SIMULATING MOVEMENT THROUGH ROAD NETWORKS: A NETWORK SCIENCE APPROACH

Sarah Rodenbeck and Mike Soukup
MSDS 452: Web and Network Data Science
December 4, 2022

Abstract

Traffic continues to be a recurrent problem in many large cities. Traffic mitigation strategies and changes are traditionally planned based on traffic studies, but these are limited in their geographic target areas and thus may not fully assess the impact of such changes across the road network as a whole. We propose utilizing network-science methods to model and analyze traffic at a network level. In doing so, one could then empirically study the utility of various network constructs along with the impact of proposed traffic solutions. For these methods to be successful though, models which mimic real-world behaviors are crucial. Therefore, our work explored various methods found in literature to determine the modeling approach which most accurately predicted real-world data. Two different data sets (an aggregated map with traffic flow metrics and a granular road network with traffic flow assessed based on actual taxi trips) are used in our analysis, and both heuristic and simulation-based approaches are used to emulate flow. Results suggest that heuristic approaches yield the most accurate results for studying network flow currently, but that simulation-based approaches are likely to be the better choice for future work due to customizability.

Keywords: traffic simulation, network science, geospatial analysis

Introduction

Road networks play an integral role in modern economies and they continue to be a predictor of economic growth (International Road Federation 2020). Such infrastructure impacts the quality of life of citizens and how cities can grow. However, traffic remains a major concern in many cities despite attempts to control it. A 2008 study estimated that traffic congestion in the Chicago region cost more than 7 billion dollars in time, fuel, resources, and environmental effects (Metropolitan Planning Council 2008). In fact, in some areas, attempts to control traffic, for example by over indexing on monolithic routes and travel methods, have actually made congestion worse (Cortright 2019). This poses the question of what methods are traditionally used to analyze traffic to suggest solutions to congestion. Are the proposed solutions the reason for failure of past traffic control mechanisms or are the methods used to identify these solutions also at fault?

Traffic studies are a core functionality of municipal and state transportation departments and are used to examine, modify, and plan the transportation systems available in a specific area (T-Square 2020). For example, a traffic study is typically conducted prior to large development projects (e.g., a new subdivision), and they may also be used to study areas with frequent accidents. Traffic studies typically consist of collecting data about existing traffic in the target area then forecasting new traffic based on expected developments and growth (T-Square 2020). However, this suffers from several key disadvantages. First, such studies are resource intensive, requiring the expertise of a team of traffic engineers. Second, they typically focus on a very small area of interest, which greatly impacts visibility of how proposed changes may impact traffic flow across the network as a whole. These are key limitations, yet there is a dearth of literature proposing methods for full road-network analysis; research offers little with regards to

improving current traffic study methodologies. The methods which are proposed in literature have their drawbacks as well, like being computationally expensive or being sub-optimally adapted from other network domains such as social or communication networks (Marshall 2016; Park 2010). We postulate that full network analysis, specific to the traffic domain, may provide key insights into traffic flow across a network.

In order to analyze road networks effectively, we explore a graph representation where intersections are nodes and roads are edges. Road networks are unique from other types of networks because of their physical constraints. For instance, road networks are largely two dimensional and are constrained by local geography and municipal resources. Additionally, road networks have a very unique structure due to the topology of cities and often include a large number of both nodes and edges, but nodes are often practically constrained in their maximum degree (e.g., a typical value would be four for a four-way stop). Additionally, while roads as a whole may cut across an entire city, each road segment has a minute scope such as a city block.

This paper is an empirical survey of data-driven and network-based approaches for analyzing networks within the context of the Chicago road network. The research question of focus being: what type of network construct and available data sources offer the best foundation to further analyze and study modern road networks? To answer this question, we explore two different data sources and three methods for doing so.

Literature Review

Our initial research provided an overview of legacy approaches which were used to study and analyze road systems for planning. This traditional approach uses sketch planning zones, which consist of multiple regional zones of land units originally surveyed in the mid-1800s, to generate an aggregated road network with a limited number of edges and nodes (Eash et al.

1983). This aims to reduce the computational complexity needed to study traffic problems. The goal of this approach was to strategically plan infrastructure investment with regards to long-range road systems, and it also facilitated the study of the equilibrium assignment of traffic through a nonlinear convex optimization problem, although this is primarily an academic exercise. The authors even note that "Project-level and corridor planning will almost always require more detailed network coding and smaller analysis zones" (Eash et al. 1983).

While the data and approach presented by Eash et al. piqued our curiosity for studying road networks, conducting further literature review revealed much has changed in the nearly 40 years since the work was published. For instance, readily available computational power has increased dramatically since Eash et al.'s paper was published, which has reduced the need for aggregated networks such as the sketch Chicago network (Sieber 2022). The increased availability of computational resources and successes achieved with the availability of big data has enabled more ubiquitous analysis of networks using algorithms from graph theory (Ornes 2021). Additionally, the road network used to generate Eash et al.'s network in the early 1980s has likely been invalidated for modern study due to road construction, development, and growth of the city in the intervening years.

However, it also became apparent through literature review that there is no standard for modeling road networks that can scale from a single intersection to entire cities. No single approach was frequently referenced or established itself as the benchmark for comparison. Many approaches were referenced such as utilizing ArcGIS (Das et al. 2019), Line Structure (Marshall 2016), and analysis concepts borrowed from social network analysis (Park and Yilmaz 2010). While many of these ideas seem theoretically interesting, there was very little evidence of the

practicality of these options as well as concerns about their ability to scale and explain complex interactions like those that occur within a large city such as Chicago.

In search for a better way to model real-world road networks we uncovered two approaches that offered a promise to offer what other methods lacked: practical outputs, the ability to scale, and able to capture the nuances of dependent traffic networks.

