Matiss M Mednis, Nikolaus Schultze
CS579 Social Network Analysis Spring 2024
Project 2
Chicago Taxi Rides Network Analysis

Github Link: https://github.com/nschultze/CS579Project2/tree/main

Introduction

Our project analyzes taxi trip data in Chicago from 2013 to 2013 self reported by taxi drivers to the transportation regulatory agency of Chicago. We aim to utilize trip pick up and drop off data to analyze which communities are popular and have high taxi demand, which communities taxis often travel between, and how these aspects of taxi use and trips have changed from 2013 to 2023. Using this temporal and spatial data we hope to identify potential trends and shifts in travel patterns. Upon starting, we hoped to analyze many different segments of the data such as season, time of day, annually, and even on different sports seasons. In comparing the metrics and network properties for each of these subsets we hoped to shed light on changes in demand and taxi routes under various network conditions and times. With serious changes in the network under different circumstances or times, our analysis might change and thus conclusions about where taxis might expect to go could change as well. We thought determining this information could be pertinent if any predictive models were to be developed using this dataset as if different conditions notably changed the underlying network, then those conditions might require different predictive models to account for the changes in the underlying network.

At the start of the project we hoped to not only analyze the data under different subsets - such as time of day, season, yearly, and even within sports seasons or uncommonly big events - fully and gain a better understanding of the ways taxis move through the Chicago area under different city conditions but to then apply our insights to inform the creation of predictive models for future taxi flow in the city. Overall, there were many approaches to segmenting and analyzing the data so we hoped to explore it and undercover key insights through our analysis. To do so, we modeled the taxi trip data as a weighted bi-directional graph where the nodes are communities in Chicago - numbered according to Figure 13 - and the weights of edges are the number of trips between two nodes such that a weighted directed edge from V1 to V2 is the number of trips taken from Node V1 to Node V2.

The project aims to gain insights into the dynamics of taxi flow in Chicago, including peak hours, popular routes, and areas with high demand. It seeks to identify seasonal variations in travel patterns and their implications for transportation planning and infrastructure development. This information would help improve decision-making processes for taxi operators, urban planners, and policymakers by providing valuable data-driven insights. Through the development of predictive models, it aims to optimize taxi operations and improve service quality for passengers and give insight into how people travel via taxis in Chicago.

All code developed for this project can be found here:

https://github.com/nschultze/CS579Project2/tree/main

Data

For our project, we utilized the "Taxi Trips (2013-2023)" dataset provided by the City of Chicago through its public data portal found at: https://data.cityofchicago.org/Transportation/Taxi-Trips-2013-2023-/wrvz-psew/about_data

This dataset comprises taxi trip records reported to the City of Chicago's regulatory agency over ten years, from 2013 to 2023. The dataset contains information such as trip start and end timestamps, trip duration, distance, fare, payment type, and geographic coordinates for pickup and dropoff locations. We chose this dataset due to its rich temporal, spatial, and transactional information, which aligns with our project objectives of analyzing taxi flow dynamics and predicting future demand patterns. The decision to use this dataset was based on its availability, relevance to our research questions, and the desire to work with real-world data reflecting urban transportation dynamics.

The data set contains 212 million rows, with each row representing a given reported taxi trip, spanning from 2013 to 2023. It contains 23 columns. For our project, we only needed four of those columns those being shown in Table 1 below

Feature	Description
Trip Miles	Distance of the trip in miles Datatype: Floating point number to 1 decimal
Trip Start Timestamp	When the trip started, rounded to the nearest 15 minutes Datatype: Floating Timestamp
Pickup Community Area	The Community Area where the trip began. This will be blank for locations outside Chicago. Datatype: Integer
Dropoff Community Area	The Community Area where the trip ended. This column will be blank for locations outside Chicago Datatype: Integer

Table 1. The Features we used for our project

To clean the dataset, we filtered out any trips that had Null values in the dropoff or community area and any trips that had miles equal to zero. This means we are not considering trips which were canceled or that started or ended outside of Chicago or that potentially were missing that data for any other reason. In removing trips with a Trip Miles value equal to 0 we reduced the dataset size from 211,670,894 rows to 167,537,019 rows. By removing the trips with Null values for pickup or dropoff community area we reduced the number of rows to 148,135,771. In downloading the entire dataset, it was over 80GB but by filtering on the dataportal by removing invalid rows and filtering to only the 4 columns we needed we reduced the dataset to about 4GB. This made it considerably more feasible to download and process for our project. The data processing code can be found here

https://github.com/nschultze/CS579Project2/blob/main/data_preprocessing.ipynb

Working with the reduced dataset size, we then also found that creating graphs from the dataset took a long time. It took an estimated 1 minute 9 seconds for every 1 million rows to create a directed graph. For 2013 only this would take about 15 minutes. We decided to sample 1 million rows randomly without replacement for each year in the dataset to further reduce the dataset size to help with our analysis. To ensure the sampling method represented our underlying distribution we compared a sample from 2013 to the full set of data in 2013 and found the overall attributes of the graph were well represented by the sample size. Figures 1 and 2 show the comparison of in and out degree distribution of the full 2013 dataset and our 1 million row sample. This process and comparison can be viewed in depth here

https://github.com/nschultze/CS579Project2/blob/main/2013_metrics_sample_test.ipynb

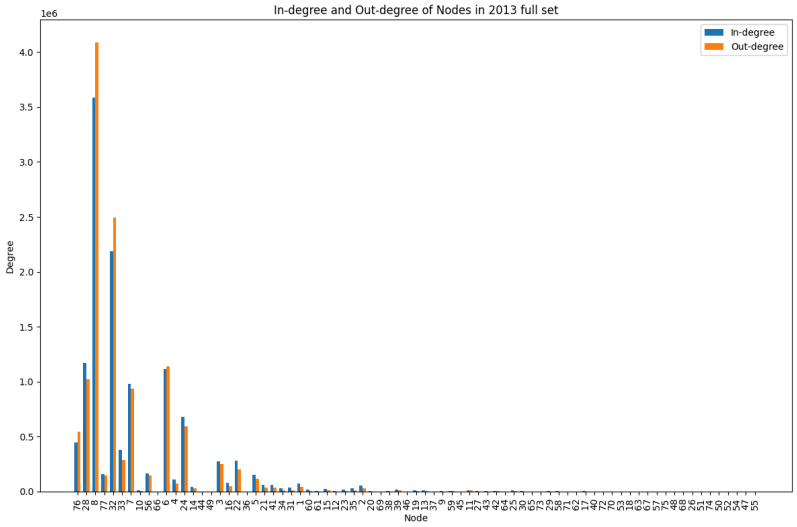

Figure 1. In and out degree distribution of full 2013 dataset

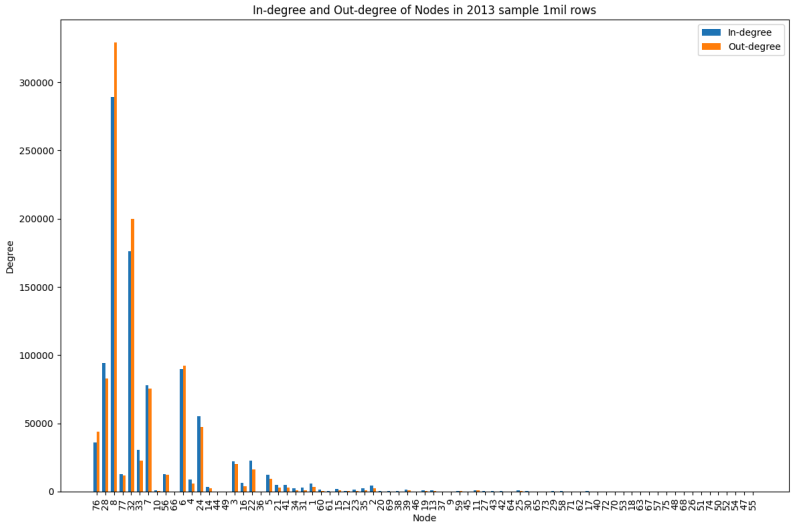

Figure 2. In and out degree distribution of 1 million randomly non-replaced rows of 2013 dataset

