Network Analysis on Four Different City Subway

Bingwei Wang
Electrical and Computer Engineering
Carnegie Mellon University
San Jose, California
bingweiw@andrew.cmu.edu

Shi Bai
Electrical and Computer Engineering
Carnegie Mellon University
San Jose, California
shibai@andrew.cmu.edu

Dezhi Yu
Electrical and Computer Engineering
Carnegie Mellon University
San Jose, California
dezhiy@andrew.cmu.edu

***Abstract*—The paper focuses on the network analysis of Four Different subways. The preliminary goal is to find out what affects the reliability and robustness of a subway station. This paper will conduct network topology characteristic research and functional characteristic analysis on the subway networks of four cities around the world, New York, Chicago, London, and Shanghai, to evaluate their reliability and robustness. We use Five parameters to measure the topological properties while discussing the proportion of removed nodes in the subway network of four different cities respectively, and comparing with the random network we get the critical threshold of this proportion. This paper concludes that the subway network is robust to random attacks but vulnerable to malicious attacks.**

***Index Terms*—Subway, Network Analysis, New York City, Chicago, London, Shanghai, Robustness, Graph Theory**

I. Introduction and motivation

In modern society, city subways serve as essential infrastructures for the public transit system to mitigate traffic pressure and provide connections among different living areas in the city, especially in metropolises. Subways are composed of complex networks of stations and routes. When subways are designed, many requirements and questions should be considered. For example, the subway network should serve as many people as possible, be fuel-efficient and time-efficient, and cost less construction material. Choosing the location of each subway station should be seriously considered, with considering many factors around the location such as the number of residents, number of event centers and office buildings, population flow and etc. By analyzing existing subway networks, we aim to reach the following goals: 1. We aim to know about what affects choosing the location of a subway station, such as population distribution near the location or other factors. 2. We want to find the factors of deciding a successful subway network, such as efficiency and robustness. 3. Based on the given data sets, we are able to design an efficient subway network for a city. 4. How to design a robust subway network to prevent station crash problems due to physical reasons or population overflow.

The motivation is that if we could reach these goals, we are able to discard the traditional costly, and complex process of designing a subway network. This could reduce a huge amount of costs for the city's public transit department.

II. Previous work

Previous studies have focused on the relationship between the city population and the stations[1, 2]. If the city has more population, more subway lines and stations tend to be established to mitigate the traffic of the city. Also, previous studies have utilized complex topological theory and network models to analyze passenger flow and transportation systems[3, 4]. However, this research can be extended to be more detailed to find the relationship between the number of stations & lines and the community population around the stations. Additionally, to design a successful network, robustness is also a significant factor to consider. Robustness is a measurement to decide whether a network can continue performing well when facing failures or attacks[7]. We will use four different city subway networks, to draw a conclusion on the subway-population relation.

In the article[3], the authors gave the example methods to analyze the robustness through *efficiency*, *betweenness centrality*, *network size*, *largest connected cluster*, and *functionality loss*. The paper showed random and intentional attacks on the graph and the impacts they have[3]. We will use the same analyzing approaches to calculate these measurements on four different city subways to get general results. Our analyzed networks are weighted graphs instead of the paper's unweighted graphs. The weights are described in **Approach** section. We are willing to see any differences from the paper results.

III. Approach

In this section, two milestone approaches are independently explained and introduced. Before going into details, the following assumptions were made:

1. Any route between two stations is undirected.
2. Assume the data of the subway network data set are up to date.
3. Assume the data of the population distribution data set generated from the 2020 American Community Survey are up to date.

Milestone 1

In milestone 1, we focus on analyzing the correlation between network betweenness centrality and resident size based on postal codes. New York City's subway was analyzed.
Based on the above assumptions, We start collecting data.

I. Data Collection:
We downloaded the subway map data of New York City from kaggle[5]. In this way, we get the station data of the subway.

It is easy to get the data for each subway line from the data set and extract the starting and ending stations of each station for each city. Each subway station is a node in the network graph. Then we save the (start, end) node pairs to a CSV file.

II. Generate Network:

Begin to build the networks in **Gephi**. We import the CSV file from step 1 into the **Gephi** software. **Gephi** generates a complete network graph (Fig.2). Finally, the betweenness of this network graph is calculated.

