Urban Simulation Assignment

This is the assignment for the UCL CASA module Urban Simulation (CASA0002)
Code & data is hosted and availble at [Github](#)
Word count: 2968

Part 1: London's underground resilience:

I. Topological network:

In this section, the data of the London tube network in csv format provided on Moodle will be used. The network model can be generated by using the data, only the topological relationships between nodes are considered, and the network is named as **G**.

1.1 Centrality measures:

Centrality measures of a network are used to measure the importance of nodes in the network in different aspects. These important aspects can include the influence of nodes in terms of information dissemination, control, influence, and connectivity. In this section, three measures will be used: Degree centrality, Closeness centrality and Betweenness centrality.

Degree Centrality (DC) is a commonly used centrality metric to measure the importance or centrality of nodes in a network. In London tube networks, degree centrality can be used to assess the interactivity of stations, i.e., the number of direct connections to other stations. DC measures the number of direct connections of a node in the network, i.e., the degree of the node. A higher degree of a node indicates more direct connections to other nodes and thus plays a more important role in the network.
For a node i in an undirected network, degree centrality can be calculated by the following equation:

$$C_d(i) = \frac{k_i}{N-1} \tag{1}$$

where $C_d(i)$ means the degree centrality of node i, k_i denotes the number of degrees of node i (i.e., the number of edges directly connected to node i), and N means the total number of nodes in the network.

Closeness Centrality (CC) is a metric used to measure the closeness between nodes and other nodes in a network. Closeness centrality measures the average shortest path length from a node to other nodes, i.e., the average distance between a node and other nodes. A higher proximity centrality of a node indicates that the distance

between the node and other nodes is shorter and the connection is tighter. In the London tube network, Closeness centrality indicates the average shortest path of a station to other stations, and the higher the value, the shorter the journey to the rest of the stations in the network and the more important the station is.

For a node i in an undirected network, closeness centrality can be calculated by the following equation:

$$C_c(i) = \frac{N - 1}{\sum_{j=1}^{N} d(i, j)} \tag{2}$$

where $C_c(i)$ means the closeness centrality of node i, $d(i, j)$ means the shortest path length from node i to node j, and N means the total number of nodes in the network.

Betweenness Centrality (BC) is used to measure the extent to which a node acts as a mediator in the network, i.e., how often a node is on the shortest path between connecting other nodes. Nodes with high betweenness centrality usually have strong information transfer and control capabilities in the network. The higher the betweenness centrality of a node, the more frequently the node acts as an intermediary in the network and plays an important role in the connectivity and information transfer of the network. In the London tube network, the value of betweenness centrality indicates the importance of the station as a transit station, and the higher the value, the more important the station is as a transit point.

For a node i in an undirected network, betweenness centrality can be calculated by the following equation:

$$C_B(v) = \sum_{s,t \in V} \frac{\sigma(s, t|v)}{\sigma(s, t)} \tag{3}$$

where $C_B(v)$ means the betweenness centrality of node i. V is the set of nodes. $\sigma(s, t)$ is the number of shortest (s, t)-paths, and $\sigma(s, t|v)$ is the number of those paths passing through some node v other than s, t. if $s = t, \sigma(s, t) = 1$ and if $v \in s, t, \sigma(s, t|v) = 0$.

Based on the network G, the DC, CC and BC ranking of each station in the London tube network can be calculated, and the top 10 of them are shown in the following table:

Rank	name	deg_cent	name	clo_cent	name	bet_cent
1	Stratford	0.922111	Stratford	0.927739	Stratford	0.098553
2	Highbury & Islington	0.806533	Highbury & Islington	0.836134	Liverpool Street	0.034307
3	Whitechapel	0.781407	Whitechapel	0.820619	Canary Wharf	0.027956

Rank	name	deg_cent	name	clo_cent	name	bet_cent
4	West Brompton	0.776382	West Brompton	0.817248	Bank and Monument	0.027956
5	Canada Water	0.771357	Canada Water	0.813906	Canning Town	0.027757
6	Canary Wharf	0.771357	Richmond	0.810591	West Ham	0.024551
7	Liverpool Street	0.768844	Canary Wharf	0.810591	Highbury & Islington	0.023023
8	Bank and Monument	0.766332	Bank and Monument	0.810591	Whitechapel	0.019682
9	Richmond	0.766332	Liverpool Street	0.808943	Canada Water	0.017898
10	Canning Town	0.763819	Canning Town	0.808943	Shadwell	0.017070

1.2 Impact measures:

In order to measure the impact of station closure on the whole London tube network, this section will use two coefficients to evaluate the overall characteristics of the network. These two coefficients are: **Average Clustering Coefficient** and **Degree Assortativity Coefficient**.

