Visualizing Transportation Networks: MTA Case Study

Paul Danifo, Jonathan Gradman, Steve Pepple, and Petrina Uhlenhopp

Abstract—This paper examines how the New York City Metropolitan Transportation Authority (MTA) could draw from data visualization techniques to enhance their business. The authors propose a purpose-based taxonomy for categorizing transportation network visualizations to identify what types of visualizations will be most helpful to the MTA. Taking advantage of newly-released datasets, they created several visualizations for this purpose.

Index Terms— Geographic Information Systems (GIS), Geographic Visualizations, Information Visualization, Network Visualization, Two-Dimensional Visualization, Transportation Quality of Service, Topographical Visualization

1 INTRODUCTION

WHEN considering a domain in which to apply a study of information visualization techniques and ultimately produce a new application based on this study, the transportation domain presents interesting aspects and unique challenges. More specifically, a focus on public transit systems allows for benefits to multiple audiences, both transit system administrators and the general public. Due to the possibilities that arise from these use cases, there is a broad range of data that can be considered for analysis. Similarly, there are different information visualization approaches and specific aspects of each approach from which value can be realized for each audience.

From an information visualization perspective, the public transit domain lends itself to analyzing different approaches to creating network visualizations. Both geographical and conceptual approaches can be employed, with advantages and disadvantages for each. Additionally, there has been an ample amount of work done and subsequent literature published with respect to applying network information visualizations to the increasingly growing store of data available for public transit systems. This provides a solid starting point from which to consider additional innovations, both technologically and analytically, in order to provide better tools and applications for the intended audiences.

The decision on which transportation system to use for the analysis and visualization work was based primarily on the availability of both data and previous work on this subject. The New York City Metropolitan Transportation Authority (MTA) was chosen, as standard transportation system data was readily available, as were additional emerging data sources related to customer satisfaction and other interesting statistics for use in analysis activities. Furthermore, the MTA was selected because of the organization's intent to release more data to the public and use data visualization as a way to analyze this data and explain the agencies services [1]. Due to both the data availability and the size and complexity of the NYC MTA system, there were also more possibilities for making choices on how to focus efforts throughout the course of the project.

Since the MTA wishes to use visualization as part of its standard metrics but has yet to release a visualization proposal, this paper address the question of what types of standard visualizations might be helpful for the MTA to create and reference on a regular basis. The authors pulled from the field of transportation network visualization literature and previous visualizations of networks similar to the MTA; considered the ways visualizations would benefit the MTA, proposing a purpose-based taxonomy for evaluation; and, using D3.js, created several proposed visualizations for this purpose.

2 CONTEXT

2.1 Information Visualization Techniques Applied to Transportation Networks

An initial review of existing literature explains benefits of visualizing transportation networks in terms of identified audiences. From a 2007 paper by Kimpel: "Transit performance measures are highly flexible in that they can encompass multiple aspects of transit service that are of interest to both agencies as well as passengers" [2]. Additionally, the importance of multivariate analysis has been suggested as a method to push information visualization in the transportation domain forward: "Visualization techniques that make use of multiple variables are able to provide additional insight into the potential causes of transit service problems" [2].

Specific conclusions have been made in conjunction with analysis of policies at the NYC MTA with respect to both data and information visualizations. In a 2012 paper, Shannon highlights a goal to: "promote improved communication between agencies and stakeholders by opening specific data and creating interactive data visualiza-

The authors are graduate students at Drexel University tions" [1]. This prioritization helps to explain the accessibility of MTA data through open data initiatives, which ultimately encourage developers to expand on the progress of information visualization in general and specifically in this case as related to network applications. There are various ways in which to represent networks visually and there are also publications that discuss these and also the specific aspects that are identifying visual characteristics and challenges of networks.