III. Population Distribution:

We download the population distribution data set of New York City[6]. This data set contains the number of permanent residents and the zip code. Combined with the zip code in the subway data in step 1, we can associate the zip code with the number of permanent residents.

IV. Analyze Relation:

Using the betweenness of step 2 and the data of step 3 to analyze the relationship between the betweenness of the subway network and population distribution in New York.

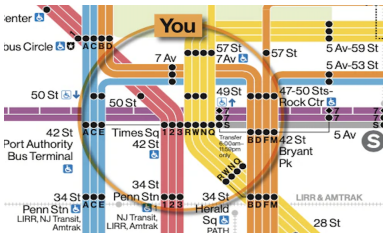


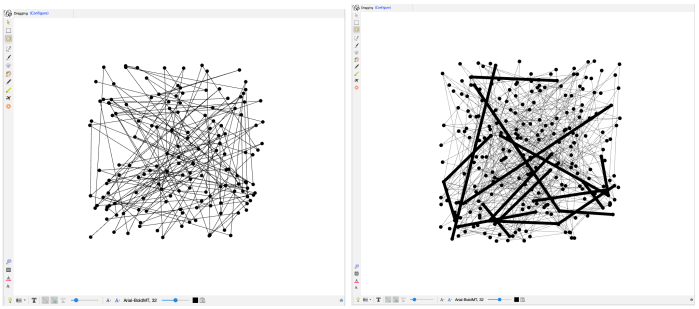
Fig. 1: New York City Network

For example, in the picture (Fig. 1), Time Square 42 street, in the middle of the picture, we define this station as a node. For the same reason, 42 Street Bryant pk and 34 Street Penn Stn are both a node. Lines 1, 2, and 3 runs from Times Square 42 Street to 34 Street Penn Stn. So there is an edge between two nodes, and the weight of the edge is 3. There are 7 and S lines from Times Square 42 Street to 42 Street Bryant Pk. But there are 3 ways connected. So there is an edge between two nodes, and the weight of the edge is 3.

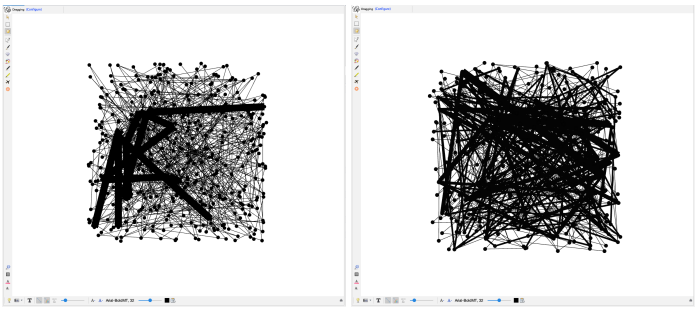
We generate the graph $G = \langle V, E \rangle$, the node set $V = \{v_1, v_2, v_3, ..., v_N\}$, the edge set $E = \{\{v_1, v_2\}, \{v_3, v_4\}, ...\}$. We consider each subway station as one node and the route between two adjacent stations as one edge. If one subway station contains different entrances and exits far from each other, this is also only one node. The graph is undirected because the route between the two stations most are the same in the world and the directed graph has too little impact on our topic.

Milestone 2

For milestone 2, we analyze and built 4 network graphs. Our main goal is to analyze the robustness of a network under different network attack protocols. Through the network analysis process, we found out that the robustness of a subway network is very important. For example, physical accidents or abrupt increasing ridership should not affect the overall abilities of a network. We built networks graph for 4 cites, which



(a) Chicago Network (b) London Network



(c) Shanghai Network (d) NYC Network

Fig. 2: Four Different City Network

are Shanghai(China), London(England), Chicago(USA), New York(USA). We will first introduce the sources and approaches to download 4 subway data-sets. Each city's is included in a subsection. Since the process of converting raw data sets into network graph is similar for 4 cities and the process is also explained in the above milestone 1 approaches, we will mainly cover how to collect data sets for each city.

1) Shanghai:

Shanghai's subway data set is from the Shanghai Metro website. This is Shanghai Metropedia, an encyclopedia of information about every line and every station on the Shanghai Metro. The Shanghai Metro system consists of 20 lines and 408 stations. We want to include and analyze Shanghai's subway because its lines are interactive. The subway network topology map is relatively complex. Analyzing Shanghai's subway network could provide us with some characteristics of large and complex subway maps. We obtain the information of each node and edge information. The weighted wireless graph is then constructed using the NetworkX Python package. The weight of an edge in the graph refers to the number of routes passing between two nodes.