The first approach consisted of utilizing OpenStreetMap (OSM) in conjunction with NetworkX to model modern, real world road networks. The second approach takes this one step further and leverages the SUMO simulation suite to simulate configurable traffic instances while providing data outputs for analysis.

While there are now considerable options for Volunteered Geographic Information (VGI), OSM is the leading example in this domain. That is because OSM has high fidelity data, a vast ecosystem of software systems and applications, tools, and web-based information stores (Mooney and Minghini 2017). OSM has been shown to compare favorably with other VGI sources in terms of data quality, and this data is updated and validated by the community on a regular basis through GPS traces and aerial imagery (Packt 2010). Developers have created a python package, OSMnx, that enables one to interact with OSM and analyze the road networks with familiar tools such as NetworkX (Boeing 2017). With OSMnx, users can download and model walkable, bikeable, and driveable urban networks with a single line of code. OSM is able to offer vast data such as points of interest, infrastructure types, and pedestrian walkways which together enable the study of large and complex urban road systems.

Another interesting road network analysis tool that came up during our research was SUMO. SUMO, or the Simulation of Urban Mobility, is a free and open source traffic simulation suite (Lopez et al. 2017). It is not based on OSM, but part of the suite enables OSM data

extraction. SUMO enables users to simulate large scale, dynamic, and complex traffic networks consisting of thousands of vehicles and pedestrians. The ability to configure nearly limitless aspects of urban transportation systems such as public transportation, waterway traffic, traffic light behavior, emissions, and hypothetical road systems offers a degree of practical insight not rivaled by any other software solution we came across.

Methods

This project consists of three phases: studying legacy road network analysis approaches by exploring an aggregated Chicago road network through traditional means, exploring the full Chicago road network with OSM and heuristic approaches, and finally, simulating traffic through the OSM network by leveraging the SUMO simulation suite.

The initial strategy utilized data from the "Transportation Networks for Research" repository (Transportation Networks for Research Core Team 2021). This repository includes road networks for several cities, including Chicago. Each city network includes information about the road network itself (e.g., a list of nodes/edges and their attributes like speed limit and capacity) as well as information about flow and trips. The sketch Chicago road network includes a relatively modest 933 nodes and 2172 edges due to the bundled format, and traffic flow data simply represents the flow of traffic between those zones rather than from/to specific points.

To analyze this data we first generated a Networkx directed graph (to preserve one-way streets) from the provided edgelist. Our analysis of this data consisted of an exploratory data analysis and visualizations of the sketch network as well as an exploration of several standard link prediction algorithms. In addition, a variety of centrality measures (overall degree, in-degree, out-degree, closeness, betweenness, Eigenvector centrality, Katz, PageRank, HITS_Hub, and HITS_Auth) were evaluated on each node aiming to determine the most central

nodes in the network. These efforts relied heavily on standard Networkx functions. However, this approach suffered from the aforementioned obsolescence of the provided road network in addition to limitations in the granularity of insights.

This led us to explore and utilize OpenStreetMaps (OSM) data for the second and third phases of this project. Our analysis of this data centered on how well heuristic and simulation-based methods could mimic actual traffic patterns in order to produce insights for traffic engineers.

In the absence of real traffic data, taxi trips were used as a proxy for traffic patterns. The taxi data was originally obtained from the City of Chicago data portal and consisted of over 16,000,000 taxi trips in 2019 (City of Chicago 2022). Of this, we conducted our study on a randomly selected subset of approximately 800 data points due to computational and data completeness limitations. The computational limitations were due to exponential runtime complexities when running SUMO simulations with marginally larger networks or marginally more traffic. Meanwhile, some of the initially subsetted taxi trip data contained null values for required fields for analysis such as trip duration, pick up coordinates, and dropoff coordinates so the corresponding entries were necessarily dropped.

While the taxi data may have bias towards certain types of travelers and suffers from using approximate and bundled locations (to preserve riders' privacy), we postulate that it can provide an estimate of travel in the network. We attempt to model flow by predicting routes for each taxi trip and use these trips in aggregate to identify more and less frequently traversed areas of the network, which is likely to mirror overall road usage. It should be noted that real traffic data would obviate the need for this approach and its inherent issues. One of these issues is that a large number of taxi trips, ideally disaggregated by time of day, are needed to provide an

accurate picture of traffic, but it is quite computationally expensive to repeatedly calculate routes for such a voluminous data set.

Phases two and three of this project centered on finding the best ways to model trips and traffic flow in order to mimic the "ground truth" trip durations captured in the Chicago Taxi Data. First, phase two of the project utilized standard network and heuristic methods to model flow across the OSM network of Chicago. Nodes nearest to each pickup/dropoff point were calculated, and Networkx was used to calculate the shortest path between these points in the Digraph using Dijkstra's algorithm. Shortest paths were determined based on travel time across each edge included in the path (as calculated based on the distance and speed limit of each road segment). To coarsely simulate traffic lights/stop signs, a penalty of 30 seconds was added for approximately 65% of intersections crossed along the route. Notably, travel times calculated through this method do not account for the impact of other cars on the road; on heavily used roads each car's speed is dependent on other cars around it, not just on the posted speed limit.

