Analysis of Uber Ridership Patterns in New York City in 2021

Fonda Hu[a,1], **Maya Garg**[a], **Annie Cheng**[a], **and Lizabeth Annabel Tukiman**[a]

[a]Department of Mathematics, University of California, Los Angeles, CA 90095

Understanding the mobility patterns of different groups of people within a city is key to ensuring more equitable access to transportation. We explore the travel patterns of Uber users in New York City in 2021 by representing the pick-up and drop-off locations as nodes in a network, with edges pointing from the pick-up to the drop-off locations. We merge the original Uber dataset with a Taxi Zone Lookup dictionary to match location ID's with their corresponding locations, and also merge the NYC weather dataset to identify trip patterns across seasons. We find that places with a high pick-up count tend to also have high drop-off count, and locations with low pick-up counts also have a comparatively low drop-off count, but the drop-off count is much greater in magnitude. We also find that there are more rides in high temperatures, especially to and from airports, but low pick-up and drop-off frequencies stay low across different temperatures. We calculate that the places with the highest eigenvector centrality are East New York, Crown Heights North, Canarsie, Brownsville, and Prospect-Lefferts Gardens. Our results demonstrate that Uber may be a preferred mode of transportation for more people on hot days (more rides in June compared to April and January), perhaps because of the personal space and air conditioning it provides. However, Uber is only accessible to those who can afford it, so much of the population must still depend on other forms of transportation that may be more crowded and hot.

transportation networks | mobility | rideshare | NYC | Uber

The physical mobility of people and goods is central to alleviating poverty (1). Poor people measurably have less access to nutrition, health care, education, and opportunities to work for money, and these inequities are strongly associated with deficits in physical mobility (1). When people cannot afford to travel to grocery stores, hospitals, school, or work, they cannot access the resources and services that these places provide. Within a given area, access to transportation differs greatly across different groups of people, especially across socioeconomic classes. Since rideshare is a relatively new form of transportation, only popularized within the last decade or so, there are not many studies on how accessible it is to people of different financial backgrounds. However, there is evidence that it improves access to certain locations in a city overall and that it affects economic activity, one study showing that adding UberX to a location's prior accessibility doubled the net creation of restaurants and increased median house prices by 4% and rents by 1% (2).

We want to better understand the places that certain groups of people travel to and from. In this paper, we study the mobility patterns of Uber users in New York City by looking at pick-up and drop-off locations of Uber trips in 2021. People who use Uber regularly are not representative of the population as a whole, since they have to be able to afford it, so this data is limited to people of a certain socioeconomic class. In our analysis, we find that there are similar commuting patterns within neighborhoods, further indicating that people who live in the same area likely share other characteristics as well, such as having similar incomes, workplaces, hangout spots, and, importantly, similar levels of access to rideshare and other forms of transportation.

Methods

Our code can be found in this Github repository.

Data Description. We obtained our dataset from a data repository on Kaggle (3). The main dataset consisted of trip data of all taxi vehicles services managed by the Taxi and Limousine Commission (TLC), the agency in NYC responsible for licensing and regulating the city's "for-hire vehicle bases". We also made use of the Taxi Zone Lookup and NYC weather datasets from the same repository.

Data Cleaning and Preprocessing. Since the original dataset comprised companies dispatching over 10,000 trips per day, including Uber, Lyft, Via, and Juno, and since Uber facilitated over 80% of the trips in the dataset, we first filtered our data to only trips with the TLC license for Uber (HV0003). Additionally, as the original dataset for each month only includes numerical IDs for the Pick-up Location and Drop-off Location, we merged this dataset with the Taxi Zone Lookup dataset to match each Location ID with its corresponding zone in NYC. We did a similar process with the NYC Weather dataset provided, as we wanted to perform a temperature

Significance Statement

Equitable access to transportation is an important area of social and environmental justice. The level of access an individual has to transportation in a given area depends on many factors, including race, gender, and particularly, socioeconomic class. To work towards making transportation more accessible, it is necessary to understand the mobility patterns of people of different incomes within a city. We look at the travel patterns of one subset of people in New York City, the people who can afford to use Uber. Our analysis provides insight into the most popular places those people take Uber to, the connections between places, and daily and seasonal trends.

Author contributions: Fonda mapped the locations from the Uber dataset onto taxicab lookup zones, conducted the zoning analysis, and wrote the corresponding section of the report. Maya cleaned and processed the data, visualized the networks, conducted the degree analysis, temperature analysis, and eigenvector centrality analysis, and wrote the corresponding sections of the report. Annabel cleaned and processed the data, conducted the seasonal and peak hours analysis, wrote the corresponding section of the report, and cited sources in the references section. Annie found sources to potentially base the analyses on; interpreted the findings from the different analyses; wrote the abstract, introduction, keywords, significance, conclusions and discussion, and limitations (with some input from the others); and cited sources in the references section.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: fmhu03@g.ucla.edu

analysis and determine patterns in Uber trips throughout the year. However, to merge the NYC Weather dataset with our previously merged dataset, we had to preprocess the datetime columns to be used as the key for merging. We also converted the units of the temperatures in the temperature column from degrees Celsius to degrees Farenheit. This resulting dataset was what we used to generate our first network (see *Network Formation*). We additionally calculated the frequency of rides from unique ordered pairs of locations, adding it as a column to the dataset, and used it to generate our second network (also see *Network Formation*). For the dataset used in our peak hours and seasonal analysis (see *Degree Analysis for Network 2* and *Temperature and Seasonal Analysis* respectively), we additionally extracted date, month, day, and hour information from the pickup_datetime column, creating new columns for them, and dropped rows with unknown zones since they made up less than 4% of our data. We also dropped the columns we deemed irrelevant, which differ for the different types of analyses we conducted.

Network Formation. After processing our data, we began to build a snapshot of our first network, a directed network. Since our data contained a high volume of trips for each month, with January itself having close to 9 million trips, we took a random sample of our trip data to generate our network and create a cleaner visualization. We first created an edgelist for our Pick-up Location ID's and Drop-off Location ID's, denoting our "source" as the Pick-up Location and "target" as the Drop-off Location for generating a directed network. Then taking a sample of our edgelist, we generated the network as seen in Figure 1.

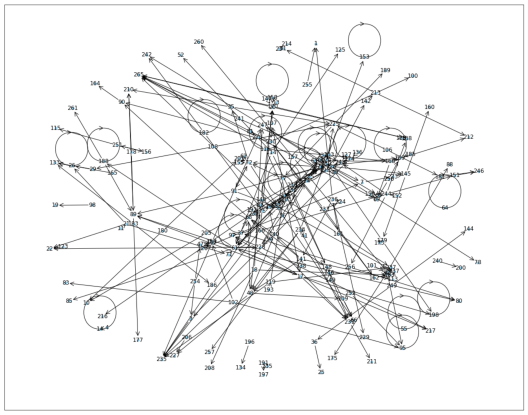

Fig. 1. Directed Network (1) of a sample of 300 Uber trips in January 2021 in NYC

Furthermore, we build a second network, as seen in Figure 2, that consists of both direction and weight. The weight of each edge between unique pairs of Pick-up Location ID and its corresponding Drop-off Location ID is denoted by the frequency, which represents the number of rides from Pick-up Location A to Drop-off Location B. While the data for this network has significantly fewer edges, about 55,000 for the month of January, we still utilize a sample of 200 to visualize our weighted network.