In the dataset deliverable we did plan to segment the data via temporal properties such as year or season. We did not anticipate how difficult it would be to work with the full dataset size even under these segmentations and thus we decided for a random sampling technique. To capture the underlying properties of the distribution we still took a fairly large sample size for each year taking 1 million rows with random sampling for each year. This allowed us to do in depth and tractable annual metrics and comparisons. We also later combined these annual subsets into seasonal subsets and one large overall dataset covering 2013-2023. This gave us 11 million sampled trips over 2013-2023 and 1 million sampled trips per year.

In exploring the dataset we found determined counts for taxi rides per year shown in Figure 3 and the counts for taxi rides by season of the year in Figure 4. Figure 5 shows the number of taxi trips taken by month over the entire dataset. We see the taxi rides drop off in recent years potentially due to competing ride share services, increased use of public transit, reduced travel in the city, or many other factors that we cannot determine without access to other datasets. As far as change in ride counts throughout the seasons and months we see uniformity.

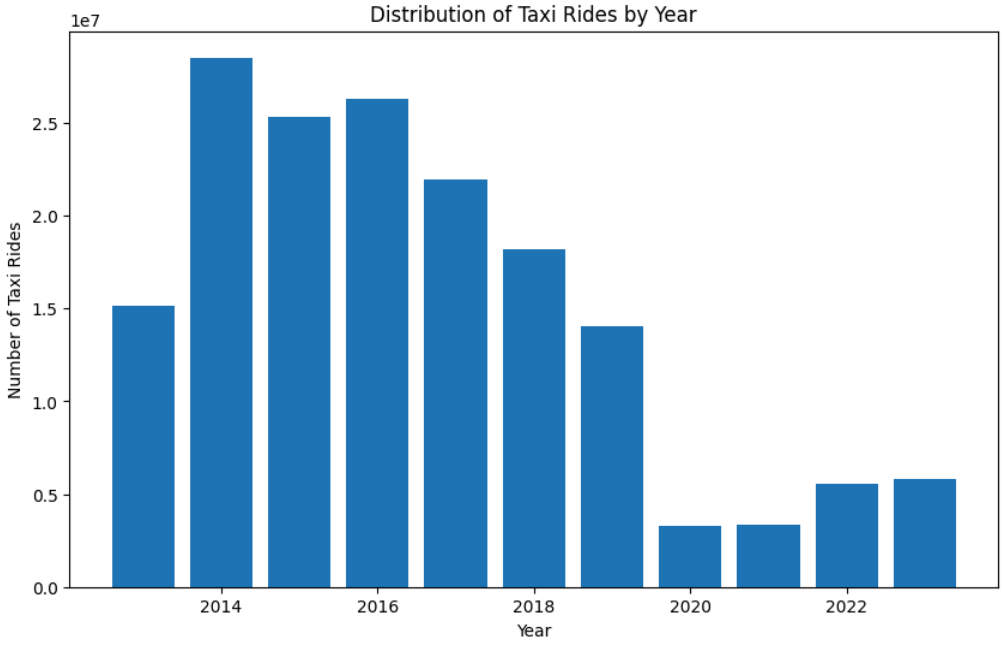

Figure 3. Count of taxi rides by year

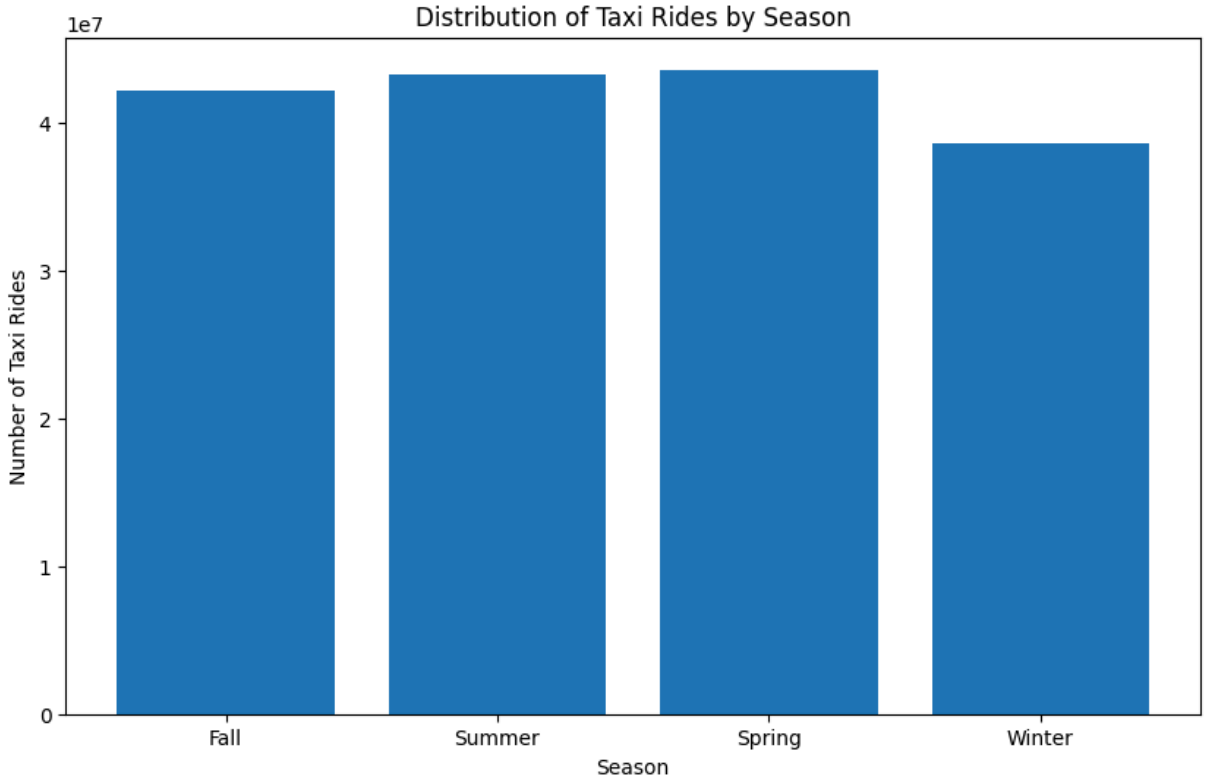

Figure 4. Count of taxi rides by season

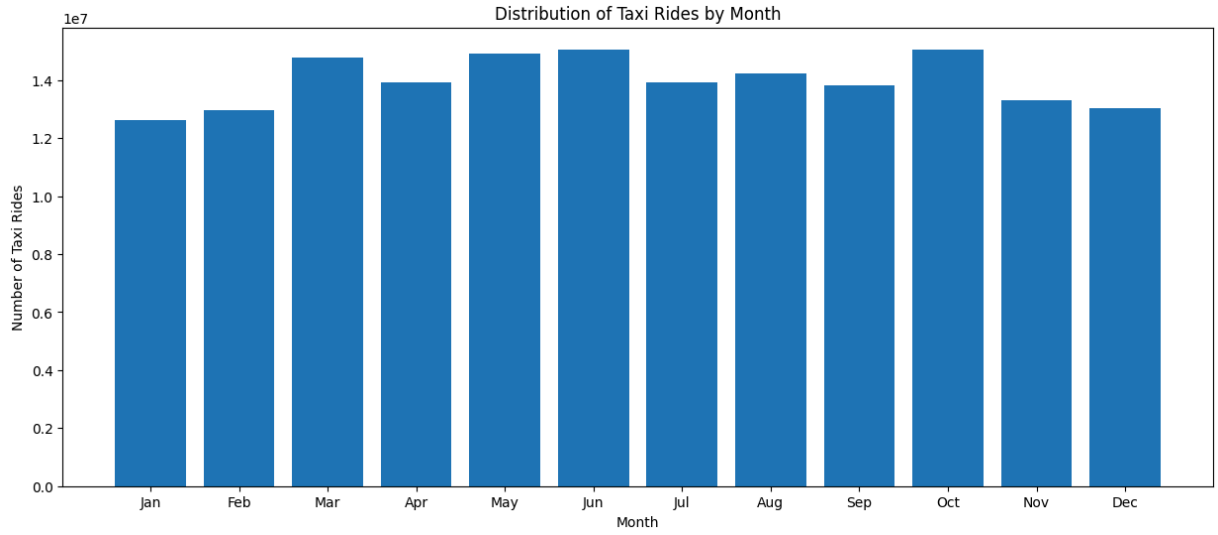

Figure 5. Taxi ride distribution by month

Project Process and Results

First we completed a lot of initial data analysis to uncover the patterns of taxi usage overtime some of this was discussed in the previous section shown in Figures 3, 4 and 5. We additionally looked into what time of day and what day of the week taxis were most often used. Shown in Figures 16 and 17 below we see that taxi demand has a peak around 6pm which may coincide with the end of the work day or just trips people need taxis for. Additionally we see that taxi usage climbs from Monday to Friday, peaks on Fridays, and drops off on the weekend. To collect this data we aggregated the start times where the day of the week was the same or the start times fell within the same hour.

