Graph Analysis Study of a City Bus Transit Network

Atiqur S.M. Rahman[1], Pritheega Magalingam[*2], Norshaliza Binti Kamaruddin[2], Ganthan Narayana Samy[2], Nurazean Maarop[2], Sundaresan Perumal[3]

[1] Information Management Department, Hilti, Petaling Jaya, Selangor, Malaysia
[2]Advanced Informatics Department, Razak Faculty of Technology and Informatics, University Technology Malaysia, Kuala Lumpur, Malaysia.
[3]Faculty of Science and Technology, Universiti Sains Islam Malaysia, Malaysia

*mpritheega.kl@utm.my

Abstract. Graph analysis approach was used to analyze a city bus network. The bus trip data were segregated into 4 smaller bus datasets (weekday-peak-hour, weekday-off-peak-hour, weekend-peak-hour, and weekend-off-peak-hour) with the main goal of studying and identifying the differences between peak hours and off-peak hours, as well as between weekdays and weekends. These differences are identified in terms of global characteristics and network evolution distinctions. Based on the analysis of the simplified network, we identified the relative shape of the networks and the centrality attributes. A comparative analysis is performed between the networks at different times to understand the impact of peak against off-peak and weekday against weekends. Significant differences are found in the analysis. In addition to the above, highly important routes and stops were also identified, where optimization initiative is planned for a more efficient bus transport network. The findings can serve as a starting point for future route optimization of the similar network.

Keywords—Bus transport network; social network analysis; betweenness centrality; shortest paths; bus route.

1. Introduction

The transportation system is an important entity that has substantial economic and social impact in a country. To explore and understand the underlying mechanisms of complex systems such as transportation systems, it is essential to use network analysis [1, 2]. The area of network or graph analysis is gaining popularity in recent years due to its effectiveness in studying complex systems and provide opportunities for improvement and optimization. Social network analysis refers to the process of investigating social structures via the use of networks and graph theory [3]. A network is normally characterized by nodes and edges, which are the individual focal entity and relationship between these entities, respectively. Apart from visualizing and simplifying a humongous dataset to be analyzed, social network analysis is useful in pinpointing crucial elements and interactions amongst players within a network. Traditionally, in most studies related to social network analysis on transportation networks, stations and stops are treated as nodes, whereas the edges link consecutive stations along specific routes [2].

Representing transportation systems in a network system with nodes and edges is an important step in analyzing the structure of the transportation system, and any subsequent optimization methods. Throughout the years, centrality measures, statistical parameters, demographic and population info, economic and social structure have been used to analyze different transportation networks. Spatial clustering is used to identify the location-based groups that are strongly or weakly knitted [4, 5]. Some analyses were done using point to point route that shows the inter-relation of stops or bus stations. Besides the network parameters, other local attributes were seemed to be important to identify the relevant route that does not impede the flow of traffic [6]. In another occasion, authors suggested statistical properties of a transport network by analyzing the dataset and adding correlation techniques to observe the network pattern [7]. The amount of travel time, waiting time, transfer time were merged with network parameters [8] to identify important transfer points and to plan more significant routes for people. It was found that the degree distribution value of each stop (bus station) and additional transfer hubs could affect the efficiency of a route in the network [9].

Transportation optimization can be achieved by simplifying travel routes, so that people can commute between destinations within shortest distance and time. However, overall connectivity should also be considered. This paper addresses the transportation optimization problem by considering the overall connectivity. In this paper, we will study the New York City bus network data by introducing an approach that reveals: (a) the trends of the bus network in the peak and off-peak hours. (b) the trends of the bus network in the weekdays and weekends. (c) the efficient routes of bus trips for people in New York City.

New York City is one of the busiest cities in the world and relies on a well-connected commute system including metro-rails and buses. To that extent, the analysis and understanding of such a busy and well-connected network can help us improve the public transport system in other parts of the world, as people are getting more reliant on them, resulting in the massive increase of traffic in those networks. In order to assess the differences between the datasets (peak versus off-peak, and weekday versus weekend), we have used a combination of multiple centrality measures such as degree, betweenness centrality, closeness centrality, and eigenvector centrality as well as global measures consisting of diameter, average path length, and edge density. On top of that, community detection, network evolution and shortest path analysis have also been applied in this research. By comparing the results obtained from the different analyses, we can get detailed insights on the differences between each of them, where it could help public transportation planners to manage their networks more efficiently by identifying the most sensitive routes and stops that potentially are bottlenecks to the networks in scope.

The organization of this paper is as follows. Section II briefly discusses the previous related work. In Section III, the dataset used is described further. As for Section IV, the methodology of this research and analysis of the dataset using social network analysis methods is explained. Next, in Section V, experiment results derived from the analysis in the previous section is discussed. In Section VI, future research opportunities have been discussed to improve on the work done. Finally, Section VII provides the conclusion.

2. Related Work

There are several studies that have been done related to the analysis of transportation. The approach is more or less the same as using network science to analyze the movement of transportation, be it in a public transport or logistic context. In one of the earliest studies, when the social network was introduced at the beginning of 1960s, Burt and Miner in their study have identified social network as a group of social actors that interrelate with one another [10]. Scott and Carrington mentioned that clique analysis was utilized as it requires every member of a given group to be connected to each other [11]. The initialization of social networks is grounded on studies of the relationship between social networks and their spatial embedment [4]. Spatial clustering is a commonly observed phenomenon. Onnela et al. concluded that only small social groups are geographically very tight, while larger groups split into clustered smaller ones over space [5].

Moller and Svahn have identified two types of value network formation in public transport that are; (i) incremental improvement in established value networks, and (ii) radical leaps in emerging value networks [12]. These two types offer a comprehensive understanding of how dynamic and operational capabilities facilitate network formations in each of these networks. Based on these two perspectives, factual information that was retrieved was how the networks are not only developed to coordinate the existing transportation capacities, but also helps in the creation of value in new forms [13].

