A method to ascertain rapid transit systems’ throughput distribution using network analysis

Muhamad Azfar Ramli, Christopher Pineda Monterola, Gary Lee Kee Khoon, and Terence Hung Gih Guang

Institute of High Performance Computing, Agency for Science, Technology and Research, Fusionopolis, 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632
{ramlimab,monterolac,leekk,terence}@ihpc.a-star.edu.sg

Abstract
We present a method of predicting the distribution of passenger throughput across stations and lines of a city rapid transit system by calculating the normalized betweenness centrality of the nodes (stations) and edges of the rail network. The method is evaluated by correlating the distribution of betweenness centrality against throughput distribution which is calculated using actual passenger ridership data. Our ticketing data is from the rail transport system of Singapore that comprises more than 14 million journeys over a span of one week. We demonstrate that removal of outliers representing about 10% of the stations produces a statistically significant correlation above 0.7. Interestingly, these outliers coincide with stations that opened six months before the time the ridership data was collected, hinting that travel routines along these stations have not yet settled to its equilibrium. The correlation is improved significantly when the data points are split according to their separate lines, illustrating differences in the intrinsic characteristics of each line. The simple procedure established here shows that static network analysis of the structure of a transport network can allow transport planners to predict with sufficient accuracy the passenger ridership, without requiring dynamic and complex simulation methods.

Keywords: rapid transit, passenger ridership, betweenness centrality, throughput, correlation.

1 Introduction
The burden of an increasing population and the severe limitations of available land continue to challenge transport planners all around the world. While mass rapid transit systems are generally successful in reducing commuter reliance on road networks, the prediction of the distribution of passenger throughput across stations is a difficult task to accomplish. The task

*Corresponding Author
is further complicated by the need for constant expansion of the transit system to accommodate a growing population. The addition of new lines, stations and interchanges cause various complex changes to the ridership thus modifying the utility of all stations within the system. It is therefore difficult to predict the impact of these changes on the robustness and economic viability of the system. Transport planners would therefore have to design based on an excess flow model instead of optimal flow model in order to accommodate these changes within the transport system. In this paper, we offer a simple procedure that uses network analysis to extract relevant structural information from a rail network and predicting to within a reasonable accuracy, the distribution of passenger throughput. Although this methodology assumes that the pattern of passenger travel would in fact conform to the capacity of the network, we show that for our purpose, this is sufficient. It is therefore hoped that by using this prediction, transport planners can have a basic understanding of the effect of structural changes to the network and adjust the capacity of the network accordingly so as to maximise the utility of the transport system.

2 The Singapore Rail Transport System Network

We validate the accuracy of our methodology by using the example of the rail transport system (RTS) network of Singapore which consists of 90 train stations connected by four distinct lines. Figure 1 shows a snapshot of the current Singapore rail transport network. In the latest Land Transport Masterplan 2013 [11], the Land Transport Authority of Singapore has also announced that by 2020, an additional five more lines will be added to the existing rail system; effectively doubling the number of interchanges in the network, thus increasing tremendously the complexity of the network.

The public bus transportation and the taxi system form the other two components of the public transportation system in Singapore; these utilize the road network infrastructure to move passengers to and from their required origins and destinations. Together, these three systems provide an economical alternative to private modes of transportation in the crowded city-state of Singapore. This is especially important since the entry cost of owning private vehicles in Singapore is possibly one of the highest in the world today, due to the tight government policy.
that aims to effectively reduce road traffic congestion.

Several articles in the literature have conducted prior analysis on various features of the RTS network of Singapore. Derrible et al compared the Singapore rapid transit network with various other subway or rapid transit systems in the world by analyzing the various network properties. However they only considered the interchanges as the nodes within the network and considered all the stations in between as being part of a single edge. This helped to unify the various features of different networks and made it easier for comparison between networks of different cities, however this is not suitable if the analysis requires us to discern the flow patterns through different stations within the lines. Soh et al conducted a comprehensive analysis of the properties of both the rail and public bus transport network, however this analysis does not take into consideration the actual physical structure of the network. Instead, they used a weighted network approach that connects various stations together by the frequency of passengers traveling between the two stations. Notably, both analyses were also carried out prior to the addition of the most recently added line (Line 4 marked in yellow in Figure 1) to the network. Our work therefore aims to extend the work conducted here by validating the use of network centrality measures to characterize flow through the transport network and ensure that a consistent methodology can be applied regardless of changes made to the network.

