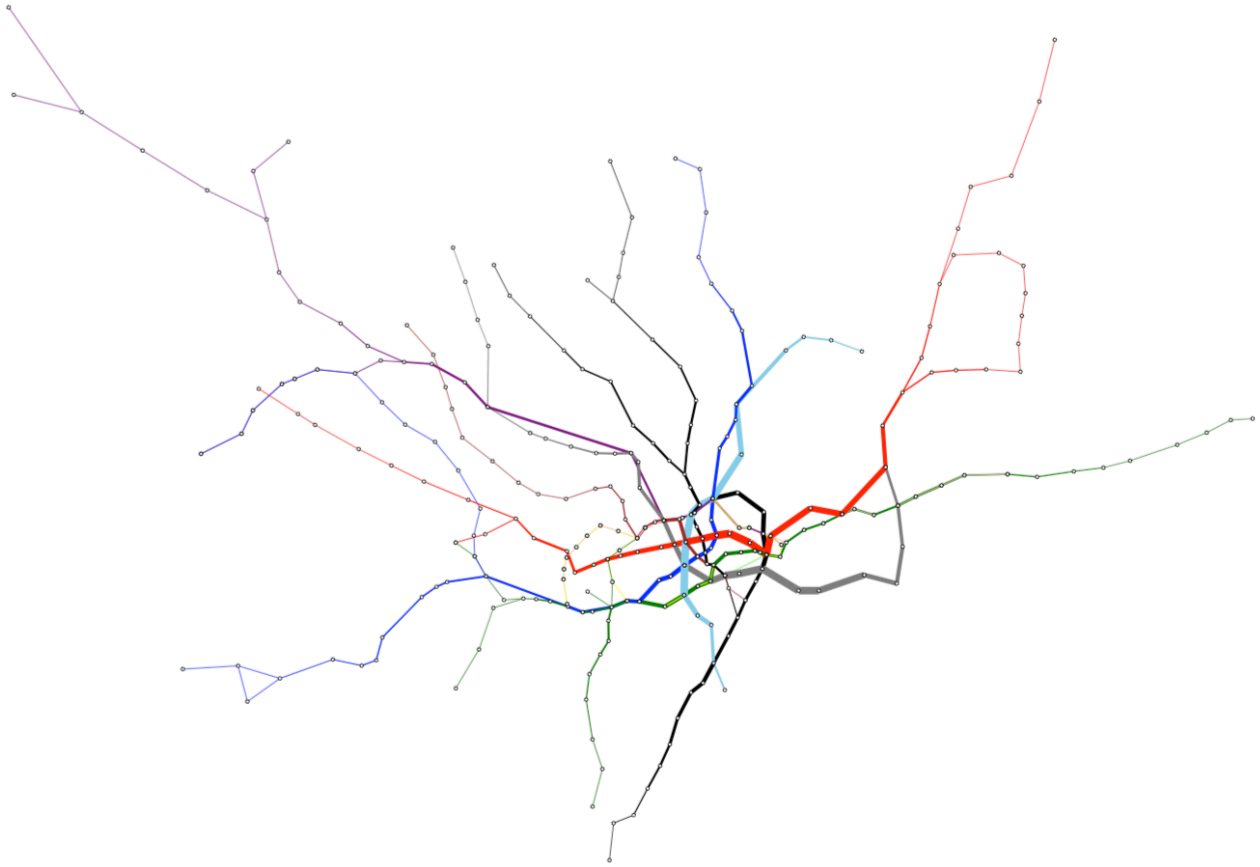
How can network analytics be applied to understand passenger journeys on the London Underground, and how can this improve operations of the underground during closures?



MSc Business Analytics
Imperial College London
August 2017

Word count: 5,322

Abstract

Closure of part of lines cause changes in passengers' route choices. This leads to changes in passenger flow distribution on the underground network. By better understanding how the passengers travel on the London Underground (LU) network, Transport for London (TfL) can plan and operate London's transport system more effectively. To predict passenger flow during closures, it is necessary to know how passengers choose their route and which factors influences their choice.

This analysis attempted to enhance the knowledge and contribute to existing knowledge on passenger flow dynamics. The passenger flow was forecasted after a hypothetical closure to inform operators of exposed links in the network, where the line between the underground stations Waterloo and Westminster was removed to exemplify a worst-case scenario. As a proof of concept, a model to predict the new passenger flow was introduced. The forecast was based on a method introduced by Li and Sue (2003) using a double search shortest path algorithm, in combination with a percentage threshold value to incorporate the number of transfers. The results show that the network efficiency did not suffer from one link removal. However, the Bakerloo line experienced an extensively increase in passenger flow, implying huge consequences for the operations of the network. As a result, insight statistics about the change is included as a guideline for TfL, both at line and link level. The results demonstrate the potential of using network analytics in improving TfL's operational services.

Table of Contents

1. Introduction

1.1 Background

The London Underground (LU) network is one of the most important transport modes that connects the city. It is the world's oldest underground network with 11 lines and 268 underground stations. Approximately 5 million journeys are made every day, where, on average, 67 % are commuters (TfL, 2017). However, typically there are times when part of the network is closed.

Closures can be either planned or unplanned, where the network is especially vulnerable to unplanned closures. There can be up to 1000 unplanned closures during a year, where reasons include fire alerts, defective lifts or escalators, or industrial action (TfL, 2009). Partial line or station closures change the travel patterns of affected passengers, and how this is handled is crucial for the network.

It is of great value to Transport for London (TfL) to know how the closure impacts the passenger flow on the network. Knowledge of the forecasted passenger flow before planned closures could improve the allocation of resources such as additional trains and staff. It is also helpful when informing affected passengers about optimal routes to enact as alternatives. Moreover, an analysis of the passenger flow in unplanned events is also of high importance. Even though it is difficult to accommodate the need for resources immediately, an analysis of passenger flow will inform the operators of which lines and stations that are most exposed, and where to reallocate resources over time.

The research question that will be addressed is *"How can network analytics be applied to understand passenger journeys on the London Underground, and how can this improve operations of the underground during closures?"*. The analysis employs network analytics to explore the passenger flow and its characteristics to get an in-depth understanding of how passengers travel on the underground. An investigation of how the passengers choose their route is also conducted. Then, a hypothetical closure is introduced to represent a worst-case scenario. Further, a model to predict the new passenger flow is introduced, where the passengers' route choice is found based on travel time and transfers. The final results are summarised to inform operators of exposed links in the network to the hypothesised closure, such that improvement of operational services can be planned and conducted.

1.2 Literature review

The main focus on research in this field has been on network structure (Lee et al 2008; Lee et al 2011; Xu, Mao & Bai, 2016), crowdedness (Ceapa, Smith & Capra, 2012), passenger flow (Jia & Chow, 2016; Li & Zhu, 2016; Lee et al, 2011; Lee et al, 2008; Xu, Mao & Bai, 2016), attack analysis (Zhang et al, 2011; Wu, Gao & Sun, 2007; Barabasi, 2014), and route choice (Dou et al, 2014; Li & Su, 2003; Prato, 2009).