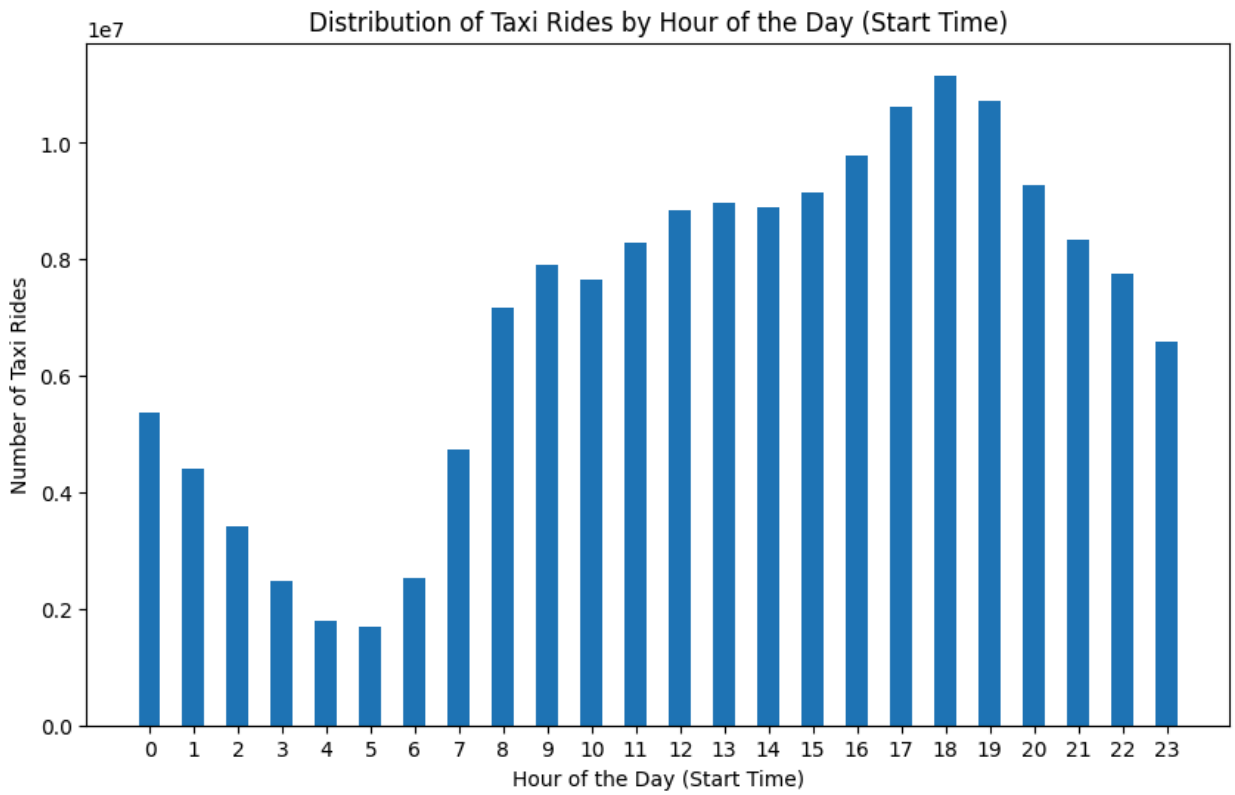

Figure 16. Taxi ride distribution by hour of the day

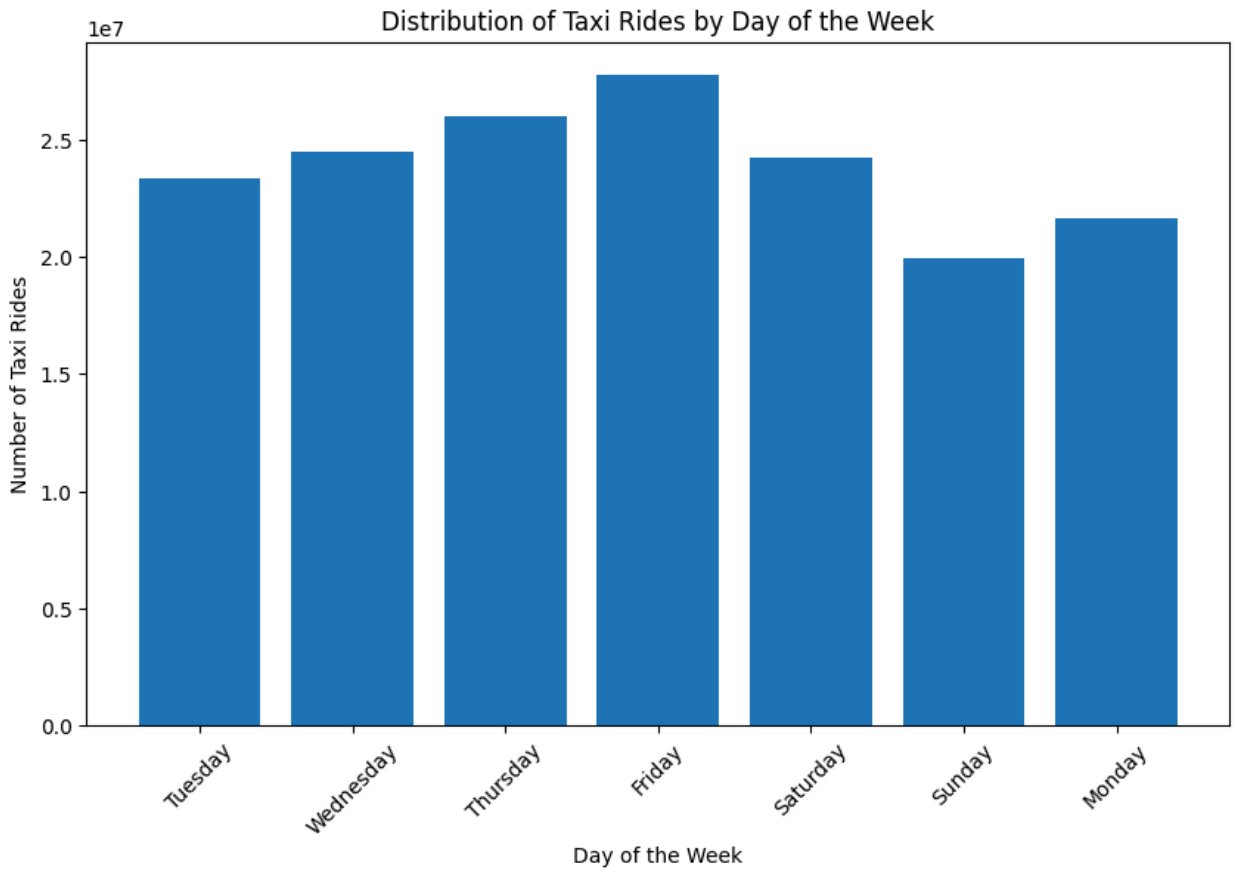

Figure 17. Taxi Ride distribution by day of the week.

The previous section describes the data preprocessing, cleaning, and sampling we did to create a smaller dataset we could work with. From there we created bi-directional graphs with edge weights - as described in the introduction - for each subset of data we desired. Throughout the process we split the data into annual subsets, warm and cold month subsets, and finally looked at the graph for our 2013-2023 sample. For each of these sets of data we found the most popular taxi routes (those with highest edge counts), plotted in descending order the in and out degree of each node, the least likely routes that occur (shortest paths in our network) that are common among all years, PageRank to determine node or Chicago community importance in the network, and community detection to determine what Chicago areas are most linked together via taxi trips. All of these steps can be found at the following GitHub notebook:

