CS591-L1
Nikolaj Volgushev

Graph Metrics for the Public Transportation Network

Introduction

The Boston public transportation system is a complex network of hundreds of stations, routes, and connections. While the T rail network only consists of seven lines there are over 100 bus routes [4] connecting different parts of the city. This complexity makes it hard to pin down shortcomings in the overall station layout or gain concrete insights into its structure. With projects on improving the transportation network by expanding and modifying existing T lines underway it is crucial to create metrics against which to measure the quality of existing (or hypothetical) stations and routes. As of now efforts [3] give a rigorous, computational analysis of the utility of existing stations as well as commuting trends. This type of analysis relies on the pre-existence of rich commuter and station usage data which does not lend itself to preliminary station layout planning when such data is not yet available. We propose to investigate two usage-data-agnostic metrics: structural station importance and route similarity. To this end, we model the transportation network as a graph and apply two popular algorithms from the domain of citation ranking: PageRank [1] and SimRank [2].

Preliminaries

PageRank is an algorithm developed and employed by Google to rank the relative importance of each website indexed by the Google search engine. These rankings are then used to improve search result quality by prioritizing sites with high rank among all sites relevant to a given search query. PageRank relies on link topology to calculate website importance; it " works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites." [1]

Underlying PageRank is a stochastic model approximating the behavior of an average web-surfer, referred to as the *random surfer model.* The random surfer model assumes that a web-surfer starts on a random website and begins following links from that website at random thus walking the web-graph. Apart from following links, the random surfer has a set probability of jumping to any website without following a direct link. In this model the PageRank of a website represents the probability that the random surfer will land on the website (or, equivalently, the expected amount of time the surfer spends on it).

Formally, PageRank can be expressed as follows:

$$\mathrm{PR}(A) = \frac{1-d}{N} + d\left(\frac{\mathrm{PR}(B)}{L(B)} + \frac{\mathrm{PR}(C)}{L(C)} + \ldots\right)$$

where d is the damping factor, N is the total number of pages, and L(X) is number of outbound links of X.

SimRank is closely related to PageRank, however instead of capturing importance, it captures the similarity of two given nodes within a graph. In the original paper [2], several concrete use cases are proposed, ranging from finding similar documents (textual as well as in the world-wide-web) to similarity-based clustering. SimRank captures the following intuitive observation: “Two objects are similar if they are related to similar objects.” [2]

The underlying stochastic model is a variation of the random surfer model. Given a random surfer starting on one node and a second random surfer starting on the other node SimRank reflects the two surfers expected meeting distance.

SimRank has the following formal definition:

$$s_1(a,b) = \frac{C_1}{|O(a)||O(b)|} \sum_{i=1}^{|O(a)|} \sum_{j=1}^{|O(b)|} s_2(O_i(a), O_j(b))$$
$$s_2(a,b) = \frac{C_2}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s_1(I_i(a), I_j(b))$$

where C1 and C2 are damping factors I(n)/ O(n) are the in/out-neighbors of a node.

The above formulation applies to the bipartite SimRank where the graph is split into two groups of nodes and edges only go from one group to the other.

Based on the above, we develop our own two metrics, *Station PageRank* and *Route SimRank* (future work includes coming up with catchier names ;). We use *Station PageRank* to measure the importance of bus and T stations under varying adjacency models and *Route SimRank* to measure the similarity of popular bus routes and T lines. Among other things, *Station PageRank* gives an intuition for the structural accessibility of a station while *Route SimRank* captures potential route redundancy as well as *route centrality*. Further, we propose a model for interpreting our results akin to the random surfer model. We call this model the *random commuter model*.

Used Datasets

Official MBTA station data. This dataset contains all existing T and bus stations along with station names, geo-coordinates, and “unique” identifiers. NOTE: there is no trivial way of deriving a mapping from routes to stations and the linkage between stations.
url: http://www.mbta.com/uploadedfiles/MBTA_GTFS.zip

Bus routes and schedules. This dataset contains bus routes, stops, and schedules. We use this dataset to derive bus route to stop mapping and stop linkage. The website is maintained independently from the MBTA. We only use data for the following bus routes because of overhead for data cleanup: 1, 9, 16, 23,

39, 47, 57, 66, 70, 83, 86, 87, 89, 101, 105, and 116.
url: http://www.mbtainfo.com/

T lines topology. This dataset contains linkage for all red, blue, and orange T stations and well as a graphical layout of the topology. We use station-network.json and spider.json.
url: https://github.com/mbtaviz/mbtaviz.github.io/tree/master/data

Green line topology. This dataset contains linkage for all green T stations and well as a graphical layout of the topology. We use station-network.json and spider.json.
url: https://github.com/mbtaviz/green-line-release/data

Derived Datasets

Station PageRank (direct adjacency, T only). PageRank of T stations based on direct connections only.

