went out to that zip code. This is calculated for "in" as follows, with ζ as the expected cost, n as our target zip code, E as all edges of n, and μ as a mapping of zip codes to average CPSF:

$$\zeta = \frac{\sum_E \mu[E_{dst}]}{|E|} \tag{1}$$

5.2.1 Results

The results once again show a strong correlation, showing that wealthier areas, where average property price is used as a proxy for wealth, interact with other wealthy areas. This trend is seen in both the "in" graph in Figure 11 and the "out" graph in Figure 12. Note that we now put CPSF on the x axis, and consider expected value of neighbors as the dependent variable. We see that the correlation is stronger for the "out" neighbors, with $p = 2.76 * 10^{-9}$ vs. $p = 6.94 * 10^{-5}$ for the "in" neighbors. The magnitude of α is also greater for the "out" neighbors, at $\alpha = -0.536$ vs. $\alpha = -0.175$ for the "in" neighbors.

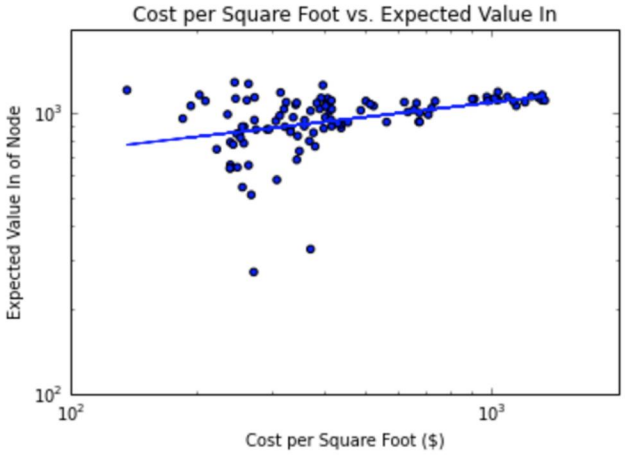

Figure 11: Log-log scatter plot of CPSF vs. Expected Value of "In" Neighbors.

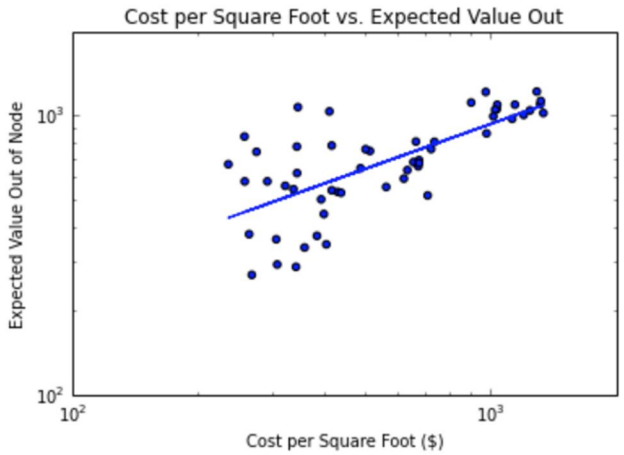

Figure 12: Log-log scatter plot of CPSF vs. Expected Value of "Out" Neighbors.

5.2.2 Analysis

The dominance of high CPSF nodes in total trips is demonstrated by the high average of these two graphs. However, we see that the expected value in is consistently higher than the expected value out for low CPSF nodes. Our community analysis helps explain this trend, as we know that there are connected communities, which may be low CPSF communities, in the graph that are both the source and destination of individual trips. For low CPSF nodes, these trips represent a large number of the trips out of the node. However, trips in from high CPSF nodes, that may have orders of magnitude greater total trip volume, dwarf the magnitude of trips within the community, causing the expected value in to exceed the expected value out, thus masking the community structure.

