The MBTA transportation network as a graph: exploring the random commuter model

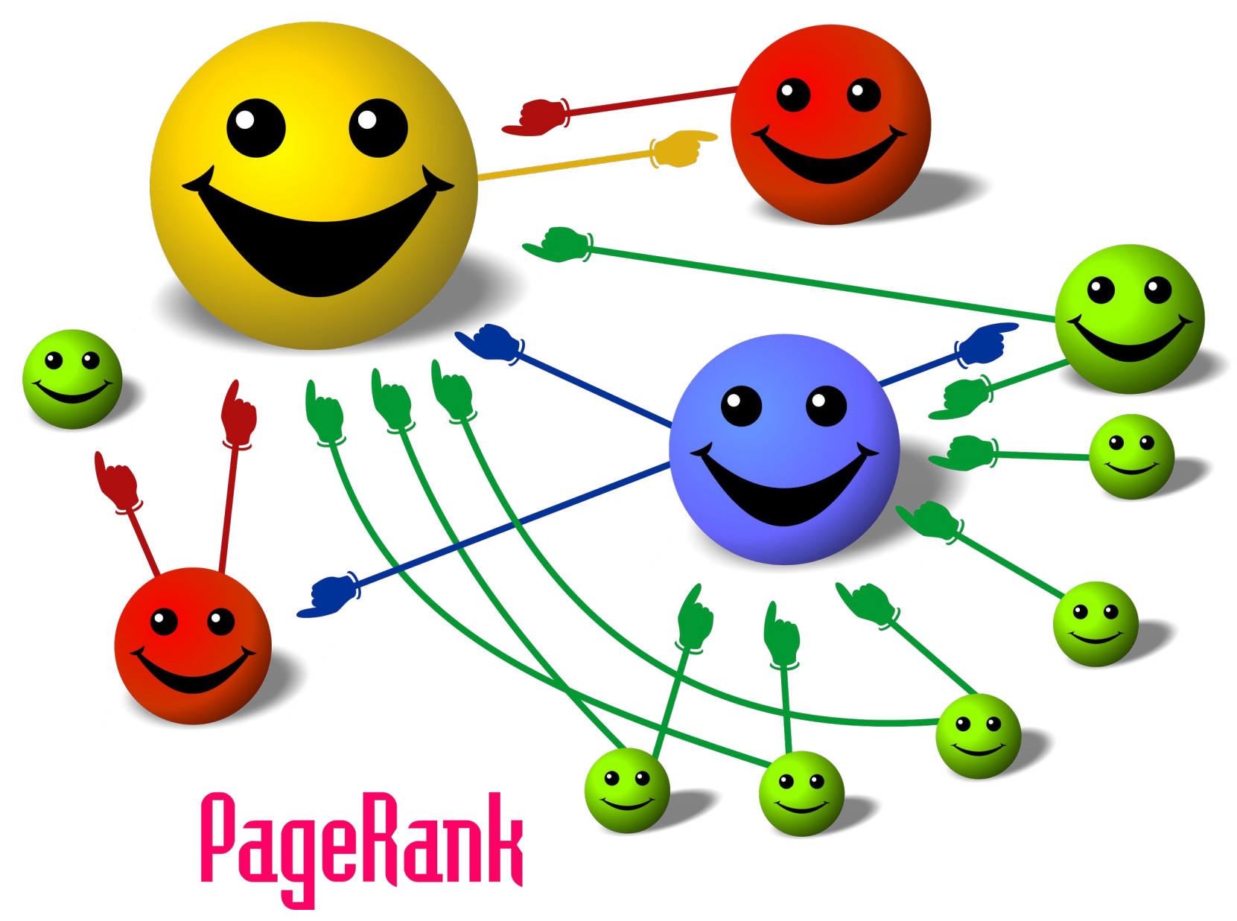
Nikolaj Volgushev

Motivation

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].