Finally, the third phase of the project centered around simulating the taxi trips via the open-source SUMO simulation suite (Lopez et al. 2018). The SUMO software enabled the execution of our taxi trips in a highly configurable simulation environment. The SUMO simulations were conducted three times. First, each taxi route was simulated in isolation meaning no other cars were on the road and the taxi only had to navigate the network and obey basic traffic laws (referred to as the series simulation). The second SUMO simulation consisted of running the taxi routes in parallel meaning that all taxi trips arbitrarily chose when to begin their route within an hour window and could interact with one another (referred to as the parallel simulation). The third and final SUMO simulation added an arbitrary amount of traffic to the parallel approach over the one hour period (referred to as the traffic simulation). The parameters

used to model traffic via the SUMO osmWebWizard tool consisted of the car-only network with a through traffic factor of one and a count of twelve. The car-only network ensures SUMO only extracts roads that permit passenger car traffic to reduce network complexities. The through traffic factor indicates how much more likely it will be for passenger routes to depart or arrive at a boundary of the network. A value of one indicates that it is equally likely for a trip to depart or arrive at a boundary or inside the simulation area. Lastly, the traffic count attribute indicates how many vehicles are generated per hour and lane-kilometer (Lopez et al. 2017).

After the heuristic and simulation method results were collected, they were compared to the Chicago Taxi Data subset to evaluate which method best resembled real-world scenarios. In order to compare our outputs to the truth set, several approaches were invoked. First, an EDA was conducted which included visualizations as well as an exploration of standard statistical metrics of the data as well as their delta arrays following the formula: truth duration - simulation duration.

Beyond EDA, the forecasted trip results for all taxi trips were compared and the most accurate was identified for each trip to reveal the proportion of time each forecast method was closest to the actual data. Lastly, the Kolmogorov-Smirnov test for goodness of fit was invoked to determine how closely the distribution of each set of forecasted trips compared to the distribution of the truth set of results.

Results

Phase one of the project yielded a high-level understanding of the Chicago road network. Visually, the aggregated network conforms with high-level maps of Chicago, both of which depict Lake Michigan to the East, dense clusters of nodes in the Loop, and more sparse node

locations towards the outskirts and suburbs of the city. Additionally, it is clear that freeways and expressways are overlaid on the grid-like network structure of the city.

The Chicago sketch road network is directed and strongly connected, which makes sense as the purpose of a road network is to facilitate travel across the network regardless of direction of travel. The diameter of the network is 32, meaning that the longest path across the network crosses 32 nodes or intersections. However, the full Chicago road network will likely have a much higher diameter without aggregation. This is because real intersections are only implicitly included within a zone on the sketch network and not directly represented in the graph. Also, routes consisting of many side streets traversing north, south, east, and west can cross many junctions and still get a traveler across the entirety of the network.

A small number of edges, primarily by the lakefront, have a very high volume of traffic. The majority of edges have average traffic volumes under 2,500 vehicles, which, if disaggregated, would likely be considered free-flowing traffic where a car's speed is not dependent on that of other cars (Jamal 2017). This raises the point that the majority of roads in a road network are likely to have low traffic volumes even in a crowded city (e.g., minor collector roads), and also that daily traffic volumes may not paint the full picture of what roads look like during rush hour. This exponential distribution of traffic volumes does not translate to degree though as degree appears to follow a normal distribution, meaning that this is not a scale-free network (Figure 1). Intuitively, this makes sense as the goal of a road network is to distribute traffic not centralize it in one place.

In an attempt to determine which nodes were most central in this network, the top 50 nodes for each centrality ranking were determined and the set intersection of each of these ten centrality metrics revealed the three most central nodes in this network (Figure 2). These nodes

all appear to lie within the heart of the city and are arranged linearly in a north/south orientation. Interestingly, these nodes are all arterial nodes and do not correspond to Freeway or Expressway junctions. Beyond centrality, we also explored various other network patterns and measures such as cliques, triangles and k-cycles, but found computational times to be very demanding. However, typical structure analysis of road networks is likely offers little value here. For instance, it offers little value to traffic engineers to know that a road network of a vast city like Chicago contains nearly 40,000 triangles.

The next step of analysis involved exploring how the network structure could help inform transportation projects in Chicago. First, a naive visualization of all edges where capacity exceeded volume suggests that there are certain areas of the city where future road construction could be targeted to have the greatest impact on congestion (Figure 3). Standard link prediction algorithms, however, are not likely to be an optimal way to study road networks for two primary reasons. First, almost all standard link prediction algorithms (e.g., Jaccard, resource allocation, preferential attachment, etc.) can only be implemented on undirected graphs (Networkx n.d.). While the Chicago network can be converted to an undirected network this removes information about one-way streets. This is detrimental to analysis because an understanding of directional flux is vital to analyzing the Chicago road network in order to account for differential in-bound/out-bound patterns throughout the day. Second, because these methods rely solely on the structure of the graph to predict edges rather than on attributes of the edges like volume or speed, they are poorly equipped to create use-informed suggestions. In particular, they yield suggestions that would often more than double the number of edges in the network which is not a practical traffic mitigation strategy (Figure 4).

The limitations of our initial findings led us to more contemporary and use-informed approaches leveraging OpenStreetMaps and taxi trip data. In contrast to the aggregated map, the OSM map of Chicago included 28,686 nodes and 76,000 edges (Figure 5). This graph is directed and a multi-graph, but is only weakly connected. Notably, while it includes high-fidelity data about each edge's attributes (e.g., number of lanes, the type of road, speed, length, etc.), it does not include any information about traffic flow.

Using the heuristic methods and SUMO software, we were able to create four models to determine which most accurately reflected real-world taxi data. Evaluating the summary statistics of the four models, it was found that the simple heuristic approach was able to match the mean taxi trip duration. The median delta value of the heuristic model was also a mere 20 seconds off the real-world taxi data and had a comparable standard deviation to that of the non-traffic SUMO simulations (Figures 6 and 7, Table 1). Of the SUMO simulations the traffic simulation had the most similar values to the ground truth data with mean and median delta values of 131.5, and 156 seconds respectively. This indicates that the best SUMO simulation still underestimated travel times by a factor of 2-3 minutes which may be a result of underestimating the number of additional vehicles needed in the simulation. While the traffic simulation yielded the most accurate of the SUMO simulation results, the delta array was also the most variable with the highest standard deviation. Visualizations of heuristic and simulation OSM maps and UIs can be found in figures 8, 9 and 10.