2) London:

London underground data set is from Transportation For London website. We used the JSON file from the website. Then, we used the python Networkx package to generate the multi-graph by parsing the JSON data. London's underground has 10 lines, 259 stations, and 340 routes. London underground is one of the oldest subway systems in the world. The subway has some old lines that may not adapt to today's complex transportation system, which is different from other cities.

Hence, we can have a more convincing result from subway networks.

3) Chicago:
Chicago's subway data set was collected from the Chicago data portal's website. We downloaded a geojson file containing Chicago subway's station and route information. Chicago subway's data set contains 7 lines, 153 stations, and 153 routes. We want to include and analyze Chicago's subway because its lines barely interact with other lines. This is unlike New York or Shanghai's subway networks, where each line could cross over other two or three lines, which makes the overall graph more connected and complicated. Analyzing Chicago's subway network could provide us with more aspects of different kinds of network graphs. From that geojson file, we were able to parse that file and create nodes/edges based on its stations/routes.

After obtaining the data sets, we see each station as a node and each route as an edge. Then we built the network based on these data. The result is shown in Fig. 2.

After building these networks, our main tasks for milestone 2 become analyzing the robustness of a network under 3 different network attack protocols. 3 protocols are random attack protocol, highest betweenness centrality attack protocol, and highest degree attack. Random attack protocol randomly selects some number of nodes and deletes them from the graph. For example, if there are n nodes in the original graph. If p percent of the network is attacked, p / 100 * n nodes are randomly selected from the graph and deleted. For the highest betweenness centrality attack protocol, if p percents nodes are attacked, we remove p / 100 * n nodes from the graph, where they have the highest betweenness centrality after sorting the original network graph. For the highest-degree attack protocol, we delete nodes with highest degree by sorting the original network graph.

To analyze the robustness of the network. We used 6 network properties to find the results. They are normalized network efficiency, normalized average betweenness centrality of nodes, normalized average betweenness centrality of edges, normalized network size, LCC, and FLN. The definition and meaning of each network property are given in the next section. We want to analyze these network properties because they all can represent some aspects of the network's robustness.

The main research question for this milestone is to find how different city subways perform their robustness under different network attack protocols.

We divide the scalability of our approach into two parts. The first part is looking for data sets of city subways. Our method is to look for the official websites of these city subways and download the JSON files. This should be done manually, which is not scalable. The second part is that after we get the JSON files, we use a Python script to build network graphs based on JSON files. The process is efficient and scalable since most network data sets have the same format so we can use only one script to build the network.

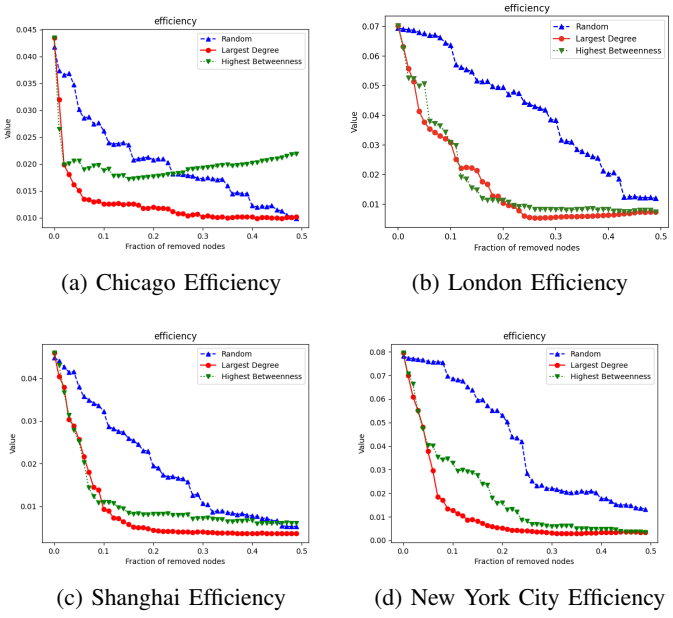
IV. Preliminary Results

A. Network efficiency

The network efficiency is given as follows:

$$E = \frac{2}{N(N-1)} \sum \frac{1}{d_{ij}} \tag{1}$$

N is the number of nodes. The shortest path between two nodes v_i and v_j is defined as the minimum number d_{ij} of links necessary to go from node v_i to node v_j. The sum is over all C_n^2 pairs of nodes. If it has a smaller shortest path, the network is more efficient. The efficiency of a network is a measure of how efficiently it exchanges information and it is also called communication efficiency. In this paper, we want to know the relationship between the efficiency of the subway and the random attack.



