Visualizing MBTA Data

A day in the life of the T

May 14, 2010

Joel Greenberg
Leslie Johnston

Tufts University
COMP 150-04: Visualization

1.0 Introduction

The Massachusetts Bay Transportation Authority (MBTA) subway system in Boston, affectionately known as the "T," has 481,300 riders each weekday. This averages to 12,666 riders per mile – the densest ridership of any subway system outside New York City. The red, orange, and blue lines alone make the T the 4th largest subway system in the United States. Added to that are the Ashmont-Mattapan high speed line and the green line; together these make up the busiest US light rail system at 255,100 riders per weekday. (APTA, 2010)

The T is undergoing many large- and small-scale changes. The blue, green, and silver lines are currently seeing updates and extended service. A new project proposes an "urban ring," providing a wheel to Boston's "spoke" subway system. It is estimated the urban ring could have as many as 170,000 riders per weekday. (EOT, 2008)

It is impossible to implement or even discuss the ramifications of these changes without a way to visualize and assess the ridership data on the MBTA. A visualization of MBTA data, whether historical or real-time, can greatly improve both an individual's understanding of Boston subway traffic, but also the MBTA's understanding of how people are using the T and what types of measures should be made to alleviate traffic issues. A good visualization can show interesting data and potentially help MBTA officials decide how to hire drivers, allocate train cars, upgrade lines and increase ridership.

2.0 Related Work

Some work with this data was completed based on the MBTA visualization challenge issued by the Massachusetts Department of Transportation (MassDOT). Many of the visualizations created in response to this challenge are able to show data for one purpose only, whether it is as a video showing trends for the entire day in just a few minutes, or as a series of graphs showing trends for specific lines. Some visualizations are more customizable than others, but few allow adequate customization to get at many different aspects of the data at once. (MassDOT, 2009)

The current MassDOT challenge looks for software applications for computers and smartphones to deal with real-time data, but it is important not to leave historical data completely behind. The historical data is certainly still relevant, and an application reading in from a historical data file can easily be adapted to show real-time data. Many of the same issues associated with dealing with large amounts of historical data also apply when focusing on real-time information.

3.0 Design Considerations

Many considerations affected our design based on qualities inherent to the data set as well as the potential uses for the data. The following questions represent some of the major considerations for designing a visualization for this data.

- Size of the data set
  - How to make large amount of data visible at once?
  - How to focus in on small sections of the data?
- Nature of the data set
  - How to best show historical data from a single day in time?
  - How to deal with timestamps that go to the exact second?
    - If showing data within a certain time range at once, what should the size of the bins be?
  - How to show data effectively during both rush hour and off-peak times?
- Use of the data visualization
  - Who will use the data?
  - How will they use the data?
  - What are some future possibilities for this visualization (e.g., expanding to real-time data or average data)?
- Application of visualization
  - Is this best implemented as a smartphone application or computer executable?

4.0 Key Features

Design highlights

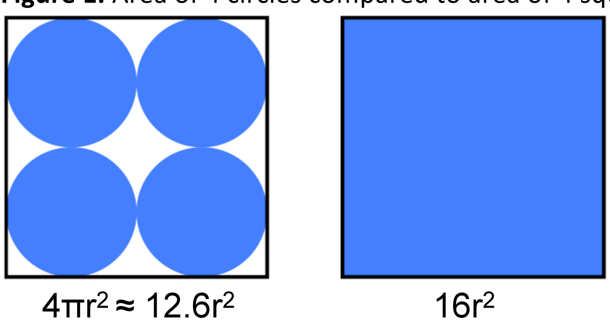
The visualization includes a number of features that are both unique and important for looking at map data.

- Customizable data views
  - The visualization is customizable in ways that most map visualizations are not. In addition to being able to turn off each of the different subway lines and the water displayed in the background, this visualization allows the user to walk through time on an hour-by-hour or minute-by-minute basis, and to change the size of a time-step. This means the user can look at data anywhere from 2 minutes at a time to several hours at a time. Using this feature, viewers can get an idea of all the traffic on the T in an entire day, or trends that are occurring on a minute-to-minute basis. All bin sizes between this are useful in different ways that only the user can determine based on their goals for using the visualization.
- Simple and accurate background data
  - The visualization is based on a drawing made from an accurate map. We chose not to use Google Maps API or another satellite or roadmap image, because these display too much additional information that isn't relevant to the subway data. We also chose not to use an overly simplified rendering of the subway system, rendered to show most

subway lines as straight as possible. We felt that this made the orientation of the subway lines and stations difficult to understand as it relates to a realistic map of Boston. Therefore, we chose to make our own image that would then allow us to turn lines on and off. The water image, which provides a little context to the city of Boston, can also be turned off to help the viewer concentrate on the relevant data. We strove to make the data-ink ratio as high as possible without oversimplifying the map.