We utilize both networks to discuss comparisons and conduct our analyses, including degree and centrality analyses.

Fig. 2. Weighted, Directed Network (2) of a sample of 200 Uber trips in January 2021 in NYC

Results and Analysis

Degree Analysis for Network 1. We start with performing a degree analysis for our first directed network (considering all 9 million edges for January, not just the 300 for the sample). For each location in NYC in the Taxi Zone Lookup file, we calculate the total pick-up count (or the out-degree) as well as the total drop-off count (or the in-degree) for the month of January. We also calculated the ratio of in-degree to out-degree as well as the absolute value of the difference between the in-degree and out-degree to help answer questions such as the following.

1. Do locations with a high pick-up count necessarily have a high drop-off count? Do we see the same for locations with a small pick-up count?

2. Are there any outliers within the locations that do not follow certain trends in the relationship between the in- and out- degree?

Through our exploratory analysis of this data, and from Figure 3, we see that locations with a relatively high pick-up count do in fact have a high drop-off count. For example, the location with the greatest pick-up count, Crown Heights North (153,016), also has a really high drop-off count (152,770). We see this trend with other locations that have both a significantly high pick-up and drop-off count, such as East New York, Central Harlem North, Bushwick South, and Washington Heights South. Correspondingly, these locations that have both high pick-up and drop-off counts have an in-degree to out-degree ratio that is close to 1, highlighting that number of Uber rides to and from these locations are of similar magnitude.

However, while it is true that many locations with low pick-up counts still have relatively low drop-off counts, it is important to note that within many of these locations, the drop-off count is much greater in magnitude than the pick-up count. For example, considering the location Saint Michaels Cemetery/Woodside, we see that its drop-off count is almost twice as large as its pick-up count (drop-off count of 1,129 vs. pick-up count of 584). We see similar trends in Flushing Meadows - Corona Park, Inwood Hill Park, Randalls Island, Prospect Park, Central Park, and Penn Station/Madison Sq West, all giving an in-degree to out-degree ratio of between 1.3 and 2.1. Thus, from this we can see that locations with a low

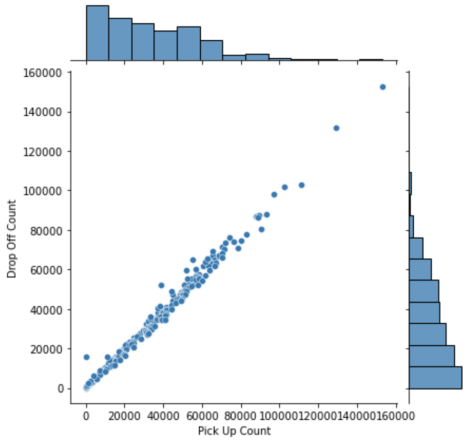

Fig. 3. Joint plot of pick-up count vs. drop-off count by location in January 2021

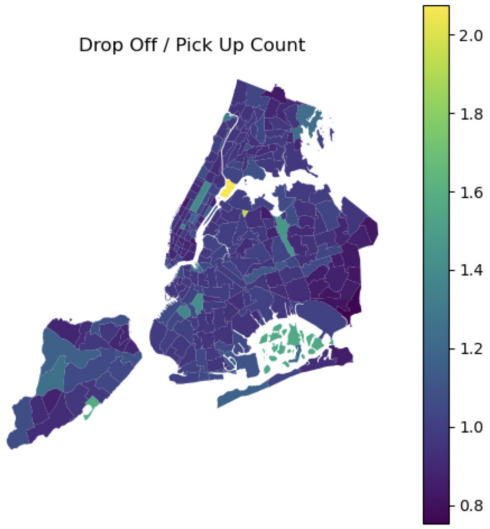

Fig. 4. Choropleth map of ratio of drop-off count (in-degree) to pick-up count (out-degree) by location, excluding Newark Airport

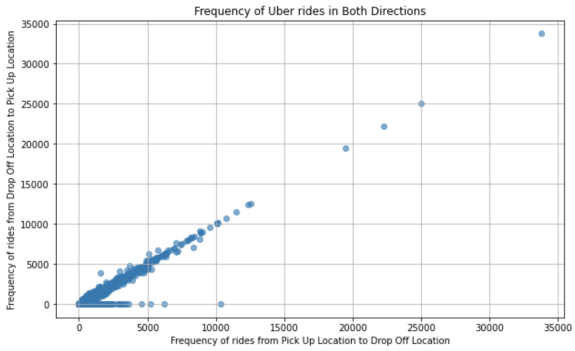

Fig. 5. Scatterplot of reverse frequency against frequency of Uber trips in NYC in January 2021

pick-up count (usually less than 10,000), often have a greater drop-off count, with the exception of Penn Station/Madison Sq West, which has a pick-up count of 38,000 but a drop-off count that is significantly greater, about 52,000.

Furthermore, we note the following two key outliers in this data.

1. Interestingly, only one location, Newark Airport, has a high drop-off count of 15,841 but a pick-up count of only 1, which is slightly unusual for a commercial zone such as an airport in NYC, especially within a duration of 30 days. We note that this could be either due to a manual error or the fact that after landing at the airport, people use different methods of transportation to travel within the city.

2. For the location, Rikers Island, we see a pick-up count of 2, and a NaN value for the drop-off count, possibly implying that there are not many or any Uber trips that go to the island.

Degree Analysis for Network 2. To perform a deeper analysis of our weighted, directed network (Figure 2), in addition to computing the frequency of rides of each unique pair of Pick-up and Drop-off Location ID's, we also calculated the reverse frequency, which corresponds to the the number of trips from Point B to Point A. In our dataframe, we made sure to avoid double-counting as the frequency column will separately calculate the number of trips from Point A to Point B as well as the number of trips from Point B to Point A. We utilize the reverse frequency column to view relationships between the number of rides from Point A to Point B and the number of rides from Point B to Point A, as in Figure 5.

While we observe a correlation coefficient of approximately 0.973 for the scatterplot in Figure 5, there are several other trends within our data when we look at it at closely. These trends are listed as follows.

1. Part of the strong linear correlation depicted in the plot can be accredited to the fact that there are loops in this network. As seen from the snapshots of both our networks (Figure 1 and Figure 2), we observe that many of our edges have the same Pick-up and Drop-off Location ID, resulting in loops. The scatterplot above includes data where the Pick-up Location ID is the same as the Drop-off Location ID, thus indicating that the frequency equals the reverse frequency. We discuss this further in *Limitations*.