Another interesting social network study on transportation network was done to analyze the challenges in supply chain. The major challenge is often found at the last step of the supply chain process, often referred to as the ‘last-mile’ problem. Ewedairo et al. found that the transport network impedance to last-mile is defined as the amount of resistance required to traverse through a route on a road network from pick-up point to the delivery point [6]. These movements are often referred to as “last-mile” delivery. Gevaers et al. described last-mile as “the last part of the supply chain” [14]. The impedance to last-mile delivery can also be calculated using a range of spatial indicators. The last-mile analysis provides us with the opportunity to analyze transportation as the inter-relation of participating nodes, and how they can affect the overall movement or information propagation. Their paper used several contributing attributes of a transport network and urban planning controls to visualize a mapped transportation network in Maribyrnong City exhibiting different levels of potential transportation network impedance to last-mile delivery [6]. The findings are important to understand how different parameters apart from the core network can affect the information propagation within the network, and thus it is an important study to understand the challenges.

The urban transportation system has been the area of interest for researchers for more than a decade. In 2002, Latora and Marchiori analyzed the small-world properties in major cities. The researchers investigated the Boston Subway Network for such properties as well as local and global network efficiency [15]. Seaton and Hackett extended that study by comparing the train line network of Vienna with Boston utilizing the usage of random bipartite graph models in addition to the small world properties analyzed in the earlier paper [16]. By considering two cities in their study, they were able to draw comparison between them based on their train-line structure. In another important research about metro and subway systems conducted in 2006, the researchers gathered data from 20 largest subway systems in the world and grouped them into two classes of a generic network that are related to an exponential distribution with simple assembly rules [17]. The researchers identified high connectivity and low maximum vertex degree to be the characteristic features of the networks and analyzed the network robustness against random failures.

In 2007, Chen et al. investigated the urban bus transport networks of four major cities in China [18]. The researchers suggested two statistical properties of the bus network: “the number of stops in a bus route” and “the number of bus routes a stop joins”. The study showed that “the number of stops in a bus route” follows asymmetric, unimodal functions, whereas, “the number of bus routes a stop joins” can be explained by a power-law function with exponential decay. Xu et al. explored scaling laws and correlation in the context of bus transportation networks that may govern various intrinsic features of the network [7]. In their study, they identified the small-world behavior in three bus transportation networks in China.

Ferber et al. studied the public transport networks of 14 major cities and performed extensive complex network analysis [19]. They found that the degree distributions of the networks follow the power-law function confirming the finding of earlier studies. Derrible and Kennedy have had a similar finding on degree distribution in their study of 33 metro systems, and analyzed how network robustness can be affected by increasing transfer hubs [9]. Zhang et al. analyzed the topological characteristics of the subway network of Shanghai, and found that the network is robust against random attacks, but vulnerable against targeted attacks [20]. They also found that the most damage was caused by the highest betweenness node-based attacks. In 2012, a group of researchers studied the temporal evolution of 14 large subway network structures and observed that the shape of the networks have similar generic features despite having differences in the geographic location or socio-economic setup [21].

The importance of central nodes in a transportation network appeared in many past literatures. In 2010, Derrible and Kennedy noted that the presence of central hubs in a transportation network leads to large scaling factors and increases connectivity [9]. Derrible followed up the research on network centrality in 2012, as he investigated network centrality in 28 metro systems worldwide [22]. He concluded that betweenness centrality is a suitable measurement for transportation centrality. He also stated that the study of betweenness centrality in the transportation network is immensely important and can aid in detecting overflowing stations and help in better network planning.

In a study conducted by Scheurer and Porta, it was indicated that the presence of highly connected transfer points and choice of routes of users are highly significant [8]. They assessed the public transportation network of Melbourne based on several measures including degree centrality. They identified travel time as an important factor of transportation, that also includes actual travel time, waiting time and transfer time. Having to make too many transfers to reach a destination can have a serious impact on total travel time. The authors proposed a GIS-based tool to access centrality and connectivity in the urban public transportation network in a subsequent paper to aid policymakers to plan and assess public transportation systems [23].

In a study conducted on the demographic and population data of Kazakhstan, the researchers identified geopolitical importance of particular cities as well as main transport routes in the country [24]. The study was done based on a mathematical model of spatio-populational principles, which is the estimation of the distance between cities and their populations. Betweenness centrality was used to determine the geopolitical importance of particular cities. Wang and Fu performed an experimental analysis on the bus network of Guangzhou and used different centrality measures such as degree centrality, betweenness centrality and closeness centrality in their study [25]. The researchers noted that these central nodes correspond to the hub stations or main transfer points in the urban transportation system, and are located in the centre of the urban network. Any optimization scheme should be started from these stations, and the transport capacity of the hub stations should be extended.

In literature, many studies have been made into analyzing and understanding of transportation networks. Despite the differences in terms of infrastructure, economic and social structure, many important similarities have been observed in the network. Network robustness is another area that has been studied as well, in conjunction with network optimization. By understanding the general structure of such networks, and eliminating weak points, we can develop a robust optimized network that aids the journey within the network with little impedance and higher reliability. Thus, this research aims to analyze the bus network of one of the largest cities in the world and identify the most efficient routes for bus trips of people in New York City.

3. The Dataset

The dataset, which was sourced from kaggle.com, is the New York City's MTA bus data stream service [26]. The information is recorded live from the buses operated in New York City and that includes various critical information about bus routes, location, schedule and absolute time. The data is collected in roughly 10-minutes increment, which includes information regarding the current bus location, bus stops, bus routes, expected time of arrival and temporal distance between bus stops. The scheduled arrival time from the bus schedule is also included, to give an indication of where the bus should be (how far behind schedule, on time, or even ahead of schedule).

In order to perform the network analysis, segregation of nodes and edges are important. The bus stops have been denoted as nodes, whereas pairs of origin and destination bus stops are treated as edges. While nodes are critical, edges play an important role. We visualized the monthly bus trip count in New York City through a line graph and found that the number of trips on a given week is well balanced on weekdays. However, passengers have to encounter every stop in a long designated route before reaching their destination. The only major difference that can be seen on

weekends is where the overall trip counts are significantly lower. Thus, this exhibits a significant difference between weekdays and weekends and can be an interesting area of study to explore the network further.

The scope of this study is to analyze the bus transit network of one of the largest metropolitan cities, New York City. More specifically, the aim is to understand the differences in transit network between peak and off-peak hours as well as between weekend and weekdays. By analyzing the bus transit network in the largest metropolitan city using the metrics and measures of social network analysis, we demonstrate a solution to optimize the public transportation system, more specifically the bus transit system in other busy cities in different countries.