- Zoom and scale functionality
    - Users can zoom in on the data, and scale the data points. Because the data points are only displaying relative traffic volumes, users can manually resize all of them to enlarge smaller data values for easier viewing or decrease larger data points to reduce the clutter on the screen. The ability to zoom and scale promotes exploration of the data in different ways.
- Square area data points
    - The symbols representing the data in the visualization are squares, rather than circles, which are more commonly used. We felt that people can better perceive and compare areas of squares than areas of circles. To illustrate this, Figure 1 shows how 4 circles together, representing an area of approximate 12.6 times the radius squared, looks similar to the area of 4 squares together, with an area that would be equal to 16 times the radius squared. To minimize the ambiguity caused by ignoring this "white space," we elected to use squares to represent relative data values.

Figure 1. Area of 4 circles compared to area of 4 squares.



Visualization controls

Although the controls of the visualization are not obvious, they allow for any computer to be able to navigate the data faithfully. There is an expectation that the user is seeking accurate data—therefore, some conditions are not 'protected.' For example, it is possible to view the map at nine o'clock and negative ten minutes, although the data rendered is meaningless.

The executable file may be run from the Terminal, and it requires the –Xmx1024M command to be passed in order to load all of the data. When the visualization is active, the following controls may be used to navigate over the two-dimensional map and through the day of August 12, 2009.

- Arrow Keys: The arrow keys allow the user to move about the map.

- Zooming in and out: The ‘M’ key zooms in to the map; the ‘N’ key zooms out. The view of the map changes but no squares are resized upon zooming.
- Adjusting global square size: A paradigm of this project was to view relative data—not exact numbers. The ‘L’ key increases the size of all squares on the screen, and the ‘K’ key decreases them. The principle of relative square areas representing traffic is maintained at any global standard.
- Adjusting the current time: The numeral ‘7’ increases the hour, while the numeral ‘6’ decreases it. The numeral ‘9’ increases the current minute, and the numeral ‘8’ decreases it.
- Adjusting the time-frame: The time-frame is defined as the plus/minus around the current time. A time-frame of fifteen minutes represents a duration of thirty minutes, fifteen in either direction of the current time. This may be incremented in steps of five with the ‘P’ key and decremented with the ‘O’ (letter) key.

5.0 Results

The following figures are a representative sample of screen shots from the interactive visualization of MBTA data.

Figure 2. Overall trend.

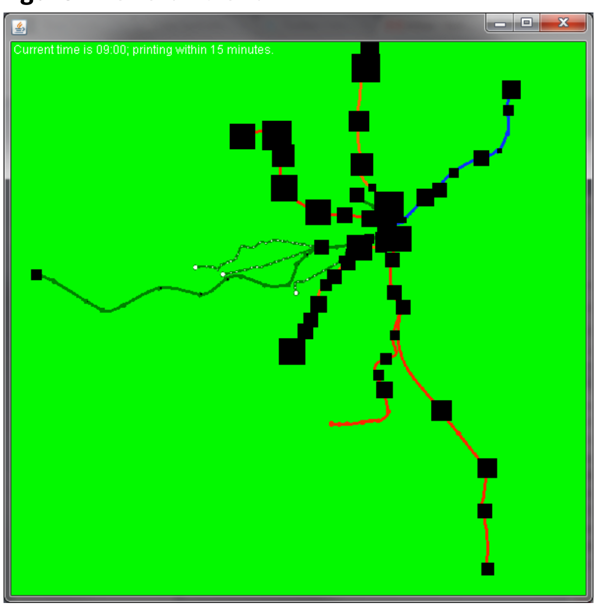


Figure 3. Point trend.