Station PageRank (direct & geo-adjacency, T only). PageRank of T stations based on direct connections as well as connections resulting from geo-adjacency (two stations are adjacent if they are within 500m for each other).

Station PageRank (direct & geo-adjacency, T and bus routes). PageRank of T stations and bus stops based on direct connections as well as connections resulting from geo-adjacency.

Route SimRank. SimRank of T lines and bus routes.

Methods

As previously discussed, we evaluate the metric of Station PageRank and Route SimRank on T and bus data. We computed three different versions of PageRank. For the simplest case, we only considered edges arising from direct connections between stations. Secondly, we considered geo-adjacency, i.e., in addition to edges resulting from direct connections we add edges between stops that are within 500m from each other. Lastly, we consider the transportation graph when several popular bus lines are included. We give the following interpretation to the PageRank of a station. Consider a random commuter, i.e., a commuter who gets on at a random stop and begins travelling the network along adjacent stations. Under this random commuter model the PageRank of a station represents the proportion of time a random commuter will spend at that station. We denote this as the *random commuter model*.

For Route SimRank we interpret our data as a bipartite graph with stations in one group and routes in the other. There is an edge from a station to a route if the station is directly associated with the route or if there is a station of that route geo-adjacent to the given station.

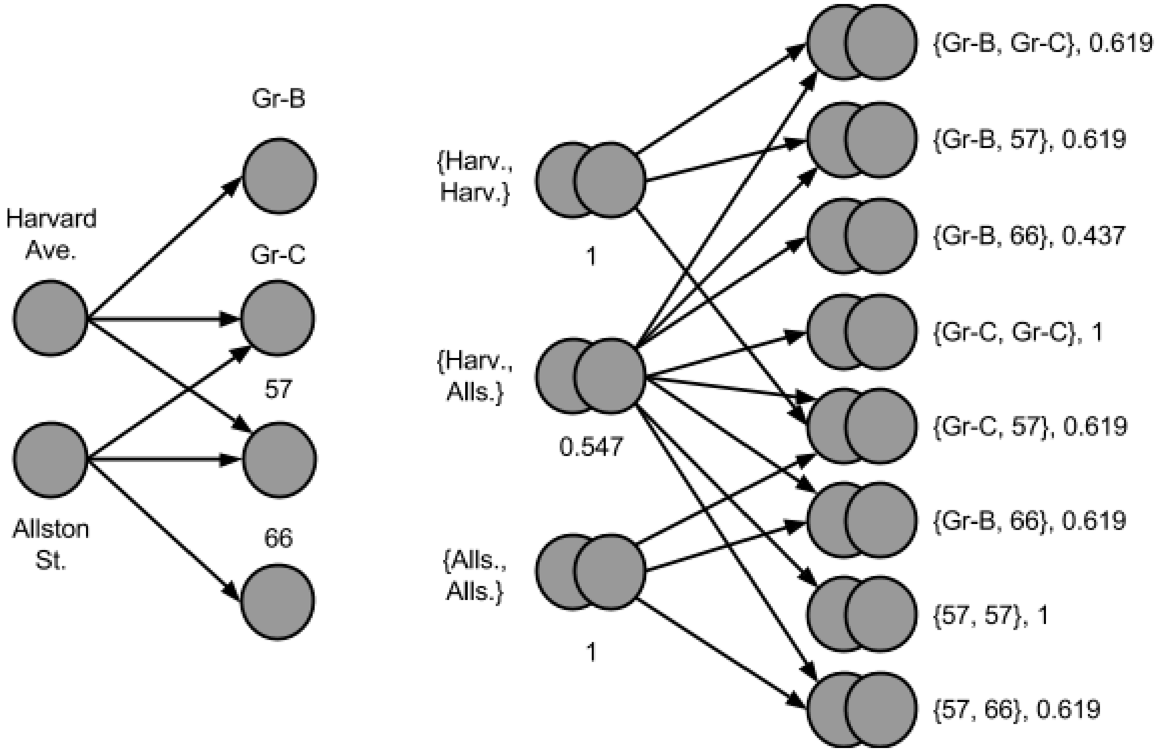


Figure 1. Example route SimRank.