We also determined that heuristic approach offered the most accurate trip duration on 88.9% of the 785 taxi trips under evaluation. The traffic model was second most accurate, offering the closest approximation 8.6% of the time followed by the parallel and series models, respectively, which together only produced the best result for 19 trips. Lastly,

Kolmogorov-Smirnov tests for equivalency of distributions were evaluated for each model type (DeepAI n.d.). Each distribution had a p-value considerably less than 0.05 indicating that while these models appear close to real taxi trip data and appropriate in theory, they are still a far way off from being statistically comparable to actual data (Table 2). Nonetheless, this test further validated our conclusion that the heuristic method produced a distribution most similar to that of the actual data as this distribution yielded the largest p-value.

Given the considerable sample size, these models are not outperforming one another due to random chance. There are considerable and meaningful differences between the models which are responsible for the difference in results. Yet, while the heuristic approach appears highly suited to initial calculations, it offers much less customizability than the simulation approach. In particular, the startup costs for this approach are lower, but the overheads would be much more significant to account for all of the factors that can be easily changed within the SUMO simulation suite. Despite these initial results, we theorize that further optimizing traffic and interactions of vehicles with the environment in SUMO simulations will lead to more accurate models of real-world systems.

Conclusions

Our work has shown that networks may be a viable way of analyzing traffic for city planning purposes, but that traditional graph methods are ill-equipped to address the complexities of modern road networks. Specialized geospatial data and methods provide a much clearer picture of the road network and use, which can inform practical transportation decisions like planning for road maintenance and closures. However, computational power is a big concern here and limited what we were practically able to accomplish on local machines – any city

official or researcher looking to use such data would be best served to use a more parallelized method.

Using raw taxi trips and likely routes may be an effective heuristic through which to assess typical traffic patterns in the absence of more precise data. However, this data may not be representative of overall traffic and heuristics are not the approach most meaningfully grounded in real-world behaviors. SUMO provides the most configurable, and likely best, opportunity to precisely simulate traffic across the network, but this is a sophisticated solution with a steep learning curve that may not be appropriate in all circumstances. The developers of SUMO allude to the fact that it can be a highly tedious process to accurately simulate traffic, especially on a large network (SUMO 2022). Therefore, while SUMO is likely the most appropriate method for assessing the impact of proposed changes to a transportation network, with better data there may still be room for other approaches with lower start-up costs. If there existed more readily available data about traffic flow or even average speed across each road segment at different times of the day, heuristic methods for analyzing the network might obviate the need for SUMO.

This project highlights the need for high-precision data and geospatial-specific methods for assessing road networks. Literature on data-informed approaches to transportation networks was limited and suggests a niche for future research. In particular, there are several next steps we propose for this analysis.

First, we propose further optimizing the SUMO simulation model as well as using a larger data sample to determine the factors that play the most significant role in creating real-world traffic congestion. Not only would such a model produce more accurate results, but it would also enable edge ablation or addition studies to identify future targets for traffic engineering. This could additionally be expanded to better model traffic throughout the day by

using time-disaggregated trip datasets and customizing the simulation to mirror traffic on the road at a particular time.

Optimizing the SUMO configurations provides more than just academic value though: there are many real world applications that could be explored if such a model existed. For instance, the model could be used to simulate a traffic study and do a preliminary forecast of the impact of a new development, construction, or any other kind of road modification. However, beyond reducing costs for current methods, this approach also has the potential for introducing a new method of traffic analysis. In particular, the simulation could be formulated into a loss metric and incorporated into a genetic algorithm in order to iteratively and automatically determine the most impactful traffic changes on the network such as road addition or removal.

References

Boeing, G. 2017. OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks. *Computers, Environment and Urban Systems* 65, 126-139. doi:10.1016/j.compenvurbsys.2017.05.004.

City of Chicago. 2022. "Taxi Trips - 2019." Last modified November 9, 2022. https://data.cityofchicago.org/Transportation/Taxi-Trips-2019/h4cq-z3dy.

Cortright, Joe. 2019. "Backfire: How Widening Freeways Can Make Traffic Congestion Worse." *City Observatory*, February 26, 2019. https://cityobservatory.org/backfire_wider_worse_traffic/.

Das, Debashis, Anil Kr Ojha, Harlin Kramsapi, Partha P. Baruah, and Mrinal Kr Dutta. 2019. "Road network analysis of Guwahati city using GIS." *SN Applied Sciences* 1, no. 8 (2019): 1-11. https://doi.org/10.1007/s42452-019-0907-4.

DeepAI. n.d. "Kolmogorov-Smirnov Test." Accessed on November 29, 2022. https://deepai.org/machine-learning-glossary-and-terms/kolmogorov-smirnov%20test.

Eash, R.W., K.S. Chon, Y.J. Lee and D.E. Boyce. 1983. "Equilibrium Traffic Assignment on an Aggregated Highway Network for Sketch Planning." *Transportation Research Record*, 994, 30-37.

International Road Federation. 2020. "Road Networks." August 10, 2020. https://worldroadstatistics.org/road-networks/.

Jamal, Haseeb. 2017. "Free Flow Speed of a Vehicle." FFS Definition, Factors Affecting FFS, April 24, 2017. https://www.aboutcivil.org/free-flow-speed-ffs.html.

Lopez, Pablo Alvarez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. 2018. "Microscopic Traffic Simulation using SUMO," *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2575-2582, doi: 10.1109/ITSC.2018.8569938.