In terms of network structure, there have been multiple studies analysing the statistical properties and distribution of the passenger flow in transit networks. Among others, Lee et al (2008) and Lee et al (2011) analyse the Seoul Subway network, and Xu, Mao and Bai (2016) use trip data from the Beijing Subway. Crowdedness has also been thoroughly analysed. Ceapa, Smith and Capra (2012) use Origin-Destination (OD) data from the London

underground to predict crowdedness at stations over time. It was found that crowdedness during the weekday is highly predictable. However, these studies have generally been focusing on trip data, where passengers route choice is unknown.

More relevant to the current research, some studies have used trip data to predict passenger flow. Jia and Chow (2014) developed a tool to estimate the passenger load and crowdedness on link level on the Metro in Washington DC. The predictions were based on OD data, without incorporating the real passenger load or route choice. Li and Zhu (2016) included a route choice in their simulation model by estimating the selection probability. The simulation model predicts passenger flow distribution on a rail transit network with train delays using a schedule-based approach. However, the actual passengers' route choices are unknown in both studies. It has also been conducted studies within mode choice under disruptions (Pnevmatikou, Karlaftis & Kepaptsoglou, 2015; Lin, Shalaby & Miller, 2016).

In November 2016, TfL conducted a data trial to collect route choices through passengers' phones using Wi-Fi connections in order to improve awareness of route choices (Irvine 2016). This data has great potential to improve the service, but is currently not publicly available for data protection reasons. TfL uses their own strategic models for planning decisions, where the Railplan public transport assignment model can simulate rerouting and crowdedness (TfL, n.d.). However, there is little public information on the methodology behind these models.

Shortest path algorithms, such as Dijkstra's algorithm, is often used to find passengers optimal route choice (Prato, 2009). However, Zhu and Levinson (2015) discovered that passengers do not always choose the shortest path. Research has identified other factors that are of importance including minimal number of transfers (Li & Su, 2003), individual preferences (Prato, 2009), congestions and train capacity (Wu, Gao & Sun, 2007; Dou et al, 2014), train fare (Dou et al, 2014), train schedules and delays (Li & Zhu, 2016), and crowdedness (Ceapa, Smith & Capra, 2012). Dou et al (2014) proposed an algorithm to find the optimal path for railway passengers using Dijkstra's algorithm in combination with residual train capacity, to give path suggestions to passengers. They also recommended adapting this algorithm in analysis of transit networks. Li and Su (2003) proposed an algorithm to find the shortest path between two bus stations, taking the passengers psychology into account by minimising transfers and distance.

Network resistance and change in passenger flow under disruptions and attacks have been of great interest. Barabasi (2014) analysed network robustness after random and targeted attack of several nodes in the network. Zhang et al (2011) did similar analysis of the Shanghai Subway using three different attack protocols. Both concluded that scale-free networks are 'robust against random attacks but fragile for malicious attacks' (Zhang et al, 2011). Wu, Gao and Sun (2007) investigated cascading failures after three different types of attacks on a link in the network. When reassigning flow to other links, they included a cost that takes the congestion effect into account.

2. Methodology

There has generally been little research on passenger flow where real passenger routes are applied. This analysis is aiming to apply analytical techniques to contribute to the understanding of how passengers travel. A model to predict passenger flow is introduced as a proof of concept to identify exposed links in the network after a hypothetical closure.

Unexpected closures impact passenger flow at stations and lines immediately. In this instance, operational manoeuvres are limited. As a result, the focus of this work is on passenger flow after closure when travellers have been informed. The analysis allows for operational services to allocate extra resources such as train and staff to mitigate the threat to the network.

First, a comprehensive overview of the London underground network is provided, including statistical properties of the network, weight and strength distributions, and a maximum spanning tree to examine the hub structure. Further, an in-depth analysis of the passengers' route choices is conducted to understand how passengers chose their route. Next, an important link is removed to present a worst-case scenario, identified through measures of centrality. Number of affected passengers are identified through the route dataset, and the passengers' route options are predicted based on Dijkstra's algorithm. The model predicting the final route choice combines both shortest paths and transfers, applying a percentage threshold for shorter paths with more transfers. The new passenger flow is then analysed by investigating changes on specific lines and links.

2.1 Data

The data was obtained from six different datasets, where five is collected from the open data source published by TfL (TfL, 2017). The data includes information on six periods during a typical day; early, AM peak, midday, PM peak, evening, and late. The numbers are based on November 2016 counts, and 'represents the number of people travelling on a typical (or average) weekday' (TfL, 2017). It is also important to emphasise that the datasets 'are adjusted to remove the effect of any abnormal circumstances that may affect demand such as industrial action or long-term closures' (TfL, 2017). In addition to the datasets, TfL's API is used to retrieve stations' names, geocodes, and associated lines.

The first two datasets contain counts for how many touches in and out at the stations during a day. The third dataset contains a route choice dataset for a typical weekday, including OD pairs and information about where the passengers changed line. The data is collected through surveys and 'reconciled to November counts' (TfL, 2017). The fourth dataset contains information of flow on each link in the network, and the fifth dataset has information about all stations with identification numbers. Moreover, a sixth dataset is collected with information of the length in kilometres of each edge and the running time (TfL, 2013).

3. Network description

3.1 Notation and topology

In this section, the LU network is described and investigated through complex network theory. The underground can be represented as a directed graph, where traffic goes in both directions between adjacent stations. The graph G is ordered with N nodes and E edges, respectively stations and lines between adjacent stations, where the edges have attributes such as passenger flow, time and distance. Figure 1 displays the topology of the network with the thickness of the edges representing passenger flow.

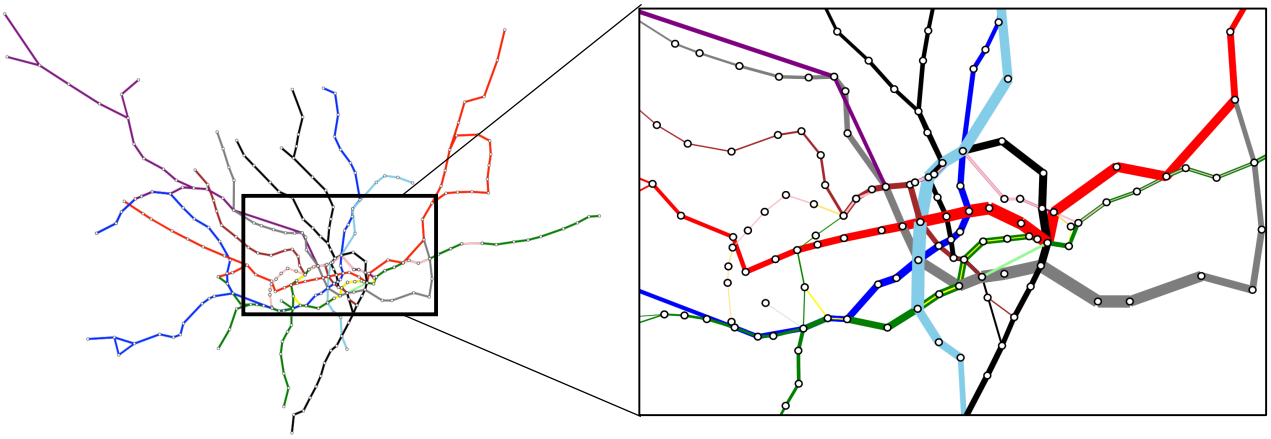


Figure 1: The passenger flow on the London Underground on a typical weekday.