(a) Chicago Efficiency (b) London Efficiency

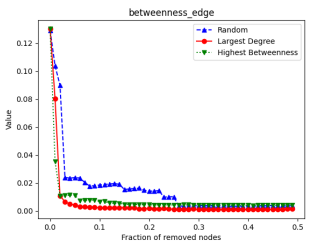
(c) Shanghai Efficiency (d) New York City Efficiency

Fig. 3: Network Efficiency

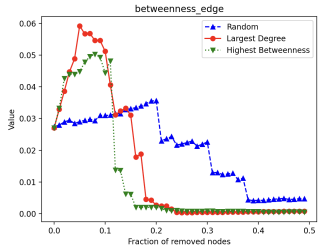
Fig. 1 shows that 3 different attack modes will lead to different damage in 4 cities. The random attack will result in the minimum damage among 3 different attack modes than any other attack. While the attacks mode of highest betweenness causes the maximum damage to the network. In this situation, the network efficiency is the smallest. Attacking the largest degree nodes impact is moderate. The connectivity of the subway is susceptible to the largest degree of node-based attacks and the highest betweenness of node-based attacks, while the attack of randomness has little impact. Because a targeted attack will lead to the destruction of a station with strong connectivity, and the shortest path connecting it will disappear, which will have a huge impact on the efficiency of the subway. In this scenario, it can be concluded that the subway network is robust against random attacks.

B. Average Betweenness Centrality

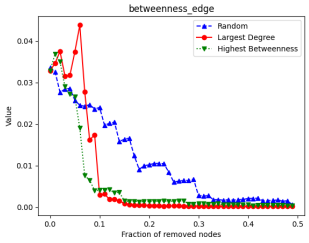
$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2}$$



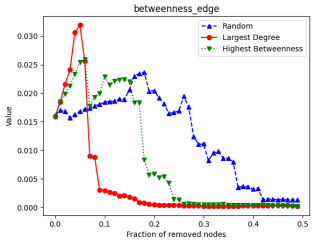
(a) Chicago Betweenness Edge



(b) London Betweenness Edge

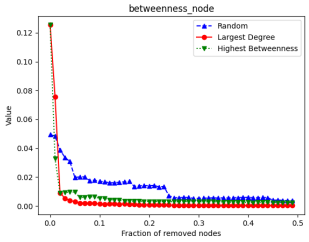


(c) Shanghai Betweenness Edge

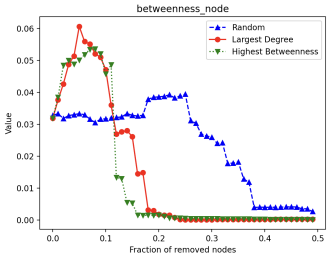


(d) NYC Betweenness Edge

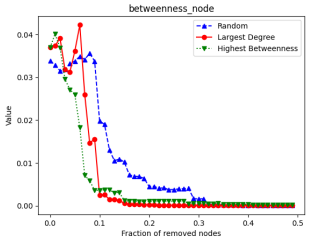
Fig. 4: Network Betweenness Edge



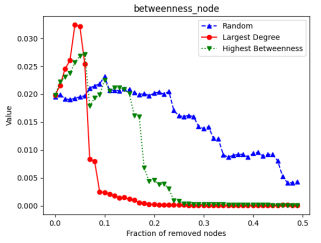
(a) Chicago Betweenness Node



(b) London Betweenness Node



(c) Shanghai Betweenness Node



(d) NYC Betweenness Node

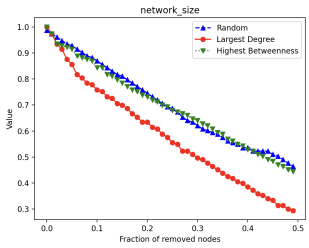
Fig. 5: Network Betweenness Node

Figure 4 shows the average betweenness centrality of edges under 3 attack protocols. Figure 5 shows the average betweenness centrality of nodes under 3 attack protocols. Betweenness centrality is a measure of the centrality of the network. In subway networks, higher average betweenness centrality usually means that the network is larger and connected.