Marshall, Stephen. 2016. "Line Structure Representation for Road Network Analysis." *Journal of Transport and Land Use* 9, no. 1 (2016): 29–64. http://www.jstor.org/stable/26203208.

Metropolitan Planning Council. 2008. "The True Costs of Traffic in the Chicago Metropolitan Area." August 2008. http://www.rmapil.org/assets/documents/mpc_report_chicago_congestion_2008.pdf.

Mooney, Peter, and Marco Minghini. 2017. "A review of OpenStreetMap data." (2017): 37-59.

Networkx. n.d. "Link Prediction." Accessed on November 20, 2022. https://networkx.org/documentation/stable/reference/algorithms/link_prediction.html.

Ornes, Stephen. 2021. "How Big Data Carried Graph Theory into New Dimensions." Quanta

Magazine, August 19, 2021. https://www.quantamagazine.org/how-big-data-carried graph-theory-into-new-dimensions-20210819/.

Packt. 2010.“OpenStreetMap: Gathering Data using GPS.” September 23, 2010. https://hub.packtpub.com/openstreetmap-gathering-data-using-gps/.

Park, Kyoungjin, and Alper Yilmaz. 2010. "A social network analysis approach to analyze road networks." In *ASPRS Annual Conference. San Diego, CA*, pp. 1-6. 2010.

Sieber, Tina. 2022. “What Is Moore's Law and Is It Still Relevant in 2022?” MUO, March 15, 2022. https://www.makeuseof.com/tag/what-is-moores-law-and-what-does-it-have-to-do-with-you-makeuseof-explains/.

SUMO. 2022. “Scenarios.” Last modified June 23, 2022. https://sumo.dlr.de/docs/Data/Scenarios.html.

T-Square. 2020. “What are Traffic Studies and Why are they Important.” August 20, 2020. https://www.t2-eng.com/traffic-studies-important/.

Transportation Networks for Research Problems Core Team. n.d.“Transportation Networks for Research. Accessed November 15, 2022. https://github.com/bstabler/TransportationNetworks.