4. Methodology and Analysis

Some social network analysis measures such as clique, centrality, community detection and network evolution are used to understand and detect the central stations that may cause bottlenecks in the network and to simplify transit between stations by reducing the number of stoppages between one particular station to another. The aim is to find an opportunity to optimize the network and reduce congestion. A brief overview of the methodology is shown in Fig. 1.

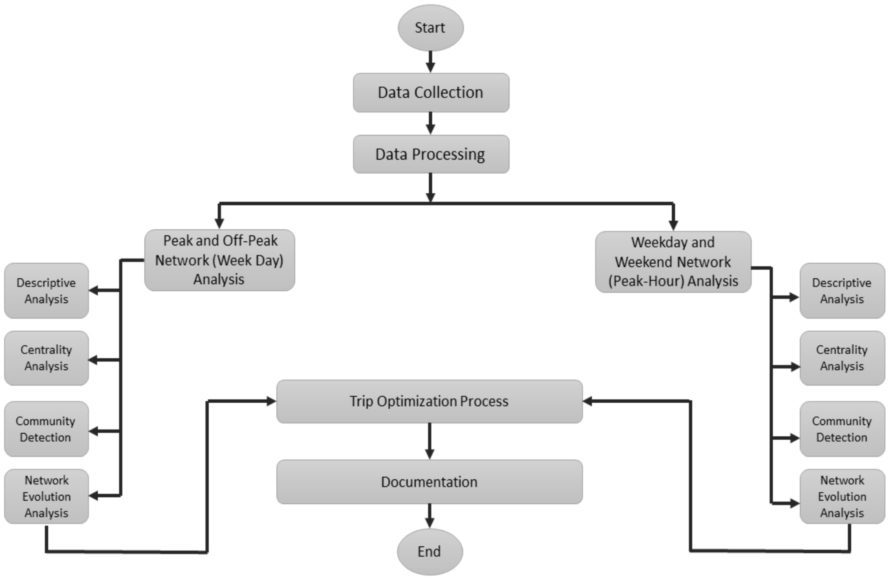
Start
Data Collection
Data Processing
Peak and Off-Peak Network (Week Day) Analysis
Weekday and Weekend Network (Peak-Hour) Analysis
Descriptive Analysis
Centrality Analysis
Community Detection
Network Evolution Analysis
Descriptive Analysis
Centrality Analysis
Community Detection
Network Evolution Analysis
Trip Optimization Process
Documentation
End

Fig.1:Methodology and Analysis Process

A. Data Collection and Pre-Processing

This dataset includes data for a few months on bus transport data for New York City which has been sourced from the NYC MTA buses data stream service. The dataset includes relevant bus trip information like Bus Location, Route, Bus Stop and other details in an approximately 10-minute increment. The scheduled arrival time according to the bus schedule is also included to give an indication where the bus is scheduled to be, and is the bus on time, late or ahead of schedule. A significant pre-processing method is required to make it suitable for network analysis. Part of the original dataset is as shown in Table I.

In Table I, “RecordedAtTime” shows the time of a particular record that has been captured. The Published Line Name denotes the name of a full trip or journey. Each bus trip may have a starting and final station. While moving towards the final station, there will be many intermediate stops. A trip has a unique name and multiple intermediate stations. The origin name and destination name are the first and last stations for a full trip. Thus, on a given trip period, the origin and destination are unchanged in the collected data. The dataset also has vehicle reference number that denotes the unique bus identification. The expected arrival time means the realistic timing that the bus will reach the next bus stations, given current time and conditions, while the “ScheduledArrivalTime” is the arrival time as per the published schedule. Without some pre-processing, this data cannot be used for network analysis. Haznagy et al. in 2015 classified bus stoppages as nodes and the trips between stoppages are represented as edges [2].

Table1: Sample Database

RecordedAtTime	DirectionRef	PublishedLineName	OriginName	DestinationName	VehicleRef	NextStop	ExpectedArrival Time	ScheduledArrival Time
1/8/2017 0:31	1	B9	50 ST/5 AV	Midwood Flatbush Av	NYCT_359	50 ST/7 AV	1/8/2017 0:31	12:30:03 AM
1/8/2017 0:31	1	B9	50 ST/5 AV	Midwood Flatbush Av	NYCT_359	50 ST/15 AV	1/8/2017 0:41	12:43:03 AM
1/8/2017 0:30	0	B9	50 ST/5 AV	Midwood Flatbush Av	NYCT_359	AV I/E 9 ST	1/8/2017 0:56	12:52:00 AM
1/8/2017 0:31	0	B9	50 ST/5 AV	Midwood Flatbush Av	NYCT_359	50 ST/7 AV	1/8/2017 1:11	1:14:00 AM
1/8/2017 0:31	0	B9	50 ST/5 AV	Midwood Flatbush Av	NYCT_359	Midwood Flatbush Av	1/8/2017 1:31	1:24:00 AM

The researcher ran into some data pre-processing challenges. One of the main challenges they faced was inconsistency in stoppage names. Same stops are often having different names, creating problems in constructing the network. They had to develop a method to merge stoppages wherever there's duplication. Though equally challenging, in our research, the data pre-processing for our study is quite different. The collected data is not suitable for network construction, as there is no node to node information. Information about only the main route that includes the first stoppage and the final stoppage of the route is available along with the approaching stoppage. No information about the departing stoppage can be found. This information must be logically and manually derived from the timestamp and vehicle registration number, as well as the route information.

As can be seen in Table 1, for a given trip or bus line, Published Line Name, Origin, Destination and Vehicle Ref are the same. This allows us to identify all the records for a single trip. Since we only have the approaching station, but not the last station information in the dataset, we need to use the record time or expected arrival time to know the order of station that the bus will stop at. Therefore, in order to build a network, we have to take all the records for a particular published line or trip, group by day, origin-destination, vehicle reference number, and order by expected arrival time, or record time. These transformations were applied to the sample data in Table 1, and the outcome is displayed in Table 2. In the transformed dataset, for a given trip, we can predict the approaching station and the last station of the bus trip, as well as the expected arrival time of the next or approaching station. This allows us to build a network. To construct a network, we need some actor or nodes, and a connection or edge between them.