3 Methodology

In this section, we introduce how the calculations of the key variables required to analyze the network; namely betweenness centrality and passenger throughput, were conducted. We use generic graph theory notation to characterize the network $G = (V, E)$ with vertices $v \in V$ and edges $e = (v_i, v_j) \in E$. We consider each of the 90 train stations as the nodes in our network while a single edge is added to the network for every portion of the network which connects two adjacent stations via a rail line.

We denote the simple path joining the nodes $s$ and $t$ by the vector $\tilde{p}(s \rightarrow t) = [s, (s,v1), v1, (v1,v2), ..., (v_i,t), t]$, which consists of the set of nodes and edges through which the path traverses. For unweighted networks, the cost of a path is typically defined by the length of the path $|\tilde{p}(s,t)|$ (i.e, the number of nodes and edges traversed) while for a weighted network, the cost is taken to be the total sum of the cost incurred for each node and edge while traversing the path: $\sum_{i \in \tilde{p}(s \rightarrow t)} weight(i)$. As there may be multiple possible paths between the nodes $s$ and $t$, the path which incurs minimum cost is known as the shortest path and this path can be obtained through various different algorithms, most notable in the literature include Dijkstra’s algorithm and the A* algorithm [1].

3.1 Computing the Betweenness Centrality of the Network

Betweenness centrality was first introduced by Freeman [8] to measure node importance in large social networks which were not fully connected. Its use in predicting traffic flow has also been previously published, with various authors [2, 3] debating its accuracy and usability. Different scenarios were also studied ranging from information traffic in communication networks [10] and road traffic networks. What is agreed upon is that betweenness is a topological property of the network and any effort in predicting traffic flow would therefore assume that the flow through the network is directly affected by the structural property of the transport network.

Betweenness centrality has also been used to analyze the resilience of a network [14]. By removing key nodes or edges with high betweenness in the network, it can be observed that the
flow patterns through different paths will be highly affected. This simulates the effect of disruption in real physical transportation networks where traffic jams or train breakdowns frequently occur and helps transport planners identify whether their system incorporates sufficient design robustness or if it requires further upgrades to its overall resilience.

For the purpose of modeling betweenness centrality for our rail transport network, we define the shortest path as the path which incurs the least amount of travel time between two stations. To do this, we add up the overall travel time cost of all the edges within the path. The travel time between each edge ranges between 2-5 minutes depending on the distance between each station and these values are determined by using the information on the travel time between stations in the network as indicated by the transport operator on their website [4]. In addition, we use the assumption that for every transfer at an interchange, a transfer cost of 4 minutes is incurred. This reflects the average walking time needed to move from one platform to another within the interchange. A modified calculation of weighted shortest path is therefore required to identify the path of least cost between two stations. This is because the cost of 4 minutes when passing through an interchange node is only incurred when the route changes to a different line. No other cost is incurred when traveling through all other nodes on the path. It is also noted that for a single origin-destination pair, there may still be multiple routes through the network which incurs the same cost. When calculating betweenness centrality, all possible routes of equally minimal cost are taken into equal consideration.

The betweenness centrality $b_c(v)$ of a node in a network is defined as follows:

$$b_c(v) = \sum_{s,t \in V, s \neq t} \frac{\sigma(s \rightarrow t|v)}{\sigma(s \rightarrow t)}$$

where $\sigma(s \rightarrow t)$ is the sum of shortest paths between the stations s and t while $\sigma(s \rightarrow t|v)$ is the sum of shortest paths between s and t that pass through the node v.

Since a single path passes through multiple edges and stations, it is trivial to realize that by adding up the counts for multiple stations, the total sum of $b_c$ for all stations will typically not add up to 1. We therefore normalize the betweenness value by dividing the total sum of betweenness for all stations so that the distribution of values can then be easily compared against the distribution of throughput.

$$\hat{b}_c(v) = \frac{b_c(v)}{\sum_{v \in V} b_c(v)}$$

By replacing the terms $v \in V$ with $e \in E$ in equations (1) and (2), we can also obtain the same definition for edge betweenness, and these values are also similarly normalized.