A station can have multiple directed edges to another station, hence it is a multidigraph. For practical reasons, previous studies have chosen to collapse the edges together to retain the directed graph (DataCamp, 2017). Here, however, additional platform nodes are added to the network to keep the detailed overview of the graph and the flow on different lines. The graph is thus kept as a directed graph, with 660 nodes, respectively 268 station nodes and 392 platform nodes.

Each station is represented with one main station node and one platform node per associated line, where the platform nodes have directed edges to and from the main station node. This makes it possible to find the travel route passengers take based on the route dataset, and incorporate additional time when the passengers change between lines. Moreover, links with walking options are also added to the network for stations that are located close to each other (< 1,000 metres), visualised as directed edges between the main station nodes[1]. For calculations at station level, such as network efficiency, all platform nodes for the station are incorporated. The topology describing how the platform nodes are applied is presented in detail in Figure 2.

[1] Part of code cited from AaronD (2013)

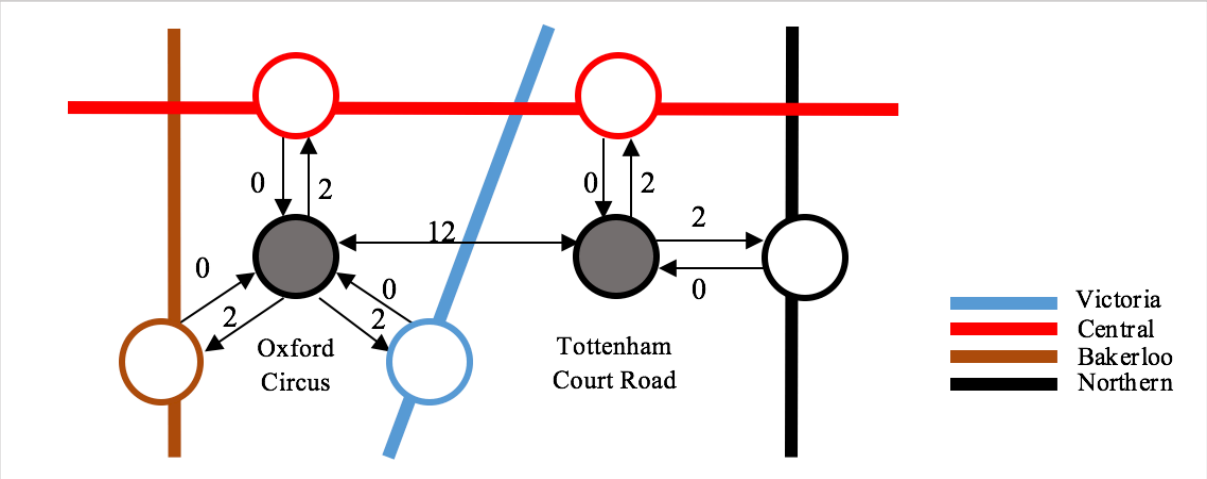


Figure 2: An excerpt of the network displaying Oxford Circus and Tottenham Court Road station including the possibility to change line and the walking option. It is assumed that changing line will take approximately 2 minutes. The walking time between Oxford Circus and Tottenham Court Road is estimated to be 12 minutes.

Three measures are included as weights for the directed network; passenger flow, distance and time. Passenger flow, denoted as f_{ij}, is the number of passengers traveling between two adjacent stations, from station i to station j. Distance is the distance in kilometres between two adjacent stations, and time is how many minutes it takes to travel between two adjacent stations on average. The time spent on a journey is dependent on direction and the time of the day, where peak-hours have a different time distribution.

3.2 Network characteristics

3.2.1 Statistical properties and centrality measures

Four measures are applied to explore the network structure and centrality; edge betweenness, closeness, clustering coefficient and network efficiency.

The betweenness of an edge can be defined as the number of shortest paths in the network that uses this edge, while it can reveal bottlenecks and crucial links to the passenger flow on the network (DataCamp, 2017). The average edge betweenness is 1619 (normalised: 0.045) in the directed graph, and the edge with highest betweenness is between Bethnal Green and Liverpool Street Station with 8633 (normalised: 0.24). The closeness centrality measures how close a node is to all other nodes (Talluri, 2016). The normalised average closeness in the LU network is 0.05.

The clustering coefficient reveals how dense the network is connected, where local clustering 'captures the degree to which the neighbours of a given node link to each other' (Barabasi, 2014:p.26). The local clustering coefficient for an unweighted graph is calculated as follows (Barabasi, 2014:p.26):

$$C_i = \frac{2Li}{k_i(k_i-1)}$$

where k_i refers to the neighbours of station i, and L_i refers to the number of links between the neighbours. The average clustering coefficient for the undirected network is 0.027. This

indicates that the connectivity is low, which is also expected given the nature of transportation network.

The underground network is a connected graph, where passengers can get from every station to every other station in the network. But more importantly, network efficiency in a transit network signals how efficient the passengers can travel on the network. The maximum number of edges in an undirected graph is N(N-1)/2, corresponding to 35,778 (268*267/2) edges in this network. However, a clique is rare in real networks and we would expect the LU network to be sparse and the efficiency to be low (Barabasi, 2014). Moreover, 'it is not conceivable for a subway network to have all-to-all connections between stations' (Lee et al, 2008:p.6232).

Consequently, to get a more realistic efficiency measure, the LU network is instead compared against an ideal transportation network based on distances on straight lines between all stations (Lee et al, 2008). The efficiency formula is defined as follows (Latora & Marchiori, 2001: p.2)[2]:

$$E(G) = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}}$$

$$E(G_{id}) = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{h_{ij}}$$

$$\text{Normalised efficiency} = E(G) / E(G_{id})$$

N refers to the total number of stations, d_{ij} represents the shortest path between two nodes in the underground network, and h_{ij} is the ideal path between two stations calculated using the Euclidean distance. The normalised efficiency measure for the LU network is 0.83. Which is similar to other underground networks such as Seoul Subway which yields in a normalised efficiency measure of 0.75 (Lee et al, 2008).

3.2.2 Passenger flow distributions

The passenger flow on a link between two adjacent stations in the directed network is the sum of passengers traveling on the link in one direction. After normalizing the data, the weight distribution of the passenger flow is plotted in a log-log plot as seen in Figure 3. In accordance with other research (Lee et al, 2008; Roth et al, 2011; Xu, Mao & Bai, 2016), we observe a heterogeneous flow distribution with power-law behaviour. Further, this power-law distribution implies that the LU network is a scale-free network, where we can expect some stations to act as hubs (Barabasi, 2014). For medium and large weights, the $P(f_{ij})$ is fitted to a power-law with an exponent of 1.421.