Table 2: Transformed Trip

Origin Name	**Destination Name**	**Expected Arrival Time**
50 ST/5 AV	50 ST/7 AV	1/8/2017 0:31
50 ST/7 AV	50 ST/15 AV	1/8/2017 0:41
50 ST/15 AV	AV I/E 9 ST	1/8/2017 0:56
AV I/E 9 ST	50 ST/7 AV	1/8/2017 1:11
50 ST/7 AV	MIDWOOD FLATBUSH AV	1/8/2017 1:31

In the case of this bus network, the stations can be considered a node, and the route between each node can be considered as an edge. Fig. 2 shows a simple bus network, constructed out of the data displayed in Table 2. Based on the approaching station and time, we can find out the order of station for a given line or trip that the bus will stop at and can construct the trip map in the form of a network. This transformation was applied on the entire dataset, in order to make it usable for network analysis. The origin and destination names were replaced with unique IDs in order to clearly display the network. The dataset was then segmented into peak and off-peak hours. Any trips made between 7AM to 9AM and 4PM to 6PM are considered peak hour traffic, as they fall during office time. The hourly bus transit volume on weekdays is drawn and analysed.

Since the data is recorded in a regular interval rather than once per station, the same station is often reported multiple times, as, within the interval, the bus may not reach the next destination. In order to mitigate this, we eliminated all the self-loops and simplified the network. This study looks at the network topology and performs an analysis of the relations between the nodes, rather than calculating the duration it takes to perform a trip. Therefore, the study was done on the simple graph form, eliminating the weight information and edge duplication. In the study conducted by Haznagy et al., the authors clustered the stations and merged similar stations, as the node names often varied for the same stops [2]. The limitation of the dataset in this study is the bus line names often changed on different days for the same route, thus the analysis was done between stations rather than on a complete trip level. Moreover, the same sub-trips are often recorded multiple times in the dataset, therefore, the dataset had to be cleaned to remove duplicate edges. At the end of the data cleaning process, the whole dataset was divided into nodes and edge lists, where edges have the connection between node IDs, and the node lists are the mapping between node IDs and node names.

B. Network Analysis

Our analysis is divided into two sections; comparative analysis between peak and off-peak hours, and between weekday and weekends. Since the behavior between weekday and weekend can be very different, and the peak and off-peak hour is primarily based on working hours. Only weekday is chosen to compare the peak and off-peak hour bus network because there was no significant difference between peak and off-peak hour that can be observed on weekends. Our pre-analysis shows that peak hour traffic has significant demands.

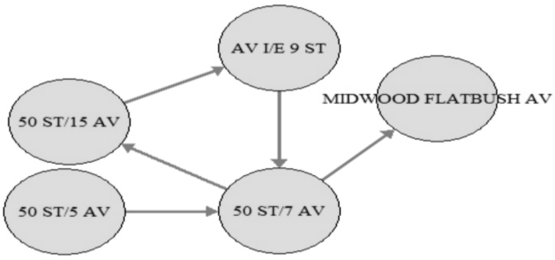

Fig.2: Simple bus network

C. Comparative Analysis for Peak and Off-Peak Hours of the Bus Transit Network

1) Descriptive Analysis:

To analyze the differences in the transportation networks between peak and off-peak hour, firstly descriptive analysis is conducted. On weekdays, during peak hours, a total of 11,986 unique stops have been observed which have 94,445 edges. The mean degree of the network is 15.8. It means that each node or bus station in the network is connected with approximately 16 other bus stations. The higher the average degree of a network, the more connected the network is. A high number of nodes are between 0 to10 degree range. To further simplify the network, and only analyze the high impact and highly connected nodes, we filter out nodes that has equal or less degree compared to the average degree. That is, any station that is connected with fewer than 16 stations are not considered for further analysis.

After filtering the network to be less than 16 degrees based on the mean degree, the number of nodes and edges have decreased to 4,689 and 51,684 respectively, retaining 55% of the network despite reducing 61% of the nodes. The diameter of the network is 16 and has an average path length of 6.7 with an edge density of 0.00235. The average path length is a measure of connectivity, with shorter path length refers to better connectivity, as it is easier to move between the nodes. In the network, it takes an average 7 hops to move from one node to another directly. Diameter is also a useful measure to compare connectivity. As per the diameter measure, in this transport network, the furthest nodes can connect to each other within 16 hops. The edge density describes the cohesion of a network and ranges between a value between 0 and 1, with 0 meaning no connectivity between the nodes, and 1 meaning all nodes are directly connected to each other. The edge density of 0.00235 refers to a low density network, which is normal for a bus network since a bus station doesn’t have a direct link to all other stations in the network. From the degree distribution of the filtered network for weekday peak-hour network, we find that there is a gradual rise in the degree value between 10 to 30 degree range, with roughly 60% of the nodes are within that range.

During the off-peak hour on the weekdays, the number of bus stations or nodes remain unchanged in the network, but the number of edges is very high. There are more transits between the stations during off-peak hours compared to peak hours. The number of edges for this network is 127,944 which is a 35% growth compared to the peak-hour. The average degree of the network is 21.35, which means, on average, each bus station is directly connected with 21 other stations. This connectivity is greater compared to the peak hours, where each station is only connected to 16 stations on average. Despite having a large number of edges or transit routes during the off-peak network, the density of peak hour is much higher. Between 12 AM to 5 AM, the demand is generally low. However, between 6AM to 12AM, peak hours have on average 52% more transits compared to off-peak-hour, despite having more transit routes.

After filtering the network to be less than 22 degrees based on the mean degree of 21.4, the number of nodes and edges reduced to 4,668 and 69,128 respectively, retaining 54% of the network despite reducing 61% of the nodes. The diameter of the network is 15 and has an average path length of 6.2 with an edge density of 0.00317. Peak-Hour Network and Off-Peak hour network has similar degree distribution. However, off-peak hour network has slightly higher degree distribution compared to peak-hour network. The majority of the off-peak nodes are connected with a higher number of nodes in comparison with peak-hour nodes.