PageRank

"PageRank 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]



Definition

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

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

Computation

$$\mathrm{PR}(p_i; 0) = \frac{1}{N}$$

$$\mathrm{PR}(p_i; t+1) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{\mathrm{PR}(p_j; t)}{L(p_j)}$$

where *t* is the iteration
M(X) are the in-neighbors of *X*

SimRank

SimRank measures the similarity between two objects. As a metric it implements the following intuitive definition: "Two objects are similar if they are related to similar objects." [2]

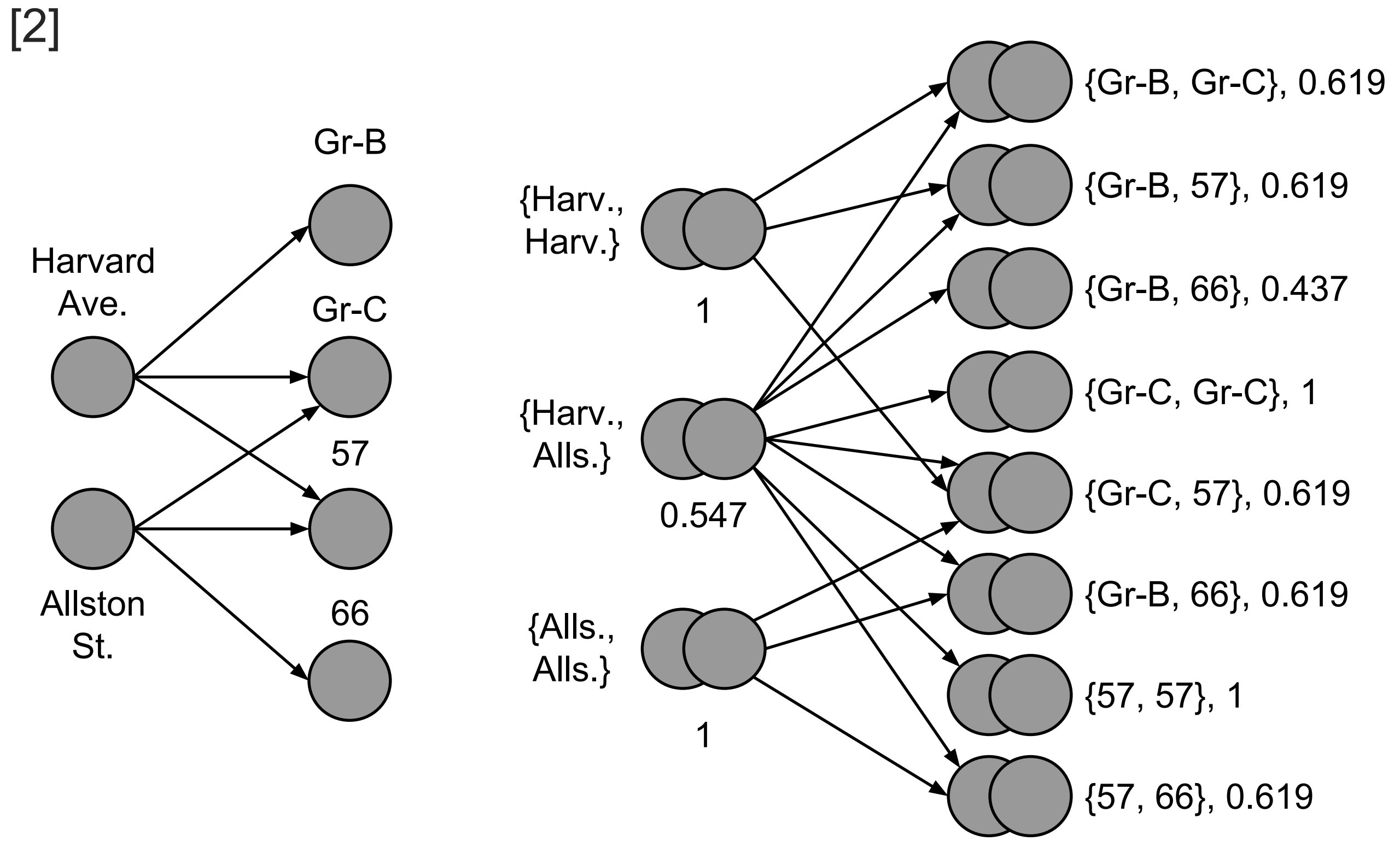


Figure 1. SimRank example.