[2] Part of code cited from Eisenman (2013)

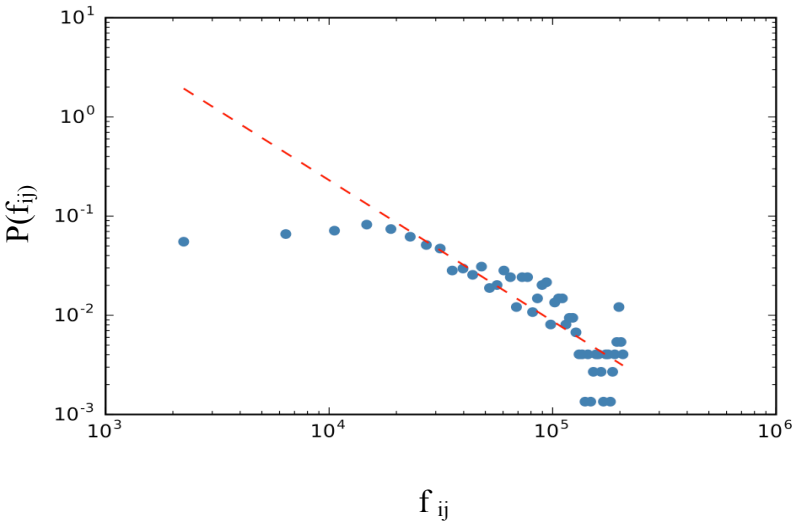


Figure 3: The probability distribution of passenger flow in the directed LU network on a log-log plot using a normalised histogram of the flows. The red stippled line follows a power-law fit with exponent 1.421.

The strength of a station in the weighted directed network is divided into incoming and outgoing passenger flow. The incoming-strength is defined as the sum of passengers on links directed to this station from other stations, while the outgoing-strength is defined as the sum of passengers on links directed from this station to other stations (Xu, Mao & Bai, 2016).

Studies of strength distribution in transit networks has led to various results. Whereas Lee et.al (2008, 2011) observed a log-normal distribution for the node strength in a transportation network, Xu, Mao and Bai (2016) found a power-law distribution. In this analysis, we observe a weak power-law behaviour ($R^2 = 0.34$), however a better fit compared to lognormal. As seen in Figure 4, the two distributions are almost identical.

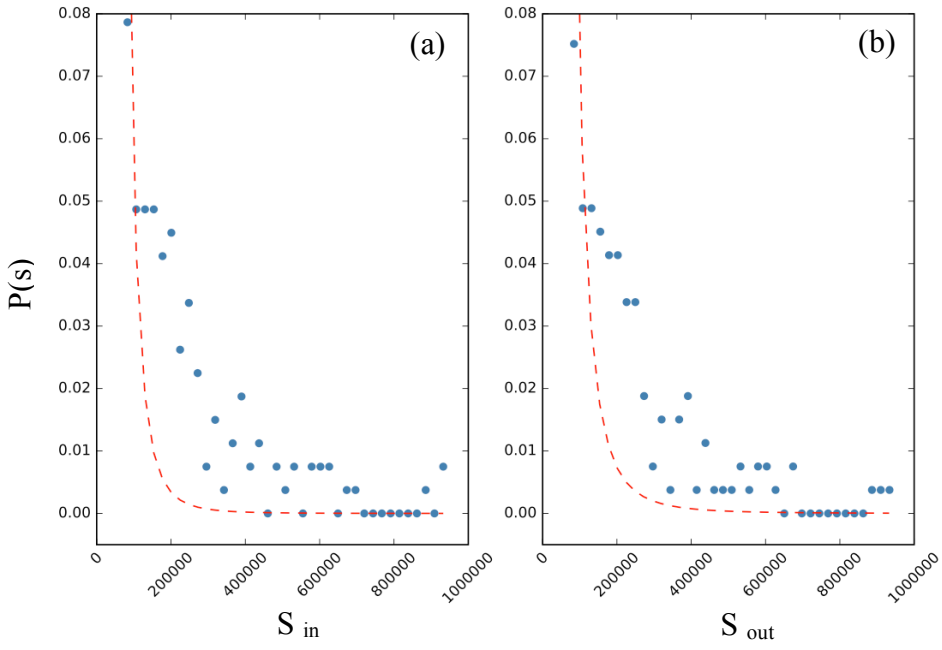


Figure 4: Two linear plots displaying the probability distribution of node strength in the directed network, including (a) distribution of incoming flow and (b) distribution of outgoing flow.

3.2.3 Maximum spanning tree of passenger flow

To provide a more detailed view of how passengers use the network and which edges are more important, a maximum spanning tree is applied using the OD dataset (Lee et al 2008; Lee et al, 2011). The links in the maximum spanning tree represents the journeys from origin O to destination D, and the weight is based on how many passengers who travel this route during a day. As suggested by Lee et al (2008) and Lee et al (2011), we also include the degree distribution for the maximum spanning tree to explore the hub structure in the network. The maximum spanning tree is shown in Figure 5a, where hubs with more than six degrees are labelled as a guidance. There are several hubs that have important links connected to them which carries many passengers during a day. This confirms the expectations from 3.2.2, that there are hubs in the network. Figure 5b displays the degree-distribution of the maximum spanning tree.

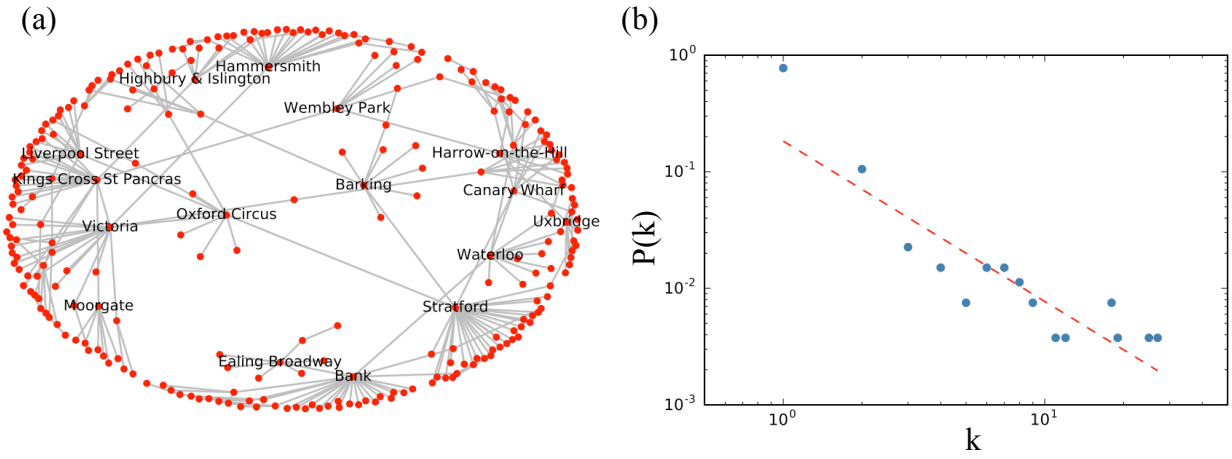


Figure 5: (a) Maximum Spanning tree based on OD journeys, and (b) degree distribution for the maximum spanning tree with degree k. The line is fitted to Power Law with an exponent of 1.376, and the goodness of fit is 0.7.