Upon exploring approaches to visualizing networks like the MTA, a starting point Kershenbaum points out is to consider geographical representations, which present challenges: "Nevertheless, the representation would be unreadable if it were necessary to draw the map to precise geographic scale" [3]. What appears more important is to focus on the goals and advantages of taking a geographical approach, particularly when considering the goals of the intended audience: "Manhattan is enlarged in order to show the many subway lines passing through it more clearly... as the user is not driving and does not need to know precise distances" [3]. In this manner, the challenge of displaying a map to scale is addressed by recognizing that the precision of the distances is not critical to the success of the information visualization. This ultimately leads to a conclusion that conceptual network visualizations are another viable option to visualize a transit system, in addition to considering other techniques to address legibility: "we would have to add additional visualization functions to keep the representation intelligible" [3]. Non-geographical network visualizations are borne out of different requirements: "there is no particular significance to the location of the nodes and the primary objectives are readability and esthetics" [3]. Labeling the nodes and relationships becomes more important, due to the lack of a map for context, as does the application of a good layout algorithm and the use of color and size to enhance readability. Incorporating user interaction capabilities is another way to improve the effectiveness of both geographical and conceptual network information visualizations, particularly for the tasks of zooming, "It is easy to do (if the controls are implemented sensibly) and it takes burden off the system", and annotation, "including enhanced annotation on a single node or edge can take a large burden off the system" [3].

After shifting the focus to the MTA specifically, a key consideration is the availability of data and the key milestones that have contributed to advancements in this area: "In 2008, the MTA took the first steps in making its transit information more accessible to the public" [1]. Other considerations include the types of data and measurements that are useful to MTA administration and the benefits that are expected as a result of gaining better insights into this data. The key performance measurements identified by the primary research in this area include: commute speeds, service frequency, and major delay frequency and recovery time [1]. The need for the data to support these is what spurred the focus to improve information availability through open data initiatives, culminating in the formulation of an Open Data Task Force in 2013: "The MTA has now created an Open Data Task Force, and continues to participate in Hackathons and promote app contests" [1]. Ultimately, these actions have a projected long term benefit to the organization, targeted at $71.7 million by 2015 [1]. Cummulativly, MTA sharing of open data and the resulting apps, visualizations, and community feedback, will improve the system in the following ways:

- Apps and visualizations can assist the public in making better decisions.
- Apps and visualizations will improve MTA understand of it's own system, both it's deficiencies and successes
- The process will teach MTA which data is most beneficial to the public, resulting in a prioritization of releasing additional similar data in an easy-to-use format.

3 Transportation Visualization Users

3.1 Taxonomy

Visualizations can serve as a window to data, assisting a particular audience with focusing on the right amount of information for their needs. As we reviewed different transportation visualization approaches, we saw three distinct target audiences. One of these groups is transportation consumers, the individuals trying to get from point A to point B in the way that will best suit their needs. Another other group is comprised of people with decision-making responsibilities for transportation, either government or private companies, who are trying to understand how transportation is used in aggregate and predict how they can impact the overall network for greater service or cost efficiency. For the purposes of this paper we refer to this group as transportation administrators. A third group is made up of interested citizens; we refer to these individuals as the civic interest group.

To ensure that data visualizations would serve the MTA, we decided to determine if visualizations could clearly serve a purpose for a proposed user. While there is a possibility of overlap between these categories (an individual might both work for the MTA and use it for transportation), for the sake of this paper we are focusing on the dominant use per visualization. We use this mini-taxonomy of visualizations by purpose/audience:

1. Transit Use, in this case by MTA customers
2. Transit Administration
3. Civic Interest

3.2 Transit Use

Transportation customers or end users are typically not concerned with understanding the overall system. Rather, they need to understand how to use the system to travel to a place at or by a particular time.

Travelers need a clear understanding of:

- Current time
- Departure time

- Arrival time
- Current location in system
- Future location in system
- Transfer points
- Cost of trip

3.3 Transit Administration

Individuals in this category would include MTA employees. Information helps them make better decisions about where to invest time and resources. An agency like the MTA has limited budget and human resources to keep an important system up and running. In addition, they must make key decision for the future of the system. All of this must be done under intense public and government scrutiny. Many decision makers often have limited experience with IS/IT systems. Visualization may be imperfect, but they allow these administrators and planners to spot interesting trend or issues in the system more quickly and then do more detailed analysis.

3.4 Civic Interest

The third category in this taxonomy does not have a clear purpose for the MTA. This is a very broad group of people who have tangential interests in transportation; they might eventually vote on issues related to public transportation or invest in private transportation companies. There are many visualization of transportation system that look beautiful and are very complex, but do not help a transit planner or end user. We recommend that the MTA places a lower priority on creating this type of visualization. However, making open source data available will allow interested parties to create these types of visualizations without requiring MTA investment.