Average clustering coefficient (ACC) reflects the average of the clustering coefficients of all nodes in the network. The clustering coefficient is a measure of how tightly connected a node is to its neighbors, and it measures the connectivity between a node's neighbors. For the London tube network, the neighbors of a station are the other stations adjacent to it, so the clustering coefficient reflects the degree of connectivity between stations in the subway network. The higher the average clustering coefficient, the tighter the connectivity between stations in the tube network.

Degree assortativity coefficient (DAC) directly reflects the correlation between the degrees of nodes in the network, and indirectly reflects the structural characteristics of the network. For the London tube networks, if there is positive degree assortativity in the network, then neighboring stations are usually connected with similar number of other stations, forming a tightly connected community structure. On the contrary, if there is negative degree assortativity in the tube network, then adjacent stations are usually connected to other stations with widely varying numbers, forming a more dispersed structure.

By removing the nodes and recalculating the two coefficients, the network's resilience can be evaluated by the change of the two coefficients. If it remains unchanged or increases, it means that the network is more resilient, which means that the network is able to retain the original characteristics as much as possible. If the coefficients decreases, then it means that the network is not resilient in the corresponding metric, or the node has a greater impact on a network.

These two measures can reflect the characteristics of the shape and clustering of the arbitrary network, which are not specific to the London tube network. For whether using these two metrics to evaluate the resilience of an arbitrary network will be significant, we need to consider factors such as the network's own characteristics. For example, if we are more concerned with the change in resilience of a network in terms of the degree of aggregation, then for an arbitrary network we would be more inclined to focus on the change in the average aggregation coefficient of that network as the nodes change.

