Washington Metropolitan Area Transit Authority (WMATA) Ridership

CMSC734 Homework #2
Allan Fong, PhD Computer Science

Introduction and Background

Every day thousands of individuals in the greater DC area use the WMATA Metro system to get from place to place. The Metro has collected a lot of information on riders including when and where they pass through turnstiles. In “the spirit of open data and collaboration,” the Metro has released data on station-to-station rider counts (1). The release was an open invitation for the public to analyze and visualize their data to find interesting insights and patterns.

Data Source

The data used for this assignment was from the WMATA released on October 31, 2012 (1). The data contained 86 vertices (one for every Metro station) and over 110,000 edges. Below are some technical notes about the data from the website:

- The data shows average ridership for each day of May 2012, excluding Memorial Day. (May is typically used as an “average” month, since it falls in the middle of seasonal swings, is relatively unaffected by extreme weather, etc.)
- Time period shows the time the rider entered (not the time they exited).
- AM Peak = opening to 9:30am
- Midday = 9:30am to 3:00pm
- PM Peak = 3:00pm to 7:00pm
- Evening = 7:00pm to midnight
- Late-Night = Friday and Saturday nights only, midnight to closing

Overview

I was very excited to analyze this dataset with NodeXL, because I thought there might be a lot of interesting insights and patterns. I had also been living in the DC area for over four years and often wondered about the Metro patterns. My first attempt to visualize all of the data in NodeXL failed because there were over 110,000 edges. I had to parse, sort, and aggregate the data in various ways to use it in NodeXL. I also color coded each station based on their respective line/lines. In addition, I chose a related color for stations with two lines. Stations with more than 2 lines were set to black. I used the default disc representation for stations that only have one line and a square representation for those that have 2 or more lines. I tried grouping by vertex attributes with the standard box layouts but the layouts generated were not intuitive. I found that in order to make the visualization more meaningful and understandable, I had to
maintain some of the geographical information for the stations. As a result, I modified a Fruchterman-Reingold layout by manually grouping the vertices with similar attributes and arranging them to be reflective of their geographical locations.

I tried to maintain some of the geographical clustering while still allowing users to read the names of the vertices and the interesting edges. I decided to use this layout for most of the remaining visualizations because a common layout allows for easier comparison and analysis. Although, I had to slightly shift some of the vertices in the various graphs to better show interesting edges and patterns, the overall spatial consistency of the vertices were kept. It is also important to note that in some graphs, vertices that did not have enough edge connections were hidden to simplify the visualization.

Headline 1: Farragut North, Farragut West, Metro Center, and L’Enfant Plaza are the work destinations for many who take part in the DC commuter’s rat race

I divided the weekly data into four time periods (AM peak, Midday, PM peak, and Evening) and displayed them in the figures below as “small” multiples. I did this to visualize and compare the different travel patterns on a typical weekday. The edge width varies between 1 and 5 while edge opacity varies between 5 and 100. The edge width and opacity correspond to the number of riders between each station (between 100 and 1900). I had to filter out edges because when they were all included, the visualization was too crowded and difficult to understand. One hundred average riders were chosen because edges with less than one hundred average riders accounted for more than half the data but less than 5% of the total riders. As a result of these changes the graphs are a lot cleaner and easier to understand.
The XY plot below shows the In-Degree versus Out-Degree metrics of the Weekday, AM-Peak travel patterns. I chose the AM-Peak rather than the PM-peak hours to understand typical commuter’s work patterns because it is less likely that people are traveling during the AM-Peak for leisure.
Although I expected Metro Center, L’Enfant Plaza, Farragut West, and Farragut North to have high In-Degrees (48, 45, 36, and 33 respectively), I was surprised by how high Union Station, Gallery Place and McPherson Square scored. This is reflected by the large count of the edges without accounting for the weight of the edges. Most people commute from Shady Grove, Union Station, and Vienna (having Out-Degrees of 24, 22, and 21 respectively).

Another interesting observation was that Farragut North, Metro Center and Union Station tend to be used more frequently during the midday period. These midday excursions could be for lunch meetings, work related activities, or other reasons. Lunch trips to Union Station, Gallery Place and Dupont Circle may also help explain the slightly darker and thicker lines, because Union Station, Gallery Place and Dupont Circle have good restaurant options. However, with the increasing popularity of food trucks in DC, it makes sense that a large number of people may not use the metro to travel for lunch.

The evening usage of the Metro during the weekday is small, except between Gallery Place and Columbia Heights. Columbia Heights has developed a lot over the past two years with a growing number of high rise apartment buildings and new shopping options. There has also been an increase in young adults moving into the area. This can help explain the edge that shows evening commutes between Gallery Place and Columbia Heights.
Headline 2: Gallery Place Chinatown, Dupont Circle, U Street, and Clarendon are Night Life Locations

The following network visualization shows the travel pattern of people using the metro late at night (between midnight and closing). Because there are significantly less people traveling during the late-night hours, I adjusted the edge width and opacity to better visualize interesting travel patterns. I also calculated the Out-Degree centrality to better understand these travel patterns.