### 3.2 Computing the Distribution of Passenger Throughput from Ridership Data

Given that $n_p$ represents the total number of passenger journeys traveling through a network for a specified amount of time and $n_p(v)$ as the number of passengers which visit the station v during their respective journeys, we define the passenger throughput $p_t(v)$ through the station v as the fraction of passengers that visit a specific station v:

$$p_t(v) = \frac{n_p(v)}{n_p}$$
Our definition of whether a passenger has visited a station is through two distinct ways; either they have passed through the station during their journey while being on board a train without entering or leaving the station, or they may have started or ended their journey at that specified station. Either way, the value of throughput obtained provides us with a simple measure for the average crowdedness and utility of a station. Notably, the quantity \( p_t \) is a dynamic variable that varies with different timings (the throughput of a station on a weekday may differ greatly from that during the weekend). In this paper, we measure the average total ridership using data obtained for a typical week from Monday to Sunday. This helps us to define a static characteristic of the average throughput that the station experiences during the week and we can now compare this measure against betweenness centrality, which is intrinsically a static structural measure.

Whenever a passenger uses the mass rapid transit in Singapore, the smart card ticketing system captures the origin time and station where he begins his journey as well as the time and station where he taps out at his destination station. Using an anonymized dataset that captures this activity for a week long period between Monday to Sunday provided by the Land Transport Authority of Singapore, we are therefore able to estimate the actual passenger throughput. The data consists of a total of 14 million journeys moving between all 90 stations thus giving us an adequate amount of samples of throughput occurring between all the stations. This gives us an accurate measure of the average throughput occurring within the stations. Unfortunately, we cannot exactly determine which route each passenger took in order to reach his or her respective destination as the data only tracks the start and end points of their individual journeys. However, in a previous study, we have shown that by analyzing the variation and distribution of travel times occurring for a specific origin-destination pair, we can obtain an accurate estimate of the fraction of passengers who travel using specific routes between a specific origin-destination pair [12].

We are therefore are able to glean the following two quantities from the data,

1. the fraction of travel that occur between a specific origin-destination pair out of all possible travels, and

2. the approximate fraction of passenger who travel through a specific station given a origin-destination pair.

The passenger throughput can therefore be calculated by using the following formula:

\[
\begin{align*}
p_t(v) &= \sum_{s,t \in V, s \neq t} \frac{n_p(s \rightarrow t|v)}{n_p} \tag{4} \\
&= \sum_{s,t \in V, s \neq t} \frac{n_p(s \rightarrow t)}{n_p} \times \frac{n_p(s \rightarrow t|v)}{n_p(s \rightarrow t)} \tag{5} \\
&= \sum_{s,t \in V, s \neq t} f(s \rightarrow t) \times f(s \rightarrow t|v) \tag{6}
\end{align*}
\]

where \( f(s \rightarrow t) = \frac{n_p(s \rightarrow t)}{n_p} \) is the fraction of journeys traveling between the stations \( s \) and \( t \) out of all the passenger travels and \( f(s \rightarrow t|v) = \frac{n_p(s \rightarrow t|v)}{n_p(s \rightarrow t)} \), the fraction of passengers traveling between stations \( s \) and \( t \) that pass through the station \( v \).
Figure 2: Comparison of Passenger Throughput (Nodes). Significant improvement in the correlation is obtained by removing 10 nodes from the dataset. However, all the outlier nodes correspond exactly to the northern portion of Line 4 (see Figure 4a).

Again, we consider the normalized form of $p_t$ given by:

$$\hat{p}_t(v) = \frac{p_t(v)}{\sum_{v \in V} p_t(v)}$$

By replacing $v \in V$ with $e \in E$ in equations (4-7), we can also consider passenger throughput through an edge and compare this to edge betweenness.

4 Correlating the Distribution of Passenger Throughput against Betweenness Centrality

In order to investigate the dependence of passenger throughput against betweenness centrality, we calculate the Pearson’s correlation coefficient $\rho_{p_t,bc}$ between the two variables calculated for all the stations and edges and analyze the fit of the data using linear regression. Figures 2 and 3 illustrate the results of the regression done for the full set of stations and edges respectively.