4. Analysis of route choice and passenger flow after link elimination

4.1 Analysis of passengers' behaviour

To predict how passengers will change their path during a closure, we first need to model passengers' route choice behaviour. To test how often passengers chose the shortest path in this dataset, an analysis of the current route dataset is conducted, where the chosen path is compared against the shortest path. The chosen path for each passenger is identified by calculating the shortest path between pairs of the *origin station*, the stations they *transfer* and the *destination station*, based on common lines serving the stations. The analysis distinguishes between the time of the day, where we could expect more people to take the shortest path, i.e. during peak commuting times in the morning.

The shortest path between all OD pairs is identified and compared against the chosen path. This results in 78.4 % of passengers taking the shortest path in total, with respectively 79.5% among AM-peak travellers and 76.9% among PM-peak travellers. On average, the passengers not taking the shortest route travelled 2.24 minutes longer than the shortest route, and had 2690 fewer passengers on their path, although with large standard deviation. Additionally, among the passengers not choosing the shortest path, 30% (7092) chose a route with fewer transfers.

To conclude, passengers often prefer to take the shortest path. However, the number of transfers is also important to many passengers, and sometimes considered more important than the shortest path.

4.2 Link elimination

Three different link removal approaches are presented to exemplify worst-case scenarios: highest edge betweenness centrality, highest edge weight and a combination of the two. As stated by Zhang et al (2011), transportation networks are fragile to attacks of stations with high betweenness centrality. Consequently, removal of an important edge with high betweenness centrality or high weight has a significant impact on the passenger flow and should therefore be of high importance to TfL.

As calculated in 3.2.1, the link on the Central line between Bethnal Green and Liverpool Street yields in the highest betweenness centrality. Further, the link with the highest weight lies on Victoria line, between Oxford Circus and Green Park, representing the link with highest number of passengers. To calculate the mixed removal, betweenness centrality (B_{ij}) are weighted based on passenger flow on the edge (f_{ij}), where passenger flow is included as the percentage difference from average passenger flow. The edge with highest mixed score is on Jubilee line between Westminster and Waterloo.

$$\text{Mixed Measure}_{ij} = B_{ij} \times \left(1 + \frac{f_{ij} - \overline{f_{ij}}}{\overline{f_{ij}}}\right)$$

The estimation of affected journeys is conducted based on the route dataset. Removal of the link with highest betweenness is estimated to affect 173,479 journeys on a typical weekday.

Total affected journeys for the highest weight is 152,176. Affected journeys for the mixed approach is 171 574. Figure 6 displays the affected journeys during a day for the three different types of link removal. In this analysis, only the combined removal (Westminster-Waterloo) will be used in the continuation of the prediction of passenger flow, where the link is removed in both directions.

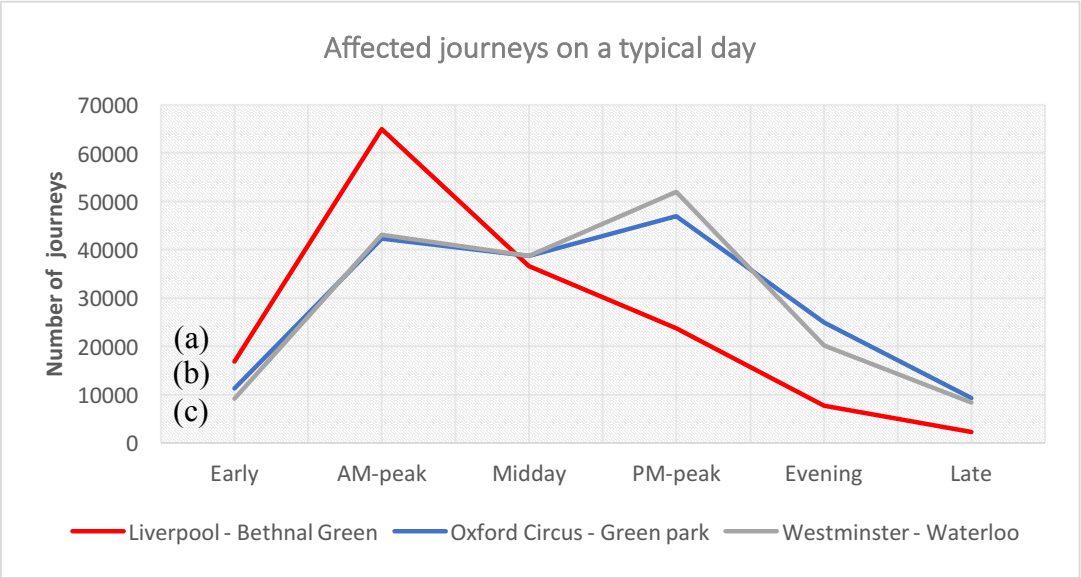


Figure 6: Affected journeys for different time periods with three removal strategies: (a) highest betweenness centrality, (b) highest edge weight, and (c) a combination of the two.

4.3 Analysis of passengers' route choice

4.3.1 Route set

After the link elimination, a route set with alternative routes for affected passengers is estimated. Here, the focus is on passenger flow after the closure is known to travellers, hence the route set is predicted based on OD data. A common issue when identifying the route set is the numerous alternative routes. However, there should be quite a few options in the route set to get accurate estimates (Prato, 2009). Furthermore, not all alternative routes are considered by the passenger due to lack of relevance (Prato, 2009).

As discovered in 4.1, both travel time and number of transfers are important factors when passengers choose their route. Consequently, the analysis is based on the approach suggested by Li and Sue (2003) where the number of transfers is combined with shortest path as an attempt to take psychological factors into account, indicating that fewer transfers are preferred. Li and Sue's algorithm works as a double search for shortest path: First, a search for the shortest path with no transfers is conducted. If this path is not shorter than a given value, a search for the shortest path with one transfer is conducted. Then, the same pattern is performed for up until four transfers (Li & Sue, 2003).

In this analysis, the approach is implemented by first finding all possible paths without overlap or loops. Then, the options are ordered based on transfers and only the shortest path is kept for each transfer number. Compared to Li and Sue, the available dataset includes the travel time for every edge in both directions, and the walking option is part of the search for

predicted path instead of a separate initial step. The resulting route set contains the shortest path for all possible numbers of transfers. An example of a route set is shown in table 1.

Table 1: The route set includes the shortest option per number of transfer. In this example, passengers travelling from Brixton to Finsbury Park have 3 options in their route set.

Number of transfers	Path	Time (minutes)
0	['Brixton', …, 'Finsbury Park']	26.4
1	['Brixton', …, 'Finsbury Park']	32.9
2	['Brixton', …, 'Finsbury Park']	38.2

4.3.2 Route choice

To predict the preferred route among the options in the route set, a percentage threshold is introduced for the number of transfers. As suggested by Li and Sue (2003), shorter options with more transfers than other options in the route set should be compared against a given value to decide if it is better than the path with fewer transfers. However, this value was not discussed in detail in the paper, and a complementary analysis is therefore conducted.

The threshold value is crucial to the analysis. Hence different threshold values are evaluated based on a training set to find the optimal value. The actual paths are identified through the route dataset, while the estimated routes are identified based on the route set from 4.3.1 and a percentage threshold value for number of transfers. The shortest path for each number of transfer is compared on travel time, where an option with more transfers needs to exceed a threshold to be chosen over a longer journey with fewer transfers. A threshold of 0.10, would indicate that a path with one more transfer should have a 10% shorter path to be chosen as the preferred route.