2. We observe that pairs with a high frequency greater than 7,000 also have a high reverse frequency greater than 7,000, which aligns with the trends observed from our first network, where locations with a higher pick-up count tend to also have a high magnitude for their drop-off count. We see this relationship between Crown Heights North and Prospect Lefferts Gardens, Crown Heights North and Stuyvesant Heights, East New York and Pennsylvania Avenue, East New York and Brownsville, Central Harlem and Central Harlem North, Bushwick South and Bushwick North, as well as Washington Heights North and Washington Heights South. As we saw earlier, many of these locations have a high pick-up and drop-off count as well. Thus, from this network data, we are able to specifically observe which trips directly contribute to the high in- and out-degree of a location.

3. We do not observe this trend for smaller magnitudes. From Figure 5, we see that the frequency of trips from the

pick-up location to the drop-off location is not perfectly linearly correlated. We see that some trips that have a high frequency (around 5,000) from the pick-up location to the drop-off location have a corresponding reverse frequency of less than 5 to 10. Taking a deeper look at these trips, we see that for most of them, their drop-off location is given by ID 265, which on our Taxi Zone Lookup file is mapped to N/A. Thus, we note that the decrease in linearity between the frequency and reverse frequency is due to N/A values in our dataset, which we discuss further in *Limitations*.

4. For about 2,200 trips, the frequency and reverse frequency of unique Uber trips is the same, but this time, the pick-up location is not equal to the drop-off location. We assume that it is likely the passengers who booked their Uber from these pick-up locations to the corresponding drop-off locations were the same ones who booked their Uber trip from the drop-off location to the pick-up location. A few trips that follow this pattern include Claremont/Bathgate and Mott Haven/Port Morris, Ozone Park and Woodhaven, and Sutton Place/Turtle Bay North and TriBeCa/Civic Center.

Inspired by Xie and Wang's analysis on the temporal distributions of bikeshare trips (4), and to check whether the observed linearity relates to work-home splits, we then focused on analyzing Uber trips during weekday peak hours. Firstly, to find the peak hours, we plotted the trip rate by hour for weekdays and weekends. We calculated the trip rate by dividing the trip count at each hour on weekdays and weekends by the number of associated days. As seen in Figure 6, the peak hours on weekdays were 8 a.m. and 6 p.m.

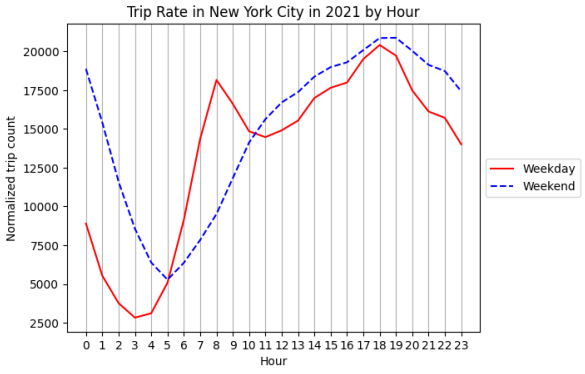

Fig. 6. Weekday and weekend trip rates in NYC in 2021 by hour

Similar to earlier, we calculated and plotted the reverse frequency against the frequency of Uber trips for each hour (Figure 7 and Figure 8), and observed that the relationship was mostly linear, supporting earlier findings.

To further explore the notion of work-home splits, we calculated the ratios of in- to out-degree and ratios of out- to in-degree of zones involved in 8 a.m. and 6 p.m. trips respectively. For the obvious outlier in ratios, namely Newark Airport's in- to out-degree ratio of infinity (since Newark Airport had a non-zero in-degree but an out-degree of zero), we replaced its ratio with the maximum of the ratios found when excluding infinity. To plot choropleth maps, we also added the missing zones (zones that were not involved in trips during the

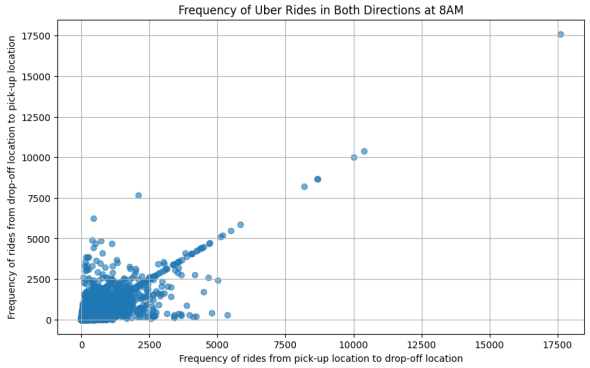

Fig. 7. Scatterplot of reverse frequency against frequency of trips in NYC at 8 a.m.

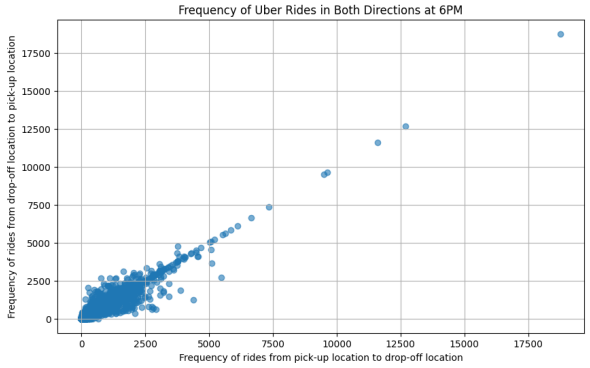

Fig. 8. Scatterplot of reverse frequency against frequency of trips in NYC at 6 p.m.

two hours) by identifying their location IDs and adding them to the dataframe with a value of 0 as placeholders for their in-degree, out-degree, ratio of in- to out-degree, and ratio of out- to in-degree. Then we plotted choropleth maps of 8 a.m. trips, with the colors of the zones based on their respective ratio of in- to out-degree (Figure 9), and of 6 p.m. trips, with the colors of the zones based on their respective ratio of out- to in-degree (Figure 10).

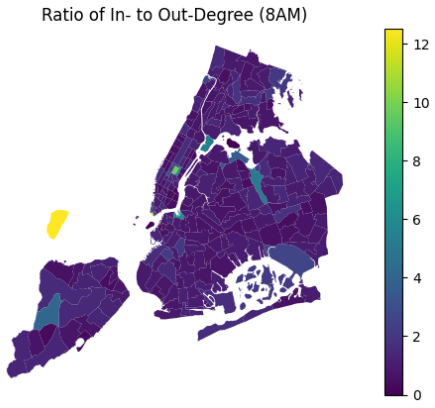

Fig. 9. Choropleth map of ratio of in- to out-degrees of zones in NYC at 8 a.m.

