HEC MONTRÉAL

Robustness of the Washington DC Metro system (WMATA)

Course:
MATH 80627A.H2023 - Complex Networks Analysis

Semester:
Winter 2023

Submitted to:
Gilles Caporossi

Submitted by:
Alexa Canuel
Daniel Tapia
Elliot Wyman

Motivation

Public transportation systems play a crucial role in enabling timely and reliable movement for important demographic groups, both in large cities and densely populated urban areas. Depending on the setting, these systems are comprised of bus, tram (light rail), ferry and metro systems (heavy rail). The metro system is one of the most important transportation systems in large modern cities, specifically because of its ability to reduce urban traffic congestion, and move large volumes of passengers. One of the busiest subway networks in the United States is the Washington Metropolitan Area Transit Authority (WMATA). Based on 2022 data, WMATA served 157 million passengers making it the third most used system in the USA. Understanding its robustness is essential for maintaining a reliable and efficient transportation system, as well as for future work by Washington civic planners. Several techniques exist to gain insight into these systems, including the use of statistical analysis, simulation and complex network analysis. Therefore, in this project, the aim is to determine the vulnerability of the WMATA system as a public transportation system with the goal of moving passengers quickly, the robustness of the Washington D.C. Metro System and whether it exhibits the scale-free property common to some public transportation systems. These questions will be investigated and answered both by conducting complex network analysis and designing and performing traffic simulation experiments.

Data

The Washington Metropolitan Area Transit Authority spans most of the District of Columbia and the Washington Metropolitan area, which are located in the states of Maryland and Virginia. As it stands today, the WMATA Metro Rail has a total system length of 208 km (129 mi), that includes 97 stations across its six lines.

Figure d-1 shows the WMATA map where the stations are denoted as points, the transfer stations in circled points and the lines in their respective colors.

Figure d-1: WMATA metro rail system map

The WMATA website provides the station level information, this consists of names, address, location (latitude and longitude), identification numbers (IDs), and connected lines (Washington Metropolitan Area Transit Authority API, 2023). To be able to use this data for the analysis, it was necessary to perform data cleaning and pre-processing. First, the columns not necessary for analysis, such as street addresses, were dropped. Second, it was necessary to create a modified list where each station was only represented once, the initial data set contained a station ID per station and line color pair.

In order to estimate travel time between stations, the WMATA API was used, this service provides real-time trip information data via JSON (Washington Metropolitan Area Transit Authority API, 2023). A set of queries were run to determine the travel time between all adjacent stations. This would allow for the approximation of travel time, as well as the determination of graph properties, such as diameter. The JSON object returned by the service contains additional information that was not used, and therefore cleaning and preprocessing was required. After removing unused fields, a lookup table was referenced to translate the line-level station IDs to the one-per station ID used in this project. The implication of this decision is that the travel time between two stations is consistent regardless of line. The data gathered from this API tool, as well as the station level data completed the data requirement for the network implementation.

In addition to the network implementation and analysis, there was a data requirement for the simulation experiments. Data World provided information related to the traffic of WMATA during the Women's March on January 21st, 2017 (Data.World, 2017). This file contains information related to the number of people that enter and exit each station at different hours during this day. Considering that this data is from 2017, and that WMATA

opened six new stations in the silver line in 2022, therefore it was necessary to define an amount of people that enter and exit these new stations. These values were generated randomly following a normal distribution, with the mean and variance calculated with the information of the new stations on the silver line.

Implementation

Given the data, it was necessary to create a representation of the WMATA system, as well as create several functions used for the preparation and simulation of experiments. This network representation would then be analyzed to determine properties and vulnerabilities. Based on vulnerabilities and network properties, the network representation could be modified to simulate these events. Using the baseline, complete network representation, wherein no nodes or edges are compromised, and then reduced or compromised representations, experiments are conducted. This led to a specific selection criterion for tooling and implementation.