2) Centrality Analysis:

Next, centrality analysis is performed on the network to find the central nodes. Part of the results are displayed in Table 3 and Table 4. In peak hour network, the node with the highest degree centrality is “ELTINGVILLE/TRANSIT CENTER” with a value of 176, which means this street is well-connected to 176 other streets. “E 23 ST/1 AV” comes at top for betweenness centrality. The bus route has to go through “E 23 ST/1 AV” to reach every other station. The maximum distance from this node (highest betweenness centrality) to all other nodes in the network is 8.

The node with the highest closeness centrality of 0.00000209 is “DOWNTOWN WORTH ST”. Meanwhile, the node with the highest eigenvector centrality value of 1.0 is “W 181 ST/AMSTERDAM AV”, which also has the third highest degree centrality. Table 4 lists the top nodes in terms of different centrality values for the weekday off-peak hour network. For the off-peak hour network, the node with the highest degree centrality is “ELTINGVILLE/TRANSIT CENTER” with a value of 181. The same station has the highest degree in Peak Hour network as well, with off-peak hour network has slightly greater connectivity.

“ARCHER AV/MERRICK BL” comes at top for betweenness centrality. The shortest paths to reach every other station passes through “ARCHER AV/MERRICK BL”. The maximum distance from this node (highest betweenness centrality) to all other nodes in the network is 6. The node with the highest closeness centrality of 0.000002238 is “SOUNDVIEW PUGSLEY AV”. Lastly, the node with the highest eigenvector centrality value of 1.0 is “BAINBRIDGE AV/E 210 ST”, which is also among the top nodes in terms of degree centrality. The output of this analysis will be further discussed in the experiment results.

Table 3: All Central Nodes (Peak-Hour Weekday)

	Degree	**Betweenness**	**Closeness**	**Eigen.Centrality**
	176	1576669.2	2.0920E-06	1.00
Node	5106	4636	4381	11276
	136	1386368.1	2.0919E-06	0.97
Node	2842	5106	8224	11277
	130	1310902.4	2.0913E-06	0.78
Node	11276	11267	5355	11470

Table 4: All Central Nodes (Off-Peak-Hour Weekday)

	Degree	**Betweenness**	**Closeness**	**Eigen.Centrality**
	181	1721021	2.2379E-06	1.00
Node	5106	2170	10188	2700
	148	1711892	2.2365E-06	0.88
Node	2842	4712	8223	2699
	142	1612749	2.2347E-06	0.78
Node	2167	3128	6732	4866

3) Community Detection Analysis:

This analysis is done based on how many cliques exist in the network. Cliques are subgraphs in a network, that includes 3 or more nodes that are adjacent to each other. First, the study looks into the number of 5,7 and 9 cliques, and lastly, it finds the largest clique in the network.

The peak-hour network has 87,453 cliques with 5 nodes, 14,422 cliques with 7 nodes and 1,385 cliques with 9 nodes. The off-peak hour network has 586,287 cliques with 5 nodes, 150,963 cliques with 7 nodes and 17,812 cliques with 9 nodes. Having a high number of cliques mean there are a group of nodes that are highly connected to each other. Comparing the number of cliques between two graphs or networks relates to the node connectivity. Next, the number of communities that exist in the network were identified.

4) The Network Evolution Analysis:

The objective of network evolution analysis is to find disjointed sets of networks by cutting off low degree nodes from the network. Fig. 3 shows the network evolution for Peak Hour network on weekdays. Peak Hour network has 90% of the nodes under degree 30. Nodes with less connectivity are identified based on degree distribution. The removal of least connected nodes result in the formation of communities. Evolution is seen by grouping nodes with several degree range into sub-graphs. Nodes start to form and can be seen clearly when the total degree chosen is above or equal to 50. However, with this degree value, the network is still being connected at large and distinct groups have not been formed yet. At degree value 60 and above, disjointed groups are noticeable.

The network evolution for the weekday off-peak network also shows that the majority of the lowly connected nodes have to be eliminated to obtain disjointed subgraphs. Disjointed sub-graphs start to develop in a network with nodes that have a degree from the range 60 to 80. Though both peak and off-peak networks displayed strong connectivity, the off-peak network requires higher elimination of low degree nodes to form disjointed graphs. That shows that the network is more connected and more small routes were formed during off-peak hours compared to peak hours.

D. *Comparative Analysis of Weekday and Weekend Bus Transit Networks*

1) Descriptive Analysis:

Firstly, a descriptive analysis is much needed to analyze the differences in the transportation networks between weekday and weekend. On a weekend, during peak hours, a total of 11,986 unique stops have been observed which have 76,043 edges. The mean of the network is 12.7. That means each bus station have a direct transit to 13 other stations on average. The network is simplified by filtering nodes with lesser degrees. Any node that having a degree count less than the average is discarded. After filtering the network to be less than 13 degrees based on the mean degree, the number of nodes and edges are reduced to 4,836 and 47,031 respectively, retaining 62% of the network despite reducing 60% of the nodes. The diameter of the network is 17 and has an average path length 7.14 with an edge density of 0.00201.

Compared to the weekday peak-hour network, weekend peak-hour network has lower degree distribution. 80% of the nodes are between the range of 15-25 degree. The majority of the weekend peak-hour network nodes are connected with a lower number of other nodes in comparison with weekday peak-hour network nodes.

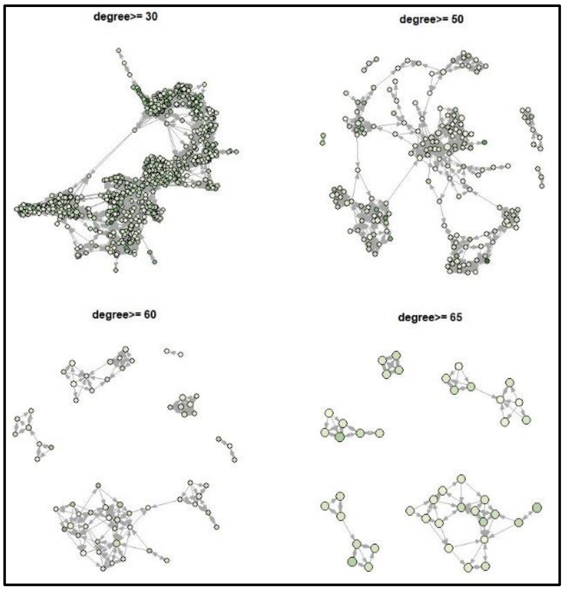

Fig.3: Network Evolution For Peak-Hour Network (Weekdays)

2) Centrality Analysis:

Next, the centrality analysis is performed on the network. For the peak hour weekend network, the node with the highest degree centrality is “W 181 ST/AMSTERDAM AV” with a value of 120 degree, which means this street is well-connected to 120 other streets. “ARCHER AV/MERRICK BL” comes at top for betweenness centrality where the routes. means that the paths between stations flow the most through “ARCHER AV/MERRICK BL”. The maximum distance from this node (highest betweenness centrality) to all other nodes in the network is 11. On the other hand, the node with the highest closeness centrality of 0.000001692 is “W 49 ST/7 AV”. Lastly, the node with the highest eigenvector centrality value of 1.0 is “W 181 ST/AMSTERDAM AV”, which also has the highest degree centrality.