Definition

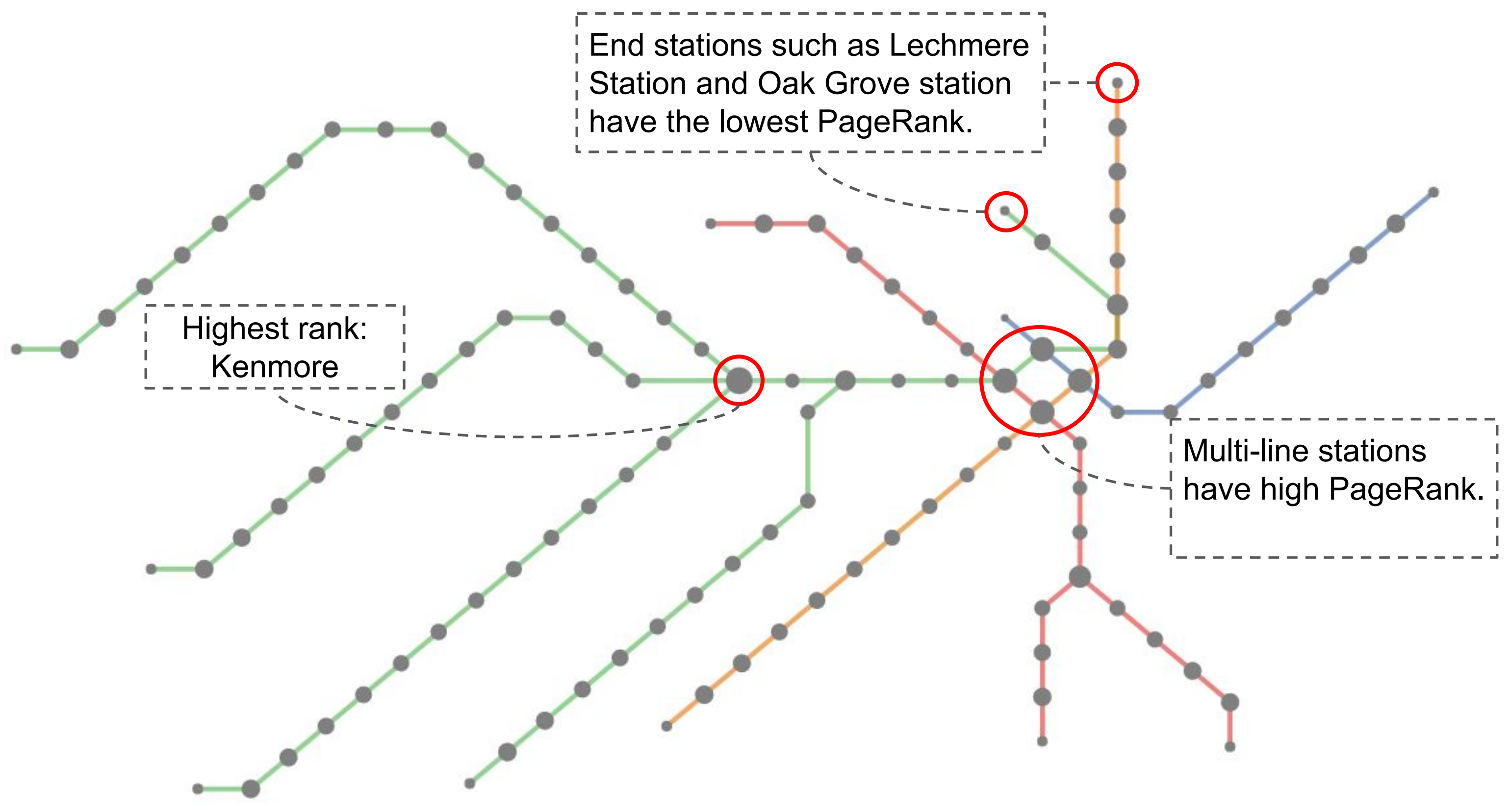
$$s_1(a, b) = \frac{C_1}{|O(a)||O(b)|} \sum_{i=1}^{|O(a)|} \sum_{j=1}^{|O(b)|} s_2(O_i(a), O_j(b))$$

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

where C_1 and C_2 are damping factors
I(n)/ O(n) are the in/out-neighbors of a node

Station PageRank

We computed three different versions of PageRank over varying adjacency semantics and data sets. 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.