Python was chosen for data cleaning, pre-processing, as well for its ability to create functions useful for simulations. The network representation was implemented via Python's NetworkX library. The network architecture consisted of nodes, denoting the metro stations, edges denoting connections between said stations, and edge weights denoting the time (in minutes) between stations. The resulting graph can be seen in *figure i-1.* In addition, colors were added as edge attributes to denote the lines of an edge, for example, between station Pentagon and Pentagon City run the lines blue and yellow, as seen in *figure i-2*. The color of the edges represents the line that connects the stations. If the color is black, it means that 2 or more lines connect the two stations. However, when creating the graph, only one edge is used to demonstrate this path. This and other simplifications or modifications to the representation were performed for different motivations.

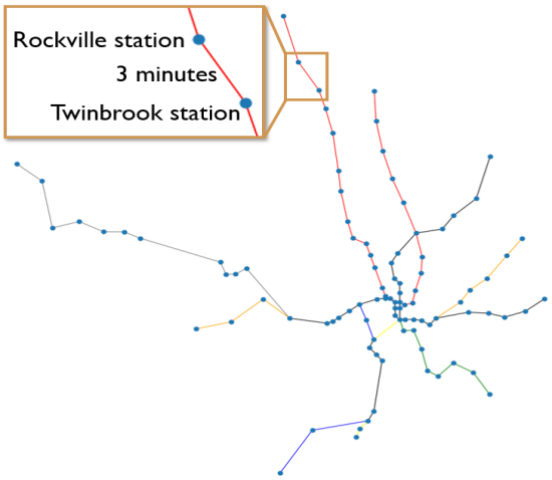

figure i-1: *Graph representation of WMATA system*

***figure i-2**: Metro lines connecting Pentagon and Pentagon City stations*

The use of simple and multidirectional (multi-di) graph representations within NetworkX were leveraged in different scenarios. A few of the relevant differences between these graphs are that simple graphs allow at most one edge between each pair of nodes, while multi-di graphs allow multiple edges, and simple graphs have undirected edges, while multi-di graphs have directed edges. Analysis was able to be conducted fully on a simple network representation. The initial intent was to include edges for each line in the WMATA system, but this was abandoned as will be discussed further in the Experiments section. The use of multi-di network representation was necessary for calculating shortest path and was therefore also created. These complete representations could now be adjusted for the different scenarios considered in the simulations.

Through all simulations, the input data needed was collected or simulated. Given the passenger entry data on the morning of the Women's Day March, simulation was carried out to determine the destination of each passenger. This was done by considering a weighted random walk for each passenger, where nodes with high degree centrality were given a higher chance of being drawn. Intuition for this approach can be made by considering users heading to hub stations, which appear to be downtown in the WMATA network, where the march was most likely to take place. This also mimics the inbound traffic pattern of commuters to city centers each morning, which is likely to be representative in the data given the march was on a Wednesday. This concludes the necessary setup and implementation for analysis and experiments, which will be outlined in the following sections.

Analysis

After having created a graph that well represents the WMATA network, its properties and characteristics are examined to better understand the dynamics of the network and identify vulnerabilities for use in the traffic simulations.

The analysis of the complete network serves as a baseline model for simulations. By examining the complete network, it is possible to gain insight to the graph's fundamental characteristics, which will allow for further understanding of how the metro system behaves. The first clear characteristic of the network is that it has one connected component. This is an obvious attribute for a metro system as it ensures all stations can

be accessed without the need of additional transportation, which maximizes efficiency and convenience for commuters. The second explored characteristic is the density of the network. The density is a measure of how “full” the graph is with edges. It is defined as the ratio of the number of edges a graph has compared to the maximum number of edges possible. The density of this network is 0.021, indicating that the WMATA is relatively sparse with few edges connecting stations. A density of 0.021 means that only 2.1% of stations have direct connections between them. In *Figure a-1* which demonstrates the distribution of the number of adjacent stations to each station, we can clearly notice that around 80% of stations are only connected to two other stations while around 8% of stations are connected to more than two. This type of degree distribution is called power law distribution where only a few nodes have many connections. In the context of the metro system, this distribution makes sense as highly connected stations represent transfer stations which allow for connecting lines. In terms of practicality in a metro system, a low density could make it more difficult for a commuter to travel quickly, however, it makes the system easier to navigate.