Accordingly, the predicted routes and the actual routes from the route dataset are compared with different threshold values. The optimal threshold is found in Figure 7. For this analysis, the optimal percentage threshold was 0.05, and the test set resulted in an accuracy of 88.63%. The predicted route choice from the route set is then found by primarily choosing the shortest path, but paths with more transfers need to save 5% travel time per extra transfer to be chosen over the longer options with fewer transfers.

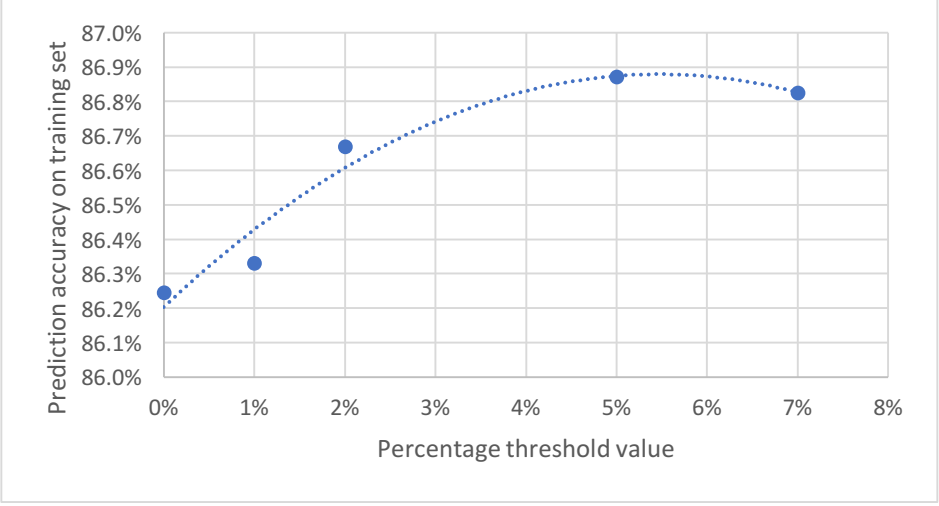


Figure 7: Optimal percentage threshold is 5 %.

4.4 Passenger flow after link elimination

The new passenger flow after closure is predicted based on the approach implemented in 4.3. The shortest path is often preferred, although a threshold value of 5% is added to account for the psychological factor that passengers prefer having fewer transfers. The option to walk short distances (<1,000 metres) is also added when predicting the new routes. The passenger flow before and after the link elimination are visualised in Figure 8.

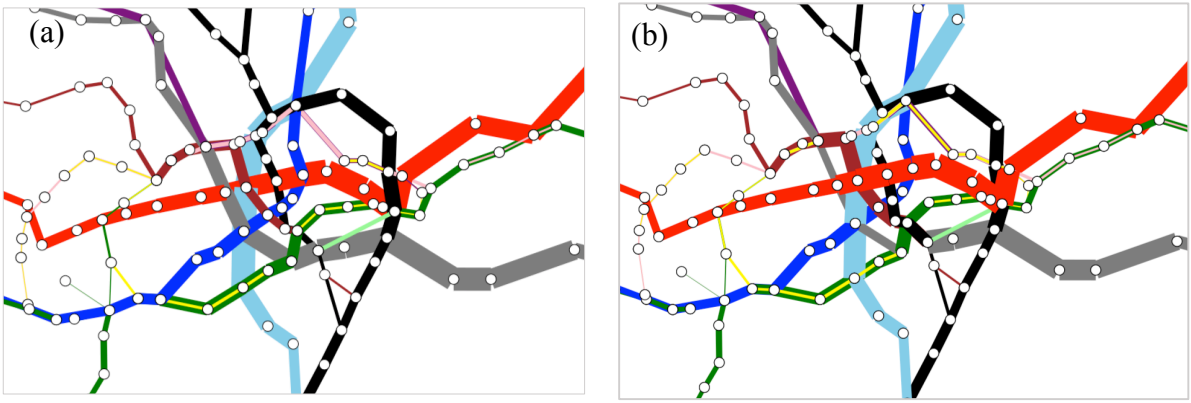


Figure 8: Passenger flow (a) before closure, and (b) after closure of the link between Waterloo and Westminster.

The new passenger flow on the LU network is compared against the passenger flow before the link removal. The elimination of the link between Waterloo and Westminster results in almost no change in either network efficiency or average betweenness centrality. To identify the exposed lines during the closure, Figure 9 displays the average change in passenger flow for the different lines during a typical day. Additionally, two tables with the absolute and percentage change is included in the appendix in table 1 and table 2.

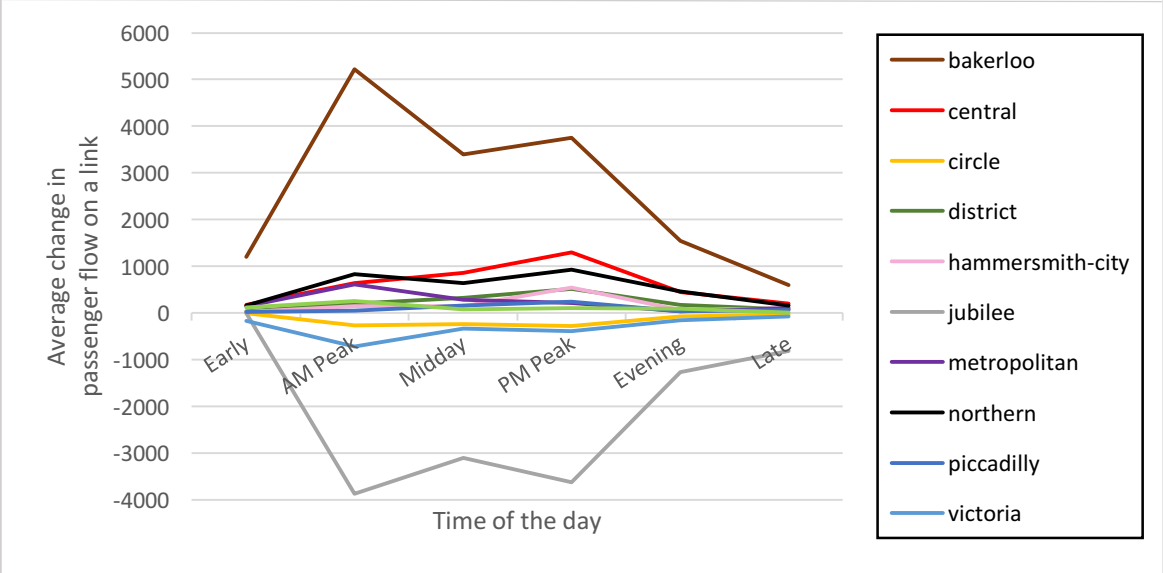


Figure 9: The average absolute change on edges for each line.

Generally, the Bakerloo line has the biggest increase of passengers and a reversed pattern compared to Jubilee line, indicating that the Bakerloo line takes a high proportion of passengers that considered taking the Jubilee line. Compared to an average of 21% increase

on each edge on the Bakerloo line, there are also other lines that are affected. The Northern line and the Central line have an average percentage increase in passengers of respectively 5.0% and 2.4% on each link.