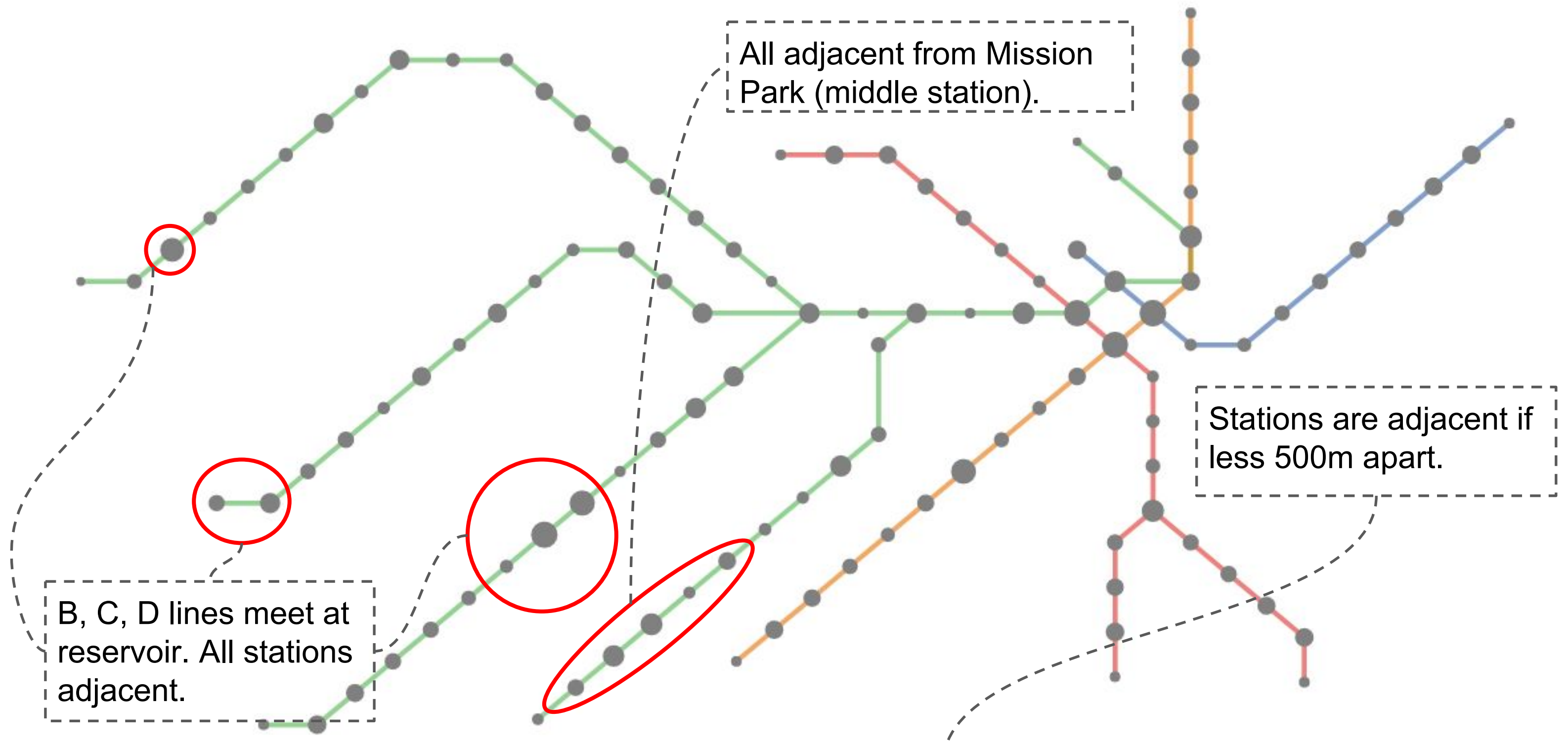


End stations such as Lechmere Station and Oak Grove station have the lowest PageRank.