We were motivated by the possibility that the previously observed linearity would give insight to whether locations with relatively high in- to out-degree ratio at 8 a.m. are the same ones with relatively high out- to in-degree ratio at 6 p.m.,

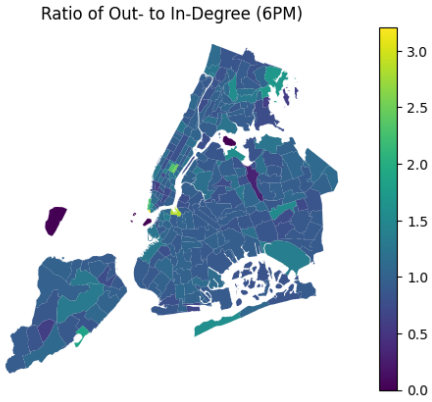

Fig. 10. Choropleth map of ratio of out- to in-degrees of zones in NYC at 6 p.m.

which would in turn allow us to label certain locations as workplaces and others as homes, assuming that those who take Uber are commuters. However, as observed in Figure 9 and Figure 10, while the relative ratio of in- to out-degrees of zones at 8 a.m. and the relative ratio of out- to in-degrees of zones at 6 p.m. somewhat match, there are some outliers. Moreover, exploring the zones with matching relative ratios, we found that those zones are not necessarily residential areas. Therefore, we are unable to conclude anything concrete about work-home splits. We note that this might be due to our faulty assumption that the majority of Uber riders in our dataset are commuters.

Temperature and Seasonal Analysis for Network 2. Unlike states such as California, New York tends to have more defined seasons and varying monthly temperatures, including rainfall and snow, allowing us to analyze how monthly temperatures might affect frequency of rides between a unique pick-up and drop-off location.

To begin our temperature analysis, we first picked 3 months with varying temperatures – January, April, and June. Applying our data cleaning and processing steps, as well as exploratory analysis, to these months, we compared the frequency of rides between unique pairs of locations and observe the following.

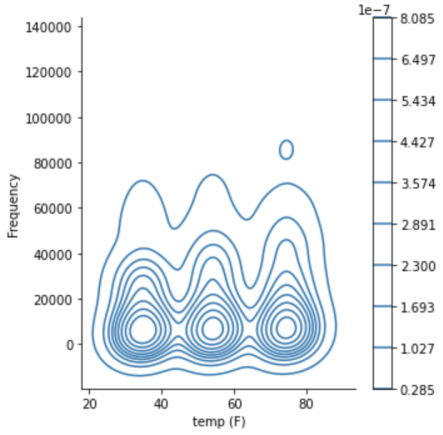

Fig. 11. Frequency of rides for January, April, June 2021

It is clear from Figure 11 that rides with lower frequencies (less than 20,000) tend to have similar patterns throughout January, April, and June. However, as the temperature increases, we see a stark increase in the magnitude of frequency of rides (between 20,000 and 60,000) with slight variations after the frequency reaches 60,000. In particular, within the cooler months, January and April, the frequency of rides for unique trips reaches a maximum of about 70,000, while for June, we see a slightly smaller value. Moreover, we note that for the hotter month of June, when the temperature is roughly 80°F, we see an outlier for the frequency of rides at about 85,000, overall indicating that as the temperature increases, the number of Uber rides tends to increase. From our exploratory analysis, we observe this phenomenon occurring especially in commercial zones such as JFK and LaGuardia Airports, where the frequency of rides to these locations is almost twice as large in hotter temperatures.

To get a bigger picture, we categorized the twelve months into four seasons consisting of three months each (March, April, and May as Spring; June, July, and August as Summer; and so on) and plotted the trip rates during each of the four seasons. As seen in Figure 12, we found that peak hours were generally consistent across the seasons and observed that Fall had the highest trip rates, possibly due to big events and holidays like Halloween and Thanksgiving. On the contrary, Summer did not have as high a trip rate as expected from our earlier temperature analysis. This might be due to the fact that there could be other confounding factors affecting the frequency of rides during the different months, which we discuss in *Conclusions and Discussion*.

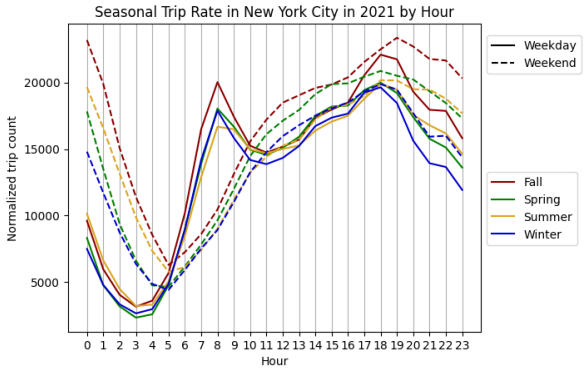

Fig. 12. Seasonal trip rates in New York City in 2021 by hour

Eigenvector Centrality Analysis for Network 2. To further define a metric that determines the importance of a specific location within NYC, we calculate the eigenvector centrality (EVC) of each location given in the Taxi Zone Lookup file. We propose that if a pick-up location is connected to multiple locations that themselves have a high in degree – meaning that the specific pick-up location has a high frequency between multiple other drop-off locations – then it is categorized as an "important" node in our network. We see this occurring for pick-up locations such as East New York, Crown Heights North, Canarsie, Brownsville, and Prospect-Lefferts Gardens, to name a few. Moreover, from our degree analysis, we see that many of these locations also had the greatest in- and out-degrees, as well as high frequencies with their corresponding drop-off

locations. We see the data evidence for this phenomenon for Stuyvesant Heights in Figure 13.

	PULocationID	DOLocationID	Frequency
46722	225	61	8346
46876	225	225	7406
46699	225	37	6173
46679	225	17	6085
46698	225	36	3775
...	...	...	...
46720	225	59	1
46719	225	58	1
46706	225	44	1
46693	225	31	1
46668	225	5	1

Fig. 13. Evidence for eigenvector centrality for Pick-up Location ID 225, Stuyvesant Heights (EVC = 0.206)

From the EVC distribution plot in Figure 14, we only see a few "important" nodes, those with a centrality between 0.17 to 0.27. We also note that there are few pick-up locations that have an EVC greater than 0.3. We discuss this further in *Limitations*, as we see that these pick-up locations tend to have greater EVC values due to the fact that they have high frequencies of loops.

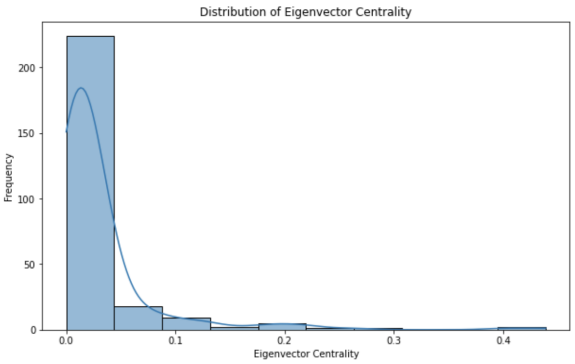

Fig. 14. Eigenvector centrality distribution

Moreover, the taxicab zones with the highest eigenvector centrality appear to spatially clustered in Brooklyn. Looking at a subset of the ten zones with the highest centrality, we notice that the ascription of high centrality is a result of the combination of two factors: a high volume of trips within the region and the localized nature of these trips. This is corroborated by two key findings. First, Figure 15 reflects significant heterogeneity in the district type, from highly residential zones to areas more dominated by commercial or residential activity. Mixed land use patterns reduce the need to travel farther distances for the same activity, which means that most Brooklyn residents are likely able to remain within the borough for day-to-day tasks, increasing the region's neighborhood-level inter-mobility. Additionally, the pick up and drop off counts incidentally are similarly concentrated in the Brooklyn and the upper Manhattan areas, as indicated by Figure 16; thus, the sheer magnitude of trips in and out of the region adds extra weight to the region from a ridership perspective.