4 Tools

4.1 Data Sources

The following primary data sources were used:

- Subway line and interchange information was downloaded via a CSV file from mta.info [4]
- Google Transit Feed data in the form of a database, which includes stops and trip times for all MTA subway and bus lines - specifications can be found at https://developers.google.com/transit/gtfs/ [5]
- Additional MTA JSON files for paths, stops, and transfers: https://github.com/daveswartz/mapperly [6]

4.2 Technology

The following computer based technologies were used for the information visualizations:

- Python to analyze the transit information as a network
- OpenRefine and Python Pandas to format and merge the datasets
- NetworkX (http://networkx.github.io/) [7] to create an object representation of the MTA network, with stops as nodes and transit lines as edges [7], [8].
- MapBox API was used for the geographic visualization and presentation of JSON polygons and nodes
- D3.js was used for the visualization of JSON network representation [9].

Raw transit data collected via CSV files was formatted with OpenRefine and converted the data into JSON.

The graph created by NetworkX was exported to a JSON file that could be visualized by D3.js or equivalent charting and visualization libraries.

5 Methods

5.1 Data Collection

In order to create appropriate visualizations of the MTA system, a number of resources and technologies were integrated. We first obtained raw MTA system data from a variety of sources. MTA subway transit data was obtained from the MTA's developer pack [4]. This package provided CSV files containing operation and performance data from all MTA operated transit lines. Google Transit Feed data [5] was obtained in the form of a database, which included stops and trip times for all MTA subway and bus lines. Additional MTA JSON files for paths, stops, and transfers in the MTA system were obtained from the Mapperly project [6]. This collection of raw data provided the basic framework for the visualization project.

5.2 Data Analysis

A number of computer based technologies to format, manipulate, and visualize the MTA transit data were used. The data was analyzed as a network using Python and NumPy. OpenRefine and Python Pandas were used to facilitate the formatting and merger of the various data sets into a standardized JSON file format. NetworkX was employed to create an object representation of the MTA network, with stops as nodes and transit lines as edges.

The primary work of creating the visualization relied mainly on the D3.js JavaScript library. Mapbox, and the leaflet JavaScript map library were also used to facilitate visualizations which represented map based locations. Geocoding was done using the Google Maps API framework, and allowed the data to be plotted more easily on maps of the New York City metropolitan area.

6 Results

6.1 Figure 1

A basic way of visualizing a transportation network is to overlay available travel routes on a geographical map.

This first figure represents MTA routes in New York.

6.2 Figure 2

This map is an example of visualizing the MTA subway lines by their relative performance. The data includes a yearly average of wait times for each MTA line (data was collected in 2009). The visualization shows the percentage of the time that lines were on schedule (based upon MTA's own transit schedule). A higher percentage means the line performed better.

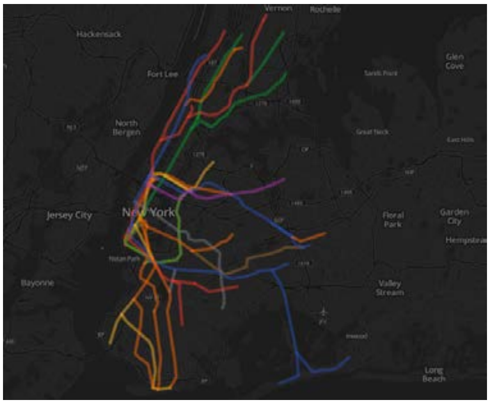

Fig. 1. This geographic map visualizes the MTA subway lines by color. The polyline data is from the Google Transit Feed Specification for MTA.

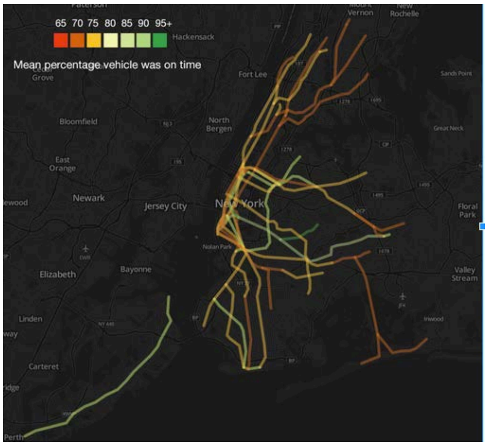

Fig. 2. This geographic map visualizes the MTA subway lines by relative performance. Color indicates mean percentage vehicle was on time.