Highest rank: Kenmore

Multi-line stations have high PageRank.

Figure 2. Direct connection adjacency. T stations only.

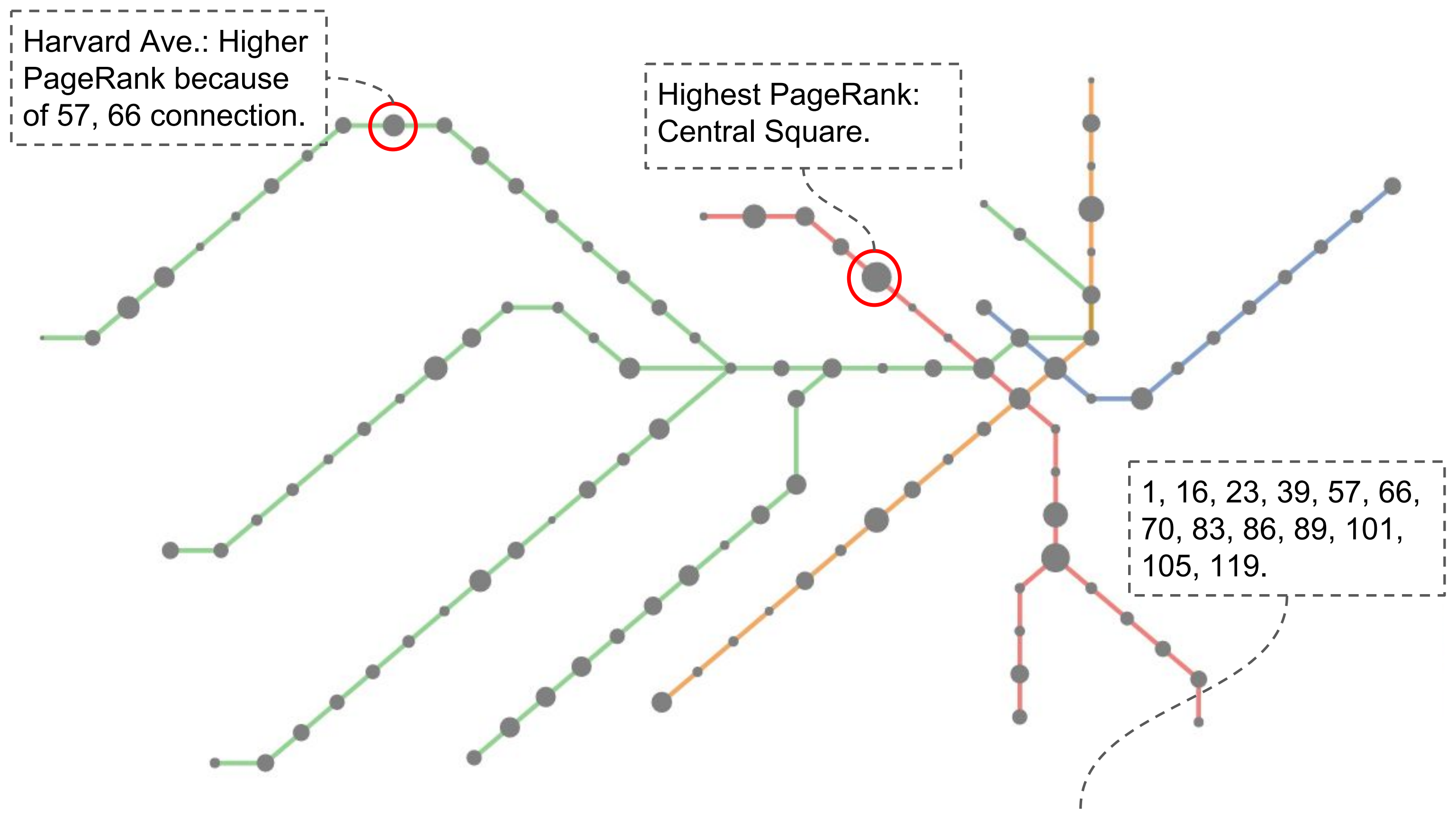


All adjacent from Mission Park (middle station).