Figure 1 gives an example station-route bipartite graph and the associated computed SimRank values.

Transformations

Geo-adjacency. For each station we compute and store all stations within a k-meter radius. This serves as intermediate data for the following PageRank computations but also indicates the vertex degree of each station.

PageRank of T network only. We computed the PageRank of nodes in the the graph induced by the green, red, blue, and orange T lines. Two stations are adjacent if there is a direct train connection between them.

PageRank of T network with geo-adjacency. In this model we include edges between directly connected stations but also when stations are within 500 meters of each other (indicating a short-walking distance).

PageRank of T and bus network with geo-adjacency. Same as the previous transformation but also includes the above-mentioned bus routes.

SimRank of T and popular bus routes. We computed the SimRank of the green, blue, red and orange T lines and the following popular bus routes: 1, 16, 23, 39, 57, 66, 70, 83, 86, 89, 101, 105, 119.

Results

Further we discuss the results of our analysis. We begin with our findings on Station PageRank.

Station PageRank

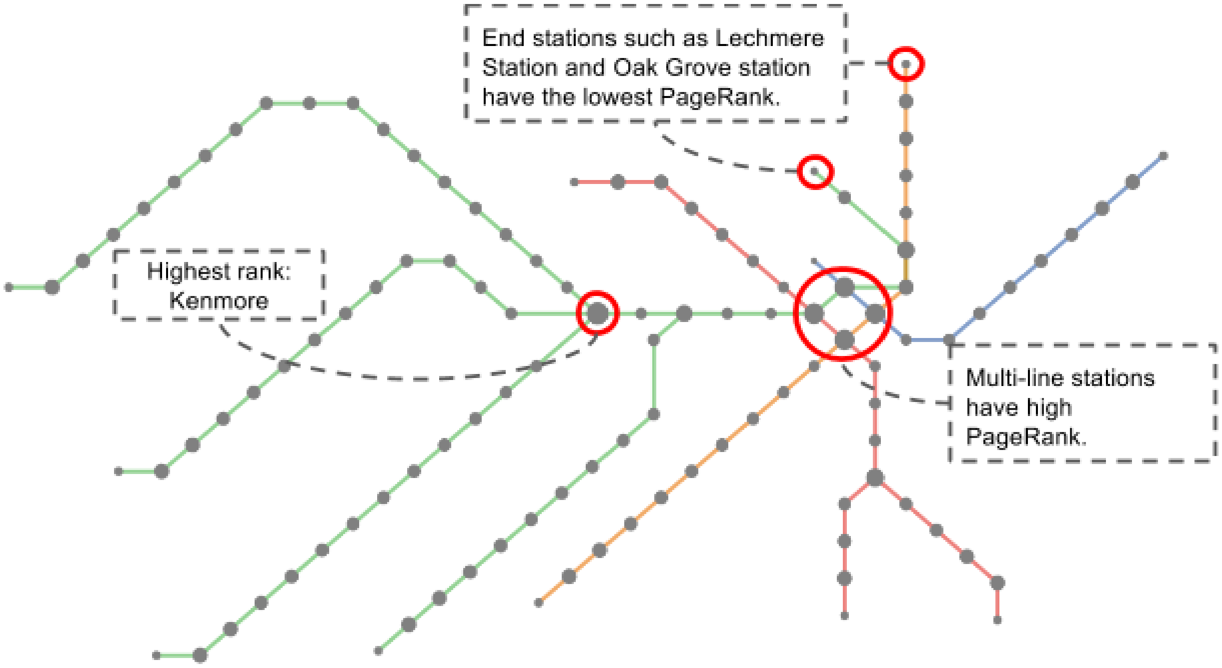

End stations such as Lechmere Station and Oak Grove station have the lowest PageRank.
Highest rank: Kenmore
Multi-line stations have high PageRank.

Figure 2. Direct connection adjacency. T stations only.

The Station PageRank based on only direct adjacency gives expected results. Stations that host multiple lines have high PageRank while end stations have lower PageRank. The station with the highest PageRank is Kenmore. We note that stations which might seem important to a Bostonian such as Central Square have low PageRank. The T network alone does not provide good access to these stations.