https://github.com/nschultze/CS579Project2/blob/main/annual_metrics.ipynb

One of the largest challenges we faced was with the dataset size, which we described our approach and solution to in the previous Data section.

For plotting the in and out degree we found that plotting those values for each node on the same graph was very visually informing. An example of this plot is shown below in Figure 6.

We created these graphs for each of the years 2013-2023 and also the warm and cold months. In Chicago we find that the weather is still quite cold for longer so we decided to split the data by cold months in Chicago - generally November to March - and the warmer months - April to October - since the weather here doesn't so strictly follow the general seasonal patterns.

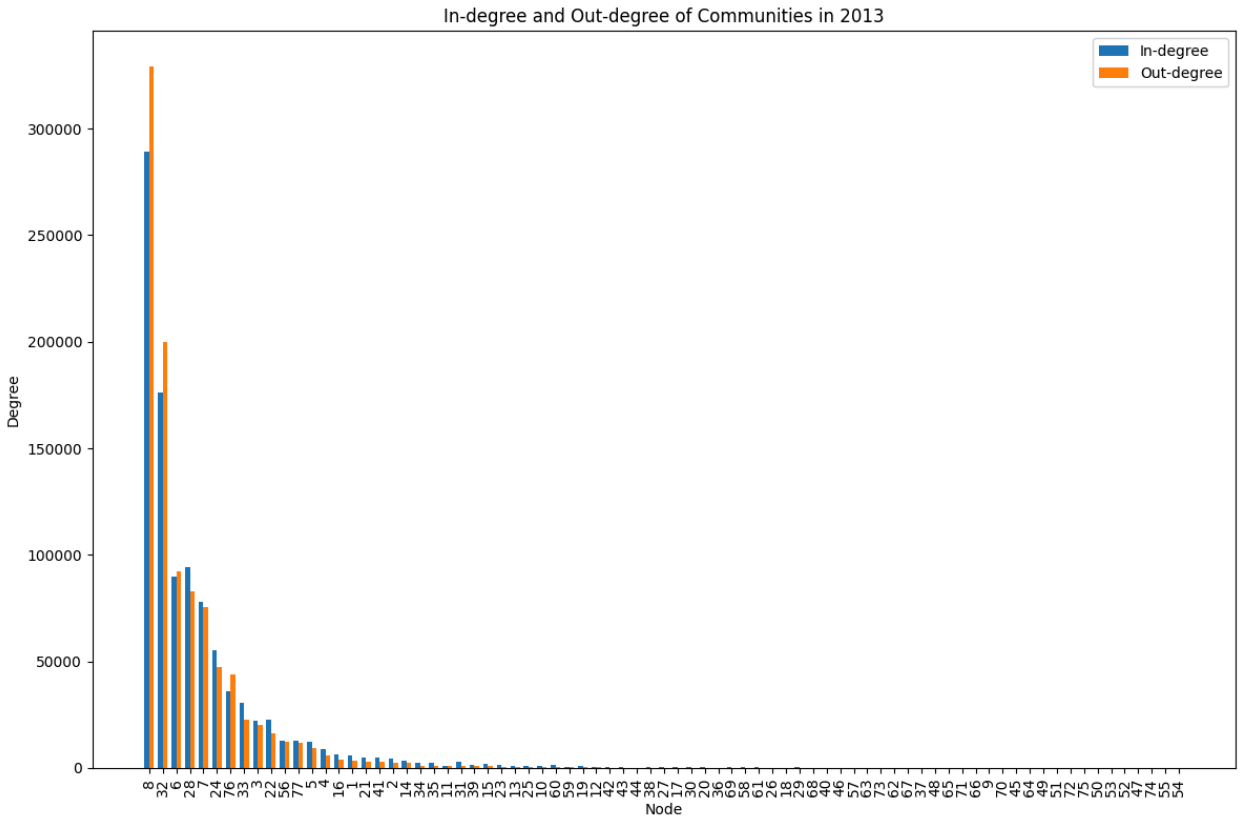

Figure 6. In and out degree of each Community number for 2013.