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.



Harvard Ave.: Higher PageRank because of 57, 66 connection.

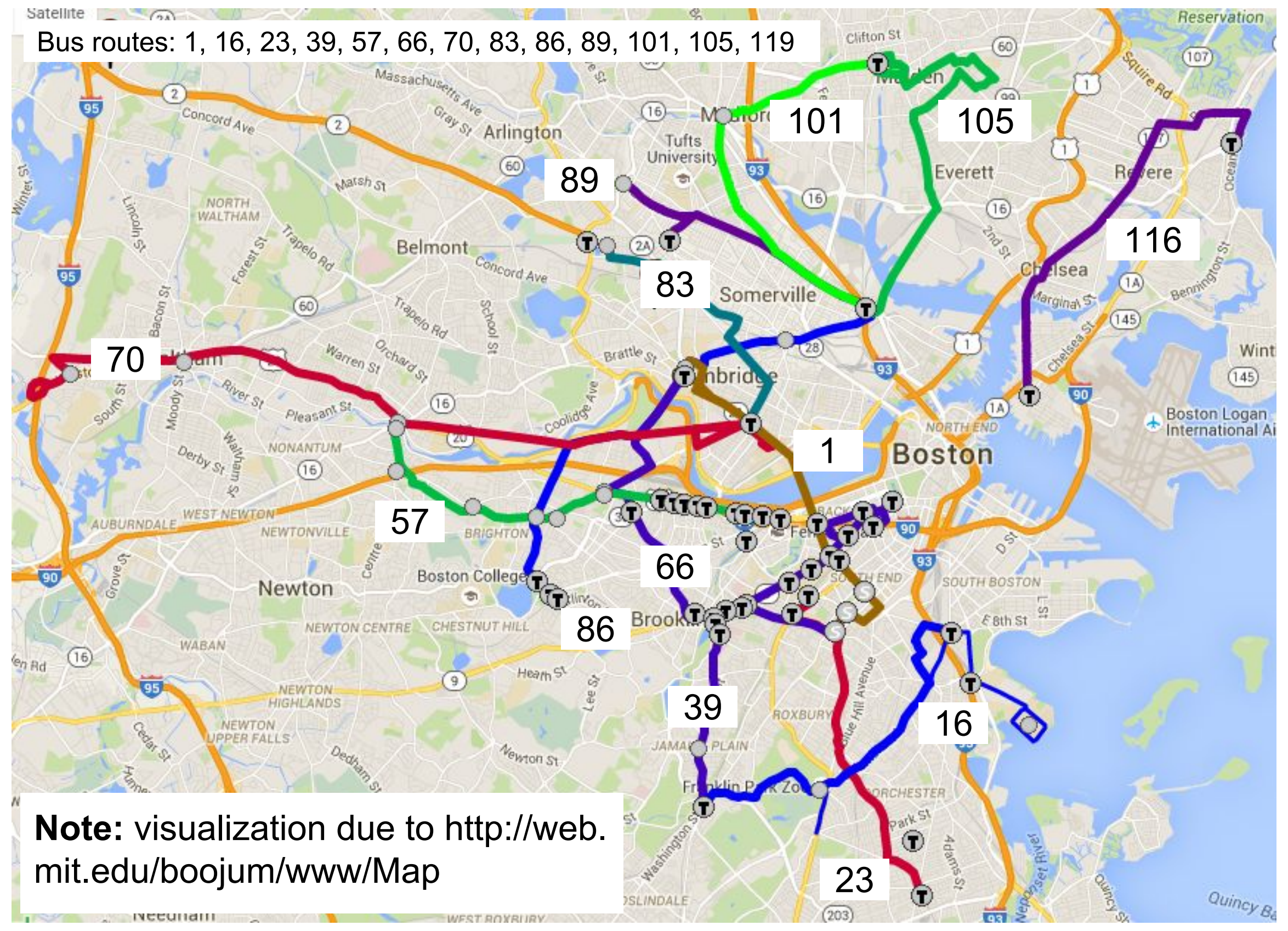
Highest PageRank: Central Square.