We find that the values of betweenness centrality and the passenger throughput for stations give rise to a correlation coefficient of 0.5345. (see Figure 2) However, most of the significant outlier nodes and edges belong to the northern portion of only one of the lines, namely Line 4 (see Figure 4a), a line that we noted was added only about 6 months prior to the collection of the ridership data. In comparison, the other next to newest line, Line 3 was added ten years prior to the data collection period. By removing these data points from the collection and repeating the correlation analysis, we found that the correlation coefficient obtained improved significantly to about 0.74076 although only 10 out of 90 (11.1%) of the data points were removed. A similar trend was observed for the correlation when considering the edges of the network. Originally, a poor correlation of only 0.3379 was obtained, however this improved tremendously to 0.7074 when 12 out of 97 (12.4%) outlier edges were removed. (see Figure 3 and Figure 4b)

We then correlated the betweenness and passenger throughput distributions separately for stations within the same line and considered interchanges as a separate classification. We now
Figure 3: Comparison of Passenger Throughput (Edges). Significant improvement in the correlation is obtained by removing 12 edges from the dataset. The outlier edges correspond exactly to the northern portion of Line 4. (See Figure 4b).

Figure 4: Outlier data points removed to significantly improve correlation between throughput and betweenness distribution. The stations and lines correspond exactly to the northern portion of the most recently added line, Line 4 in the rail network.

find a significant improvement in the individual correlations when they are considered separately without outlier removal, as the values of now range between 0.6 to 0.9713 (see Figure 5). The results also reveal that the values for the interchanges featured poorer correlation as compared to the other classifications.

Interestingly, the Line 4 group of stations showed a division between the northern portion of the line as compared to the southern portion, thus creating the group of outlier nodes. However, the correlation obtained as a single line was the higher than that obtained from any other line. The gradient obtained from the line of fit for the Line 4 group of stations was also significantly higher (3.4715) as compared to the gradients obtained for the other lines (0.621 - 1.116) which caused the combined correlation to be poor. This shows a difference in relationship between passenger throughput and betweenness centrality for this new line which may suggest that the passenger throughput through this line is characterized differently as compared to the other
Figure 5: Correlation between Passenger Throughput and Betweenness Centrality for Nodes split into the different lines. Interestingly, the Line 4 stations c) are separated into two portions, the high centrality stations correspond to the northern stations while the remaining stations have low centrality. Also notably, the interchanges f) display poor correlation and spread of values, in particularly those with high centrality and throughput.

5 Conclusion

Despite the fact that not all the passengers use shortest paths to travel between stations and that the distribution of passengers traveling between various origin-destination pairs is not uniformly distributed, we demonstrate that betweenness centrality can still predict to a reasonable accuracy, the level of utilization between different portions of the network. This is shown by the high level correlation between betweenness centrality and passenger throughput when most
of the stations and edges of the network are being considered.

In addition, we found that the outliers of the correlation correspond exactly to the stations and edges of a new section of the network. This suggests that although the passengers generally conform to the structure of the network when utilizing the transportation system, this conformity is reduced when dynamic changes are made to the structure. There may therefore exist some lag time before passengers adapt to the new structure and adopt more efficient routes which did not exist previously. Confirmation of this hypothesis would however, require a comparative analysis with the current data against a different set of ridership data which include periods before and after the addition of these lines.

We have also shown that correlation improves significantly when distinct lines are correlated separately. The different lines exhibit variation in slopes, indicating that the unique structural characteristics of lines impacts the pattern of ridership within the network. This is expected since the distribution of passenger ridership gradually evolves with the structure by adjusting to the availability of routes and convenience to individual passengers. With increasing complexity planned to be added in the network in the future, we hope that this work will serve as a standard methodology in capturing some base line information on the expected utility of a specific segment of the rapid transit system.

5.1 Acknowledgements

We acknowledge Erika Fille Legara and Vicknesh Selvam for useful discussions and for processing of the datasets used for validation in the paper. We would also like to thank the Land Transport Authority of Singapore for providing passenger ridership data that helped us validate our method. This research is supported by the Science and Engineering Research Council of the Agency for Science, Technology and Research (A*STAR) of Singapore. (Complex Systems Programme grant number 122 45 04056).

References