Next we determine and rank the edge betweenness for the 20 highest edge betweenness centralities. Since this method determines the shortest paths via the lowest edge weights it means the higher the edge centrality in our case the less used that taxi route is. To assist in understanding how these unpopular routes have persisted from year to year we found the most frequent unpopular routes and sorted those. At first, we found it a bit confusing and against our expectations as to why nodes which were often visited or had little to no taxi traffic - as informed by our in and out degree graphs - were ranked with high betweenness centrality. But in analyzing the formula and understanding how we gave edges weights we understood that we were actually ranking paths that received few taxi trips. This made our results make more sense in the context of the data we had already collected and analyzed. The highest edge betweenness for 2013 is shown in Figure 7 and Figure 8 shows the most common least popular routes through all 11 years.

```
2013 Top 20 Edge Betweenness Centralities - Paths with least taxi traffic:
         (From, To)
1. Edge: ('47', '41'), Betweenness Centrality: 0.017927046012068862
2. Edge: ('32', '47'), Betweenness Centrality: 0.015742023229189765
3. Edge: ('47', '46'), Betweenness Centrality: 0.015287310979087355
4. Edge: ('74', '8'), Betweenness Centrality: 0.013953611747729397
5. Edge: ('41', '11'), Betweenness Centrality: 0.011904903903195397
6. Edge: ('28', '47'), Betweenness Centrality: 0.011173155597141262
7. Edge: ('46', '25'), Betweenness Centrality: 0.010369219730694313
8. Edge: ('53', '38'), Betweenness Centrality: 0.00984482259802864
9. Edge: ('33', '57'), Betweenness Centrality: 0.009679088662045567
10. Edge: ('52', '30'), Betweenness Centrality: 0.009571261079647805
11. Edge: ('55', '48'), Betweenness Centrality: 0.00956691573714637
12. Edge: ('54', '30'), Betweenness Centrality: 0.009251825603374597
13. Edge: ('8', '50'), Betweenness Centrality: 0.009241772066102131
14. Edge: ('50', '2'), Betweenness Centrality: 0.009216600942889832
15. Edge: ('48', '30'), Betweenness Centrality: 0.009197533470885219
16. Edge: ('8', '52'), Betweenness Centrality: 0.00894281773895581
17. Edge: ('70', '64'), Betweenness Centrality: 0.00876566992799119
18. Edge: ('56', '26'), Betweenness Centrality: 0.008541435138967487
19. Edge: ('66', '30'), Betweenness Centrality: 0.008046391551835223
20. Edge: ('56', '54'), Betweenness Centrality: 0.007883011174118286
```

Figure 7. Top 20 edge betweenness centralities for 2013 where edge (v1, v2) is a trip from v1 to v2

```
('8', '55'): 4
('28', '74'): 4
('74', '8'): 3
('55', '48'): 3
('32', '55'): 3
('8', '54'): 3
('56', '47'): 3
('76', '54'): 3
('32', '74'): 3
('76', '47'): 3
('74', '28'): 3
('47', '41'): 2
('32', '47'): 2
('47', '46'): 2
('28', '47'): 2
('54', '66'): 2
('28', '50'): 2
('55', '33'): 2
('46', '29'): 2
('47', '44'): 2
('48', '61'): 2
('54', '49'): 2
('1', '17'): 2
('54', '70'): 2
('33', '20'): 2
('56', '18'): 2
('8', '74'): 2
('56', '55'): 2
```

Figure 8. Repeated unpopular taxi routes from 2013 - 2023 where (v1, v2) is from Node v1 to Node v2

To contrast the unpopular routes we also found and printed out the most popular routes. This gave us insight into which communities were often having people picked up by taxis or where taxis were often sent to and how those Chicago communities interacted with each other. Figure 9 shows the most popular routes for the cold months of Chicago and Figure 10 shows the most popular routes for the warm months of Chicago.

```
Top 20 Edge Weights Cold Months:
8 -> 8: 509945
32 -> 8: 361243
8 -> 32: 323260
32 -> 32: 268682
32 -> 28: 173104
8 -> 28: 157028
28 -> 8: 148318
28 -> 32: 137559
76 -> 8: 101758
8 -> 7: 88071
28 -> 28: 79917
8 -> 6: 67317
76 -> 32: 64947
6 -> 6: 58673
8 -> 24: 58551
7 -> 8: 48443
32 -> 33: 48308
8 -> 76: 47271
32 -> 7: 44708
8 -> 33: 42424
```

Figure 9. Most popular to -> from routes in the cold months of Chicago. November to March

```
Top 20 Edge Weights Warm Months:
8 -> 8: 653799
32 -> 8: 449355
8 -> 32: 428369
32 -> 32: 306704
8 -> 28: 208844
32 -> 28: 199917
28 -> 8: 184495
76 -> 8: 181249
28 -> 32: 152437
8 -> 7: 120495
76 -> 32: 119101
8 -> 6: 101784
8 -> 76: 96247
28 -> 28: 87506
8 -> 24: 80756
32 -> 76: 78093
32 -> 33: 77599
6 -> 6: 75740
8 -> 33: 71621
7 -> 8: 69498
```

Figure 10. Most popular to -> from routes in the warm months of Chicago. April to October

We calculated the PageRank score for each node for each subset of data. We then ranked those in decreasing order. This helped us further understand the important and influence of given areas of Chicago in the taxi trip network. We found that those nodes with highest in and out degrees to be ranked higher consistently across all subsets of data and those with low usage to be consistently ranked lower. An example of some of the top PageRank scores are shown below for all trips in 2013-2023 and winter weather trips in Figures 11 and 12 respectively.

2013-2023 PageRank stats

Node	PageRank
8	0.17862010367949546
32	0.12215850648374003
28	0.0797245837241568
6	0.04885651285559985
7	0.03937635556195189
24	0.03230051732831897
76	0.03180019347516924
33	0.030865653793152266
3	0.02197434426669773
22	0.016605019942311307
77	0.016514425947577203
56	0.013385152050923733
41	0.012104198125251946

Figure 11. 2013-2013 top PageRank Scores in decreasing order

Cold Months PageRank stats

Node	PageRank
8	0.17705395525676382
32	0.1231727764519575
28	0.08140797306162442
6	0.048661600562786435
7	0.039376903280721916
24	0.03257474038812452
33	0.02873506446049184
76	0.027945775022395713
3	0.022587915600168415
77	0.016981289270122827
22	0.016829017491253435
56	0.012112628388666525
41	0.01176326071325895

Figure 12. Cold months or wintery weather top PageRank Scores in decreasing order

Finally, to understand how communities in Chicago are connected to each other via taxi trips taken we used the Louvain method of community detection to find communities in the graphs. We utilized this method as it worked well with how we defined edge weights. One drawback is it does not allow for nodes to belong to more than one community which makes it difficult to represent the nuances or community complexities that our taxi graphs might have underlying them. Either way, we did this community detection analysis to compare various subsets of the data to see if communities - taxi trips between Chicago communities - have changed over the years or changed under certain conditions. After obtaining the communities detected, we labeled the Chicago community map for a few select years to help visualization. This is shown in Figures 13 and 14 for years 2013 and 2013-2023 respectively.