![Network Visualization Diagram]

<table>
<thead>
<tr>
<th>Edge Properties</th>
<th>Riders</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>125.4</td>
</tr>
<tr>
<td>2.5</td>
<td>WLDW</td>
</tr>
<tr>
<td>3.0</td>
<td>125.4</td>
</tr>
<tr>
<td>Riders</td>
<td>Opacity</td>
</tr>
</tbody>
</table>

Gallery Place Chinatown, U Street, Dupont Circle, and Clarendon have the highest Out-of-Degree metrics, suggesting that these places have the highest degree of night life in the city. This trend is very evident in the visualization. Furthermore, those spending time in Dupont Circle will usually return to one of the stations on the Red line. Similarly, people from Gallery Place-Chinatown will typically disperse to Columbia Heights, Silver Spring, and Fort Totten. Fort Totten was initially a surprising destination for late-night wanderers but on reflection this may be because it is the first northeast station on the Green/Yellow line that has parking for metro riders.

Furthermore, it was quite interesting to see a VA and DC/MD divide. One might hypothesize that people living in DC prefer to gather in DC and people living in VA prefer to stay in VA. This is especially true at the Clarendon stop, which is the location of the social/bar scene for
Arlington. There is a clear contingent of people living in and around Vienna that like to stay out late in Clarendon. Although there is a separation between VA and DC/MD, the edges from Gallery Place to Crystal City and Pentagon City are still fairly strong. This is not surprising because it is easier for people living in Pentagon City and Crystal City to access Gallery Place more easily than Clarendon via Metro.

**Headline 3: To avoid crowded metros while sightseeing and shopping on the weekends, go out on Sundays**

The figures below display all the Metro travel patterns for Saturday and Sunday. I used the same edge width and edge opacity scaling between the days to better visualize and compare changes between Saturday and Sunday.

It was really interesting to see the contrast between Saturday and Sunday, primarily, the decrease in riders at almost all the stations. This could be reflective of free street parking on Sunday. Furthermore, the amount of people traveling to the Smithsonian from Virginia metros such as Pentagon City, Crystal City and Vienna remained consistent on Saturday and Sunday while other connecting edges, such as the one from New Carrollton, decreased drastically.

Furthermore, Pentagon City, Foggy Bottom, Dupont, Metro Center, Gallery Place, Columbia Heights and Union Station are heavily traveled most likely because of the large amount of shopping malls and dining options around these stations.

Not surprisingly, the Smithsonian and Navy Yard stations have a much larger percentage of riders on the weekends than on the weekdays. This makes sense because the Smithsonian station is a central location with easy access to a lot of museums; and the Navy Yard station is the closest one to the National’s baseball stadium. It is surprising, however, to see the relatively low
amount of visitors to both the L’Enfant and Archives stops, even though they are as close as, or even closer to some of the popular museums than the Smithsonian station.

**NodeXL Experience and Critiques**

Overall, I thought NodeXL was an extremely useful and helpful tool. With some data preprocessing, I was able to upload my data and start exploring the data set fairly quickly. I was able to color code the vertices and arrange them based on geographical location and attributes. This additional spatial coding helped me understand the data much better and faster.

I’ve tried visualizing networks before using UCINET. NodeXL is several magnitudes better than the free UCINET version I have used. I was very excited to see a free network visualization tool with so many capabilities and functionalities. Many of the interfaces were intuitive and easy to implement. Most importantly, I am thankful that I can download it as an Excel template. While working for the government, it was extremely difficult to install software on the computers. Even when I found helpful, free network visualization programs, it was extremely difficult to obtain permission to install them. That is why I was quite excited when I learned about NodeXL.

Even though NodeXL has many benefits, I do have a few critiques and suggestions for improvements beyond those already discussed in class.

- My data started off with 110,000+ edges and when I tried running some of the layout algorithms, Excel would freeze. Other times, I would run the clustering algorithms and the same thing would happen. I wasn’t sure if the program crashed or if it was still processing. I didn’t know if I should wait or restart. It would be helpful to provide some indication to the user on the program’s progress or the expected calculating time.
- It would also be helpful to have a filter option to hide or remove self-referencing edges. For this project, I wrote a short macro that did most of this cleaning. However, it would be beneficial to the user if such an option existed in NodeXL.
- Often times during this assignment, I wanted to compare multiple layout visualization of the data. I had to save my current layout as an image before applying any changes. It would be helpful to be able to create different coordinated layout windows of the same data. However, this would probably take a lot of work, and I would consider the benefits to be moderate.
- The legend options are also very limited and difficult to edit and manipulate. It would be helpful to have more legend controls available, such as changing the size and location of the legend or adding and removing descriptions in the legend.

**References**

(2) [http://www.wmata.com/](http://www.wmata.com/)
(3) [http://nodexl.codeplex.com/](http://nodexl.codeplex.com/)