1, 16, 23, 39, 57, 66, 70, 83, 86, 89, 101, 105, 119.

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

Route SimRank

As a metric for route similarity we propose SimRank. We extend the intuition behind SimRank to routes and stations: routes are similar if similar stations are associated with them and vice-versa. A station is associated with a route if (a) the station is directly part of the route or (b) a station of the route is adjacent to the station in question. This yields a bipartite graph akin to what is presented in Figure 1. For the adjacency model we use direct connections in addition to geo-adjacency.



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

89 and **101** are the most similar routes with SimRank 0.30.

39 route is most similar to a T-line. SimRank(**39, green-e**) is 0.20.

66 has highest cumulative SimRank with 2.25.

116 has lowest cumulative SimRank with 1.26.

Figure 5. Bus route locations.

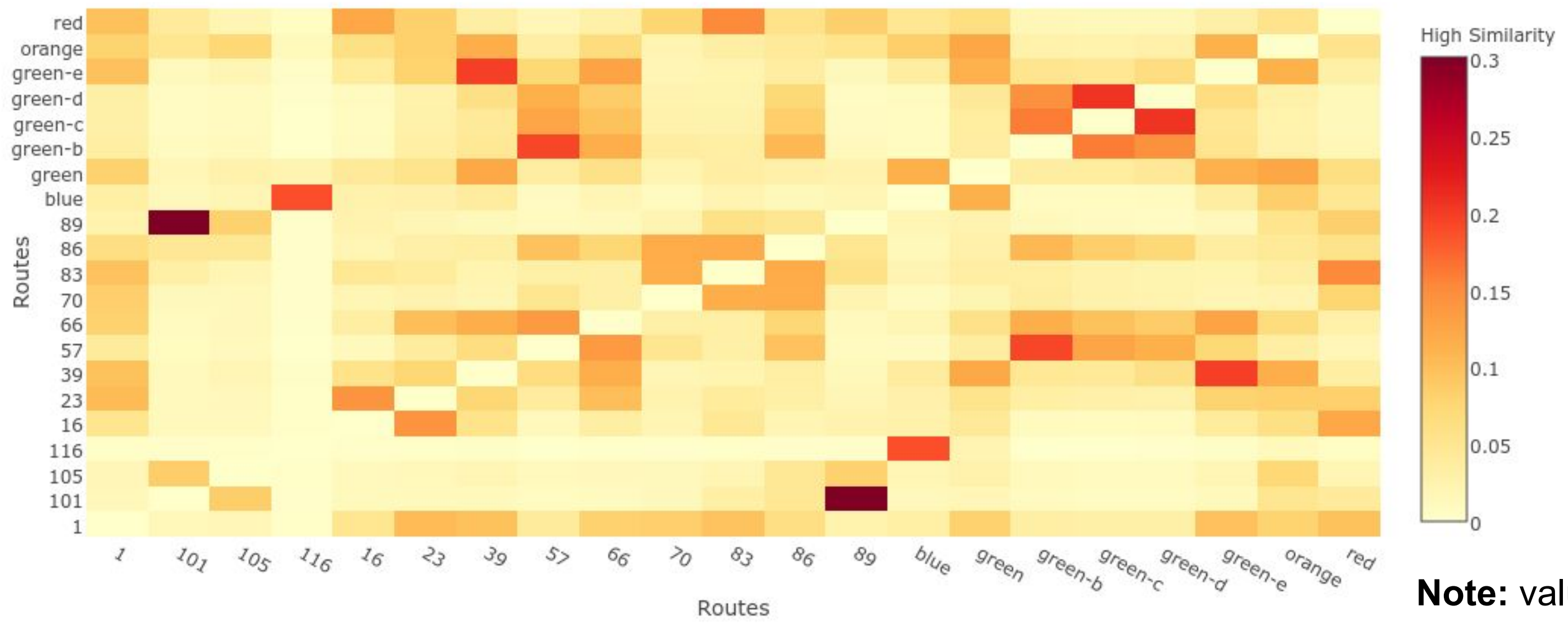


Figure 6. Route SimRank heatmap.

Note: values on diagonal are set to 0.0.

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/