3) Community Detection and Network Evolution Analysis:

The study also analyzes the number of cliques that can be found in the network. The weekday peak-hour network has 32,677 cliques with 5 nodes, 4,809 cliques with 7 nodes and 614 cliques with 9 nodes. The largest clique is 12 in the network, and there are 2 cliques with the largest clique setup. Within the community detection, the network evolution analysis is done with the objective to find disjointed sets of networks by cutting off low degree nodes from the network.

It is found that disjointed networks start to form at degree range above or equal to 50. By cutting further at degree range 54, more disjointed networks start to form. Above this range, the networks start to grow smaller. In comparison, Weekday Peak hour network forms disjointed networks at degree range of 60 or above, indicating a more cohesive network.

5. Results and Discussion

In this section, we discuss the analysis results related to the trends of the New York City Bus Network during the peak and off-peak hours, and on weekdays and weekends. The analysis has been done based on social network measures and aims to identify some differentiating patterns between the transport network at different periods. Besides, the best routes for the bus trip using shortest paths method will be proposed.

A. *Differences Between Peak-Hour and Off-Peak-Hour Networks*

The major output of the comparative study done on the peak and off-peak networks are:

- During weekdays, peak-hours observe higher degree nodes in the bus network.
- There are more edges or paths between nodes in the off-peak than peak hours on weekdays. It means that during peak hours, more trips are made between specific areas (office, commercial areas, etc).
- The degree distribution of peak-hour and off-peak hour network also shows that stations during off-peak hour have higher connectivity.
- After filtering by mean degree, we found that the average path length and diameter of off-peak hours is smaller despite having a larger number of edges and nodes. On the other side, during peak hours, a smaller number of bus stops are more connected to each other, leading to a higher diameter.
- Similarly, the central node (based on betweenness centrality) in peak hours has a higher distance from every other node in the network than in off-peak hours. This means that some of the prominent stations have more incoming and outgoing trips in peak hours.
- The number of cliques (5,7,9) are much higher in off-peak hours. There are more stoppages within short reachability in off-peak hours. Thus, the largest clique size is higher in off-peak hours as more stoppages are traversed.

- In the network evolution for peak hour, we can find a small number of disjointed sets having a small number of nodes in subgroups. While in the off-peak-hour network, more node elimination is required to derive disjointed sets that denotes the off-peak-hour network has stronger ties compared to peak-hour.
- The off-peak-hour network has higher edge density compared to peak-hour network on weekdays. That reflects the bus stations are directly connected to more other stations during off-peak hours compared to peak-hours.

From the analysis, it is found that there are some fundamental key differences between the peak and off-peak bus networks, as peak hours have a smaller number of stations, with a high number of trips in between, compared to off-peak hours, where trips are dispersed between the high number of stations. Peak hour bus network is concentrated in particular areas like office and commercial areas. More trips pass through the important central stations during peak hours, resulting in greater distance from central nodes.

B. Differences Between Weekday and Weekend Networks

The analysis between weekday and weekend has been done considering peak hours, as peak hours have higher density transit. The following are the comparison done between weekday and weekend networks:

- The average degree is higher on weekdays. After filtering by mean degree, we found that the average path length and diameter of the weekday network are smaller despite having a larger number of edges and nodes. This indicates that during weekdays despite having smaller number of trips, they are highly connected. On weekends, more hops are traversed to reach the same distance.
- Similarly, the central node (based on betweenness centrality) on weekends has a higher distance to other nodes than weekdays. This means during weekends some of the routes allow connection to many distant stations. On weekdays, more sub-trips has smaller routes between stations in order to research the destination sooner.
- The number of cliques (5,7,9) is much higher on weekdays. There are more stoppages within short reachability on weekdays.
- In the network evolution for weekends, we can identify many disjointed sets of subnetworks. Weekday networks display higher connectivity between the stations.

The study found that during weekdays the stations are highly connected with small routes compared to the weekend. On weekends, it takes more stops to reach the same destination, as can be identified through the higher average length. This causes the transit system to be much busier and congested during weekdays. In this research, network measures have been applied to introduce more unique trips between stoppages. This would allow network traversal easier and faster on weekdays.

C. Network Optimization

Transport Network Optimization is an important task to make sure that unnecessary redundant bus travels are minimized. Although bus trips are short during the weekdays, it is highly connected with many stations. Whether those stations are important for travellers depends on the necessity of them. Unnecessary bus trips are costly, bad for the optimized trip environment, and creates congestion. Having short but well-connected routes can minimize such unnecessary trips, and balance good connectivity with congestion reduction. We found a solution using SNA measure to solve this problem.

D. Shortest Path Identification

The shortest path algorithm was conceived by Dijkstra in 1956, that finds the shortest path between two given nodes in a graph [27]. The shortest path is calculated based on the least number of hops needed to reach a particular destination. This can be weighted to include the distances between nodes to calculate the actual shortest distance rather than least number of hops. Since we are performing an unweighted study, the absolute distance between nodes is not considered, and rather least number of hops are identified between nodes.

For optimization, the shortest path algorithm has been applied between two stations. In this study, the distance between a randomly picked stations; “ST PAULS AV/CEBRA AV” (node 10440) and “BRICKTOWN MALL” (node 3200) of the peak-hour network has been considered for this analysis. The original network between the stations is shown in Fig.4 and it depicts that the bus needs to visit 34 stations to reach the intended destination. However, the same trip can be made by visiting only 2 stations as per Fig 5.