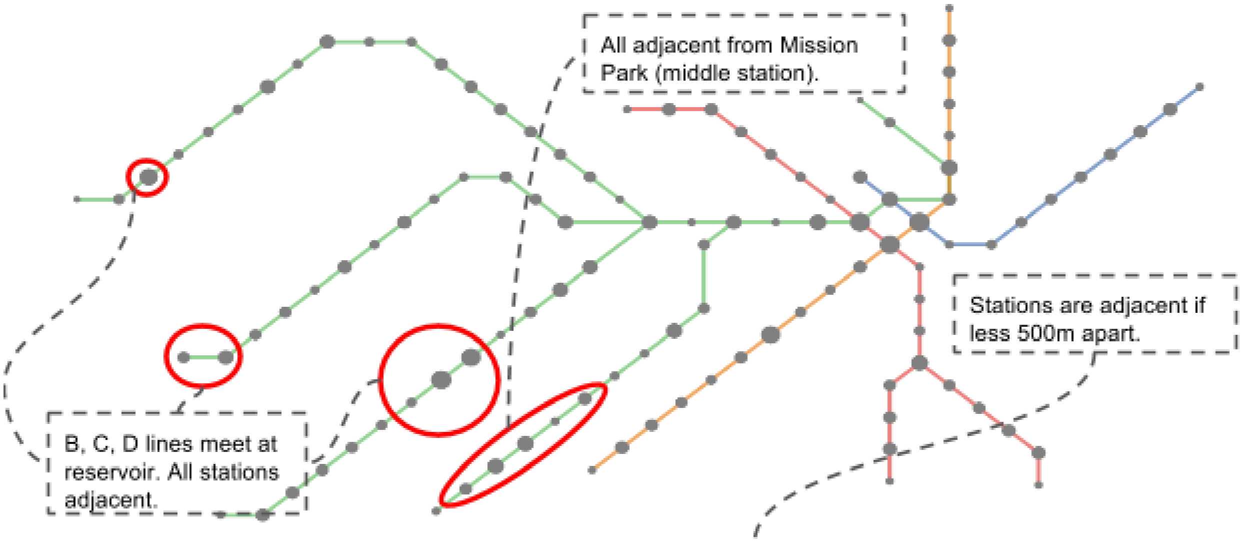

All adjacent from Mission Park (middle station).
Stations are adjacent if less 500m apart.
B, C, D lines meet at reservoir. All stations adjacent.

Figure 3. Direct connection and geo-adjacency. T stations only.

Geo-adjacency based PageRank exposes areas, especially along the B and E lines of stations that are very close together. Mission Park on the E line for example has high PageRank since there are two stations two hops over which are all within 500 meters of the station. We also note that the stations on the B, C, and lines by the Chestnut Hill Reservoir have very high PageRank for the same reason.

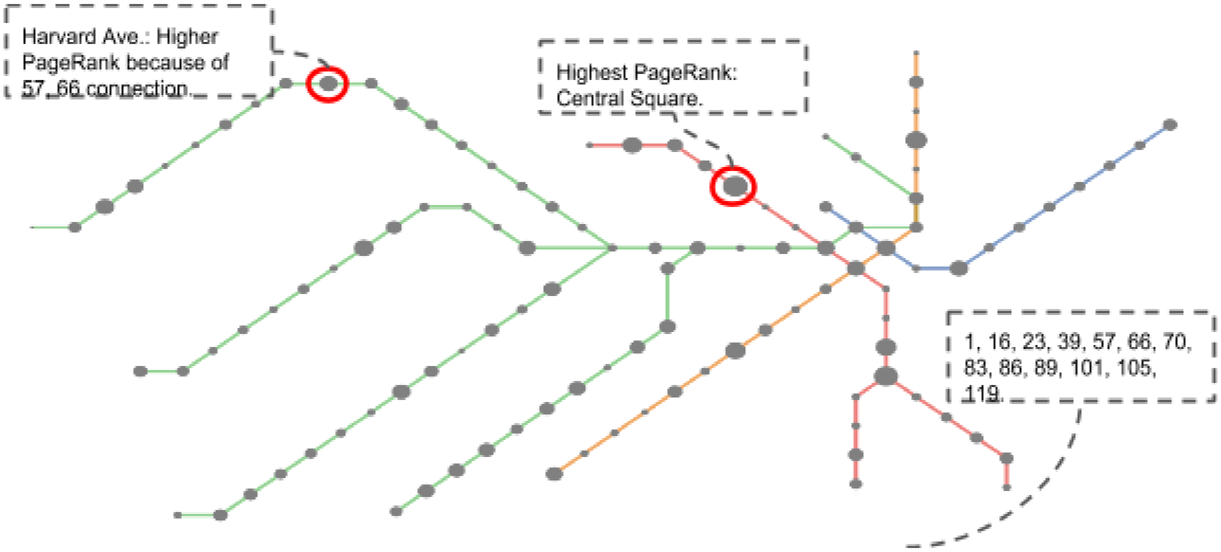

Harvard Ave.: Higher PageRank because of 57, 66 connection.
Highest PageRank: Central Square.
1, 16, 23, 39, 57, 66, 70, 83, 86, 89, 101, 105, 119.

