Network Analysis on New York City Subway

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

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

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

***Abstract*—The paper focuses on the network analysis of the New York City subway. The preliminary goal is to find out what affects the site selection of a subway station. It gives approaches to generating the subway network graph based on existing data sets and explores the relationship between graph properties and resident population distribution. The paper concludes that for New York City subway network, betweenness centrality is not correlated to population size.**

***Index Terms*—Subway, Network Analysis, New York.**

I. Introduction and motivation

In modern society, city subways serve as essential infrastructure 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.

This paper is organized as follows. In section 2, some previous approaches and formulas to help us analyze the data are presented. In section 3, we will first list the main approaches to reach the goal of Milestone 1. Then the main approaches to reaching all goals are also listed after that. In section 4, we present our preliminary results with tables and figures.

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. We will first concentrate on the NYC subway networks, and then compare the result to other major cities all around the world, to draw a conclusion on how to choose the location and what matters efficiency.

To design a robust subway network, we need to consider the network's properties, including the betweenness, closeness, and degree. In the paper on the analysis of the Shanghai subway network [3], the authors introduced the theory of functionality loss and connectivity to determine the robustness of the network.

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

$$C(i) = \frac{1}{N_i(N_i - 1)} \sum_{j \in \Gamma_{ij}, j=1}^{m} N_{ij}(N_{ij} - 1) \tag{2}$$

We will use the formula to determine how to make the subway network robust. Additionally, we will also use the modularity of the graph to determine the robustness, since the communities detected can decide the strength of the network.

III. Approach

For milestone 1, before researching this topic, the following assumptions were made:

1. Any route between two stations is undirected.

2. Assume the data of the subway network data set posted on March 05 2021 on MTA's website 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.

Based on the above assumptions, We start collecting data.

A. setup 1

We download 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.

B. setup 2

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.

C. setup 3

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.

D. setup 4

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.

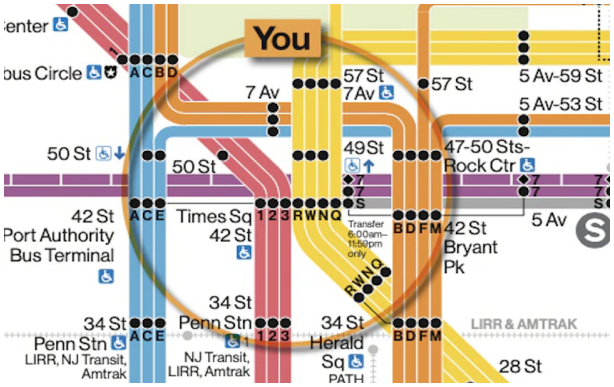

Fig. 1. New York City Network

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.

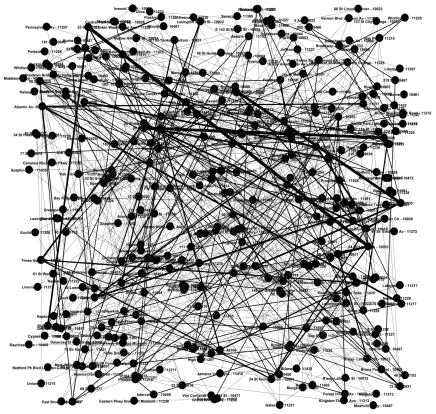

Fig. 2. New York City Subway Network Graph

The basic characteristics of New York City's subway network will be introduced in light of graph theory and complex network theory. The network (Fig. 2) is constituted of 343 nodes and 487 edges and the average degree of the network is 2.84. The clustering coefficient of the network is 0.0078, which shows that the local connectivity of the subway network is very bad. The betweenness of the edge is defined as the number of shortest paths between two nodes passing through the edge over the whole network and the highest betweenness of the edges equals 19482.275693.

IV. Preliminary Results

Show some preliminary results. Explain the significance of your results. Are they what you expected? Do they make sense?

