The code used for this assessment is hosted in this github repository

1 Part 1: London's underground resilience

1.1 Topological network

1.1.1 Centrality measures

Three centrality measures were selected to identify most important nodes for the London's Underground network, including degree centrality, betweenness centrality, and closeness centrality.

1. Degree centrality

Degree centrality is a measure to assess the importance of a node within the network. It is defined as the number of links attached to a node. Nodes with more connections is considered as more important to the network. The degree centrality of a node i in an undirected graph can be mathematically expressed as:

$$c_i = \sum_{j=1}^{n} A_{ij}, \tag{1}$$

where n is the number of nodes in the whole network, and A_{ij} is the element of the adjacency matrix A corresponding to the edge between nodes i and j. If there is an edge between nodes i and j, $A_{ij} = 1$; otherwise, $A_{ij} = 0$.

For a directed graph, there are two types of degree centrality measures: in-degree centrality and out-degree centrality. In-degree centrality measures the number of edges pointing to the node itself, while out-degree centrality measures the number of outgoing edges from the node. The mathematical notations for them are:

$$c_i^{in} = \sum_{j=1}^{n} A_{ij}, \tag{2}$$

$$c_j^{out} = \sum_{i=1}^{n} A_{ij}, \tag{3}$$

where A_{ij} in in-degree centrality is the element of the adjacency matrix A corresponding to the edge from node j to node i, and A_{ij} in out-degree centrality is the element of the adjacency matrix A corresponding to the edge from node i to node j

In the context of the London Underground, if a station has more connections, than it means passengers can go to more stations through it. If a station with high degree centrality is closed, it can result in reduced connectivity within the network, making it more difficult for passengers to travel between certain stations.Therefore, degree centrality can serve as the measure of importance in terms of how many stations are linked to a station.

2. Betweenness centrality

Betweenness centrality measures the importance of a node within a network based on the position of the node. It quantifies how many shortest paths between all pairs of nodes in the network pass through the vertex or the edge. Nodes with higher betweenness centrality can serve as bridges between other parts of the network. Therefore, more information will pass through it, allowing them to have more control over the flow of information. Mathematically, the betweenness centrality can be expressed as:

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}}, \tag{4}$$

where x_i is the betweenness centrality of node i, g_{st} is the total number of shortest paths between nodes s and t, and n_{st}^i is the number of shortest paths between nodes s and t that pass through node

i.

In the context of London Underground, a station with high betweenness centrality means it is part of many shortest paths within the network. Therefore, it can act as a transfer hub, connecting different areas of the city. In terms of network resilience, if a station with higher betweenness centrality is closed, it is more likely to cause disruption as passengers may need to take longer or more complex routes to reach their destinations.

3. Closeness centrality

Closeness centrality measures the importance of a node within a network based on its proximity to other nodes, quantifying how close a node is to all other nodes in the network. Nodes with higher closeness centrality means they are closer to all other nodes, making it more quickly for them to spread information throughout the network. The mathematical formula of closeness centrality is:

$$C_i = \frac{1}{l_i} = \frac{n}{\sum_j d_{ij}}, \tag{5}$$

where d(i, j) is the geodesic distance between nodes i and j, and n is the total number of nodes in the network.

In the context of London Underground, a station with high closeness centrality is closer to other stations, making it an important part of the network's overall connectivity. If it is closed, the underground network's resilience may be affected negatively, making the network more vulnerable to further disruptions, as passengers may have fewer alternative routing options if further disruptions occur.

The following table is the first 10 ranked nodes for each of these 3 measures. These are computed by using functions: `nx.degree_centrality(G)`, `nx.betweenness_centrality(G, normalized=False)`, and `nx.closeness_centrality(G)`. G is the unweighted network that is created using the data in `london_flows.csv`.:

Table 1: Top 10 ranked nodes for 3 meausres

Rank	degree_t		betweenness_t	
1	**Stratford**	0.9221	**Stratford**	7785.9671
2	**Highbury & Islington**	0.8065	**Liverpool Street**	2710.3772
3	**Whitechapel**	0.7814	**Canary Wharf**	2208.6279
4	**West Brompton**	0.7764	**Bank and Monument**	2208.6279
5	**Canada Water**	0.7714	**Canning Town**	2192.8480
6	**Canary Wharf**	0.7714	**West Ham**	1939.6418
7	**Liverpool Street**	0.7688	**Highbury & Islington**	1818.9040
8	**Bank and Monument**	0.7663	**Whitechapel**	1554.9659
9	**Richmond**	0.7663	**Canada Water**	1413.9777
10	**Canning Town**	0.7638	**Shadwell**	1348.5889
Rank	**closeness_t**			
1	**Stratford**	0.9277		
2	**Highbury & Islington**	0.8361		
3	**Whitechapel**	0.8206		
4	**West Brompton**	0.8172		
5	**Canada Water**	0.8139		
6	**Richmond**	0.8106		
7	**Canary Wharf**	0.8106		
8	**Bank and Monument**	0.8106		
9	**Liverpool Street**	0.8089		
10	**Canning Town**	0.8089		

All three measures gives the same first ranked node: Stratford station, and degree centrality measure and closeness centrality measure give the same top three ranked nodes.

1.1.2 Impact measures

Global efficiency and average clustering coefficient are the impact measures chosen to evaluate the impact of the node removal on the underground network.

Global efficiency is defined to be the inverse of the average shortest path length in a network (Latora & Marchiori 2001). It can be mathematically expressed as:

$$E_{glob}(G) = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}}, \tag{6}$$

where N is the number of nodes in the network, and d_{ij} is the shortest path length between nodes i and j. In python, the function `nx.global_efficiency()` can be used to compute global efficiency. This global measure gives an overall assessment of the network's efficiency in transmitting information. When a node is removed, the global efficiency is likely be affected, for it could potentially increase the shortest path length between certain pairs of nodes. By comparing drops in the global efficiency before and after nodes removal, it will show which nodes are the most crucial one to the network.

The clustering coefficient measures how much are nodes within the network clustering together. The average clustering coefficient is the mean of the clustering coefficients of all nodes, measuring the extent of clustering of the entire network. Its mathematical formula is:

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i, \tag{7}$$

where N is the number of nodes in the network, and C_i is the clustering coefficient of node i. In python, it can be computed by function `nx.average_clustering()`. A higher average clustering coefficient value indicates a more tightly interconnected network. By comparing drops in the average clustering coefficients before and after nodes removal, it can show which nodes affected the interconnectedness of the network mostly.

The global efficiency and average clustering coefficient measures are not specific to the London Underground because they are general network analysis measures that can be applied to any other network, such as social networks, other types of transportation networks, and communication networks.

1.1.3 Nodes removal

In this section, nodes were removed non-sequentially (Strategy A) and sequentially (Strategy B) ten times according to the ranking order given by three centrality measures. In each strategy of each measure, global efficiency and average clustering coefficient, along with their drop after each removal were computed and plotted respectively. The criteria for assessing which strategy and measure are better in this research is to compare which of these gives the greatest difference in the impact measures, since greater difference indicates more influence on the network.

In Figure 1, the Strategy A and B give exactly the same results for degree centrality and closeness centrality, while for betweenness centrality they give different outcomes. In this figure, the red line, which is the global efficiency under strategy B using betweenness centrality, has the largest drop. Therefore, it suggests that betweenness centrality better represents the significance of a station in maintaining the operation of the underground, and strategy B reflects resilience more effectively.

In Figure 2, the Strategy A and B still give exactly the same results for degree centrality and closeness centrality, while for betweenness centrality they give different outcomes. In this figure, degree centrality, closeness centrality, and betweenness centrality under strategy B have similar amount of decrease, with betweenness centrality under strategy B having the largest one. Therefore, average clustering coefficient gives similar result as global efficiency in the aspect of representation of importance and reflection of resilience.