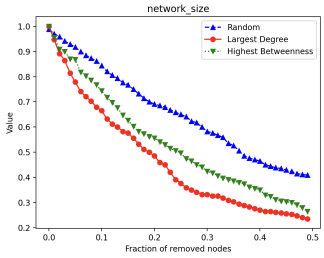
For all 4 cities, their average betweenness centrality is decreased after attack protocols most of the time for both avg betweenness centrality of nodes and edges, which could be caused by fewer connected points and a smaller number of nodes. Another finding is that the blue line is above the red and green line almost all the time. This implies that the random attack protocol has smaller impacts on the graph's betweenness centrality compared to the other two protocols. It is obvious that the highest betweenness centrality attack protocol always has a strong impact because the protocol aims to decrease the graph's centrality. We can conclude that the network is more robust due to the random attack protocol but is fragile due to the other two protocols.

C. Network size

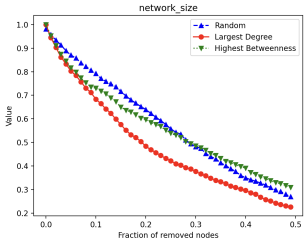
$$A_{normalized} = \frac{N}{N_0} \tag{3}$$



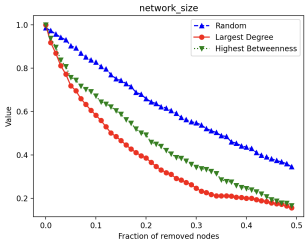
(a) Chicago Network Size



(b) London Network Size



(c) Shanghai Network Size



(d) NYC Network Size

Fig. 6: Network Size

Figure 6 shows the change in network size due to 3 different attack strategies. The network size here is defined as the number of edges in the graph. Under different scenarios, high network size represents that the graph has a larger connected component and more edges, which also means that the network is more connected and robust according to network attacks. For subway networks, if the subway has more edges, stations are connected to each other and the subway is robust.

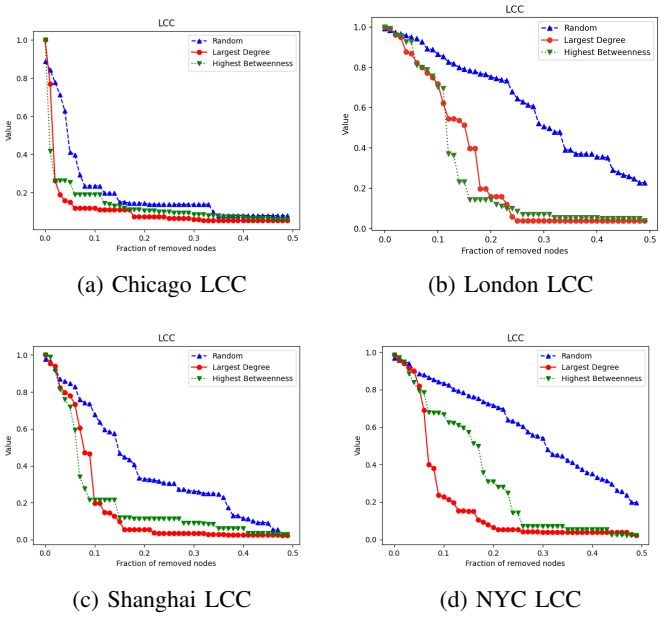
For all 4 cities, we can tell that the blue line is above the red and green lines. The huge differences mean that random attack protocol has smaller impacts on the network size of the graph. Removing the same amount of nodes, the largest degree protocol and the highest betweenness protocol remove more edges. This could imply that these two attack protocols have large impacts, which could make the graph more broken

and less connected. The reason behind it could be that nodes with larger degrees and higher betweenness centrality are more responsible to make the graph connected and have larger degrees (more edges). Therefore, removing these nodes makes the graph broken. For subway networks, if the network has a smaller network size, it is less possible for a station to reach another station, which also means that the network is not robust due to the largest degree protocol and highest betweenness protocol.