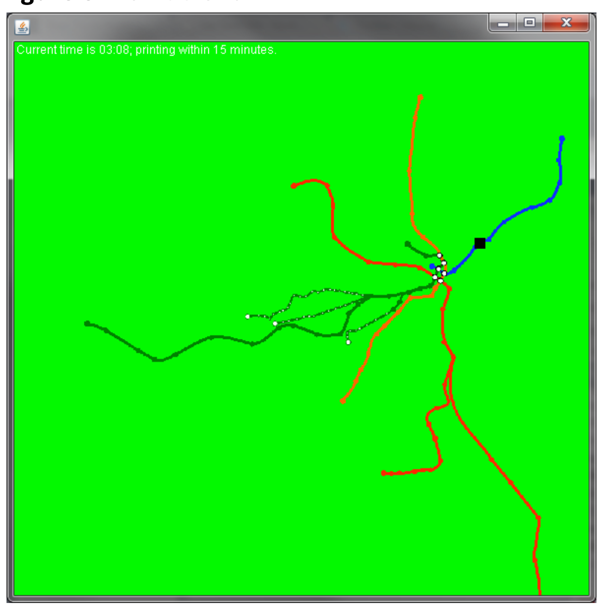
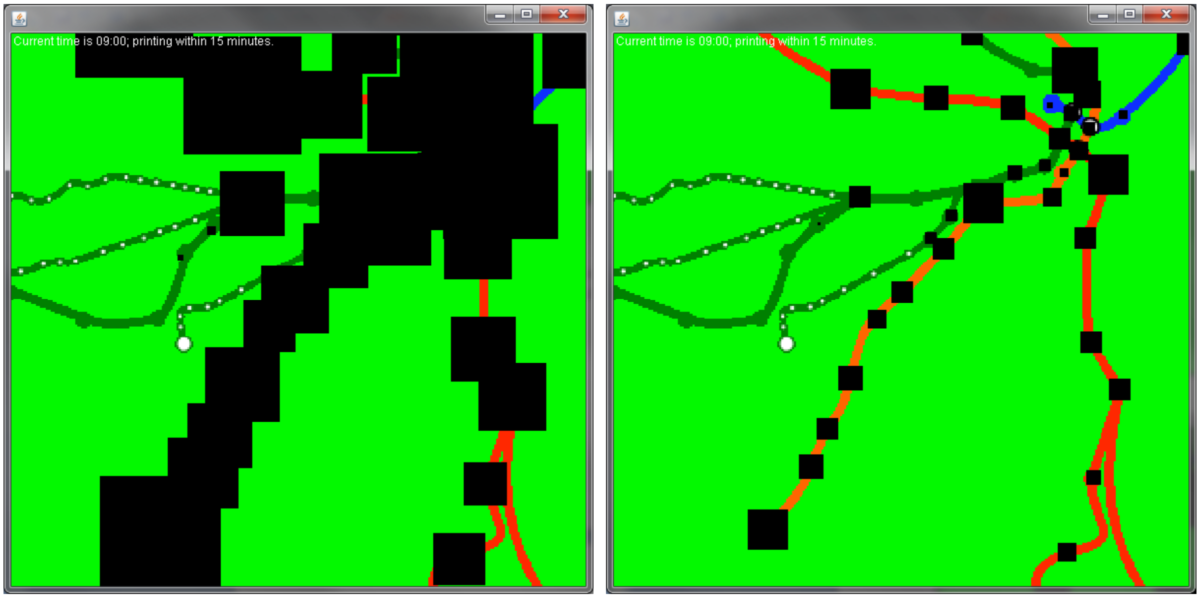


Figure 2 shows how this visualization could potentially show overall trends in T ridership. Each black square represents relative data for traffic volume entering the station on the map. Meanwhile, Figure 3 shows an example of a “point trend,” or a trend that may be unique to that particular day and time. In this case, there is one lone “rider” at 3:18am, when the T is closed.

Figures 4 and 5. The advantages of scalable data points.



Figures 4 and 5 show how the data can be scaled to reduce the clutter and help the user to better judge relationships between data points.

6.0 Discussion

The MBTA visualization as it current exists has a number of strengths that make it a useful visualization tool as well as several weaknesses that may need to be addressed in future work.

Strengths

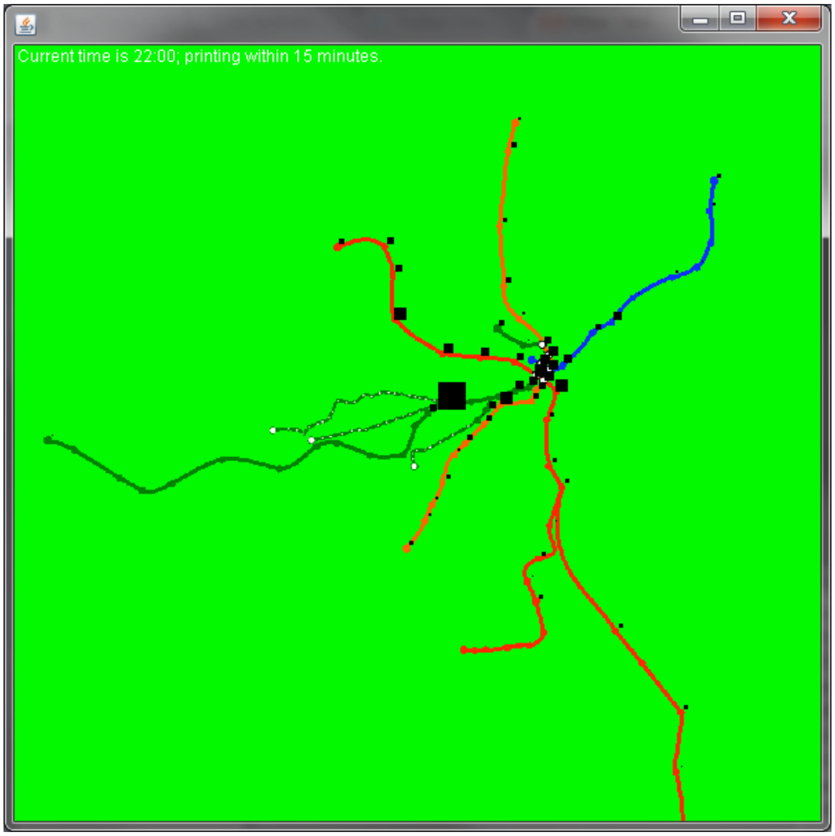
- Viewing overall trends
  - The visualization shows overall trends for the day. As the user plays through each hour of the day, they can see total traffic expand and diminish around rush hour and lunch breaks. With a larger window, they can see which stations see the most traffic throughout the day.
- Viewing point trends
  - Because the visualization allows for such precise movement through time, a variable time window, and scalable data points, users can see everything from the spike in traffic at Kenmore at the end of the Red Sox game, to the single "rider" on the T at 3:18am.
- Viewing relative information
  - The visualization displays relative data, making it simple and straightforward to compare data points to each other. A data point that has twice the area of another data point represents a traffic volume that is twice as high.
- Exploring data