5.3 Predictive Analytics

In this section, we show the ability of our network to provide predictive insights into housing prices. We use taxi and housing data from late 2012 and early 2013 to predict housing prices at the beginning of 2014. Our known variable is the error of our housing price expectation function, which is developed based on the degree of zip codes in the taxi network. Our goal is to predict the percentage change in CPSF in that zip code. Our equation for log error, LE is below, where θ is our expectation function, γ is the true (2013) CPSF, and d is the degree of our target node.

$$LE = ln(\theta(d)) - ln(\gamma) \tag{2}$$

5.3.1 Results

We find a correlation of $r = -0.232$ between LE_{out} and percent change in CPSF with $p = 1.39 * 10^{-2}$. We find higher confidence in correlation with expectation based on the total degree, with $p = 7.26 * 10^{-5}$ for $r = -0.350$.

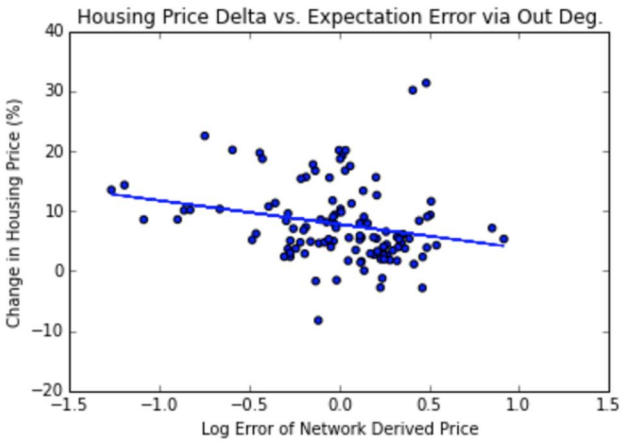

Figure 13: Scatter plot of LE of network derived price based on out degree vs. the true change in housing price from 2013 to 2014 of zip codes in the NY taxi network.

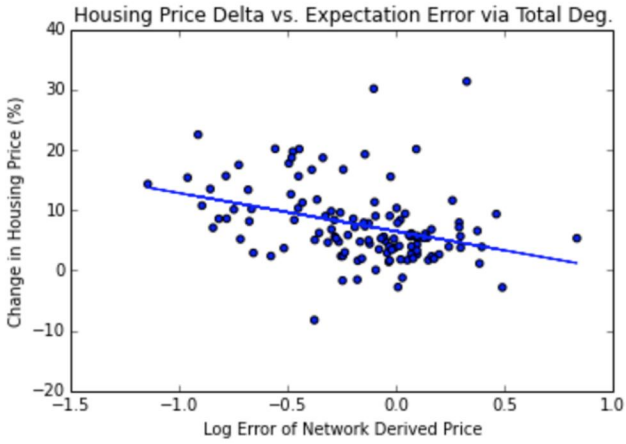

Figure 14: Scatter plot of LE of network derived price based on total degree vs. the true change in housing price from 2013 to 2014 of zip codes in the NY taxi network.

5.3.2 Analysis

These results demonstrate the potential for the NY taxi data to provide predictive insight into seemingly unrelated events in human society. While correlation doesn't imply causation, and stronger correlations exist (e.g. raw CPSF in 2013 has a stronger correlation with percent increase by 2014 than taxi data does), this result provides a proof of concept that there are latent features of the NY taxi network that can be used alongside other metrics to make meaningful predictions about the direction of important economic and societal metrics.

Analysis of the correlation above suggests that zip codes with a higher CPSF than expected based on degree in the network grew their CPSF over the 2013 year more than zip codes that underperformed their expectation. There are many possible reasons for this, although assigning causality would require substantially more data and analysis. For example, it is possible that regions with higher CPSF than expected have good access to public transportation, reducing the need for taxis and contributing to rising home prices.

6 Conclusion

In this paper we have shown how taxi data can be used to discover important characteristics of cities, from communities to housing prices, and how taxi trends can be used to predict changes in those phenomena. Future work on this topic will likely reveal even deeper insights with greater predictive power; we feel strongly that taxi data can also be predictive of crime statistics, or at the very least reflective of them. Analysis of transportation, a fundamental building block of society, can be used to help everyone from city planners to investors find better solutions for humanity's problems.

References

[1] Latora, Vito, and Massimo Marchiori. “Is the Boston subway a small-world network?.” Physica A: Statistical Mechanics and its Applications 314.1 (2002): 109-113.