Appendix

Figure 1: Degree Distribution of Nodes in Sketch Chicago Network

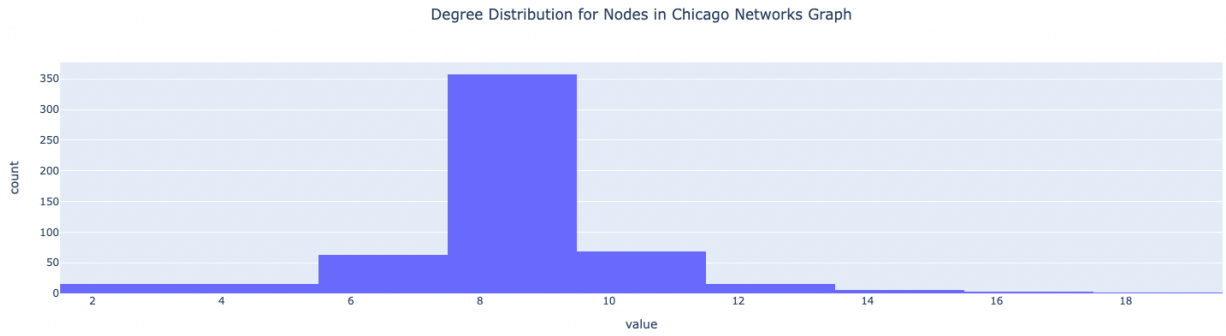

Figure 2: Most central nodes (in red) on Chicago Sketch network

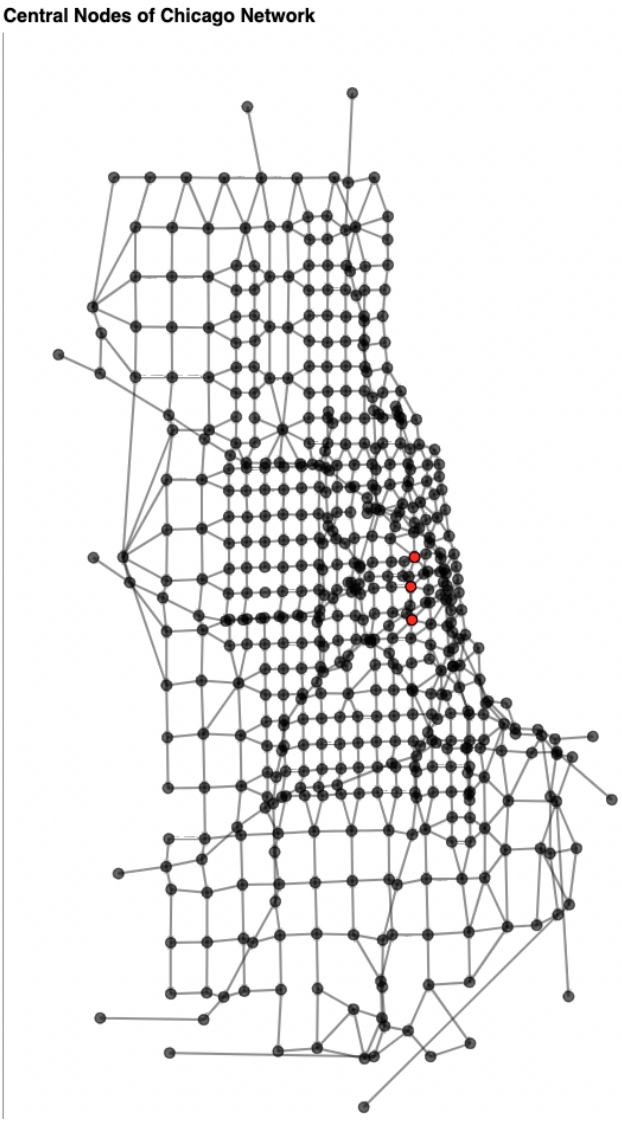

Figure 3: Edges where volume exceeds capacity on sketch Chicago Network

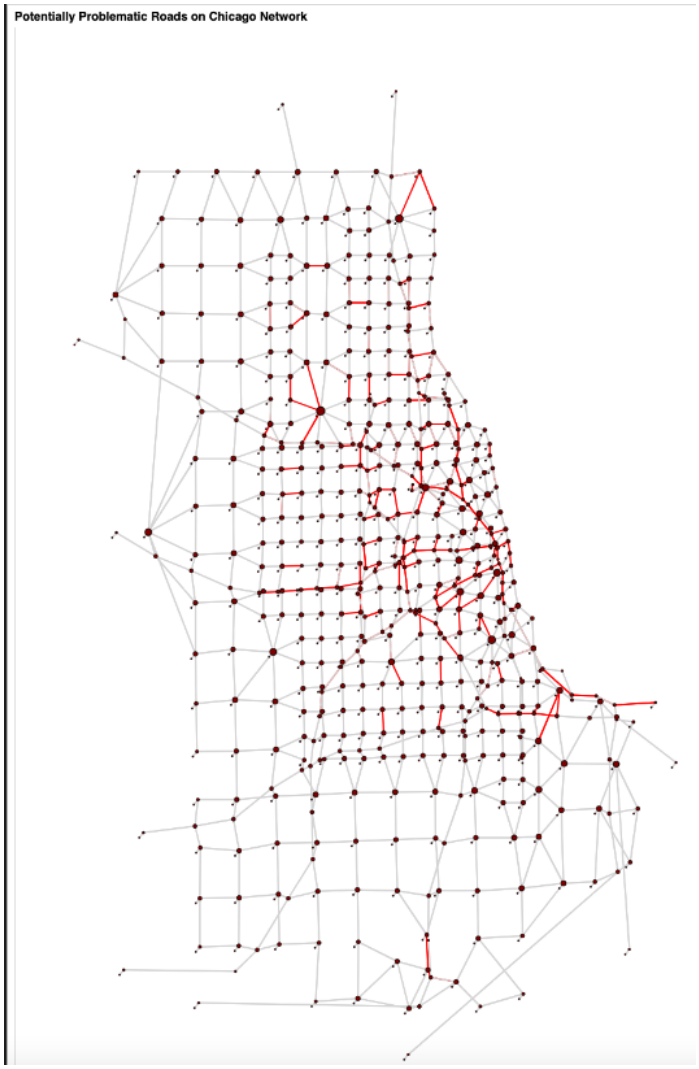

Figure 4: Results of running Jaccard link prediction algorithm on Sketch Chicago network

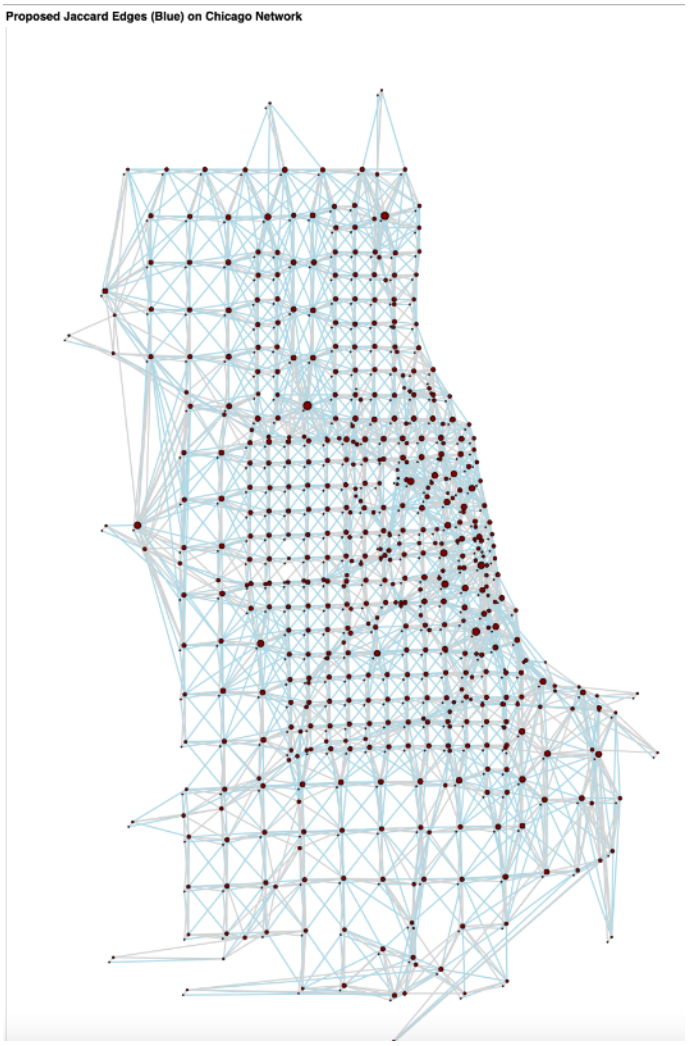

Figure 5: OSM Chicago network

Figure 6: Histogram of simulation and heuristic distributions compared to ground truth