1.3 Node removal:

Two strategies based on three central measure rankings, using sequential and non-sequential deletion of nodes, generated two results affecting the variation of the coefficients, as shown in the following figure:
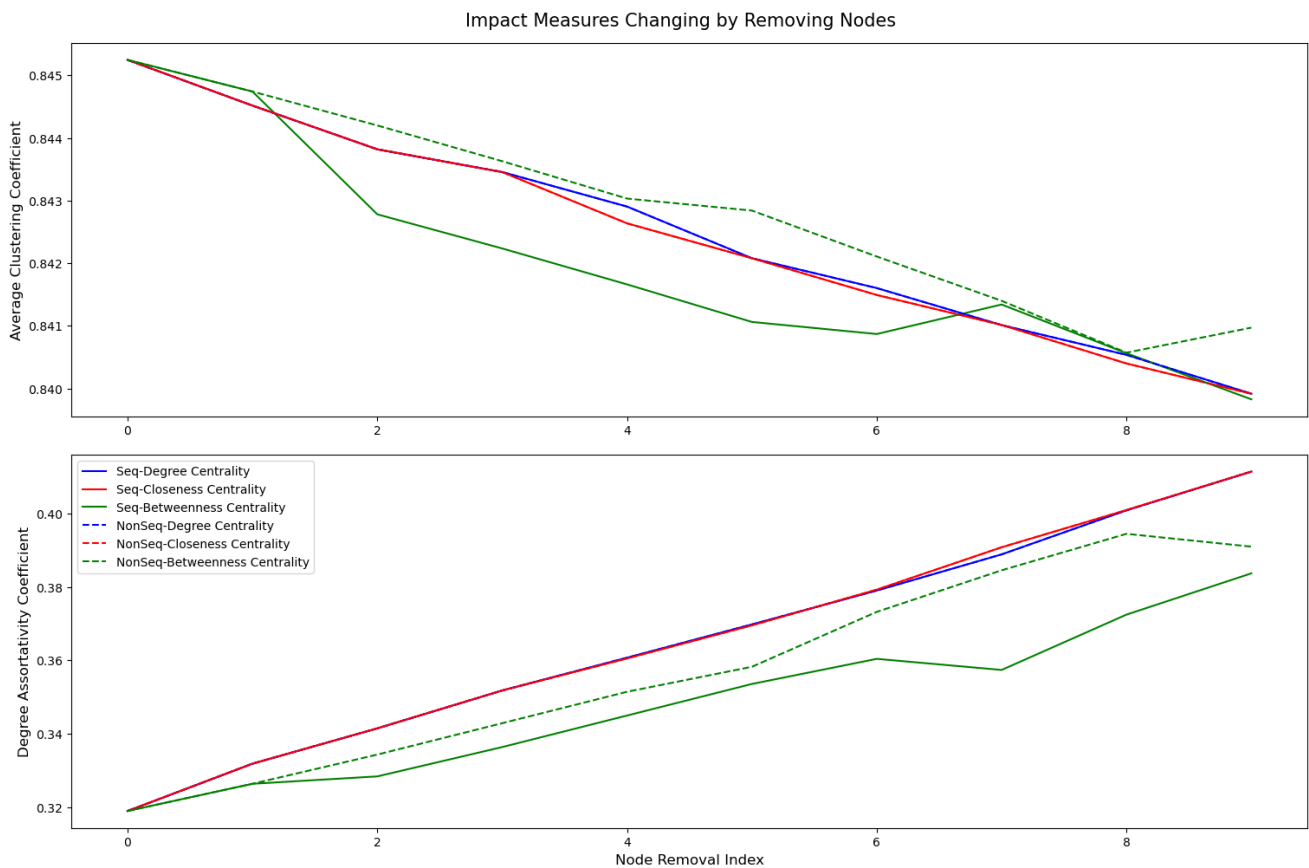
For which centrality measure reflects better the importance of a station for the functioning of the underground, in the figure, BC can better reflect the importance of the station for the London tube network. This is because the variation of BC is more discrete with the removal of nodes at both ACC and DAC measurement scales. The discrete nature of BC variation better reflects the different magnitudes of impact of

different station closures on the network as a whole than the approximately linearly varying DC and CC. Therefore, BC can better reflect the importance of stations for underground operation. However, because the method only considers topological networks, there are still limitations in using this method to calculate the importance of real stations.

For which deletion strategy is more effective at studying resilience, in the figure, it can be seen that Non-sequential and Sequential deletion strategies have little effect on the change of two types of centrality measures for both DC and CC, but the effect on BC is more significant. Overall, the Sequential deletion strategy is more likely to reflect the size of the destructiveness of different site deletions on the original network, which in turn can help to study the resilience of the network.

For which impact measure is better at assessing the damage after node removal, In my opinion, ACC is better than DAC to assess the damage of the network after node removal. The ACC directly indicates the average density of interconnection between stations in the tube network and their neighbors, which also directly reflects the flexibility of passengers to use the network, and indirectly reflects the functionality of the network. If the value of ACC decreases, it indicates that the aggregation of the network decreases and the cost of passengers using the metro network increases. the ACC is more explanatory and can be better related to real life.

II. Flows: weighted network:

II.1. Old vs new measure:

The three Centrality measures described in 1.1 will need to be reconsidered when we consider passenger traffic into the London Tube network. First, DC is unchanged because the DC calculation method only considers the degree of the nodes and does not consider the weights of the network, so DC is unchanged. Secondly, for CC and BC, the normalized result of the foot traffic is used as the weight because it is added to the calculation as the weight. Meanwhile, because the node importance of CC and BC is calculated based on the distance cost, the meaning of distance cost is opposite to the meaning of pedestrian flow, so the difference between the normalized result and 1 needs to be taken as the weight cost in the end. After calculation, the results of CC and BC change as follows:

Rank	name	cc	name	cc_w	name	bc	name	bc_w
1	Stratford	0.928	Stratford	0.945	Stratford	0.099	Stratford	0.132
2	Highbury & Islington	0.836	Liverpool Street	0.870	Liverpool Street	0.034	Bank and Monument	0.125

| Rank	name	cc	name	cc_w	name	bc	name	bc_w
3	Whitechapel	0.821	Highbury & Islington	0.851	Canary Wharf	0.028	Canada Water	0.063
4	West Brompton	0.817	Canary Wharf	0.847	Bank and Monument	0.028	Liverpool Street	0.061
5	Canada Water	0.814	Bank and Monument	0.839	Canning Town	0.028	Waterloo	0.046
6	Richmond	0.811	Waterloo	0.839	West Ham	0.025	Highbury & Islington	0.029
7	Canary Wharf	0.811	Whitechapel	0.828	Highbury & Islington	0.023	Seven Sisters	0.025
8	Bank and Monument	0.811	Canada Water	0.827	Whitechapel	0.020	Barking	0.021
9	Liverpool Street	0.809	Canning Town	0.823	Canada Water	0.018	Canning Town	0.019
10	Canning Town	0.809	West Brompton	0.822	Shadwell	0.017	Canary Wharf	0.017

The table shows that when flow is added as a weight to the importance calculation of the network nodes, the importance ranking of the network nodes changes. For example, in the change of closeness centrality, the importance of Liverpool Street site moves from the 9th to the 2nd. This indicates that the consideration of pedestrian flow will change the importance ranking of the station. Meanwhile, Stratford station is still at the top status, indicating that the station is very important, regardless of whether flow is considered.

II.2. Impact measure with flows:

For weighted networks, the two imapct measures used in 1.2 can still be used, since both ACC and DAC calculations support the computation of weighted networks. In addition, for weighted networks, we can also use Average Shortest Path Length (ASPL) to evaluate the closeness of the network.ASPL represents the average of the shortest path lengths between all node pairs. In general, the smaller the Average Shortest Path Length, the closer the nodes are to each other and the faster the information is disseminated. For the London Underground network, the smaller the value, the fewer the average number of stations a passenger needs to pass through to get from the departure station to the destination station. If a station is closed, the ASPL value increases, indicating that the closure of that station increases the average number of stations that passengers pass through. The larger the change in ASPL value, the greater the impact of the removal of that station on passengers' experience of using the subway.

II.3. Experiment with flows:

Using the best performing Betweenness Centrality found in I.3, the station importance of the weighted tube network can be evaluated and the results show as following:
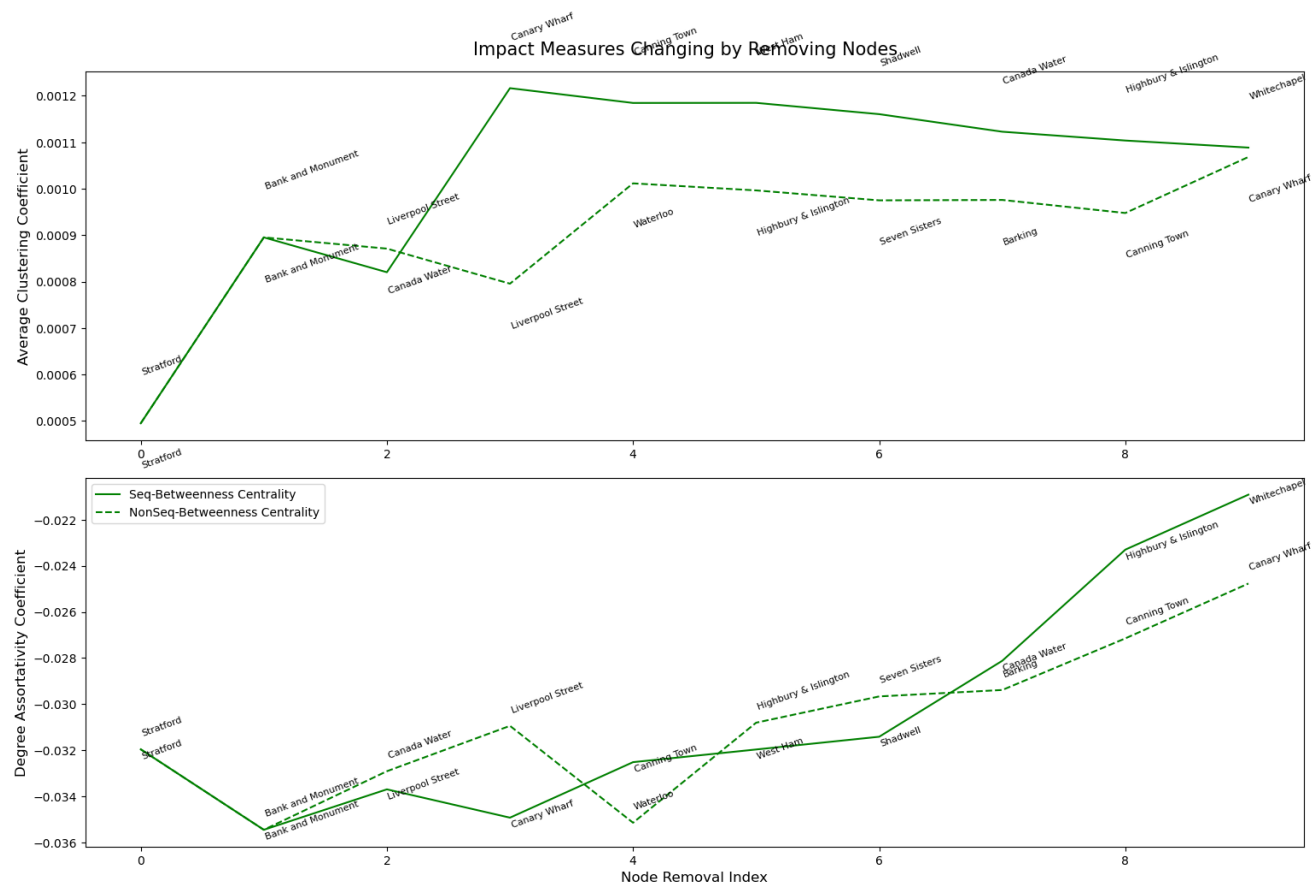
The results in the figure show that the deletion of the Stratford station has the greatest impact on the overall network. In this case, the deletion of Stratford station decreases the DAC value, which indicates that the network is more discrete, while the increase of ACC value indicates that the closure of Stratford station as a transit station makes the network more connected, which indicates from the side that the closure of Stratford station has a significant impact on the ridership. This shows that the closure of Stratford station has a significant impact on passengers. A similar effect is observed for Liverpool Street station.

Part 2: Spatial Interaction models:

In this section, the data of the London tube OD matrix in csv format provided on Moodle will be used.

III. Models and calibration:

III.1

Gravity Model

Gravity model is the spatial interaction model that was highlighted in the lecture. The gravity model is commonly defined as:

$$T_{ij} = K\frac{O_i^\alpha D_j^\gamma}{d_{ij}^\beta} = KO_i^\alpha D_j^\gamma d_{ij}^{-\beta} \tag{4}$$

- Where T_{ij} is the transition or flow T, between origin i and destination j.
- O is a vector of origin attributes which relate to the emissiveness of all origins in the dataset, i - means the index of the origin area.
- D is a vector of desination attributes which relate to the attractivenss of all destinations in the dataset, j - means the index of the desination area.
- d is a matrix of costs relating to the flows between i and j.
- K, α, γ and β are all the model parameters to be estimated

This model shows that the flow between origin and destination is proportional to the product of the masses of origin and destination and inversely proportional to the distance between them. As the quality of the origin and destination increases, the flow increases, but as the distance increases, the flow decreases, and vice versa.

On this basis, constraining the parameters can lead to a constrained gravity model. For example, the Production (orign) Constrained Spatial Interaction Model, which is used in practical, is a single constrained model with constraints on O_i^α. In addition, there are also Attraction (destination) Constrained Spatial Interaction Models that constrain the destination value.