A. Network Properties Analysis

TABLE I
NYC Subway Network Data Lab

Station Name	Zip Code	Degree	Weighted Degree	Eccentricity	Closeness Centrality	Betweenness Centrality
Atlantic Av- Barclays Ctr	11217	11	20.0	17.0	0.202	19482.276
59 St	11220	14	20.0	16.0	0.198	13509.76
36 St	10012	7	10.0	17.0	0.201	9084.59
7 Av	11217	8	14.0	17.0	0.199	8657.94
125 St	10035	8	20.0	15.0	0.182	8583.28
86 St	10028	9	18.0	16.0	0.184	8393.94
14 St- Union Sq	10003	14	32.0	15.0	0.191	8236.53
149 St- Grand Concourse	10451	6	8.0	14.0	0.173	8135.97
3 Av- 149 St	10455	4	6.0	14.0	0.183	7546.99

Using **Gephi** with our data, we list the top 9 *betweenness* centrality of our stations (Table 1). We define the *betweenness* of a node to measure how important the node is to the flow of information through the network. We find if the node's betweenness is higher, the *closeness*, *degree*, and the *weighted degree* are tend to be larger. This relationship is **positive correlation**. As a result, if a station connects with more stations (*degree*), has more routes going through (*weighted degree*), or is in the middle of the whole network (*closeness*), the station will be more important to the subway and makes more impact when crashing or population overflow.

B. Community Detection

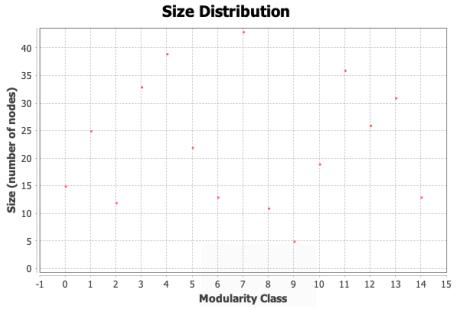

Fig. 3. New York City Subway Size Distrbution

Given by Size Distribution Analysis (Fig. 3), the *modularity* of the NYC subway network is 0.75, and we find there are 15 communities in NYC subway network. Since the *modularity* is close to 1, the whole network will stay strong if meeting some problems (crash).

C. Population Distribution VS. Station

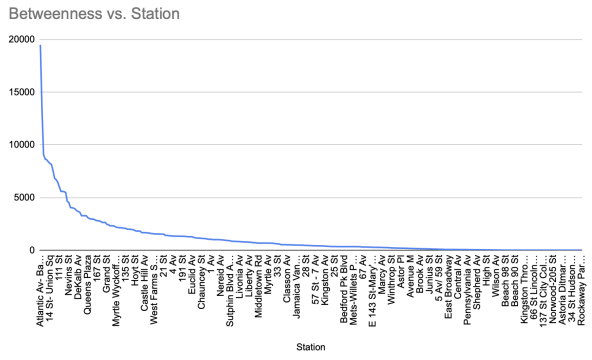

Fig. 4. New York City Subway Station vs. Betweenness

We sort the betweenness and plot the station vs. betweenness graph (Fig. 4). Then we look up the residential population [6] in the same zip code as the station's to plot the station (in the same order) vs. station graph (Fig. 5). The graph is random and not gradually reduced as shown in Fig. 4. Thus, we can conclude that there is no direct relationship between *betweenness* and population. The population size of

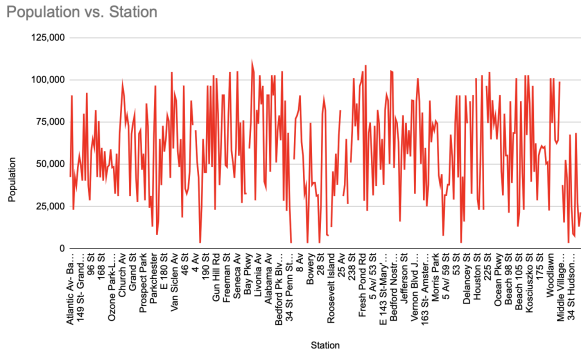

Fig. 5. New York City Subway Station vs. Population

the community around the subway is **not** the deciding factor for building subway stations.

V. Conclusion and Short-Term Plans

Conclusion: 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. For milestone 2 and milestone 3, we will seek more factors that may affect the site selection of subway stations, such as tourist flow, etc.

Short-Term Plans: 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 will dig into robustness in the future. Furthermore, in milestone 1, we focus on New York City. In the next milestone, we will do more analysis of different cities to collect more data. We may also use node2vec in the future to analyze more aspects of the subway network.

In this paper, New York City subway network is investigated with network analysis and Graph Theory.

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