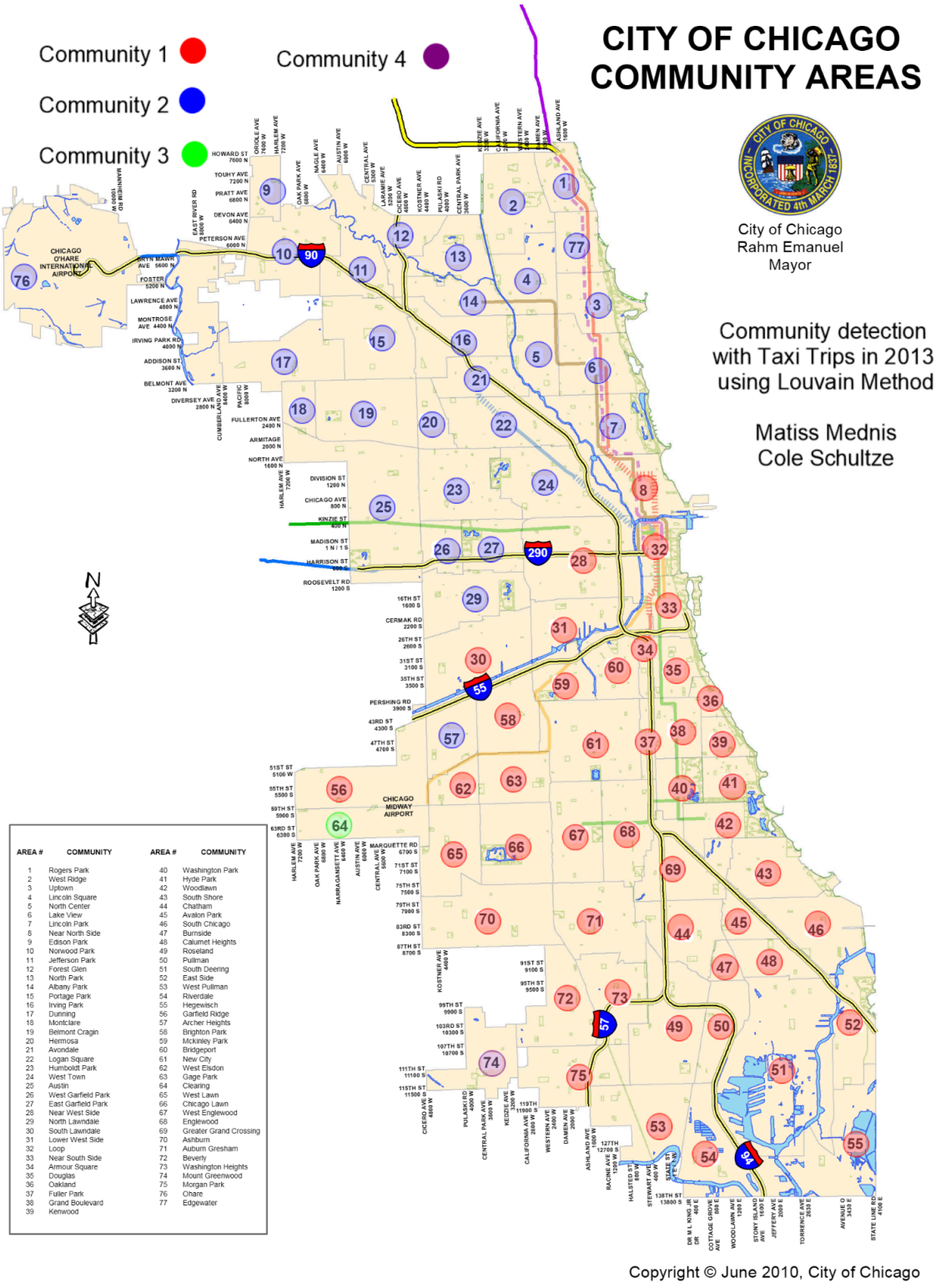

Figure 13. Community detection using taxi trip data in 2013

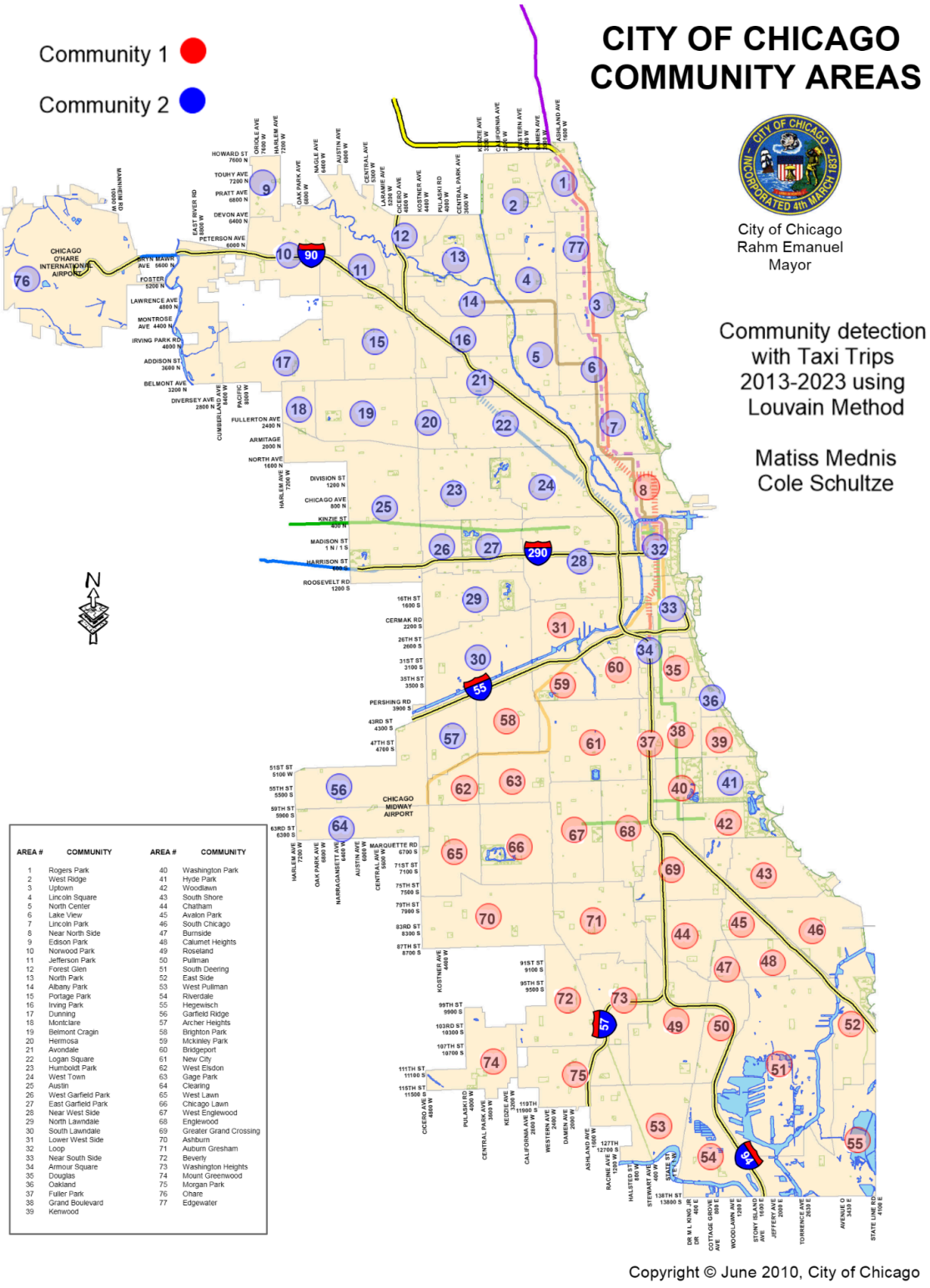

Figure 14. Community detection using taxi trip data from 2013-2013

Finally, Figure 15 combines our community and popular path data to give an idea of the intercommunity and intra community movement of taxi trips along the top 20 popular paths obtained in the previous step.