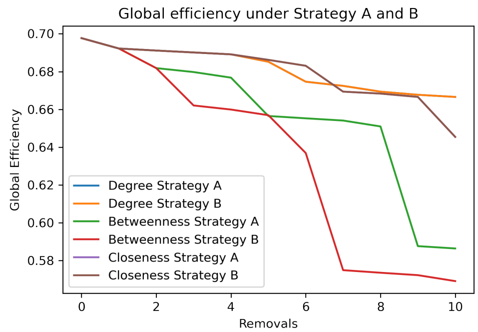

Figure 1: Global efficiency

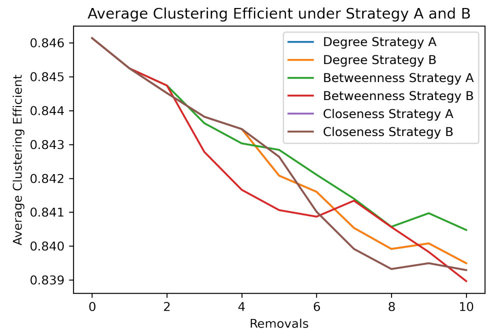

Figure 2: Average Clustering Coefficient

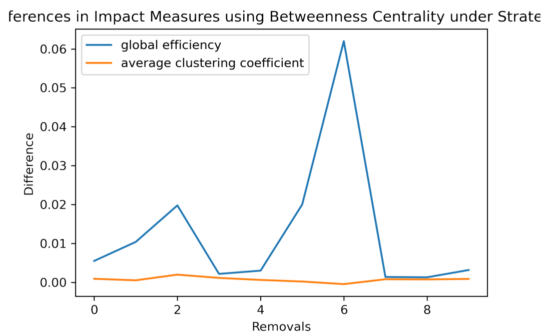

Figure 3: Drops of each node removal in global efficiency and average clustering coefficient

To identify which impact measure can better assess the damage after each node removal, drops in the two impact measures after each node removal are plotted in Figure 3, which are computed by using betweenness centrality measure under strategy B. This figure suggests that global efficiency is better than average clustering coefficient in assessing the damage, as it fluctuates much more, making it easier to observe which node removal caused the greatest impact. From this graph, the difference in global efficiency peaked at the 6_{th} removal, corresponding to the removal of Shadwell station. Therefore, Shadwell should be the station whose closure will have the most serious impact on this underground network.

1.2 Flows: weighted network

1.2.1 Old vs new measure

In weighted network, betweenness centrality is computed by taking flows into account as weight. In python, the function needs to be modified as `nx.betweenness_centrality(G,weight='flows',normalized=False)`. The result of the top 10 important nodes according to this function is:

Table 2: Top 10 nodes using weighted betweenness centrality

Rank	betweenness_w	
1	**West Ham**	9.69E+93
2	**West Brompton**	5.75E+93
3	**Shepherd's Bush**	2.83E+93
4	**Kew Gardens**	1.64E+93
5	**Surrey Quays**	8.27E+92
6	**Kenton**	6.37E+92
7	**Richmond**	6.09E+92
8	**Willesden Junction**	2.21E+92
9	**Kentish Town West**	1.42E+92
10	**Stratford**	1.08E+92

Comparing the Table 2 with Table 1, only West Ham station and Stratford remained in the ranking, with West Ham jumping to the first important node, and Stratford being the 10_{th}.

1.2.2 Impact measure with flows

Global efficiency would be adjusted for a weighted network by including flows as the weight. As shown in Equation 6, global efficiency is calculated based on shortest path length, so it can be done by including weight in the calculation of shortest path length using the function `nx.shortest_path_length(G, weight='flows')` and writing a function called `global_efficiency_weighted(G)`. An alternative measure is to calculate the average shortest path length with weight. Since global efficiency is inversely proportional to shortest path length, and its scale is always $0 \leq E_{glob}(G) \leq 1$, average shortest path length may result in larger difference after each node removal, making the impact more obvious and noticeable. The equation is as follows (*Average_shortest_path_length — NetworkX 3.1 Documentation* n.d.):

$$a = \sum_{s,t \in V, s \neq t} \frac{d(s,t)}{n(n-1)}, \tag{8}$$

where V is the pair of nodes in network G, d(s, t) is the shortest path from s to t, and n is the number of nodes in G. A larger value of this measure would suggest a greater impact on the network and passengers because of the closure of a station.