6.3 Figure 3

The third figure shows the MTA as a network where nodes represent individual stops and the size of the nodes represents the number of lines passing through each stop. Edges represent connections between stations. The color of each edge represents the line it is part of. Colors align to the MTA style guide [11].

6.4 Figure 4

The fourth figure combines node representations with geographical distribution. Node size again represents number of lines passing through each station. Nodes are placed on a New York map to align with their physical locations.

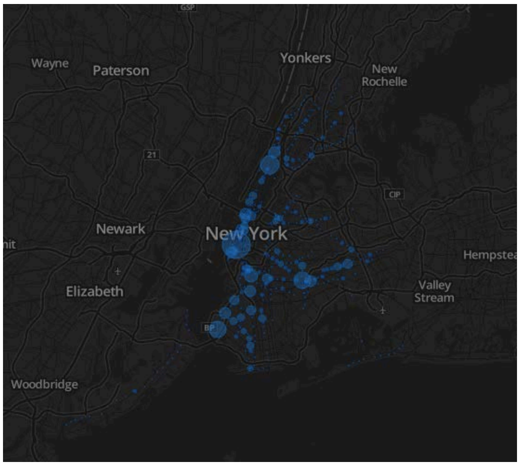

Fig. 4. Represents nodes geographically—an undirected graph. The node size is base upon its degree, or the number of MTA lines that pass through the stop.

6.5 Figure 5

The fifth visualization shows traffic patterns of vehicles on MTA bridges. This is again a geographical representations with nodes centered on bridge locations and traffic volume represented by both size and color.

7 Interpretation

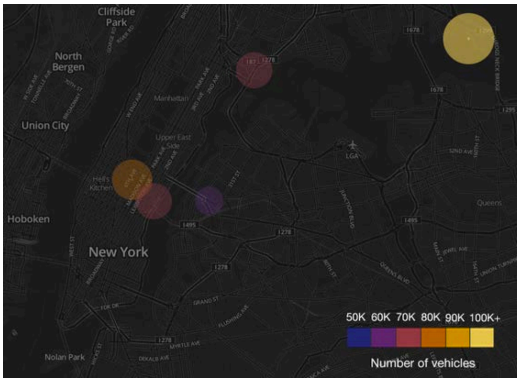

Fig. 5. Visual elements represent the number of vehicle entering New York via MTA bridges.

7.1 Evaluation Approach

This section contains evaluations of the visualizations, using a user-centered task-focused methodology detailed by Winckler in 2004 [12]. Each visualization will be assessed against the tasks identified by the methodology in order to determine the effectiveness of the visualizations with respect to the broadest audience. This is accomplished by considering the end user specifically and the typical tasks that one would perform with an information visualization. As discussed below, there are that tasks do not apply to some of the visualizations and others that are relatable to some visualizations more than others. After considering other taxonomies and evaluation methodologies, the one chosen was deemed the most applicable to the types of visualizations that we created.

7.2 Figure 1: Geographical View of Lines

This first visualization gives a user a quick overview of how MTA lines cover the geography of New York. This figure could be used by any of the audiences we described, however would be more useful if augmented will additional layers of information. For example, a transit user could decide which line goes between two trip end points, but the visualization does not provide information about what stops the traveler should use.

7.3 Figure 2: Mean Wait Time by Line

The second visualization builds from the first, but this time layers in detail about wait times. A transit user could use this visualization to decide which lines to avoid using. Transportation administrators could use this visualization to quickly see patterns among line wait times. This visualization would be particularly useful for monitoring any changes over time (either expected or unexpected).

Using Winckler's methodology [12] this visualization assists with the following tasks in these ways:

- Locate and Identify: locating subway lines is accomplished by using the map, while additional identification is done with respect to the chosen measurement
- Distinguish and Categorize: the colors associated with the average wait time measurement also serve to distinguish the lines from each other and categorize them appropriately
- Cluster and Distribution: clustering does not apply to this visualization but distribution can be detected by following the subway lines through the network using the color designations
- Rank: ordering of lines by the average wait time measurement is done by using the color spectrum at the top left portion of the visualization, where orange signifies a relatively low average percentage of on-time vehicles and green signifies a higher percentage of on-time vehicles
- Compare/compare within and between relations: comparisons among the subway lines are achieved by color differences with respect to the key given at the top left
- Associate and Correlate: associations and correlations are implied in this visualizations, existing only through geographical interpretation