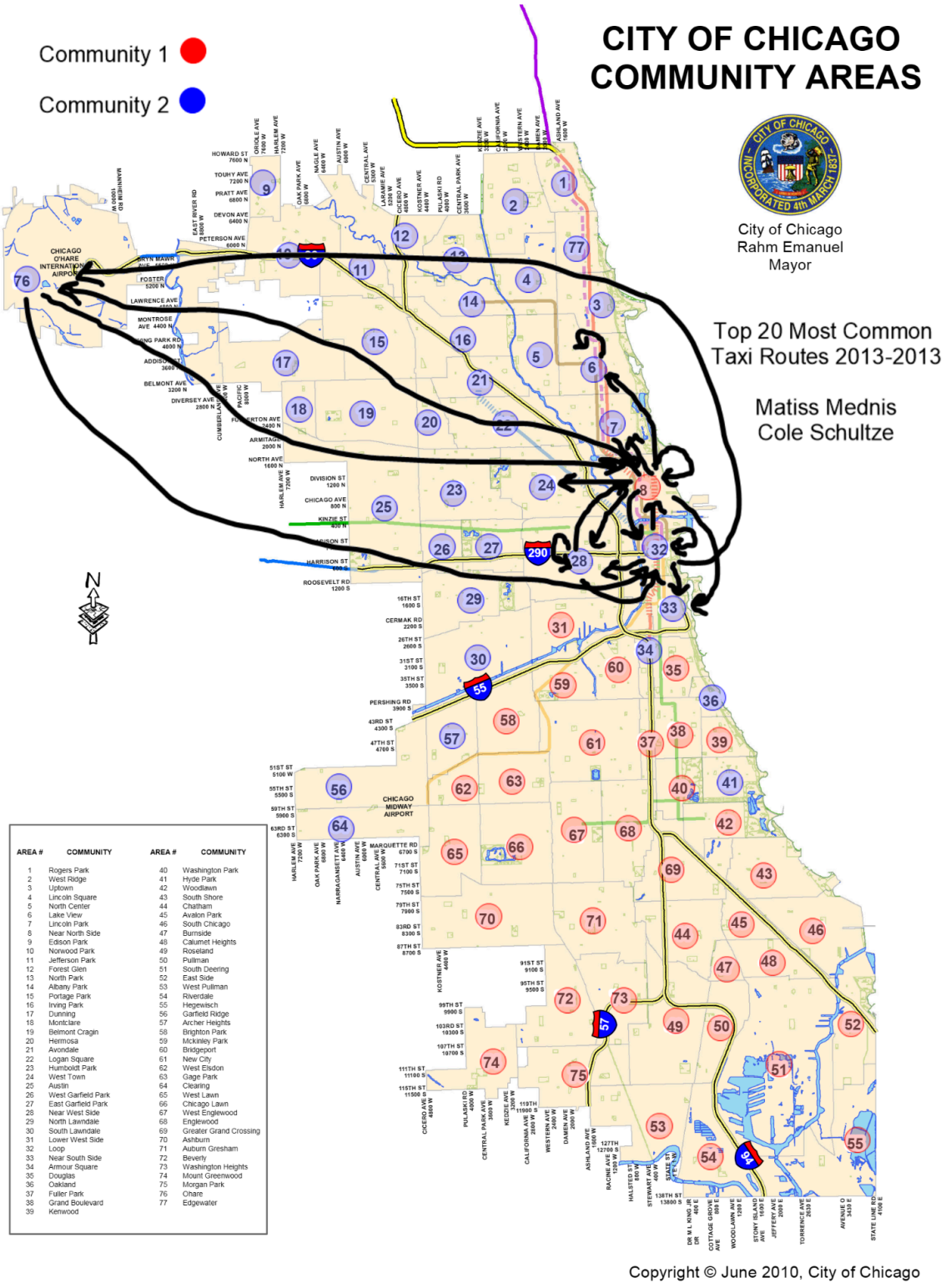

Figure 15. Top 20 most popular to and from paths for taxis in chicago from 2013-2023. Black arrows represent the path of the trip between nodes where an arrow pointed to itself is a cycle

within a community. Communities are also shown in red and blue according to the diagram legend.

Once we were able to process the data and sample it to create usable sizes we did not run into too many challenges with the processes completed in our project. It was important for us to think critically about the results we obtained at each step to ensure we were interpreting them according to the formula or algorithm definitions and aligned them with how we represented our graph. Coming up with a good visualization method was challenging as well and took some manual work to create those graphs, though they helped tie all the information together in a single understandable manner. It was also challenging to get all of the graphs and metrics for each subset of data but with proper planning we were able to store those in data structures that allowed for easy iteration of the graphs through the metric functions we created. Additionally, by downloading the graphs locally after creating them for the first time we were able to speed up processing time by loading in the graphs rather than having to begin every new coding session by reading in the data files and creating the graphs. This saved us a lot of time everytime we wanted to run our code or try something new. Our own automation and reduced processing certainly helped us overcome the challenge of the dataset size.

For all of our graph creation and calculations we used NetworkX and referenced the documentation found here

https://networkx.org/documentation/stable/reference/index.html

Discussion of Results

As a result of this project we were able to identify popular taxi routes, unpopular or unlikely taxi routes, communities within Chicago defined by where taxis are going and coming from, popular starting and ending destinations for taxi trips, and how all of those factors interact and inform a general understanding of taxi trips within Chicago. We did not dive as deeply into all of the different ways to segment the taxi data - such as by hour of the day, day of the week, month, sports season, or large events in the city - but we did gain a larger picture of the ways taxis are used in the city. Although we did not create a complex predictive model we certainly learned enough to understand what factors would be important for that model and even learned enough to have a very basic - and maybe good enough - understanding of the dynamics of taxi traffic over the last 11 years and across seasons to make some informed decisions.

Firstly when considering how taxi usage has changed from 2013 to 2023. We ultimately found that essentially taxi route distributions and data look about the same over the 11 year span. The number of taxi rides did drop significantly in years following 2018 compared to 2014 as we see in Figure 3. This may be due to various external factors that our dataset doesn’t capture as

discussed in previous deliverables and earlier in this report. Even so the number of trips made have changed but the underlying properties of where those occur has not changed significantly. The most popular routes, the most popular start and end destinations, and the most popular routes in 2013 were also the most popular routes in 2023. Even across different overall weather conditions we find the same routes and nodes to be important. This is shown in Figures 11 and 12 as we see the top PageRank scores from 2013-2023 are ranked in the same order as those in wintery weather months.

We see the communities detected throughout the years to be fairly similar as well. Notably in 2019 there are three communities where some of the more central communities of Chicago are considered within their own community. This is likely because these areas are popular for members of all communities to visit and thus these Chicago community areas realistically could belong to more than one community in a community detection at a time. We see this in Figures 13 and 14 as the North and South portions of Chicago are fairly consistently split but the more central regions move between the South and North communities. This would show us that taxi drivers which start in the North may expect to continue to complete trips in the North or go to more central downtown and loop areas. Similarly, those in the South could expect the same. Taxi drivers in the loop or downtown area might then expect to stay there or end up taking trips both North and South.

Overall, in combining all of the information we obtained via the metrics and process discussed in the previous section for all 2013-2023 data combined, separated by year, and the data separated by seasons we see an overall pattern in the taxi trip data that holds true for all of these subsets. We find Figure 15 to be the most informative and greatest summary of what we observed among our data. We see the most popular routes in any subset of the data we analyzed to be within downtown areas and to the airport from those areas. Additionally, going from the airport taxis most often are dropping people off in the Loop or Near North Side. This makes sense as these are not only popular destinations for attractions the city offers but also areas with high population density in the city. Since we did not normalize the counts of taxi trips we see highly populated areas receive a lot of demand as we might expect. Also these are popular areas with many attractions for tourists and residents alike.

Also informative was our analysis of the lowest PageRanks and least popular edges. We see that communities far from the city such as number 74, 55, 9 all consistently receive few taxi rides. This may be due to the fact that people in those areas don’t use taxis or that those areas simply don’t have taxis. But maybe those taxis aren’t there because people don’t use them. Without more data we cannot be certain but we can say that those are not popular destinations for taxis to come from or go to. We find many of the communities on the outskirts of Chicago Land do not have a high influence on the network and are not common trips to make when taxis are starting in more high influence community areas on the network.