1.2.3 Experiment with flows

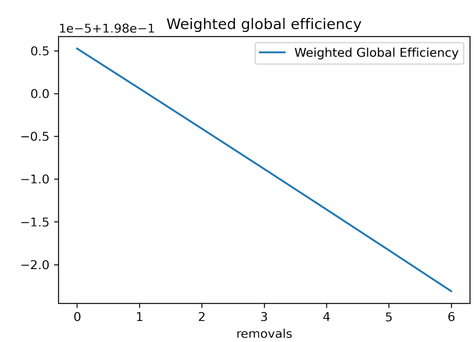

Figure 4: Weighted global efficiency

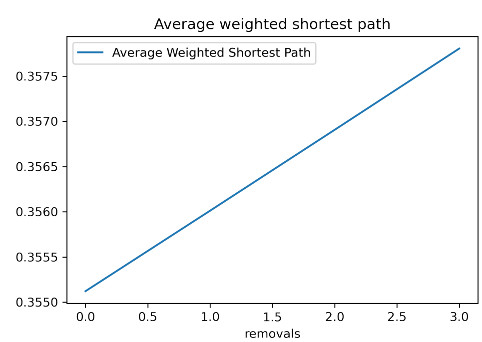

Figure 5: Average weighted shortest path

In this section, weighted global efficiency was computed before and after each removal. Specifically, the three highest ranked nodes according to weighted betweenness centrality were removed by adopting strategy B in 1.1.3. Then this same experiment was repeated on the remaining nodes, resulting in totally 6 removals, and 7 weighted global efficiencies including the first one when no nodes was removed. From Figure 4, the weighted global efficiency kept dropping steadily, but the differences were all too small, so it is better to look at average weighted shortest path instead. Average shortest path was computed after each of the three removals. From Figure 5, this impact measure increased steadily, giving the same information as weighted global efficiency that the impact on passengers was growing. The differences in average weighted shortest path were not as small as the weighted global efficiency, suggesting it may give a clearer understanding of the importance of the stations.

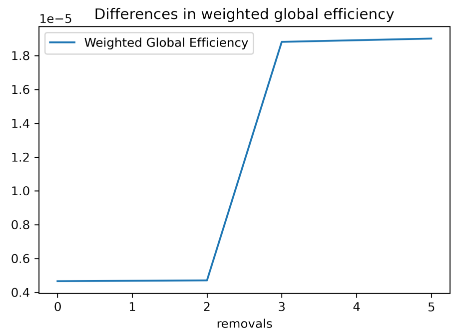

Figure 6: Drops in weighted global efficiency

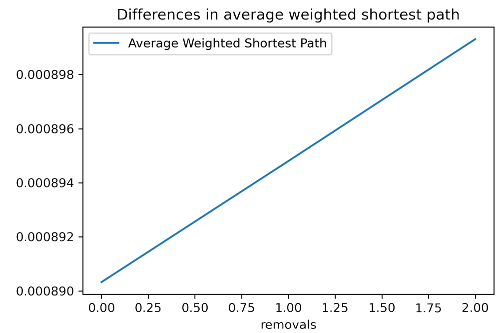

Figure 7: Increments in average weighted shortest path

Figure 6 and 7 shows the drops in weighted global efficiency and increments in average weighted shortest path. Both plots shows the differences are increasing, meaning that the last station removed should be the most influential one. In the first three removals, both impact measures give the same result that Shepherd's Bush should be the most influential station on passengers. In the first six removals, weighted global efficiency gives that West Hampstead should be the station with the greatest impact. However, these analysis and results should be verified with further investigation. If this increasing trend of differences is correct, it means the station with lower betweenness centrality would have greater impact on passengers, which is unlikely to be true. Therefore, further research on these measures and results should be conducted to verify this relationship.