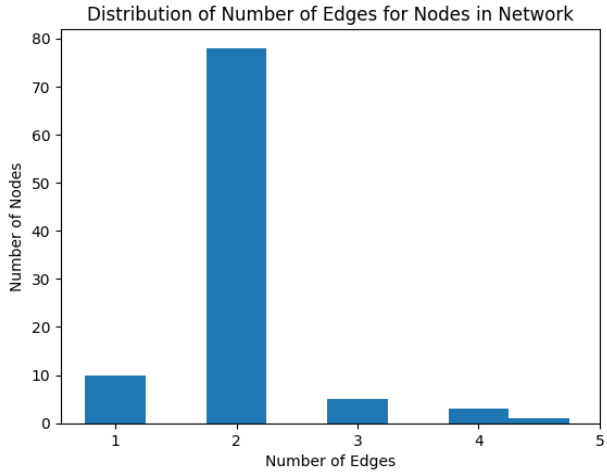

Figure a-1: *Distribution of the number of edges for nodes in the network*

The diameter of the network, which represents the length of the shortest path between the most distanced stations, is 34 minutes. In practical terms this means that, in the given representation, the longest possible journey between two stations is 34 minutes, assuming that the journey takes place entirely within the metro system and without considering the time waiting for the metro to arrive, or time spent at stations.

Aside from exploring the graph's fundamental characteristics, analysis must be performed to identify vulnerabilities. In the case of the experiments, vulnerabilities will be leveraged to identify targeted attack points. A targeted attack is a deliberate action aimed at a specific node with the intention of disrupting the system. In the context of the WMATA system, a targeted attack would disrupt one or more stations, or lines. In contrast to

targeted attacks are what are known as random failures. A random failure could be looked at as the loss of a node due to chance as a non-routine outage or accident occurring at a station or metro-line level. Random failure should ignore any bias identified by analysis of the network; however, targeted attacks will likely be very effective when found via complex networks analysis.

Analyzing the Fidler value and vectors, as well as the centrality measures, will help us identify the most important nodes in the network. This information will be used to perform targeted attacks in the simulations to observe how the system would be affected. This can be performed independently of data known about the system, such as station or line popularity traffic.

The Fiedler value is a key parameter in graph theory that helps characterize the connectivity of the graph. In the complete graph, the Fidler value was measured to be 0.016, indicating that the graph is not well-separated into clusters. *Figure a-1* demonstrates the use of the fiddler vector to divide the graph into 2 subgraphs and it is obvious that the separated graph does not demonstrate distinct clusters or provide any useful insight. Since the Fidler value did not yield a consistent result in identifying targeted attack points, centrality measures will be investigated to determine targeted attack points.

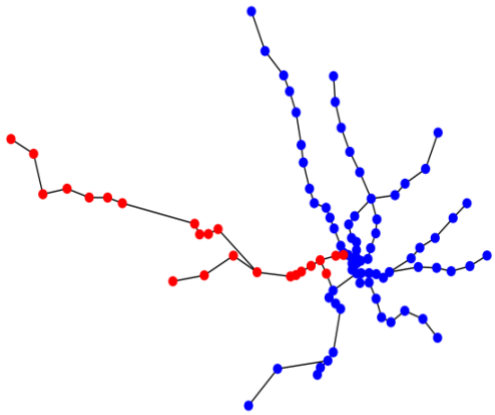

Figure a-2: *Network divided into two subgroups*

Centrality measures are used to measure the importance of a node in a network. For this study, degree centrality, eigenvector centrality, betweenness centrality and closeness centrality are evaluated. Degree centrality measures the number of edges indecent to a node, eigenvector centrality measures a nodes importance based on the importance of its adjacent nodes, betweenness centrality measures the number of shortest paths that pass through the node and closeness centrality measures the average distance between a node and all other nodes in the network. Through the analysis, it was found that the most central node in the WMATA based on all four centrality measures is *L'Enfant Plaza*.