Figure 7: Histogram of simulation and heuristic deltas

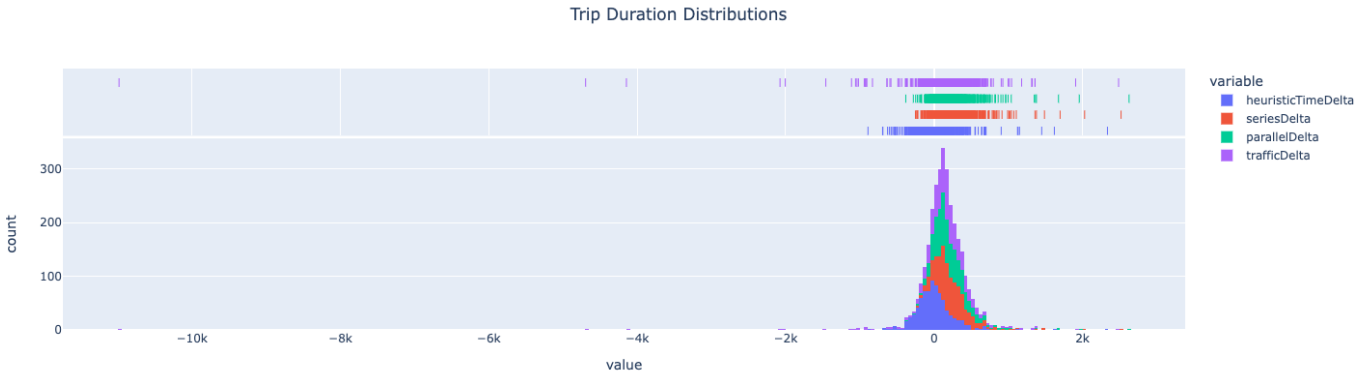

Figure 8: Frequency at which nodes are crossed during taxi trips (heuristic approach) for a small portion of the Chicago road network