2 Part 2: Spatial Interaction models

2.1 Models and calibration

2.1.1 Spatial interaction models

A classic spatial interaction model is the gravity model. The equation for it is:

$$T_{ij} = kO_i^{\alpha} D_j^{\gamma} d_{ij}^{-\beta}, \tag{9}$$

where T_{ij} is the flow between location i and location j, O_i is a vector of origin attributes such as population, D_i is a vector of destination attributes such as population, d_{ij} is the matrix of cost of the flows between i and j, and k, α, γ and β are model parameters to be estimated. Specifically, k is a constant that ensures the total flows are conserved. It can be computed by:

$$k = \frac{T}{\sum_i \sum_j O_i^{\alpha} D_j^{\gamma} d_{ij}^{-\beta}}, \tag{10}$$

where T is the sum of the observed flows. α and γ capture the impact of the origin attributes and destination attributes, and can control their influence. β determines the impact of the cost.

Depending on the restrictions in specific contexts, there are unconstrained and constrained model. In unconstrained mode, there is no restriction on the total inflows or outflows. In singly constrained model, either the inflows (attraction-constrained) or outflows (production-constrained) are fixed, and in doubly constrained model, both inflows and outflows are fixed.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                  flows   No. Observations:                61413
Model:                            GLM   Df Residuals:                    61409
Model Family:                 Poisson   Df Model:                            3
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:            -1.2785e+06
Date:                Wed, 26 Apr 2023   Deviance:                   2.3848e+06
Time:                        17:03:32   Pearson chi2:                 4.76e+06
No. Iterations:                     6   Pseudo R-squ. (CS):              1.000
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept         -3.7475      0.014   -273.078      0.000      -3.774      -3.721
log_population     0.7325      0.001   1048.145      0.000       0.731       0.734
log_jobs           0.7608      0.001   1163.936      0.000       0.759       0.762
log_distance      -0.6228      0.001   -674.846      0.000      -0.625      -0.621
==================================================================================
```

Figure 8: Regression Result

2.1.2 Calibration of model

Unconstrained model is used in this research because of its simplicity and flexibility. This model is the simplest spatial interaction model, with no constraints and easy to understand and implement. It can be advantageous when we have limited data and do not have constraints requirements. This model also allows for more flexibility in the distribution of flows between stations. Since it doesn't impose any constraints on the total inflows or outflows, it can better capture the variations in the flows between different stations, which can help when we explore different scenarios, such as changes in jobs or transportation costs. However, this model does not ensure to conserve the total flows, and is sensitive to parameters. Inaccurate parameters can lead to unreliable results.

The parameters are calibrated by taking log transformation on data and perform Poisson Regression. The result is in Figure 8. The calibrated parameter β for the cost function is -0.6228. The R-Squared and the Root Mean Squared Error are 0.3212 and 108.334 respectively, which suggests the model accounts for about 32.12% of the variation of flows. It means this model is not good and accurate enough. It is possible that constrained model would be better for this network.

2.2 Scenarios

2.2.1 Scenario A