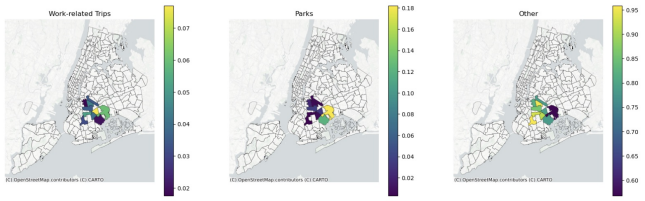

Fig. 15. 10 Zones with Highest Eigenvector centrality by District Type

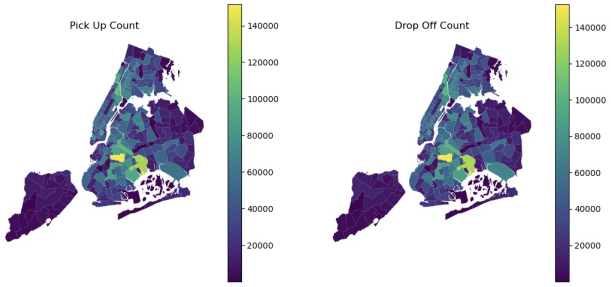

Fig. 16. Pick Up and Drop Off Counts by Taxicab Zone

Zoning Analysis: Inferences on Trip Purpose. The built environment, embodied by land use and transportation infrastructure, exerts significant influence on movement across New York City. The quality of the built environment is commonly assessed using the 4 D's: density, design, diversity, and destination accessibility. Land use patterns affect the first three categories, as they determine how far residents have to travel for different spheres of activity. Mixed use patterns, embodied by the 15-minute walkable city, typically alleviate the burden of travel-related costs. Transportation infrastructure impacts the accessibility of a panoply of activities, and especially in New York City, where bus, rail, and subway networks have enjoyed lasting eminence, although they continue to remain spatially heterogeneous from an accessibility lens. Ride-hailing apps such as Uber occupy a relatively small niche within the modalities of transportation around the Big Apple, given that the caliber of the city's public transit network juxtaposed with the vehicular congestion that plagues most road networks erases a lot of the traditional socioeconomic stratification in commute mode.(5) Both high-income and low-income residents may choose public transit over a private vehicle as the gains in efficiency far outweigh the costs of discomfort.(6) This finding is corroborated by the fact that transport entropy (the mixed degree of transport variables) has historically had little to no correlation with taxi ridership. Thus, a focus instead on land-use, which does bear influence on ridership patterns, is utilized in the rest of this section.

Uber riders represent a pretty small segment of the population of New York City, self-selected by characteristics such as means to pay for transportation, familiarity with the app's platform, and travel preferences. From a mix of informal surveys and online travel websites, we infer that most of the

Uber customer base in New York City employs the service for tourism, commuting, and intermediary travel (e.g. traveling from home to the airport). Land use patterns can shed more insight on how the first two motivations, tourism and commuting, impacts whether a person uses Uber and if so, how far they will travel.(7) To establish a few quantitative comparisons, we created two measures of land use entropy to assess the degree of mixing of different types of activity in a taxicab zone. Given the granularity of NYC's zoning data, two different entropy measures were created: one that agglomerates the zones on the four main categories of residential, commercial, manufacturing, and parks, and another that also accounts for the presence of different subgroups.(5) Looking at the first index measure g, for each taxi zone i, the entropy index h_i^g is calculated as follows:

$$h_i^g = -\sum_{k=1}^{K} p_{ik}^g \log_2 p_{ik}^g$$
$$p_{ik}^g = \frac{n_{ik}^g}{n_i}$$

where g represents the general classification, K refers to the total number of district types - residential, commercial, manufacturing, or parks - in the taxi zone, n_{ik}^g refers to the total area within the taxicab zone covered by the district k, and n_i refers to the total area of the taxi zone. The formula for the second index measures follows the same structure:

$$h_i^s = -\sum_{k=1}^{K} p_{ik}^s \log_2 p_{ik}^s$$
$$p_{ik}^s = \frac{n_{ik}^s}{n_i}$$

but instead of the general classification of districts, we account for the presence of different subgroups, e.g. different types of commercial districts, within a taxi zone as well. Figure 17 displays the relationship between the two indexes, with the second index giving greater weight to a larger proportion of taxi zones. To test our hypothesis of the inverse relationship

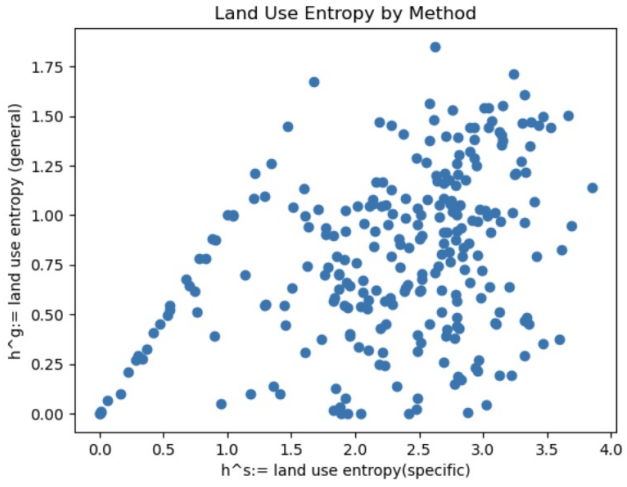

Fig. 17. Land Use Entropy Measure

between the degree of land-use mixing and distance traveled, we first need to calculate the average distance traveled for each trip from one taxicab lookup zone to another. To do so, we obtain the coordinates of the centroid of each taxi zone geometry, and then calculate the distance between the centroid of the pick up zone and that of the drop off zone. The distance was then averaged by pick up location and drop off location, producing two significant findings. First, the average distance of a trip with its destination within a taxicab zone is weakly inversely related ($r = -0.315$) with the average distance of a trip starting from the same zone, a consistently negative correlation which is strongest during the evening. The loss of weak correlation corresponds with a sharp spike in average trip distance around 5 AM. Incorporating our previous finding of a positive correlation between the pick up and drop off counts by taxi zone, this indicates that activity flowing in and out of the zone might be more anchored to the land use patterns within the taxi zone rather than the cyclic mobility patterns of regular customers.(8)