Figure 4. Direct connection and geo-adjacency. Bus routes included.

Lastly, we include bus stations of the following popular bus routes in our analysis: 1, 16, 23, 39, 57, 66, 70, 83, 86, 89, 101, 105, 119. The physical locations of the routes are depicted in Figure 6. We note that including the bus routes changes the PageRank drastically. Central Square and Harvard St in Allston for example now have a much higher PageRank.

Route SimRank

Further, we discuss our evaluation of Route SimRank.

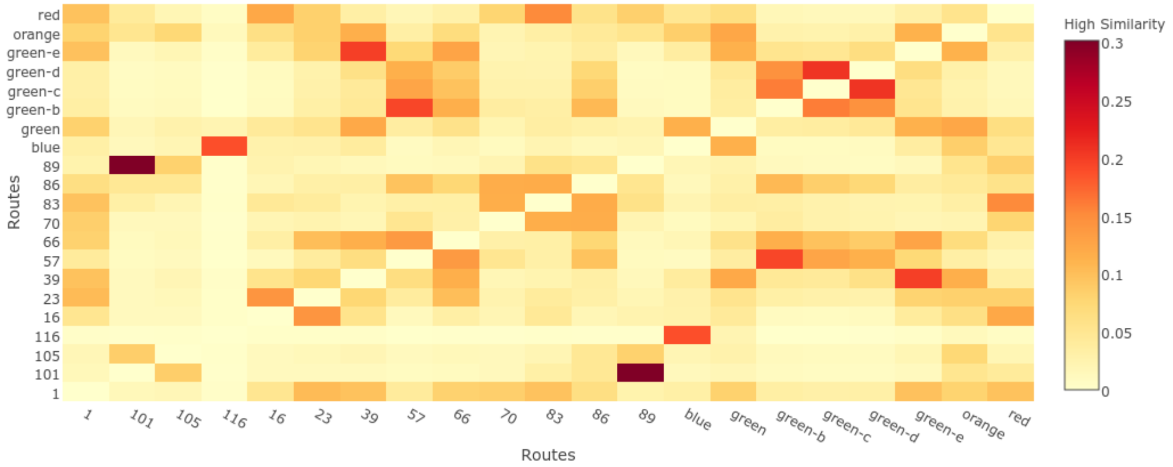
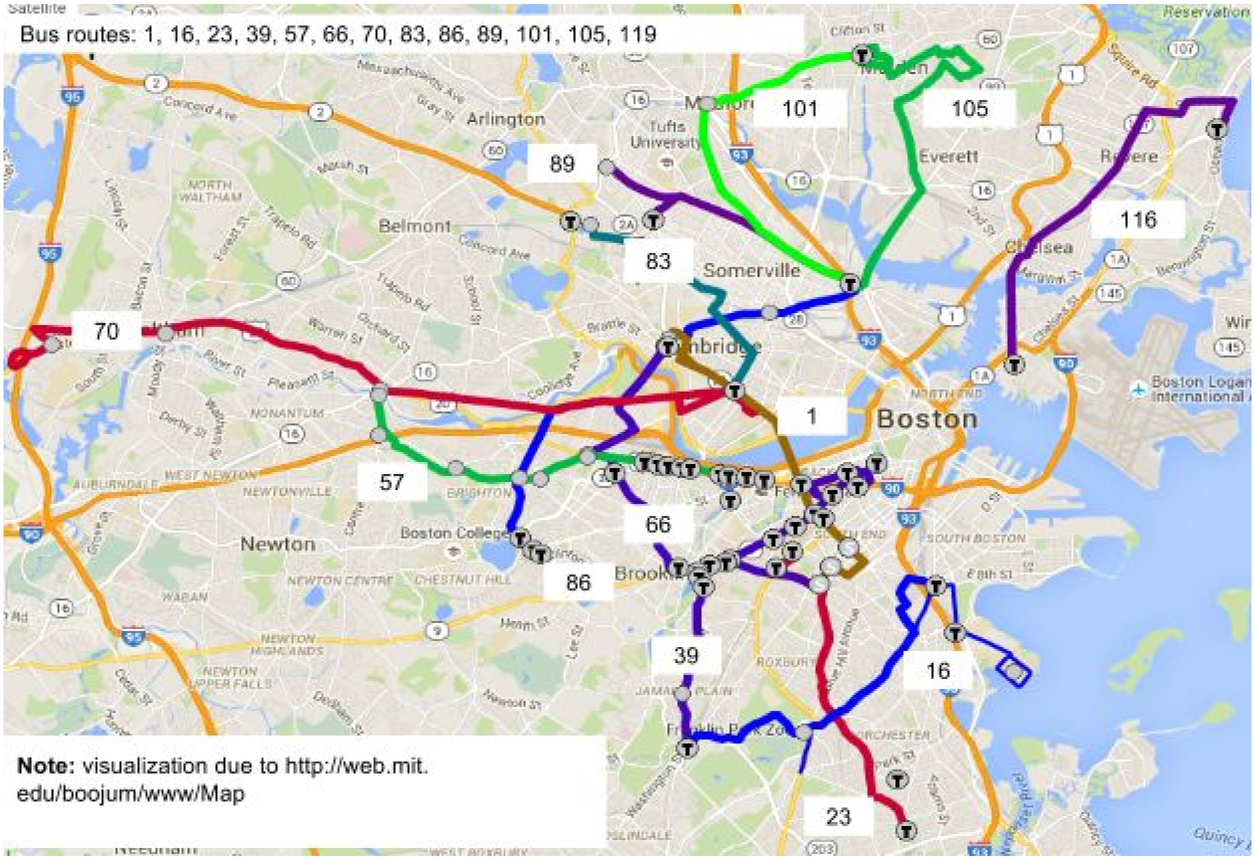