[2] Guimera, Roger, et al. “The worldwide air transportation network: Anomalous centrality, community structure, and cities’ global roles.” Proceedings of the National Academy of Sciences 102.22 (2005): 7794-7799.

[3] Colizza, Vittoria, et al. “The role of the airline transportation network in the prediction and predictability of global epidemics.” Proceedings of the National Academy of Sciences of the United States of America 103.7 (2006): 2015-2020.

[4] “2014 Taxicab Factbook.” New York City, n.d. Web. 20 Oct. 2016.

[5] “NYC Taxi Limousine Commission - Trip Record Data.” NYC Taxi Limousine Commission - Trip Record Data. N.p., n.d. Web. 21 Oct. 2016.

[6] Altman, Alon, and Moshe Tennonholtz. Ranking Systems: The PageRank Axioms (n.d.): n. pag. Web. 20 Oct. 2016.

[7] West, Robert, and Leskovec, Jure. Automatic versus Human Navigation in Information Networks. Web. https://cs.stanford.edu/people/jure/pubs/navigating-icwsm12.pdf

[8] Girvan, Michelle, and M.E.J. Newman. “Community Structure in Social and Biological Networks.” N.p., n.d. Web. http://snap.stanford.edu/class/cs224w-readings/girvan01community.pdf.

[9] Brandes, Ulrik. “A Faster Algorithm for Betweenness Centrality.” (n.d.): n. pag. Web. 20 Oct. 2016.

[10] Newman, M. E. J. “Scientific Collaboration Networks. II. Shortest Paths, Weighted Networks, and Centrality.” Physical Review E 64.1 (2001): n. pag. Web.

[11] Benson, Austin. “Spacey Random Walks.” GitHub. N.p., 29 Apr. 2016. Web. 12 Dec. 2016. https://github.com/arbenson/spacey-random-walks/blob/master/scripts/form _taxi _trajectories.py.

[12] Opsahl and Panzarasa, “Clustering in weighted networks,” Social Networks, Volume 31, Issue 2, May 2009, Pages 155-163, ISSN 0378-8733, http://dx.doi.org/10.1016/j.socnet.2009.02.002.

[13] Clauset, Newman, and Moore. “Finding community structure in very large networks.” Web. http://snap.stanford.edu/class/cs224w-readings/clauset04community.pdf 18 Oct. 2016.

[14] Lu, Sun, Wen, Cao, La Porta “Algorithms and Applications for Community Detection in Weighted Networks.” N.p., n.d. Web. http://mcn.cse.psu.edu/paper/zongqing/tpds-zongqing15.pdf.

[15] “Data - Zillow Research.” Zillow Research. N.p., n.d. Web. 18 Nov. 2016.

[16] “NYPD - Office of the Chief of Department.” NYPD - Office of the Chief of Department. N.p., n.d. Web. 18 Nov. 2016.

[17] Laboratory, Oak Ridge National. “Long-Term Daily Climate Records from Stations Across the Contiguous United States.” Long-Term Daily Climate Records from Stations Across the Contiguous United States. N.p., n.d. Web. 18 Nov. 2016.

[18] “Financial District, Manhattan.” Wikipedia. Wikimedia Foundation, n.d. Web. 11 Dec. 2016. https://en.wikipedia.org/wiki/Financial_District,_Manhattan.

[19] “East Village, Manhattan.” Wikipedia. Wikimedia Foundation, n.d. Web. 11 Dec. 2016. https://en.wikipedia.org/wiki/East _Village, _Manhattan.

[20] “100 Million Trips in New York City.” New York. Uber Newsroom, 11 Aug. 2016. Web. 11 Dec. 2016. https://newsroom.uber.com/us-new-york/100million/.

[21] Bialik, Reuben Fischer-Baum and Carl. “Uber Is Taking Millions Of Manhattan Rides Away From Taxis.” FiveThirtyEight. N.p., 08 Mar. 2016. Web. 11 Dec. 2016. http://fivethirtyeight.com/features/uber-is-taking-millions-of-manhattan-rides-away-from-taxis/.