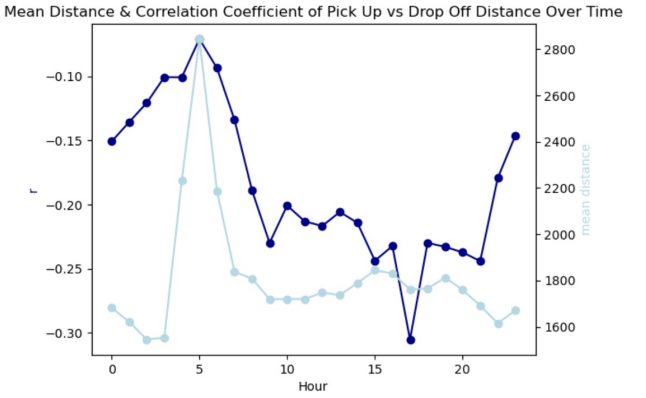

Fig. 18. Average Distance

Now, we wanted to examine the effect of mixed land use patterns on the average trip distance to and from a specific region. We used the two measures identified earlier, with the "entropy_g" column referring to general district classification-based land use entropy measure h_i^g and the "entropy_s" column referring to the district subgroup-based index h_i^s. We find no correlation between the general classification of entropy and average distance traveled, while the more granular index yields a slightly negative correlation with the average trip distance based on pick-up location, and slightly negative correlation with the other measures. Thus,the availability of a wider variety of types of activity in a certain area is indeed inversely related with travel distance from the origin point, and we hypothesize that the slight positive relation with travel distance to the destination point may be a result of the positive externality of an increased appeal of the location to visitors from farther away, which is why total trips both to and from the area increase with the mixing of land use patterns. Another key trend from the correlation matrix was the strong positive correlation between the number of trips and the average trip distance to a destination. We hypothesize that this may result from the fact that public transportation options get less efficient as the total trip distance increases, especially extending towards the outskirts of the metro where transit coverage is sparser. Thus, with longer trips, the comfort and efficiency of using direct end-to-end means of transportation services such as Uber begin to outweigh the cost-saving nature of public transit, thus reducing the substitution effect between the two modes of travel.

A spatial comparison of average distance of trips to and

	entropy_g	entropy_s	avg_distance_pu	avg_distance_do	total_trips_from_pickup	total_trips_to_dropoff
entropy_g	1.000000	0.358564	-0.011072	0.043568	0.138793	0.099368
entropy_s	0.358564	1.000000	-0.180185	0.121969	0.213328	0.204068
avg_distance_pu	-0.011072	-0.180185	1.000000	-0.314568	-0.051546	-0.369135
avg_distance_do	0.043568	0.121969	-0.314568	1.000000	-0.142122	0.736949
total_trips_from_pickup	0.138793	0.213328	-0.051546	-0.142122	1.000000	0.182812
total_trips_to_dropoff	0.099368	0.204068	-0.369135	0.736949	0.182812	1.000000

Fig. 19. Correlation matrix (Land Use Entropy)

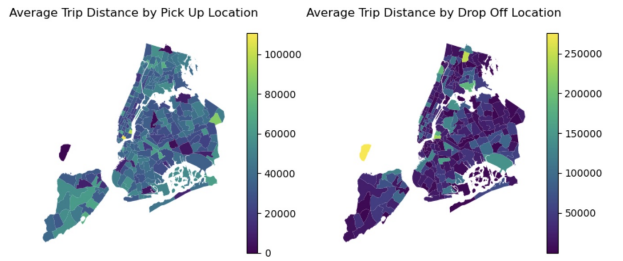

Fig. 20. Average Trip Distance To and From TAZ

from taxicab zones (Figure 20 provides more insight to this nuance in origin-destination heterogeneity. Both maps show that larger distances typically correlate with regions mostly on the periphery of the metropolitan area, with the exception in both cases of a small taxi zone near the heart of the city. The Financial district has by far the largest average trip distance originating from the zone, and Williamsburg trails the Newark Airport in having the second longest average trip distance dropping passengers off within its geography. The factors inflating the number of miles travelled to and from these distances cannot be deduced solely using Uber data, but comparing these findings to maps on transit accessibility published by other researchers creates a more cohesive narrative. In particular, the neighborhoods of Williamsburg and South Williamsburg score highly in mixed land use patterns on the weekend but lower in rail density, bus stop accessibility, and road network navigability. Consequently, visitors who live farther away may opt for ordering an Uber to avoid the hassle of switching among multiple inadequate modes of public transit to reach the same destination.(9) Thus, although transport entropy might not play a huge role in determining ridership patterns, transit accessibility still exerts tremendous influence in transit preferences to vibrant destinations.

Notably, the two zones with the highest Uber ridership activity also are adjacent to three important bridges which as of June 30, 2024, will administer a toll to crossing vehicles under NYC's new congestion pricing plan in an effort to attenuate the heavy flows of traffic into Manhattan. Given our knowledge of ridership trends and the quality of transportation infrastructure, we recommend that the policy's disruption of mobility patterns be accompanied by efforts to improve access to efficient transit in the area, providing the incentive to substitute car-centric transportation for public transit alternatives.(10)

Trips Grouped by District. To make inferences on the purpose behind Uber trips in New York City, we classified the taxi zones as falling under one of the land use districts on the assumption that the bulk of the trips in 2021 were motivated by the characteristics of the destination's built environment. Given its historical context of dense urbanization, diverse population, economic vitality, and an active urban planning department, most of New York City's land can be categorized as mixed-use development, with a plethora of housing, businesses, and recreational spaces all within close proximity.(11) Nevertheless, forces of industrial agglomeration as well as widely recognized need for more public green space prevailed in certain areas, leading to some spatial clustering of commercial and park districts.(9) This presence of spatial clustering for commercial/manufacturing and parks allowed us to categorize certain taxicab zones based on the criteria listed in Figure 21. The conditions allowed the capture of all outliers (see Figure 23) in the distribution as abnormalities, in this case areas with high levels of clustering. Taxi zones that satisfied multiple criteria were assigned the "mixed(m/c)" or "mixed(all)" categories which in total accounted for 29 of the 263 zones. There was one zone that satisfied both the park and commercial criterion which was assigned to the commercial classification given the underestimate of area coverage in the category, seen in Figure 22.

Criteria	Classification	# Zones
>20% commercial	commercial	34
>20% manufacturing	manufacturing	40
>20% comm. and >20% manuf.	mixed (m/c)	8
>20% park	park	29
>median (76%) residential	residential	131
other	mixed(all)	21

Fig. 21. Criteria for Land Use Classification

Classification	Estimated Coverage	Actual Coverage
commercial	3.20%	4.97%
manufacturing	18.64%	14.79%
mixed(all)	7.22%	na
mixed(m/c)	0.75%	na
park	21.35%	16.99%
residential	48.83%	63.21%
battery park city	na	0.04%

Fig. 22. Comparison of Estimated to Actual District Area Coverage

After the classification of taxicab zones as representations of a single zoning district, zone groupings were created based on the assigned type of activity. Using the set of groups, the modularity of 0.1719 was produced for a directed network of Uber trip data in July 2021. The groupings were manipulated then to amalgamate (1) commercial and manufacturing activity and (2) "mixed(all)" and "parks" zones to create three broadly defined categories of work, home, and leisure, which proceeded to increase the modularity of the community structure to 0.1904. These scores were less than the more spatially intuitive community structure of borough, a partition which produced a modularity of 0.4812. This indicated a strong presence of trips across districts with different forms of land-use compared to trips across longer distances for a mixed variety of purposes. Figure 24 shows the change in modularity based on time of day. All modularities experience a sharp dip at 5 AM which corresponds with the sharp spike in average distance traveled. This consistent deviation is sensible if we account for the fact that there is significant variance in the type of Uber rider over time, and those who utilize the service at 5 in the morning are

likely motivated by different purposes than those using the app hailing a ride in the middle of the day.