We certainly gained a lot of insight into how taxi trips occur in Chicago and we observed a sort of power law in the degree distribution. A few communities in Chicago receive a lot of the taxi trips.

As far as how our results can be applied, we find that having taxis along these popular routes is probably a good idea as even when overall taxi use drops, the proportion of traffic between nodes stays the same. The important communities in this network - those that receive a lot of trips - as well as the most important routes have not changed considerably over the years and don't change considerably by overall weather conditions either. In the big picture view we see the network keeps around the same properties. The full results for all of the data, code, and all of the graphs which helped us reach these results and conclusions for both annual and season subsets can be found hosted here:

https://github.com/nschultze/CS579Project2/blob/main/annual_metrics.ipynb

Also it was interesting to see the distribution of ride lengths shown in Figure 16 below as we learned that most trips are short which was further explained by the most popular edges we see visualized in Figure 15. Aside from trips to the airport, most trips are within neighboring Chicago communities.

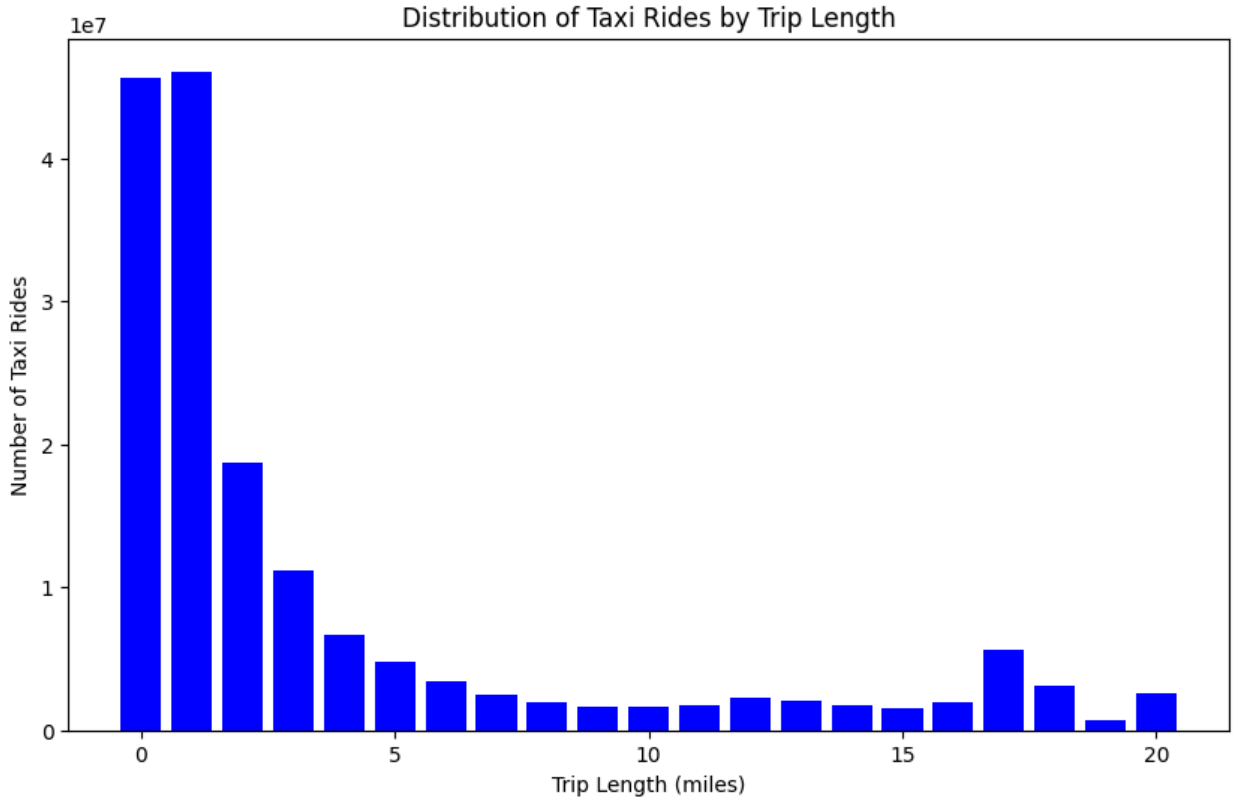

Figure 15. Distribution of number of taxi rides by trip length

Future Work

We observed no clear changes in the network with the subsets we had time to develop and analyze. Additionally, we were unable to apply more sophisticated methods of determining smaller changes in the network or create more sophisticated predictive models that pick up more nuances and patterns that emerge between subsets of our data. We did observe a power law distribution in our dataset, where the top few popular nodes held most of the importance and information in our network but maybe it would be of use to dig deeper into the less popular nodes behavior overtime to determine growing Chicago communities or those which may have higher taxi demand or become more important routes in the future. We could analyze these behaviors and shift our focus on up and coming neighborhoods that received more traffic over time.

We may also implement more sophisticated community detection algorithms which allow for membership to multiple communities to capture the deeper complexities of how the given nodes in our graph interact with each other through taxi trips.

In our EDA we saw how taxi demand changed depending on the time of day and day of the week. It could also be of interest to combine this data and segment taxi network information based on the time of day and day of the week. By creating more granular and specific time periods for analysis we might find particular combinations of data which give unique behavior in our underlying taxi network. This could be informative at finding potential anomalies or significant changes in the network from what we see on a larger annual basis such as we did in our project. These patterns might also be picked up by predictive models which are fed this data as features. Similarly we could dedicate resources to researching major festivals and events in the area covered by the taxi ride data. We could explore potential correlations between these events and taxi ride patterns by identifying and cataloging major festivals, concerts, sporting events, conferences, and other significant gatherings. This analysis could provide valuable insights into the impact of events on transportation demand and help transportation authorities, event organizers, and ride-sharing companies better anticipate and plan for increased transportation demand during major events.

Continuing to collect taxi data and see how it changes on a larger time frame would be more informative as well as it might help network analysis draw patterns of how popular neighborhoods are formed overtime via taxi traffic. Additionally, additional datasets such as rideshare transit, public transportation, and personal transportation all might help inform this network or help us create alternate transportation networks.

We might also elect not to drop rows with missing community pickup or dropoff data as we might learn more insight about which communities often leave the Chicago area and which communities people are visiting from outside the Chicago area. This may be more difficult as missing drop off or pick up data could also be due to some missing information or errors in the trip row.

Additionally, with more time, we could further refine our visualization techniques to communicate our analysis results effectively. This may involve exploring advanced visualization libraries, creating interactive dashboards, and developing dynamic visualizations that allow stakeholders to explore the data and gain deeper insights interactively.

We could explore additional advanced analytical techniques and machine learning algorithms to uncover hidden patterns, correlations, and trends within the dataset. This could include implementing predictive modeling approaches to forecast taxi demand, identify hotspots for ride-sharing services, or optimize transportation routes for efficiency and cost-effectiveness. Overall, with more time, more complex methods, models, and more granular data segmentation could all be implemented to further inform taxi networks and usage overtime and under different conditions in Chicago.

References

https://data.cityofchicago.org/Transportation/Taxi-Trips-2013-2023-/wrvz-psew/about_data
https://networkx.org/documentation/stable/reference/index.html

https://python-louvain.readthedocs.io/en/latest/

https://neo4j.com/docs/graph-data-science/current/algorithms/louvain/

https://stackoverflow.com/questions/49151996/meaning-of-weights-in-community-detection-algorithms

Fast unfolding of communities in large networks: https://arxiv.org/abs/0803.0476