In scenario A, the number of jobs at Canary Wharf is updated by reducing it by 50%. Then the flows between stations are recomputed using the unconstrained gravity model and the calibrated parameter beta in the last section. However, the sum of the new flows is smaller than the sum of the original flows. To conserve the total number of flows, the parameter k in the unconstrained model equation 9 is adjusted by using equation 10. Then the new flows are recomputed and summed using the new k and the previous calibrated β. Now the sum of the new flows is approximately the same as the sum of the original flows, with slight rounding error.

2.2.2 Scenario B

The parameter β of the cost function controls how cost impacts the flows. A large β will raise the cost of commuting, therefore decrease the flows. In scenario B, the transport cost has increased, meaning the value of β should increase as well. The two values selected to reflect this are 0.8 and 1. After the calculations using these new β values, it can be observed the flows have decreased dramatically as the β value increases. This is consistent with the model equation 9.

2.2.3 Analysis

To observe how the overall distribution of flows has changed, the summary statistics of the three scenarios and the original flows was calculated and presented in Table 3.

According to the summary statistics results, the overall distribution of flows did not change when jobs at Canary Wharf decreased by 50%, while when the transport cost increases, the flows significantly decreased. When the parameter β rises from the original 0.62 to 0.8, the mean flows decreased by

Table 3: Summary statistics for different scenarios

Scenario	Mean	Median	Standard Deviation
Original	25.1145	8.0000	73.1065
Job Decrease	25.1145	8.0000	73.1065
Transport Cost 1	5.3572	1.3768	18.6248
Transport Cost 2	0.9698	0.2019	4.2339

approximately 78.67%, and the standard deviation also decreased from 73.11 to 18.62. When β rises from 0.62 to 1, the mean flows decreased by around 96.14%, and the standard deviation decreased from 73.11 to 4.23. This result means when transport cost rises, the flows across most stations are reduced dramatically. Comparing the two scenarios where jobs at Canary Wharf decreased by 50% and the cost of transport increased, the Scenario B has much more significant impact in the redistribution of flows.

To observe more clearly the flows, each scenario is plotted in a heat map. The locations of stations are from TransportForLondon (2018).

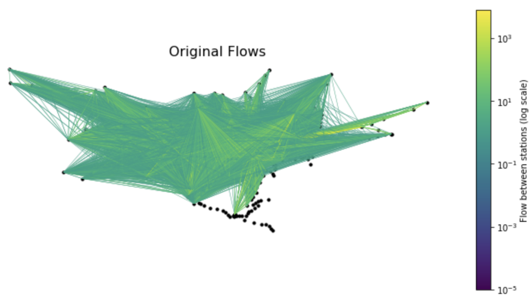

Figure 9: Original Flows

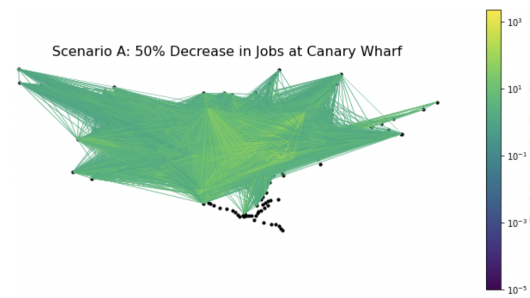

Figure 10: Scenario A

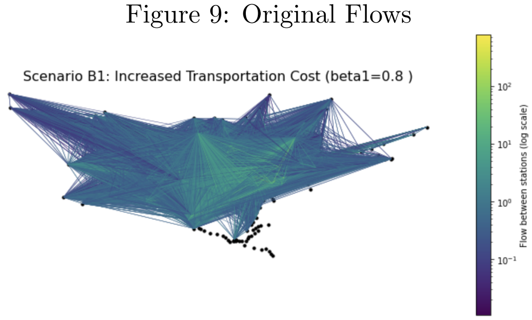

Figure 11: Scenario B beta 1

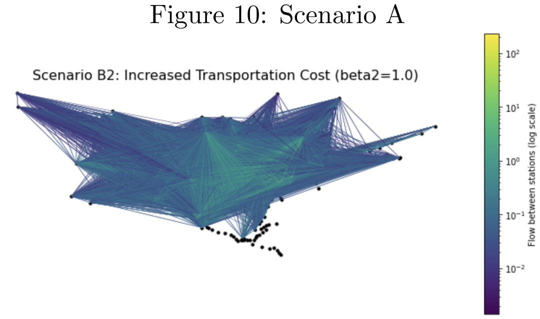

Figure 12: Scenario B beta 2

The scales of flows are transformed into logarithm scales, since a few values are very large with most of the values are much smaller, so using the logarithm scale can help better visualize the differences. As flows decrease, the color of the line in the heat map gets darker. These maps suggest flows within the center part of London are less affected than other parts in either scenarios. Basically, the maps give the same results as the summary statistics that Scenario B has much more significant impact on flows.

References

Average_shortest_path_length — NetworkX 3.1 Documentation (n.d.).
URL: *https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.shortest_paths.generic.average_shortest_path_length.html*

Latora, V. & Marchiori, M. (2001), 'Efficient behavior of small-world networks', *Phys. Rev. Lett.* **87**, 198701.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.87.198701*

TransportForLondon (2018).
URL: *https://foi.tfl.gov.uk/FOI-1451-1819/Stations*$_2$0180921.*csv*