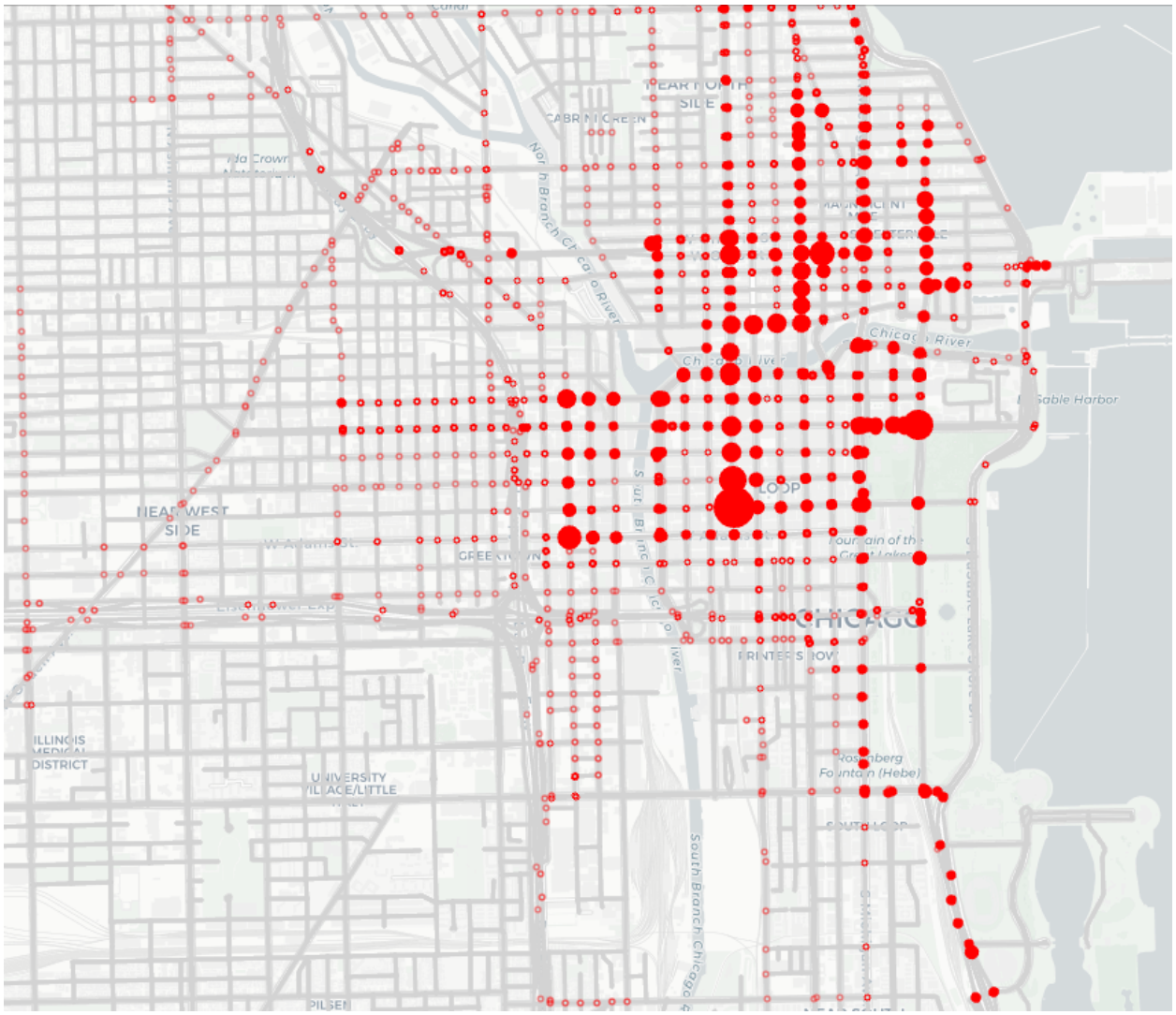

Figure 9: Image of SUMO GUI showing the Chicago OSM road network data

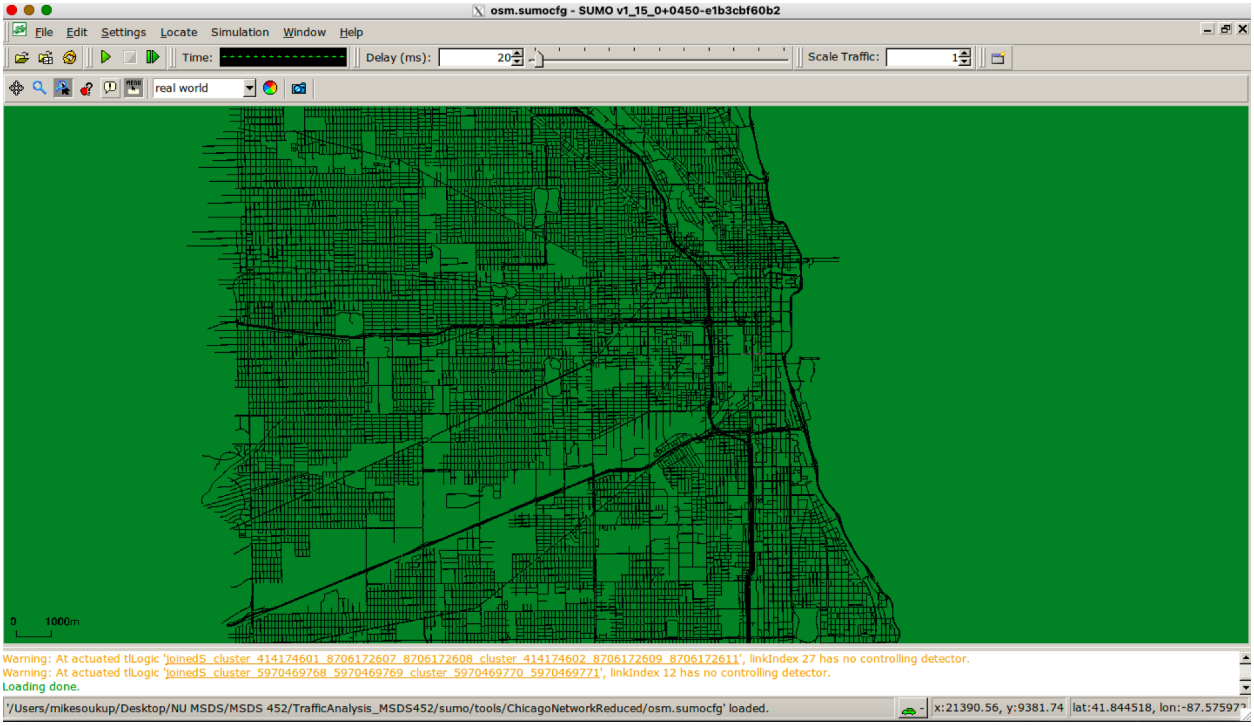

Figure 10: Close up view of SUMO GUI during a simulation run depicting five vehicles stopped at a red light of a four-way intersection.

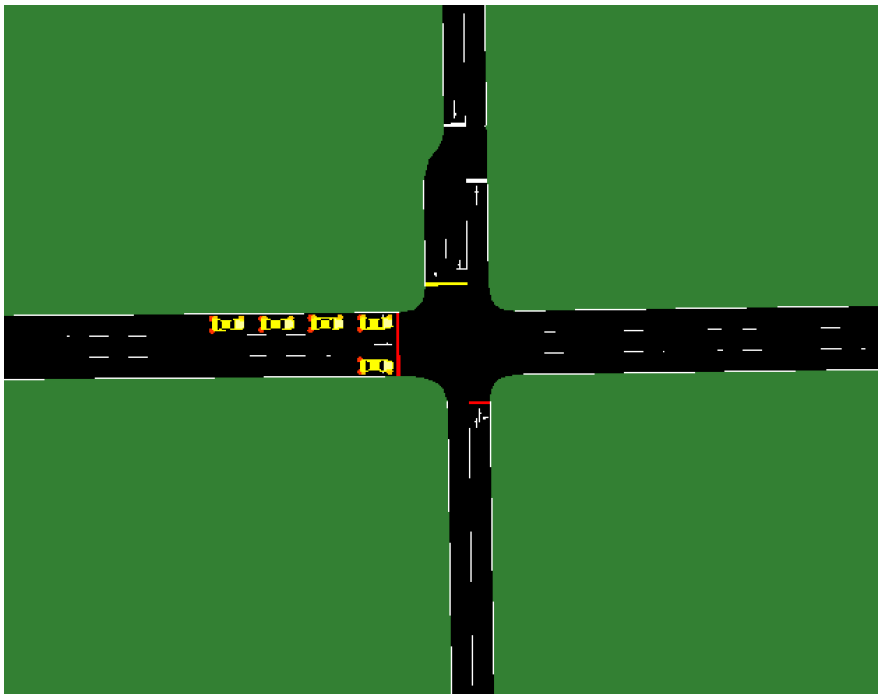

Table 1: Statistical summary of estimated versus actual trips through various methods

	heuristicTimeDelta	seriesDelta	parallelDelta	trafficDelta
count	785.000000	785.000000	785.000000	785.000000
mean	0.027771	235.203822	216.094268	131.504459
std	257.752256	249.680738	246.390792	558.309740
min	-890.400000	-251.000000	-381.000000	-10984.000000
25%	-144.400000	93.000000	63.000000	25.000000
50%	-20.700000	193.000000	174.000000	156.000000
75%	111.300000	336.000000	322.000000	289.000000
max	2336.800000	2520.000000	2628.000000	2487.000000

Table 2: Statistical test results

	Method	Kolmogorov-Smirnov P-Value
0	Heuristic	1.744441e-02
1	Traffic	4.894536e-37
2	Parallel	2.148562e-55
3	Series	1.567308e-68