  - The ability to zoom, scale, and play through time steps at hour- or minute-long intervals promotes exploration of the data. Users can move through the data as they wish, searching for expected trends (such as rush hour), exploring the lack of expected trends (such as the deficit of riders on the green line), and finding unexpected trends in the process (such as the big lead up to rush hour beginning around 3pm).

Weaknesses

- Viewing exact data values
  - The visualization does not show exact data values, or even give a scale to help viewers determine ballpark data values. While the viewer can see trends in the data, it isn't possible to get an idea of exactly how many people are entering the T. Although such data is available elsewhere and probably easier to see in a table form, viewers may still wish to have access to this data while looking at the overall trends.
- Guiding viewers toward trends
  - The visualization does not guide users toward finding these trends, but rather leaves it open for them to discover (or not discover) them on their own. One example of a guide to these trends could include a timeline of the day, marking things like rush hours, lunch hour, and the start or end of any unique events (such as the Red Sox game, located near the Kenmore T stop, ending around 10pm).
- Showing data for a single day in time
  - It is easy to look at this data and assume certain general trends about T usage in Boston. However, the data is only reflective of one day in history - a day which may not accurately represent usage. For example, this particular day (and the other day's worth of data available from MassDOT) had a Red Sox game. This may be a relatively frequent occurrence in the summer or early fall, but is certainly not representative of all days throughout the year. Knowing this, we can easily account for the trend in the data showing heavy traffic through the Kenmore station, but there may be other flukes in the data based on this particular day that we are not aware of. Things like the start of a rainstorm, particularly cold weather, or a large event may also affect the traffic for this day, so showing several days or an "average" of days might be beneficial as well.

Figure 6. Ridership at Kenmore stop at 10pm (end of Red Sox game)



7.0 Future work

Future work for this application can include expanding the functionality and the data set to better suit the viewer's needs.

- Timeline (with trend data)
    - A timeline showing some possible causes of trends, such as the start and end of the Red Sox game, could better guide users toward the potential trends in the data.
- Turn lines on and off
    - While it is currently possible to turn off the image of the line itself, it would be helpful in the future to allow users to turn off the data for certain lines to help them focus in on the information that is most relevant to them.
- Add mouse over effects
    - Mouse over effects could display the actual traffic volume value for a given data point, as well as the station name. This extra information could help the user better understand the magnitude of certain trends.
- Real-time data (with Twitter integration)
    - In the future, this application could be run with real-time data and include tagged Twitter updates about the status of certain stations, lines, events or weather going on in

Boston. Condensed to a smartphone application, this could be incredibly useful for T riders and MBTA officials alike as it can give people a better understanding of trends as they are occurring.

REFERENCES

American Public Transportation Association (APTA) (2010). Public transportation ridership report, fourth quarter 2009. Retrieved on May 10, 2010, from <http://www.apta.com/resources/statistics/Documents/Ridership/2009_q4_ridership_APTA.pdf>

Executive Office of Transportation (EOT) (2008). Urban ring phase 2: proposed circumferential transit project. Retrieved on May 10, 2010, from <https://www.commentmgr.com/projects/1169/docs/FenwayJune9_PublicMeetingFinal_06-09-08.pdf>

MassDOT (2009). Visualization challenge home. Retrieved on May 10, 2010, from <http://www.eot.state.ma.us/default.asp?pgid=content/developer_VizChallenge&sid=about>

It should be noted that some code from a project for COMP 50-GD was used to create motion and a canvas. That project is also available open-source, and may be found at http://code.google.com/p/tuftscoregame/.