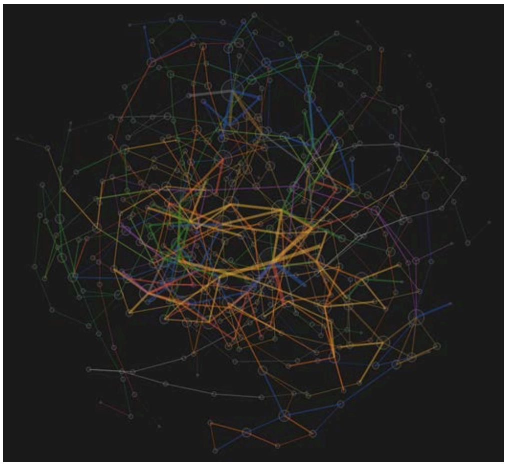

Fig. 3. Network visualization of MTA lines (edges) and stops (nodes). Stop location, number of MTA lines, and stop number is on node hover. Line number shows on edge hover.

7.4 Figure 3: Directed Graph

The directed graph visualization provides information that is conducive to comparing against "classified operations that a user might need to execute to analyze data" [12]. Highlights of this evaluation are:

- Locate and Identify: the overview of the visualization does not necessarily allow a user to quickly perform these tasks, but simple hover-over operations help to construct the mental model of a subway system
- Distinguish and Categorize: the colored lines to represent subway lines and the circles to represent stops or stations help to distinguish both lines from other lines and lines from stops and also categorizes the properties of the visualization into these two groups
- Cluster and Distribution: clustering does not apply to this visualization but distribution can be detected by following the subway lines through the network using the color designations
- Rank: ordering of nodes through parts of the network is achieved by finding the endpoints, which is made easier by the animation feature of the visualization, as the starting and ending points of each subway line are moved to the outer edges of the visualization. The size of the nodes also serves to rank the stations by the number of lines to which they are connected - the larger the node that represents a station, the more lines pass through it.
- Compare/compare within and between relations: comparisons among the nodes are achieved by size differences
- Associate and Correlate: associations and correlations are drawn simply, since the visualization consists of two primary objects and these circular nodes and lines are sorted out via an animation that obviates their organization

Of all the visualizations in this paper, this is the least accessible and useful for transit users. There is a small learning curve involved for understanding and using this figure, especially among an audience unfamiliar with directed graphs. Transit administrators who learned to use the visualization would be able to determine junction points and the areas where users are trying to accomplish the most diverse sets of transportation tasks. More options for line changes at a stop increase potential for confusion. Transit administrators could focus signage and station experience design resources at stations with the most connections.

7.5 Figures 4 and 5: MTA Stations and Bridges by Network Degree

The visualization for MTA stations and bridges by network degree is evaluated using the same criteria:

- Locate and Identify: the ample size of the bubbles, the contrasting color scheme, and the quality of the map serve to enhance the location and identification operations
- Distinguish and Categorize: these operations do not apply to this visualization
- Cluster and Distribution: clustering occurs naturally in this visualization due to the geographical nature and the rendering of the bubbles and distribution is based on this same geographical property
- Rank: the size of the nodes serves to rank the stations by network degree, where the largest bubbles represent the station locations with the highest network degree
- Compare/compare within and between relations: comparisons among the nodes are achieved by size differences
- Associate and Correlate: associations and correlations are implied in this visualizations, existing only through geographical interpretation

Of our three visualization user types, figures 4 and 5 are most useful to administrators. This group is able to determine high and low use areas, informing decisions about resource allocation. These figures do not directly help an individual in navigating the city.

8 Discussion

Various visualizations have been developed to address use cases for MTA data, both for transit system administrators and the general public. While some, like the "mean wait time by line" and "stations by network degree" use a geographical approach, others like the "as a directed graph" use alternate approaches that in many ways convey information more effectively than those that use maps as a central component. The directed graph has multiple features that add to its effectiveness: colored lines, circular nodes, nodes of varying sizes, an animated layout, meaningful edges (to signify the endpoints of subway lines), and hover-over text descriptions. These aspects of the visualization in different instances enhance the clarity of the components, make the visualization more readable, and ultimately provide the most comprehensive overview of the network of any of the visualizations. The clarity of the overview would appear to have the most value to the general audience, but the manner in which the node size and centrality properties highlight the most highly trafficked stations would likely have value to transit system administrators as well.