Figure 5. Route Simrank.

We present all pairwise Route SimRank values as a heatmap in Figure 5. Note that while the SimRank of a route with itself is 1, we present it as 0 in the heatmap to maintain a better contrast for the SimRank values of the other tuples. The two routes with the highest SimRank are 89 and 101 with a SimRank of 0.30. In Figure x we see that these routes do in fact overlap for about a quarter of the 89 route. The 39 route is most similar to a T-line where SimRank(39, green-e) is 0.20. The 116 has lowest cumulative SimRank with 1.26. The route with the highest cumulative SimRank is the 66 bus route with 2.25 even though its individual SimRank values are not extreme. This indicates that the 66 bus overlaps with many routes without sharing too many stations with any one other route. We note that identifying routes with high cumulative SimRank could be valuable because it gives indication to that route's centrality in the overall transportation system. Intuitively, if a route with high cumulative SimRank is experiencing delays many other routes should be affected also. This provides a potentially valuable heuristic for traffic analysis. It would be interesting to evaluate this claim empirically.



Bus routes: 1, 16, 23, 39, 57, 66, 70, 83, 86, 89, 101, 105, 119

Note: visualization due to http://web.mit.edu/boojum/www/Map

Figure 6. Bus routes locations.

We have discussed several findings based on Station PageRank and Route SimRank. To draw any further conclusions regarding the potential utility of these metrics a more thorough analysis should be conducted. We address this and more future work in the following section.

NOTE: for the raw PageRank and SimRank values we refer the reader to the derived datasets available as part of the datamechanics repository.

Conclusion and Future Work

We have developed two new metrics to evaluate the quality of the existing MBTA transportation network. Our metrics, Station PageRank and Route SimRank do not rely on rich commuter data and only on graph topology which makes them appealing for preliminary station planning. We have proposed an interpretation framework, the random commuter model for these values. Lastly, we have developed several new insights such as the effect of bus routes on T station importance. Interesting future work includes running the algorithms on all available MBTA data, as well as on a "role-model" city with a good transportation network such as Tokyo and studying the resulting values. Comparing the PageRank values of stations to the relative importance, i.e., availability of services, employment levels, population density, etc., of their physical locations might further uncover shortcomings in the overall transportation system.

References

[1] Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (1999) The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab.
[2] Glen Jeh and Jennifer Widom. 2002. SimRank: a measure of structural-context similarity. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02). ACM, New York, NY, USA, 538-543. DOI=http://dx.doi.org/10.1145/775047.775126
[3] Visualizing MBTA Data. http://mbtaviz.github.io/
[4] Massachusetts Bay Transportation Authority. http://www.mbta.com/