This provides a great opportunity to optimize trips in the transit network. At the same time, there are some major limitations in this shortest path analysis. Since this is a bus network, the simple network optimization may not serve a practical purpose, as some of the stations in between might be important stations that must be visited along the route. Moreover, the trip is limited to a finite group of stations, and passengers of this trip have limited travel opportunity. Thus, we tried to combine the shortest path and between centrality value to identify the optimized network.

E. Shortest Path and Betweenness Centrality

Betweenness centrality is an important centrality measure for graphs based on shortest paths. It is a measure of the influence of a node over the flow of information between every pair of nodes, under the assumption that information is primarily transmitted via the shortest paths. The first formal definition of the measure was provided by Freeman in 1977 and found a wide range of application in network theory [28]. The significance of hub nodes with high betweenness centrality has been discussed extensively in the literature [8, 23-25, 29]. Any optimization effort on transportation network system need to consider the hub stations.

We can extend our shortest path findings with betweenness centrality, that will improve the network connectivity greatly. If the trip visits some bus stations with high betweenness centrality, then the passengers will be able to easily change bus and take a preferred route. This way, there will be a high number of shorter trips, as passengers can travel to preferred destinations via major hub stations, indicated by high betweenness centrality.

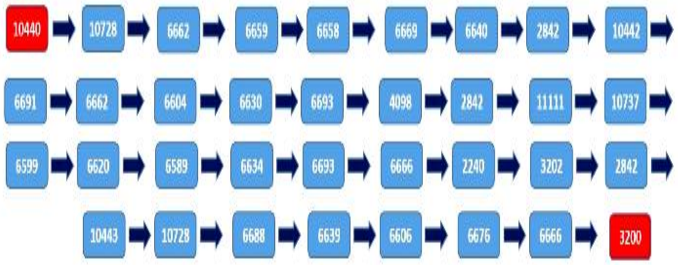

Fig.4: Original Trip

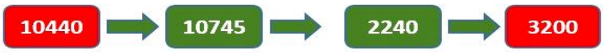

Fig.5: Optimized Trip

In order to optimize the trip with the help of hub stations, we followed the following guidelines:

a) Select trips with a high number of stations
b) Calculate the shortest path with the high betweeness centrality nodes
c) Calculate the distance for each route for the high central nodes. Pick the shortest route.

In Table 5, we can see the nodes with highest betweeness centrality, and the length between the stations “ST PAULS AV/CEBRA AV” (node 10440) and “BRICKTOWN MALL” (node 3200) while going through those central nodes. As we can observe, all the optimized routes are much shorter than the original route. The maximum distance is 15 opposed to 34 stations. The shortest route involves one of the most central nodes, 5106. When the bus touches the central station “ELTINGVILLE/TRANSIT CENTER” in its path, the bus is able to reach its destination within five stations. On the other hand, despite the shorter route, the passengers have the opportunity to utilize the high connectivity of the central station “ELTINGVILLE/TRANSIT CENTER” and reach other destinations as desired.

Table 5: Shortest Distance via Central Nodes

Node	**Length**
4636	6
5106	5
11267	11
2170	14
8091	15
4908	15
4183	15
5256	11
7747	7
4711	7

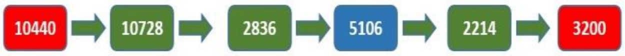

Fig. 6. Optimized Trip with Central Node

Fig. 6, shows the optimized trip, that includes the central node 5106. This route is a little longer than the most optimized route that is shown in Fig. 5, but provides passengers with greater connectivity and flexibility. Therefore, to perform a proper optimization, we need to know not only the source and destination but also all the important stations in between. The result found in Fig. 5 uses Dijkstra's shortest path algorithm [27], that affectively shortened the route. However, that doesn't result in the most useful public transportation route, as that reduces overall connectivity. By integrating social network concepts such as betweenness centrality, we are able to find a shorter route, while preserving connectivity.

Our study here can act as a guide for such optimization tasks. Knowing the important and destination nodes based on passengers' needs through a survey can also help us to have a proper optimized route that will not only reduce journey time and traffic congestion but will also provide better connectivity and flexibility to the passengers. In this study, we have only considered a simple network. In future, we will also take into account the edge weight, capacity, distance and travel time, that can also add an important element to perform effective network optimization.

6. Future Works

The transportation network optimization presented in this paper for New York bus transportation system is an important step towards an efficient bus route system. In the study conducted by Scheurer and Porta, the authors argued that one of the most important factors of a transportation network is time taken between source and destination stations [8]. Unlike private transportation, passengers have little control over choosing the path, and duration to reach the destination. This becomes the most important factor rather than actual distance or number of stations in between.

The total travel time is the sum of Waiting Time, Actual Travel Time and station Changing time. While reducing the number of stations, it can reduce the actual travel time. However, having to change stations too many times can increase the changing and waiting time, and in a result increase the overall travel time. Thus, we need to look into these factors too in the future study. Another important factor to consider is the capacity of stations. Having too much reliance on hub stations can put much pressure on the station and potentially create bottlenecks and slow down overall transportation. Capacity is needed to consider for both the number of passengers that can wait in the stations and the number of buses that can be served by the station in a given time. Extending the capacity of such important stations can be one outcome of the study.

This paper only looks into an unweighted simple graph. We need to know not only the number of buses between stations but also the number of passengers who are riding the buses. Knowing the number of real passengers commuting between stations will provide us with valuable information in order to optimize the network. We can add capacity on high demand trips and reduce capacity where less passenger movement occurs.

Another future area of research could be studying the suitability and effectiveness of the optimization process presented in this paper by comparing against popular optimization and pathfinding algorithms such as Dijkstra's algorithm and A* search algorithm [27,30]. Overall, the study done in this paper sets up some important avenues of future research that can aid urban bus transportation planning.