The edge with the highest percentage and absolute increase in passengers is on the Bakerloo line between the stations Embankment and Waterloo with approximately 142,000 additional passengers, corresponding to a 286% increase in passengers on this edge. The edge with the biggest decline of passengers in absolute numbers is from the stations Green Park to Westminster, with approximately 149,000 fewer passengers. The link with the biggest percentage decrease in passengers is on the Jubilee line between the stations Bond Street and Green Park.

4.5 Discussion of results

There is no considerable increase in prediction accuracy by introducing transfers. This can either be due to transfers not having as great impact on route choices as first assumed, or the assumption that a change takes 2 minutes might be too low and is therefore already incorporating partial error.

The reason why there is not a big change in network efficiency or average betweenness centrality, is most likely because the network is resistant to closures and one removal will not result in much damage, as stated by Barabasi (2014).

As for the passenger flow prediction, the results emphasise how important it is for TfL to focus on the Bakerloo line if the link between Waterloo and Westminster station is closed. The results are as expected, where the most affected links are close to the removed link. Further, resources should be reallocated from stations connected to edges with decreasing passenger flow to edges with increasing passenger flow.

A more realistic view would be to incorporate train capacity since all passengers cannot choose the best option due to congestions. However, this analysis is mainly done to inform TfL of links that are exposed in a hypothetical scenario, as a result of where passengers most likely plan to travel during the closure. This analysis is therefore primarily a proof of concept rather than a complete analysis, and requires further research.

5. Limitations

The relevant datasets available from TfL are only provided for six periods of the day. A more detailed analysis requires a more detailed dataset with shorter time intervals. The data consists of only one typical weekday, where all abnormalities and closures are removed. If a dataset with actual passenger paths was released and closures was included, it would enhance the potential of the analysis and provide more accurate results. Moreover, the route data is based on surveys. Survey data is not always fully accurate, and passengers can provide answers incoherent to the route they actually take.

By predicting passenger flows based on OD data, all passengers with the same origin and destination will have the same route set and route choice, which is highly unlikely (Prato, 2009). As mentioned by Prato (2009), it is also important to include enough options in the route set. In this analysis, the route sets consist of up to four different routes, and a limitation is that relevant routes are excluded in favour of the shortest alternative with the same number of transfers. This is especially applicable for trips close to the city centre where the network is denser. Additionally, there will also be times when passengers choose a less optimal route due to the lack of information or change in plans during the travel. This introduces an unpredictable error to the model.

There are also other important factors that were not included in the analysis to find the optimal path. Due to lack of detail in the data, it was not possible to conduct an analysis based on train capacity or train schedules, which is hypothesised to have an impact on route choice. With respect to crowdedness, there were no clear patterns found in this dataset, however including crowdedness in the prediction of paths could improve accuracy and should be considered if more data is provided. Further, there is no information about the passengers in the route dataset due to privacy considerations. Therefore, passengers' characteristics or individual preferences cannot be included in the analysis.

6. Conclusion and recommendation

In both unplanned and planned closures, it is of high value to know where passengers are more likely to travel during the closure. There has been little research on the subject due to lack of data on passengers' route choice. However, this analysis has attempted to enhance the knowledge and contribute to existing knowledge on passenger travel dynamics. The hub structure and power-law distribution of the edges was identified, and the importance of this structure appeared when measuring the change in network efficiency after the link removal. The network is robust against one closure, but lines and stations get crowded and reallocation of resources is necessary.

When TfL knows which edges and stations have more or less passenger flow after a closure, reallocation is easier and the service can be improved. The removal strategy intended to illustrate a worst-case scenario, of which the stations Westminster and Waterloo was removed. However, while this analysis focused on predicting passenger flow and only investigated one edge removal, further analysis should include additional edges and conduct a deeper analysis of how these impact the network.

Furthermore, in addition to the transportation options underground and walking, additional analysis is recommended to include bus options. Additionally, crowdedness would be an interesting variable to include in the route choice model if more data is provided. With more detailed route information, it would also be possible to include train schedules and capacity. If data with passengers' route choice over time (e.g. data TfL collected through their Wi-Fi tracing) becomes publicly available, further analysis could incorporate weekends and seasonal patterns. It would be of high relevance to test the predictions on real passengers' route choice during closures, and conduct a more thorough analysis of how paths are chosen.

7. References

AaronD (2013) How can I quickly estimate the distance between two (latitude, longitude) points? [code] *Stackoverflow*. 1th April. Available from: https://stackoverflow.com/questions/15736995/how-can-i-quickly-estimate-the-distance-between-two-latitude-longitude-points [Accessed 21th August 2017]

Barabasi, A. (2014) *Network Science*. Cambridge, Cambridge University Press. Available from: http://barabasi.com/networksciencebook/ [Accessed 21th August 2017]

Ceapa, I., Smith, C., & Capra, L. (2012) Avoiding the Crowds: Understanding Tube Station Congestion Patterns from Trip Data. *UrbComp '12 Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. 134-141. Available from: doi: http://dx.doi.org/10.1145/2346496.2346518 [Accessed 21th August 2017].

DataCamp (2017) [video]. *Network Analytics in Python (Part 1)*. Available from: https://campus.datacamp.com/courses/network-analysis-in-python-part-1/bringing-it-all-together-4?ex=2 [Accessed 21th August 2017]

Dou, F., Yan, K., Huang, Y., Wang, L. & Jia, L. (2014) Optimal Path Choice in Railway Passenger Travel Network Based on Residual Train Capacity. *The Scientific World Journal*. 2014. Available from: doi: http://dx.doi.org/10.1155/2014/153949 [Accessed 21th August 2017]

Eisenman, L. (2013) Calculation of local/global efficiency. [code]. *Google Groups: networkx-discuss*. 30th October. Available from: https://groups.google.com/forum/#!topic/networkx-discuss/ycxtVuEqePQ [Accessed 21th August 2017]

Irvine, S (2016) Wifi Data Trial – Understanding London Underground Customer Journeys. *TfL Digital Blog*. Weblog. Available from: https://blog.tfl.gov.uk/2016/11/23/wifi-data-trial-understanding-london-underground-customer-journeys/ [Accessed 21th August 2017]

Jia, W. & Chow, M. (2014) *How crowded is crowded? A practitioner's tool to assessing rail congestion*. 2015 Transport Research Board Annual Meeting. Available from: http://docs.trb.org/prp/15-0789.pdf [Accessed 21th August 2017]

Latora, V. & Marchiori, M. (2001) Efficient Behaviour of Small-World Networks. *Physical Review Letters*. 87(19). Available from: doi: https://doi.org/10.1103/PhysRevLett.87.198701 [Accessed 21th August 2017]

Lee, K., Goh, S., W., Park, J.S., Jung, W. & Choi, M.Y. (2011) Master equation approach to the intra-urban passenger flow and application to the Metropolitan Seoul Subway system. *Journal of Physics A*: *Mathematical and Theoretical*. 44(11), 115007. Available from: doi: https://doi.org/10.1088/1751-8113/44/11/115007 [Accessed 21th August 2017]