This suggests that this station is a key station and hub in the system with many connections to other stations and lines, therefore a perfect targeted attack point.

In summary, the analysis of the complete network provides us with a better understanding of the network and the first attack point, which is the station L'Enfant Plaza.

Experiments

The analysis of the network structure underlying the WMATA system has provided us with baseline measures for centrality, which can be used to determine targeted attack points as well as for comparison with graphs modified through failures and attacks. Simulations can now be conducted, starting with the baseline network.

The experiment design is to route the full passenger list through the network under various scenarios. These scenarios include the system in a full working state, i.e. with no outage stations or lines, and scenarios where nodes or lines are experiencing outages. From an analysis point of view, outages occur at the node or station level, the associated graph will have this node removed, and centrality and other measures can be recalculated. When performing simulation, outages occur at the edge or line level, and lines connected to the chosen station are penalized by a factor, in all experiments of 10x. To illustrate this example, let's assume the travel time between Pentagon and Pentagon City is 4 minutes, then during an outage at Pentagon station, the travel time between these stations is increased to 4 times 10 which is equal to 40 minutes. This was chosen for a few reasons. Firstly, it allows for the passenger list to not be rerouted, i.e. a passenger will still enter the network at Pentagon even during an outage. Secondly, it allows for passengers to continue their commute, even if stations are on branches of the network. Intuitively here, it can be considered that passengers walk to the next station when an outage occurs, hence the magnitude of the factor chosen. It is now possible to fully perform simulations through single and successive random failures or targeted attack scenarios, as detailed in the following scenarios.

Scenario 1-5: single random failure

Scenarios 1 through 5 observed the traffic simulation during a single random failure. These scenarios were replicated five times in order to obtain a more accurate representation of a random failure, attempting to prevent the scenario in which critical nodes or more redundant nodes were chosen only.

Scenario 6-11: successive random failures

Scenario 6 through 11 used multiple, two and three, successive random failures. In order to accomplish this, the metrics for single random failures were averaged, and the node with resulting impact most closely resembling the average was selected. Successive random attacks would then include this node, and one more randomly selected node. The same process was performed for successive attacks of three nodes.

Scenario 12: single targeted attack

Scenario 12 included a single targeted attack. L'Enfant Plaza was selected as it is the most central node. After this targeted attack, three new nodes were identified by the analysis of centrality measures. These three nodes corresponded to the stations Fort Totten, Gallery Place-Chinatown, Metro Center, which are all transfer stations themselves.

Scenario 13-16: successive targeted attacks

Scenarios 13 through 16 used multiple, two and three successive targeted attacks. Again, the node pair most closely resembling the average for targeted attack pairs was chosen for the successive attack scenario with 3 nodes. For example, during successive targeted attacks of two nodes, L'Enfant Plaza and one of the three new central nodes were attacked one at a time with L'Enfant Plaza, in order to observe the effect on the network.

Overall, 16 experiments were designed and simulated, with total time, total stops, and average shortest path length captured after each scenario. These results will now be explored to determine the system vulnerability, robustness, and scale-free property of the network.

Results

Through the results of the simulated experiments, various metrics were obtained. As mentioned previously, the complete (without outage) network will act as the baseline for comparison of results obtained through the experiments. There are three questions our study seeks to answer.

1. Is the system vulnerable to outages?
2. Is the underlying network robust?
3. Does the network exhibit a scale-free property?

To answer these questions, the analysis of three measures is used; average trip time per passenger in minutes, average shortest path length, and average number of stops per passenger.

Table 1 presents the results of simulations conducted on the WMATA network. The first row represents the baseline measurements of the network, including the average trip time per passenger, the average shortest path length, and the average number of stops per passenger. The subsequent rows represent the average measurements after single and multiple random and targeted attacks on the network.