D. Largest Connected Cluster

$$LCC = \frac{S}{S_0} \tag{4}$$

S is the number of nodes on the largest connected subnetwork after attacks and S_0 is the number of nodes on the largest connected graph of the initial network before attacks. When the subway network is attacked randomly or deliberately, a small number of nodes and their connections will be deleted, and the integrity of the network may be compromised. The largest Connected Cluster reflects the severity of connectivity damage after the subway was attacked.



(a) Chicago LCC (b) London LCC

(c) Shanghai LCC (d) NYC LCC

Fig. 7: Network Largest Connected Cluster

From Fig. 7, the random attack has the least impact on Largest Connected Cluster. The blue one is the curve of a random attack, and the blue curve is basically higher than the other two attack methods. This shows that the random attack method has the least impact on the largest Connected Cluster of the subway. Corresponding to the attack methods of largest degree and highest betweenness, different subways have different states after receiving the attack. Attacks on the largest degree or highest betweenness nodes would destroy more subway routes with a high probability. After these important nodes are deleted, it will make the Connected Cluster split and become smaller, and even cause some nodes to be completely disconnected. From Fig. 7, based on the largest degree or highest betweenness nodes attack mode, after removing 20% - 30% of the nodes, the Largest Connected Cluster basically tends to the horizontal line. In summary, the Largest Connected Cluster property of the subway is robust to random attacks.

E. Functionality Loss of the Network

In this section, we will do an experiment about the characteristics of the subway functions using the interdependence analysis of the subway network. We use a parameter defined in the paper[3] which is called the functionality loss of the subway network. Supposing that initial functionalities of the nodes equal 1.

$F_0(v_h)$ = 1, h = 1, 2, . . . , N. For each attack, the degree of the node and the shortest path length between any two nodes can be used to measure the functionality loss when one node is attacked. When the node v_i is attacked, and this attack is supposed as the m th attack, the functionality loss of the node v_j(j $\neq$ i) is defined as follows:

$$FL(v_j) = \sum_{l=1}^{m} F_{l-1}(v_j) - F_l(v_j) \tag{5}$$

$F_l(v_j)$ is the transport functionality of node v_j after the node vi has been deleted. The iteration on the functionality of node v_j is defined as:

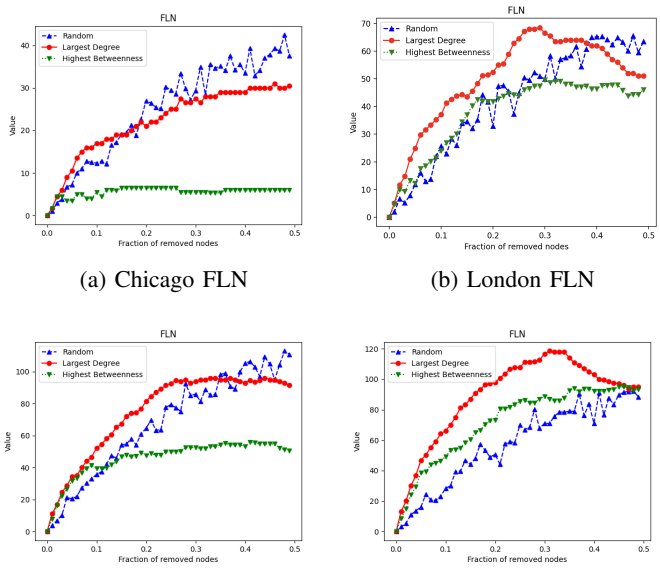
$$F_l(v_j) = F_{l-1}(v_j) - \frac{1}{d_{ij}^2 k_j} F_{l-1}(v_j) \tag{6}$$

The meaning of d_{ij} is the shortest path length between node v_i and v_j after the (m-1) th attack. k_i is the degree of the node v_i after the (m-1) th attack. When node v_i is deleted, the functionality loss of the network is calculated as follows:

$$FLN = \sum_{j=1, j\neq i}^{N} FL(v_j) \tag{7}$$

From Figure 7, we can see that in less complicated subway networks like Chicago, the random attack makes the most loss compared to Highest Betweenness or Largest Degree attack. In more complicated subway networks like New York City, the random attack makes the least loss compared to the intentional attacks. Moreover, attacks of Highest Betweenness will usually have more functionality losses than attacks on Large Degree.