The geographically focused visualizations derive much of their value from this defining property itself, as the location feature will help riders determine which lines are more reliable in terms of wait time by where they would want to travel. Similarly, the nodes with the highest network degree would help riders determine the most useful stations in terms of connecting lines by location, implying that they could more easily get to an ultimate destination from these stations. Higher network degree also indicates to administrators where to direct the most resources and which stations will require general maintenance more frequently.

There are multiple aspects of the visualizations that support ideas covered in the research sections above, most notably the properties of network visualization discussed in the Kershenbaum paper as they relate to the "as a directed graph" [3]. Using the hover-over labels to add information and the use of color, size, and a superior layout algorithm are all mentioned as ways to enhance a network visualization and the directed graph in this paper incorporates all of these effectively. The shading of the nodes and the comparatively large node sizes in the "stations by network degree" visualization recall the idea of less stress on actual scale in a geographically based graphs noted by Kershenbaum [3]. Summarily, the visualizations covered in this paper use various techniques, all adding value in different ways to transportation system information visualization use cases. As the MTA makes even more data available, the possibilities for developers in terms of approaches that they can employ and results that they can produce will increase significantly, adding even more value to the intended audiences.

9 Conclusion

We recommend the MTA consider regular use of all the visualizations created for this paper. Thinking about these visualizations and the people who would be using them focuses our evaluation of their usefulness. Based on our research and findings, we believe that these visualizations would all be useful primarily for transit administrators to review on a regular basis to monitor overall system health. They would also be useful for point in time analysis if decisions had to be made about resource prioritization, including potential opening or closing of services.

These visualizations also have usefulness for MTA users while planning their rides, and for general use by third party or amateur analysis of the MTA system. With regular use of visualizations such as these, we feel that the MTA can increase transparency of the transit system, improve user satisfaction, and better monitor the overall performance of its operations.

Looking forward, further release of new data and quality-of-service statistics could bring added improvements to the range and quality of available visualizations. As shown by this paper, making data available for analysis engages the academic community and other interested parties to contribute to visualization creation. While most of the currently available statistics were about line performance, we would be interested in layering on information about experiences at individual stops. Metrics speaking to safety and accessibility would be particulary interesting in creating visualizations appealing to both the transit administrator and transit user audiences.

References

[1] E. Shannon and A. Bellisio, *The MTA in the Age of Big Data: Transforming the Wealth of MTA Data into Accessible, Meaningful, Visual, Interactive Information.* Permanent Citizens Advisory Committee to the MTA. March 2013.

[2] T. Kimpel, *Data Visualization as a Tool for Improved Decision Making within Transit Agencies.* Center for Urban Studies, College of Urban and Public Affairs, Portland State University. 2007.

[3] A Kershenbaum and K. Murray, "Visualization of Network Structures," *Journal of Computing Sciences in Colleges*, Volume 21 Issue 2, pp. 59-71, December 2005.

[4] Developer Data Downloads, [Online] http://web.mta.info/developers/download.html (Accessed: 16 March 2014).

[5] Google Developers Transit (GTFS), [online], https://developers.google.com/transit/gtfs/ (Accessed: 16 March 2014).

[6] D. Swartz, GitHub, [online] https://github.com/daveswartz/mapperly (Accessed: 16 March 2014).

[7] NetworkX, [online] http://networkx.github.io/ (Accessed: 16 March 2014).

[8] Python NetworkX library for node centrality and degree. http://networkx.lanl.gov/reference/algorithms.centrality.html

[9] M. Dewar, *Getting started with D3*. Sebastopol, CA: O'Reilly, 2012.

[10] MTA map colors http://web.mta.info/developers/resources/line_colors.htm

[11] M. Winckler, P. Palanque, P. Sabatier, and C. M. D. S. Freitas. "*Tasks and scenario-based evaluation of information visualization techniques,*" TAMODIA '04 Proceedings of the 3rd annual conference on Task models and diagrams, pp. 165-172, 2004.

APPENDIX

Digital versions of the visualizations in this paper available at https://github.com/stevepepple/nyc-mta-network/blob/gh-pages/README.md

Network Analysis of New York MTA for INFO 633 Information Visualization course.

- New York MTA Transit with colored lines
- New York MTA as a directed graph
- New York MTA mean wait time by line
- New York MTA stations by network degree
- MTA Bridge traffic