Simulation	Average Trip Time per Passenger (minutes)	% change	Average Number of Stops per Passenger	Average Shortest Path Length (minutes)	absolute change (minutes)
Baseline	26.3	-	10.9	12.49	-
Average single ***random failures***	35.1	+33.8%	10.9	41..84	+29.35
Single ***targeted attack***	38.9	+48.0%	11.8	46.4	+33.91
Average successive (2) ***random failures***	38.7	+47.4%	10.9	44.6	+32.11
Average successive (2) ***targeted attacks***	51.3	+95.1%	11.7	54.4	+46.14
Successive (3) ***random failures***	44.0	+67.4%	10.9	49.8	+37.31
Successive (3) ***targeted attacks***	64.1	+144.2%	11.5	71.6	+59.11

Table 1: *Statistics on Simulation*

Overall, the results show that random and targeted attacks increase the average trip time per passenger and average shortest path length. The average number of stops per passenger is constant through random attacks but increases after targeted attacks. This

is a result of decreasing a node's connectivity and therefore forcing passengers to have to take longer routes to reach their destination.

Regarding the vulnerability of the transportation system, the results show that trip times increase under both outage types, though the increase is higher for targeted attacks than it is for random failures. It can also be seen that random failures did not increase the average number of stops, i.e., did not force passengers to reroute, however, this could be due to simulation approximations. Overall, the network does not appear to be resilient to either type of outage or does appear to be vulnerable.

The results illustrate that the average shortest path length increases more under targeted attacks than random failures, however it increases significantly in both cases. This forms the conclusion that the network is not robust.

The analysis of the degree distribution of the WMATA network showed that it does not follow a pure power-law distribution. However, it exhibits a heavy-tailed degree distribution, which is a characteristic of scale-free networks. The experiments conducted on the network further refute the notion that it has a scale-free characteristic as it was not significantly less susceptible to random failure than it was to targeted attack.

In conclusion, the simulations and analysis suggest that the WMATA network is a vulnerable transportation network and does not clearly exhibit robustness or scale-free properties. These findings can help transportation planners and engineers to improve the design and management of subway networks.

Further exploration

Through this set of experiments, various techniques were used to both approximate and analyze the model. There are a variety of potential improvements that could be made to make the experiments more data-driven, more realistic of a public transportation system, or to draw further information from analysis. Firstly, additional data is available and could be leveraged, for example, it is possible to calculate the stop time at stations, by querying the API for time between two stops, as was done during the data collection stage, and between successive stops, to find the stop time at a station. For example, if the time between station A and B is known to be 5 minutes, as is the time between B and C, but the time between A and C is 12 minutes, then the stop time at station B is 2 minutes. Secondly, additional complex network analysis techniques could be employed, such as modularity. This measure quantifies the degree to which a network can be partitioned, in our example, into regions. Networks with high modularity have a clear division between different groups of nodes, while networks with low modularity are more homogenous.

Thirdly, the simulation of passenger itineraries was done using degree centrality; it could also be weighted by the distribution of exits by each station, to use a more data-centric approach. Finally, the experiments could also incorporate multiple edges per station, denoting metro lines, with transfer times between lines, allowing for lines to be down independently of stations. Lastly, the design of this study could be replicated on other public transportation networks, including those for buses, trains and automobiles.

References

[1] Washington Metropolitan Area Transit Authority API. (2023). JSON - Station Information. https://developer.wmata.com/docs/services/5476364f031f590f38092507/operations/5476364f031f5909e4fe3310?

[2] Washington Metropolitan Area Transit Authority API. (2023). JSON - Station to Station Information. https://developer.wmata.com/docs/services/5476364f031f590f38092507/operations/5476364f031f5909e4fe3313

[3] Data.World. (2017). WMATA Ridership Data. https://data.world/transportation/wmata-ridership-data

[4] Michael Fensterheim. (2023). Transit in Washington, D.C. - A brief look at transportation in the nation's capital. https://storymaps.arcgis.com/stories/e998d9fa9ae3497998cb537a02230819