Significance: By building a workflow to find subway data sets, build network graphs, and analyze efficiency, average betweenness centrality, network size, LCC, and FLN, we easily convert the complex process of robustness analysis into an easy process. With all the approaches and result analysis above, people easily find the robustness of a city subway network under different network attack protocols. If more people and more cities can apply our paper to other



(a) Chicago FLN (b) London FLN

(c) Shanghai FLN (d) NYC FLN

Fig. 8: Network Functionality Loss of the Network

cities, they can efficiently improve their subway networks.

Limitations: One limitation we noticed is that our 3 attack protocols maybe not be enough. In the real world, these attack protocols maybe not be realistic. For example, in the real world, if there is a big event near a subway station, that station could be paralyzed due to population traffic. This kind of attack is not based on the random, highest degree, or highest betweenness centrality. It is based on population size or population flow near the station. Another limitation is that in this milestone, we focus our work on big cities and international cities. The result may not represent small cities.

V. Conclusion

Milestone 1

By analyzing the modularity of the New York City subway network, we found that its robustness is overall okay. Through generating the network graph and comparing its betweenness centrality and population size, we found out that betweenness centrality is not necessarily correlated to population size, which also means that the population size of a position may not affect the site selection of a subway station.

Milestone 2

In milestone 2, we analyzed the robustness of 3 subway networks. By using network efficiency, average betweenness centrality, network size, LCC, and FLN, we are able to generate some results about the robustness of these networks. A common finding is that all 4 subway networks are more robust under random attack protocol and are more fragile under malicious attack protocols (largest degree attack and highest betweenness centrality attack). Throughout this finding, these subways should improve their robustness under these attack protocols. Since our network studies the weighted graphs instead of the unweighted graphs presented in the paper[3], our results on FLN and betweenness centrality are different. However, the overall results of random and intentional attacks are similar. According to Graph Theory and network analysis, all three attack protocols can change the performance of the subway networks, and malicious attacks can make more impact than random attacks.

VI. Short-Term Plans

In milestone 3, we would like to continue to work on network robustness analysis. We may want to introduce more network attack protocols based on population resident size or population flow. Additionally, we may build more subway networks based on more cities. Last but not least, we would like to find factors of successful subway networks. For example, we would like to find out if there is a threshold for betweenness centrality or other properties, which can represent the success of a network.

VII. Member Contribution

In milestone 1, all 3 members worked on the whole process together, which included data collecting, graph building, network analysis, and presentation preparation. In milestone 2, Dezhi Yu collected Shanghai's subway data set and built the network. Bingwei Wang collected London's subway data set and built the network. Shi Bai collected Chicago's subway data set and built the network. Bingwei Wang implemented the methods to generate a network efficiency graph and normalized average betweenness. Shi Bai and Dezhi Yu implemented the methods to generate network size graphs, LCC graphs, and FLN graphs. All 3 members worked on the paper together.

References

[1] Jeyapragasan, K., & Maniyar, G. K. Y. (2019). An Analysis of Subway Networks using Graph Theory and Graph Generation with GraphRNN.

[2] Feng, J., Li, X., Mao, B., Xu, Q., & Bai, Y. (2017). Weighted complex network analysis of the Beijing subway system: Train and passenger flows. Physica A: Statistical Mechanics and its Applications, 474, 213–223. doi:10.1016/j.physa.2017.01.085

[3] Zhang, Jianhua et al. "Networked Analysis of the Shanghai Subway Network, in China." Physica A 390.23-24 (2011): 4562–4570. Web.

[4] Derrible, S., & Kennedy, C. (2009). Network Analysis of World Subway Systems Using Updated Graph Theory. Transportation Research Record, 2112(1), 17–25. https://doi.org/10.3141/2112-03

[5] Text maps for subway lines. MTA. (n.d.). Retrieved October 13, 2022, from https://new.mta.info/maps/subway-line-maps

[6] New York ZIP codes by population. New York Outline. (n.d.). Retrieved October 13, 2022, from https://www.newyork-demographics.com/zip_codes_by_population

[7] W. Ellens and R. E. Kooij, "Graph measures and network robustness," arXiv.org, 07-Nov-2013. [Online]. Available: https://arxiv.org/abs/1311.5064v1. [Accessed: 16-Nov-2022].