7. Conclusion

In this paper, we have discussed different research efforts using social network analysis, more specifically in the public transportation context. Differences between peak and off-peak trips, as well as between weekend and weekday trips in the New York City bus network have been studied. We found that some stations have higher preference level on weekdays, which suggests that the weekday peak hour network are more optimized to support working people in primary business areas, opposed to more entertainment-oriented areas on weekends. We proposed a simple method of network optimization using major hub stations to provide better connectivity to passengers while reducing their bus trips. This allows public transportation companies and urban planners alike to deep dive into data to uncover patterns and insights on root causes of currently inefficient and ineffective services, hence enabling the provision of better quality services in the future.

References

[1] Y. Hu and D. Zhu, 2009, "Empirical analysis of the worldwide maritime transportation network," *Physica A: Statistical Mechanics and its Applications,* vol. 388, no. 10, pp. 2061-2071.

[2] A. Háznagy, I. Fi, A., 2015, London, and T. Németh, "Complex network analysis of public transportation networks: A comprehensive study," in *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pp. 371-378: IEEE.

[3] E. Otte and R. Rousseau, 2002, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of information Science,* vol. 28, no. 6, pp. 441-453.

[4] Y. Wang, R. Kutadinata, and S. Winter, 2019, "The evolutionary interaction between taxi-sharing behaviours and social networks," *Transportation Research Part A: Policy and Practice,* vol. 119, pp. 170-180.

[5] J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási, and N. A. Christakis, 2011, "Geographic constraints on social network groups," *PLoS one,* vol. 6, no. 4, p. e16939.

[6] K. Ewedairo, P. Chhetri, and F. Jie, 2018, "Estimating transportation network impedance to last-mile delivery: A Case Study of Maribyrnong City in Melbourne," *The International Journal of Logistics Management,* vol. 29, no. 1, pp. 110-130.

[7] X. Xu, J. Hu, F. Liu, and L. Liu, 2007, "Scaling and correlations in three bus-transport networks of China," *Physica A: Statistical Mechanics and its Applications,* vol. 374, no. 1, pp. 441-448.

[8] J. Scheurer and S. Porta, 2006, "Centrality and connectivity in public transport networks and their significance for transport sustainability in cities," in *World Planning Schools Congress, Global Planning Association Education Network*.

[9] S. Derrible and C. Kennedy, 2010, "The complexity and robustness of metro networks," *Physica A: Statistical Mechanics and its Applications,* vol. 389, no. 17, pp. 3678-3691.

[10] M. J. Minor and R. S. Burt, 1983, *Applied network analysis: a methodological introduction*. Sage Publications.

[11] J. Scott and P. J. Carrington, 2011, *The SAGE handbook of social network analysis*. SAGE publications.

[12] K. Möller and S. Svahn, 2003, "Managing strategic nets: A capability perspective," *Marketing theory,* vol. 3, no. 2, pp. 209-234.

[13] H. Gebauer, M. Johnson, and B. Enquist, 2012 "The role of organisational capabilities in the formation of value networks in public transport services," *Management Research Review,* vol. 35, no. 7, pp. 556-576.

[14] R. Gevaers, E. Van de Voorde, and T. Vanelslander, 2014, "Cost modelling and simulation of last-mile characteristics in an innovative B2C supply chain environment with implications on urban areas and cities," *Procedia-Social and Behavioral Sciences,* vol. 125, pp. 398-411.

[15] V. Latora and M. Marchiori, 2002, "Is the Boston subway a small-world network?," *Physica A: Statistical Mechanics and its Applications,* vol. 314, no. 1-4, pp. 109-113.

[16] K. A. Seaton and L. M. Hackett, 2004 "Stations, trains and small-world networks," *Physica A: Statistical Mechanics and its Applications,* vol. 339, no. 3-4, pp. 635-644.

[17] P. Angeloudis and D. Fisk, 2006, "Large subway systems as complex networks," *Physica A: Statistical Mechanics and its Applications,* vol. 367, pp. 553-558, 2006.

[18] Y.-Z. Chen, N. Li, and D.-R. He, 2007, "A study on some urban bus transport networks," *Physica A: Statistical Mechanics and its Applications,* vol. 376, pp. 747-754.

[19] C. Von Ferber, T. Holovatch, Y. Holovatch, and V. Palchykov, 2009, "Public transport networks: empirical analysis and modeling," *The European Physical Journal B,* vol. 68, no. 2, pp. 261-275.

[20] J. Zhang, X. Xu, L. Hong, S. Wang, and Q. Fei, 2011, "Networked analysis of the Shanghai subway network, in China," *Physica A: Statistical Mechanics and its Applications,* vol. 390, no. 23-24, pp. 4562-4570.

[21] C. Roth, S. M. Kang, M. Batty, and M. Barthelemy, 2012, "A long-time limit for world subway networks," *Journal of The Royal Society Interface,* vol. 9, no. 75, pp. 2540-2550, 2012.

[22] S. Derrible, 2012, "Network centrality of metro systems," *PloS one,* vol. 7, no. 7, p. e40575, 2012.

[23] J. Scheurer, C. Curtis, and S. Porta, 2007, "Spatial network analysis of public transport systems," in *Australasian Transport Research Forum, Melbourne, Australia*, vol. 19.

[24] M. Goremyko, *2018,* "Betweenness centrality in urban networks: revealing the transportation backbone of the country from the demographic data," in *IOP Conference Series: Earth and Environmental Science*, vol. 177, no. 1, p. 012017: IOP Publishing.

[25] K. Wang and X. Fu, 2017 "Research on centrality of urban transport network nodes," in *AIP Conference Proceedings*, vol. 1839, no. 1, p. 020181: AIP Publishing.

[26] M. Stone, 2018, *New York Bus Data*. Available: https://www.kaggle.com/stoney71/new-york-city-transport-statistics

[27] E. W. Dijkstra, 1959, "A note on two problems in connexion with graphs," *Numerische mathematik,* vol. 1, no. 1, pp. 269-271.

[28] L. C. Freeman, 1977, "A set of measures of centrality based on betweenness," *Sociometry,* pp. 35-41.

[29] R. Puzis, Y. Altshuler, Y. Elovici, S. Bekhor, Y. Shiftan, and A. Pentland, 2013, "Augmented betweenness centrality for environmentally aware traffic monitoring in transportation networks," *Journal of Intelligent Transportation Systems,* vol. 17, no. 1, pp. 91-105.

[30] W. Zeng and R. L. Church, 2009, “Finding shortest paths on real road networks: the case for A*.