Lee, K., Jung, W., Park, J.S. & Choi, M.Y. (2008) Statistical analysis of the Metropolitan Seoul Subway System: Network structure and passenger flows. *Physica A: Statistical Mechanics and its Applications*. 387(24), 6231-6234. Available from: doi: https://doi.org/10.1016/j.physa.2008.06.035 [Accessed 21th August 2017]

Li, S. G. & Su, Y. M. (2003) Optimal transit path finding algorithm based on geographic information system. *Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE*. 2, 1670-1673. Available from: doi: https://dx.doi.org/10.1109/ITSC.2003.1252767 [Accessed 21th August 2017]

Li, W. & Zhu, W. (2016) A dynamic simulation model of passenger flow distribution on schedule-based rail transit networks with train delays. *Journal of traffic and transportation engineering*. 3 (4), 364-373. Available from: doi: https://doi.org/10.1016/j.jtte.2015.09.009 [Accessed 21th August 2017]

Lin, T., Shalaby, A. & Miller E. (2016) Transit user behaviour in response to subway service disruption. London. CSCE Annual Conference, Resilient Infrastructure. Available from: http://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1239&context=csce2016 [Accessed 21th August 2017]

Pnevmatikou, A.M., Karlaftis, M. G. & Kepaptsoglou, K. (2015) Metro service disruptions: how to people choose to travel? *Transportation*. 42(6), 933-949. Available from: doi: https://doi.org/10.1007/s11116-015-9656-4 [Accessed 21th August 2017]

Prato, C. G. (2009) Route choice modelling: past, present and future research directions. *Journal of Choice Modelling*. 2(1), 65-100. Available from: doi: https://doi.org/10.1016/S1755-5345(13)70005-8 [Accessed 21th August 2017]

Roth, C., Kang, S. M., Batty, M. & Barthélemy, M. (2011) Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. *PLoS ONE*. 6(1), e15923. Available from: doi: https://doi.org/10.1371/journal.pone.0015923 [Accessed 21th August 2017]

Talluri, K. (2016) BA Network Analytics Week 2, Lecture 4. [Lecture] Network Analytics. Imperial College London, 25th November.

Transport for London (2009) Statistics on station closures. *What Do They Know*. 30th December. [Excel spreadsheet] Available from: https://www.whatdotheyknow.com/request/statistics_on_station_closures [Accessed 21th August 2017]

Transport for London (2017) *TfL Rolling Origin and Destination Survey*. Available from: https://data.london.gov.uk/dataset/tfl-rolling-origin-and-destination-survey [Accessed 21th August 2017]

Transport for London (n.d.) *Strategic impact assessment*. Available from: https://tfl.gov.uk/info-for/urban-planning-and-construction/transport-assessment-guide/transport-assessment-inputs/strategic-impact-assessment [Accessed 21th August 2017]

Wu, J.J., Gao, Z.Y. & Sun, H.J. (2007) Effects of the cascading failures on scale-free traffic networks. *Physica A: Statistical Mechanics and its Applications.* 378(2), 505-511. Available from: doi: https://doi.org/10.1016/j.physa.2006.12.003 [Accessed 21th August 2017]

Xu, Q., Mao, B.H. & Bai, Y. (2016) Network structure of subway passenger flows. *Journal of Statistical Mechanics: Theory and Experiment.* 2016(3), 033404. Available from: doi: https://doi.org/10.1088/1742-5468/2016/03/033404 [Accessed 21th August 2017]

Zhang, J., Xu, X., Hong, L., Wang, S. & Fei, Q. (2011) Networked analysis of the Shanghai subway network, in China. *Physica A: Statistical Mechanics and its Applications.* 390(23-24), 4562-4570. Available from: doi: https://doi.org/10.1016/j.physa.2011.06.022 [Accessed 21th August 2017]

Zhu, S. & Levinson, D. (2015) Do people use the shortest path? An empirical test of Wardrop's first principle. *PLoS ONE.* 10(8): e0134322. Available from: doi: https://doi.org/10.1371/journal.pone.0134322 [Accessed 21th August 2017]

Dataset sources

Transport for London (2017) *Busiest times on trains and in stations* [Excel spreadsheets]. Tfl.gov.uk. Available from: https://tfl.gov.uk/info-for/open-data-users/our-open-data?intcmp=3671 [Accessed 21th August 2017]

Transport for London (2017) *London Underground passenger counts data* [Excel spreadsheets]. Tfl.gov.uk. Available from: https://api-portal.tfl.gov.uk/docs [Accessed 21th August 2017]

Transport for London (2017) *Rolling Origin & Destination Survey* [Excel spreadsheets] Tfl.gov.uk. Available from: https://api-portal.tfl.gov.uk/docs [Accessed 21th August 2017]

Transport for London (2013) Distances between adjacent Oyster stations [Excel spreadsheet]. *What Do They Know.* 29th July. Available from: https://www.whatdotheyknow.com/ request/distances_between_adjacent_oyste [Accessed 21th August 2017]

Appendix

Table 1: Average percentage change on each edge on each line, after closure of link between Waterloo and Westminster

LINE	EARLY	AM PEAK	MIDDAY	PM PEAK	EVENING	LATE	TOTAL
BAKERLOO	80.5%	32.9%	20.1%	20.6%	17.2%	16.1%	23.7%
CENTRAL	6.4%	1.9%	2.1%	2.4%	2.0%	2.5%	2.1%
CIRCLE	-1.0%	-5.3%	-4.1%	-4.7%	-2.8%	-4.4%	-4.6%
DISTRICT	1.6%	0.7%	1.6%	1.8%	1.7%	1.5%	1.1%
HAMMERSMITH-CITY	6.6%	2.6%	2.4%	7.5%	2.0%	6.7%	4.4%
JUBILEE	-2.7%	-11.0%	-12.2%	-12.8%	-12.1%	-19.9%	-12.7%
METROPOLITAN	6.3%	2.4%	2.9%	2.3%	1.5%	21.6%	2.7%
NORTHERN	9.4%	5.1%	3.8%	4.5%	4.3%	3.0%	4.2%
PICCADILLY	0.9%	0.2%	0.5%	0.8%	0.3%	1.3%	0.6%
VICTORIA	-1.1%	0.1%	-0.4%	-1.0%	-0.9%	-1.3%	-0.9%
WATERLOO-CITY	36.5%	7.1%	1.3%	0.4%	1.3%	1.6%	1.5%

Table 2: Average absolute change on each edge on each line, after closure of link between Waterloo and Westminster

LINE	EARLY	AM PEAK	MIDDAY	PM PEAK	EVENING	LATE	TOTAL
BAKERLOO	1134	4822	3120	3459	1439	549	14526
CENTRAL	170	634	934	1715	510	206	4168
CIRCLE	-9	-238	-145	-174	-35	-45	-679
DISTRICT	45	278	328	512	167	79	1410
HAMMERSMITH-CITY	-13	-183	-108	95	-8	27	-190
JUBILEE	-32	-3871	-3222	-4060	-1359	-799	-14078
METROPOLITAN	151	608	273	212	39	104	1384
NORTHERN	179	969	737	1034	492	177	3588
PICCADILLY	-20	-127	97	196	26	58	229
VICTORIA	-97	-368	-207	-304	-151	-74	-1202
WATERLOO-CITY	114	245	75	108	87	12	640