Agent Based Model

In addition, the agent-based model(ABM) is introduced in the course as a spatial interaction model. It has good simulation effect for urban phenomena such as traffic congestion.Futhermore, the course also introduces the cellular automata model. Cellular automata models are a special class of ABMs that can well represent the self-similarity of urban sprawl based on fractal geometry.

III.2

The target of spatial interaction model is to models and predicts the number of occurrences of events in a certain time period. Thus, Poisson regression is chosen to build the prediction model. Because Poisson regression models assume that the response variables are Poisson distributed and model and predict count data better. As a result, we can build a Production-constrained Model by using the formula:

$$T_{ij} = A_i O_i D_j^\gamma d_{ij}^{-\beta} \tag{5}$$

Where

$$O_i = \sum_j T_{ij} \tag{6}$$

and

$$A_i = \frac{1}{\sum_j D_j^\gamma d_{ij}^{-\beta}}$$ (7)

Thus, we can get the formula as:

$$\lambda_{ij} = \exp(\alpha_i + \gamma \ln D_j - \beta \ln d_{ij})$$ (8)

Where α_i is the equivalent of the vector of balancing factors A_i, D_j is the jobs count in the destination area and d_{ij} is the distance between OD.

Using Python, the Poisson regression model was built and the distance cost coefficient is calculated as: $\beta = 0.1229$
By verifing with the original flow data, the R2 of the result is 0.2899, which shows that the model need to be improved.
And the RMSE of the reult is 111.276, which seems good as as result.

Futhermore, although the definition of the model is theoretically correct, the result of R2 is not very satisfactory and the result does not support the model well. Presumably, the reasons for this situation are:

1. the existence of anomalies or noise points in the OD network. When there are anomalies or noise points, the gravity model is easily interfered, which leads to a bad model performance.
2. The scale of data distribution in the OD network is too large. The gravity model relies on the physical distance between the OD, and if the data are not evenly distributed or there are discontinuous areas of density, the model will become less effective.
3. The distribution of the data does not match the Poisson distribution. Because we assume the source data are Poisson distributed when the model is built, no Poisson test is performed. If the source data is not Poisson distributed, the regression result of the model will not be very good.

IV. Scenarios

In this section, the code is hosted at at [Github](#)

IV.1

Scenario A can be based on the Production-constrained Model of III.2, which calculates the change in the overall network traffic after a 50% reduction in jobs at Canary Wharf.
In order to make the number of commuters is conserved, the parameters need to be

fine-tuned based on Equation (7), i.e., the A_i is recalculated based on the A_i before the job reduction. Finally the flow is calculated using the same method as in III.2.

IV.2

Define the distance cost function as

$$d_{ij} = \mu \cdot D_{ij} \tag{9}$$

Where μ is the cost coefficient, D_{ij} is the distance between origin i and detination j. Thus, we can get the new constrained gravity model as

$$\lambda_{ij} = \exp(\alpha_i + \gamma \ln D_j - \beta \ln(\mu \cdot D_{ij})) \tag{10}$$

Using the equation (10), we can regulate the change in transportation cost by controlling the size of μ. Suppose the traffic cost increases, then take the values $\mu_1 = 1.1, \mu_2 = 1.3$. Eventually the change of flow can be calculated.

In the code, the result is displayed as a 400*400 grid, the coordinates of which indicate the stations flow, and the color of the grid indicates the change value. Thus the scale of change can be obtained in two variation cases:
μ	scale
1.1	(-71.56, 0.93)
1.3	(-194.24, 0.83)

IV.3

By comparing the changes in the OD matrix before and after the changes in the two scenarios, we obtain the table of changes:

Scenario	scale
A	(-1013.0, 164.0)
B_1.1	(-71.56, 0.93)
B_1.3	(-194.24, 0.83)

The table shows that Scenario A has a greater effect on the flow of OD, which means the reduction of jobs has a greater effect on the flow.
In addition to the change table, the correlation coefficient and P value of the regression model can also show that jobs have more influence on the flow of OD than distance. In the results of the regression model the coefficient of job is 0.8156 while the coefficient of distance is -0.1229, while the P value of both converge to 0, indicating that both are extremely significant, but the effect of job is stronger than distance.
As conclusion, the job decreasing has more impact in the redistribution of flows.