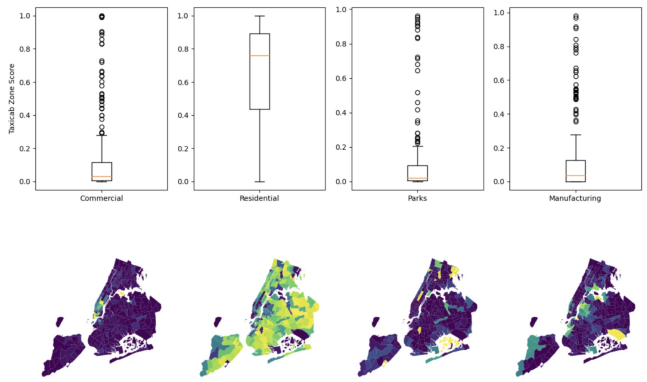

Fig. 23. Distribution of District Coverage within a Taxi Zone

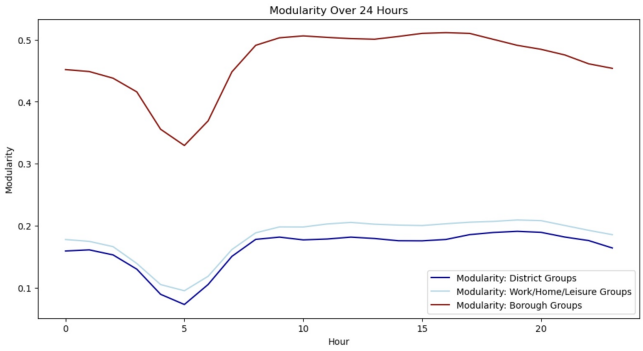

Fig. 24. Modularity Over Time

Conclusions and Discussion

In our degree analysis, we find that places with a high pick-up count tend to also have high drop-off count. It makes sense that people using Uber to get to a place would probably also take Uber to leave it. We also find that locations with low pick-up counts also have a comparatively low drop-off count, but the drop-off count is much greater in magnitude. This may be because people going to less popular places take an Uber to get there, but then walk or take some other form of transportation to a different area nearby, and then perhaps take an Uber to leave from there.

In our peak hours analysis, we find that the peak hours for Uber rides on weekdays are 8 a.m. and 6 p.m., which seems to correlate with the start and end times for many jobs. The peak hours on weekends are 6 to 7 p.m. and 12 a.m., which are popular times for leisure activities, such as going out for dinner, or to parties or clubs. Therefore, we infer that most people take Uber to travel to and from work during the week, and for fun on the weekends.

In our temperature analysis, we find that there are more rides in higher temperatures (June compared to January and April), especially to and from airports. This is likely because more people want to avoid walking, biking, or public transportation, when it's hot out and opt for an AC-cooled Uber instead. However, there may be confounding factors contributing to this as well, such as increased tourism in certain seasons, adding to the amount of people who are in NYC, and therefore the amount of people available to take Uber. We also find that low frequencies stay low across different temperatures. Less popular places are still not popular when it is hot, and people may choose to avoid going there altogether if the place is not very important, instead of more people taking Uber to get there.

In our eigenvector centrality analysis, we find that the places with the highest EVC values are East New York, Crown Heights North, Canarsie, Brownsville, and Prospect-Lefferts Gardens. These places could be considered the most "important" for Uber users in NYC, although there are some limitations to this, discussed in *Limitations*.

Finally, zoning analysis using a land use entropy measure found a weak inverse relationship between land-use mixing and travel distance, particularly in the evening. Higher land-use diversity within a zone tends to decrease travel distances for trips originating from the zone, while attracting travelers from farther away. A strong positive correlation was found between the number of trips and average trip distance, especially to peripheral areas with limited public transit coverage. This indicates a preference for Uber over public transit for longer trips due to comfort and efficiency. Trip distances were generally longer to and from peripheral areas, with the exception of the Financial District and Williamsburg, which both had notably high average trip distances and were located adjacent to key bridge crossings. Improved transit access in these areas could mitigate reliance on Uber, ameliorating New York City's persistent road network congestion problems.

Implications. We already know from existing research that poor people and people of color disproportionately face adverse effects of climate change. Our finding that there are more Uber rides in June than in January or April could be another indication of that – they may not be able to afford cool transportation in summer. They may have to deal with hot subway stations with no air conditioning, crowded subway cars, or walking or biking in the heat while people with more money can opt for more comfortable modes of transportation.

Some places in NYC may also be more easily accessible by rideshare or personal cars, and therefore more accessible to people who can afford those. This causes a disparity between the abilities of different groups of people to access places in the city, and the level of difficulty they face in reaching those places.

Future Work. We could normalize trip counts by the population density of each neighborhood, as there may be more rides going to and from an area just because more people live there, and not because that area is more popular or important. Additionally, we could find datasets that include rider demographics for more in-depth analysis of ridership patterns. We could also conduct similar analyses on other modes of transportation, such as taxis, buses, the metro, et cetera, and see how they compare.

To account for the diversity of land use in the city, we could also refine our zoning analysis to assign trip origin and destination points more granular zoning districts based on the urban planning department's zoning and land use map. A configuration model would need to be applied that

assigns a trip to a particular residential, commercial, park, or manufacturing district based on the proportion of coverage of a particular land use type as well as the land use type of the origin point. More granularity can also be applied to the entropy measure to account for the presence of commercial overlays and transit accessibility, to ensure a more accurate representation of the built environment. Quantifying the amount of amenities in a particular area, such as the amount of third places such as entertainment venues, cafes, or bars, can also allow for more accurate inference on the personal motivations influencing a rider's mobility patterns. A variety of papers have utilized geographically and temporally weighted regressions (GTWR) as well as binomial models to account for these additional factors, which could provide a more accurate picture on how land use and transportation infrastructure plays a role in shaping mobility patterns, especially in an area as diverse as New York City.

However, the aforementioned measures of land use diversity must be coupled with land use density to identify target areas that are most vulnerable to traffic-related congestion in mobility networks. With changing zoning laws allowing for more office-residential conversions, the density of New York City's neighborhoods is about to undergo significant transformations. Areas such as Williamsburg that underwent urban revitalization after a period of industrial decay reclaimed their economic vitality through mixed use development but exerted additional strain an increasingly outdated public transit network that failed to keep up with travel demand. Thus, creating a comprehensive overview of the effects of housing and transportation policy (e.g. property development or congestion pricing) on multimodal transportation networks can ensure that the city's system maintains its status as the vanguard of transit efficiency in an increasingly dynamic environment.

Limitations

As previously discussed, our data does not capture the travel patterns of the entire population of NYC, but only a portion of it – the people who can afford to use Uber. Also, this analysis is limited to NYC in a single year – there may be different patterns in other cities and in other years. Furthermore, the COVID-19 pandemic likely affected transportation patterns in 2021, though the scope and time limit of this project were not broad enough for us to compare this data to pre-pandemic data. Finally, the drop-off points in our data may not be the final destination of where people are traveling, because they could take a different form of transportation afterwards, such as a plane, or the metro, or walking. Therefore, we cannot make exact conclusions on where people go, and the centrality measures we calculated may not be entirely accurate methods of determining which places are the most "popular" or "important".

As mentioned in *Degree Analysis for Network 2*, item 3, there are some locations in the Taxi Zone Lookup file that are mapped to N/A, which decreases the linearity of the correlation between frequency and reverse frequency. Thus, we would need more information to make conclusions about this correlation.

The self-loops in our network, referring to people who take an Uber from a place to the same place, may be because the place is very large so people are actually going from one part of it to another, and it may not be convenient or possible to walk. We think the presence of self-loops in our network is likely not an error in data entry because there are so many of them, tens of thousands in just one month, in the dataset. There were also self-loops at airports, and some airports may be big enough for this to make sense, although we would have to look further into the structure and transportation options of New York airports to verify this.

The self-loops also skew the eigenvector centrality. When the pick-up location ID equals the drop-off location ID, the EVC gets skewed to higher magnitudes. For example, location ID 76, which corresponds to East New York, has nearly 34,000 self-loops in January, and has an EVC of 0.439. The EVC is skewed because it is saying the location is important because it has a lot of edges pointing to it from itself, instead of from its neighbors. Therefore, we do not have enough information, using only eigenvector centrality, to conclude that locations with many self-loops, such as East New York, are important, even if they appear to have a high EVC. Location ID 225 (Figure 13) has a smaller proportion of its outgoing edges pointing to itself, and it has an EVC of 0.206. For locations with no self-loops, more representative values of relatively high EVC are 0.17 to 0.27.

Self-loop frequency was also a strong indicator of spatial bias in the dataset in favor of taxicab zones with larger areas. As indicated in Figure 25, the areas with the most self-loop trips tended to be on the outskirts of the city, where the area tends to be larger. A strong linear correlation is found when plotting area against self loop frequency Figure 26. Thus, a more spatially homogeneous trip aggregator should be used for future analyses to overlay more meaning to these trips. Comparisons to other zoning district assumptions by other researchers found that our taxicab district classification overlooked a significant amount of neighborhoods in Manhattan with high residential population densities.(12) Thus, the use of alternative methods of zoning such as the use of user-generated online data to design more representative activity spaces might yield more accurate insight on ridership trends.(13)

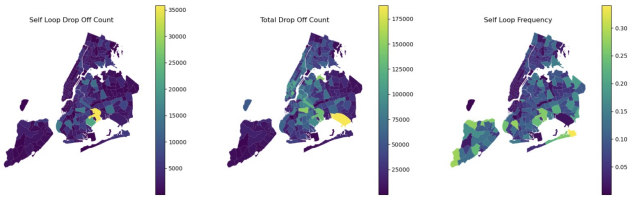

Fig. 25. Self Loop Frequency

ACKNOWLEDGMENTS. We thank Professor Mason Porter and Michael Johnson for referring us to relevant sources, discussing ideas with us, helping us with our code, and answering our questions.

1. M Wachs, Transportation policy, poverty, and sustainability: History and future. *Transportation Research Record* **2163**, 5–12 (2010).
2. C Gorback, Your uber has arrived: Ridesharing and the redistribution of economic activity. *Job Market Paper* (2020).
3. S Mo, Uber nyc for-hire vehicles trip data (2021) (https://www.kaggle.com/datasets/shuhengmo/uber-nyc-forhire-vehicles-trip-data-2021/data) (2023) [Accessed 31-May-2024; originally from the NYC Taxi and Limousine Commission, https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page].
4. XF Xie, ZJ Wang, Examining travel patterns and characteristics in a bikesharing network and implications for data-driven decision supports: Case study in the washington dc area. *Journal of Transport Geography* **71**, 84–102 (2018).
5. C Xie, D Yu, C Lin, X Zheng, B Peng, Exploring the spatiotemporal impacts of the built environment on taxi ridership using multisource data. *Sustainability* (2022).

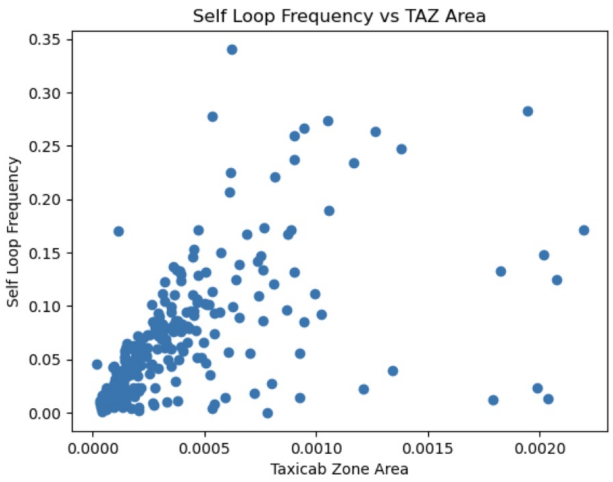

Fig. 26. Self Loop Frequency

6. JP Schwieterman, Uber economics: Evaluating the monetary and travel time trade-offs of transportation network companies and transit service in chicago, illinois. *Transportation Research Record* (2019).
7. R Grahn, S Qian, HS Matthews, C Hendrickson, Are travelers substituting between transportation network companies (tnc) and public buses? a case study in pittsburgh. *Transportation* (2021).
8. RB Noland, MJ Smart, Z Guo, Bikesharing trip patterns in new york city: Associations with land use, subways, and bicycle lanes. *International Journal of Sustainable Transportation* (2019).
9. L Wolf-Powers, Up-zoning new york city's mixed-use neighborhoods: Property-led economic development and the anatomy of a planning dilemma. *Journal of Planning Education and Research* (2005).
10. B Schaller, New york city's congestion pricing experience and implications for road pricing acceptance in the united states. *Transport Policy* (2010).
11. J Carlen, et al., Role detection in bicycle-sharing networks using multilayer stochastic block models. *Network Science* (2022).
12. JG Yunda, J Jiao, Zoning changes and social diversity in new york city, 1990–2015. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability* (2019).
13. O Nenko, M Kurilova, A Konyukhov, Y Bogomolov, Defining the real structure of the city through spaces of everyday activity based on user-generated online data. *Springer International Publishing* (2023).
14. R Kellermann, DC Sivizaca, D Rößler, N Kliewer, HL Dienel, Mobility in pandemic times: Exploring changes and long-term effects of covid-19 on urban mobility behavior. *Transportation Research Interdisciplinary Perspectives* **15** (2022).
15. T Hossmann, T Spyropoulos, F Legendre, A complex network analysis of human mobility. *IEEE Conference on Computer Communications Workshops* (2011).
16. M Newman, *Networks.* (Oxford University Press), (2018).