DSM160 Coursework 1: Network Science - London Transport Network

Submitted for partial fulfilment for the Social Media and Network Science course.

By

Hendrik Matthys van Rooyen

University of London

December 2023

CONTENTS

CHAPTER 1 CONTENT

QUESTION 1

Is this network weighted or unweighted? Is it directed or undirected? Please justify.

The network is unweighted and undirected. This is derived from the information given in the "README.pdf" document, which describes the London Multiplex Transportation Network dataset. The dataset includes nodes representing train stations in London and edges for existing routes between stations.

The network is unweighted because the edges, representing the routes between stations, do not have weights associated with them. While the raw data includes details such as the train line and the stations connected, it does not include weights that could represent factors like the frequency of trains, the distance between stations, or travel time.

The network is undirected as indicated by the README file's description of the dataset, stating that the multiplex is "undirected (with only one direction specified)". In the context of a transportation network like London's, this means that if there is a route from station A to station B, it is implicitly understood that there is also a route from station B to station A, thus making the network undirected.

QUESTION 2

Import the data into Python using the pandas library and convert it into a networkx graph.

```
df =
pd.read_csv('London_Multiplex_Transport\Dataset\london_transport_raw.edges',
sep=' ', header=None, names=['line', 'station1', 'station2'])

G = nx.from_pandas_edgelist(df, 'station1', 'station2')
```

QUESTION 3

What is the number of nodes and links in the network?

Nodes: 369

Edges/Links: 430

QUESTION 4

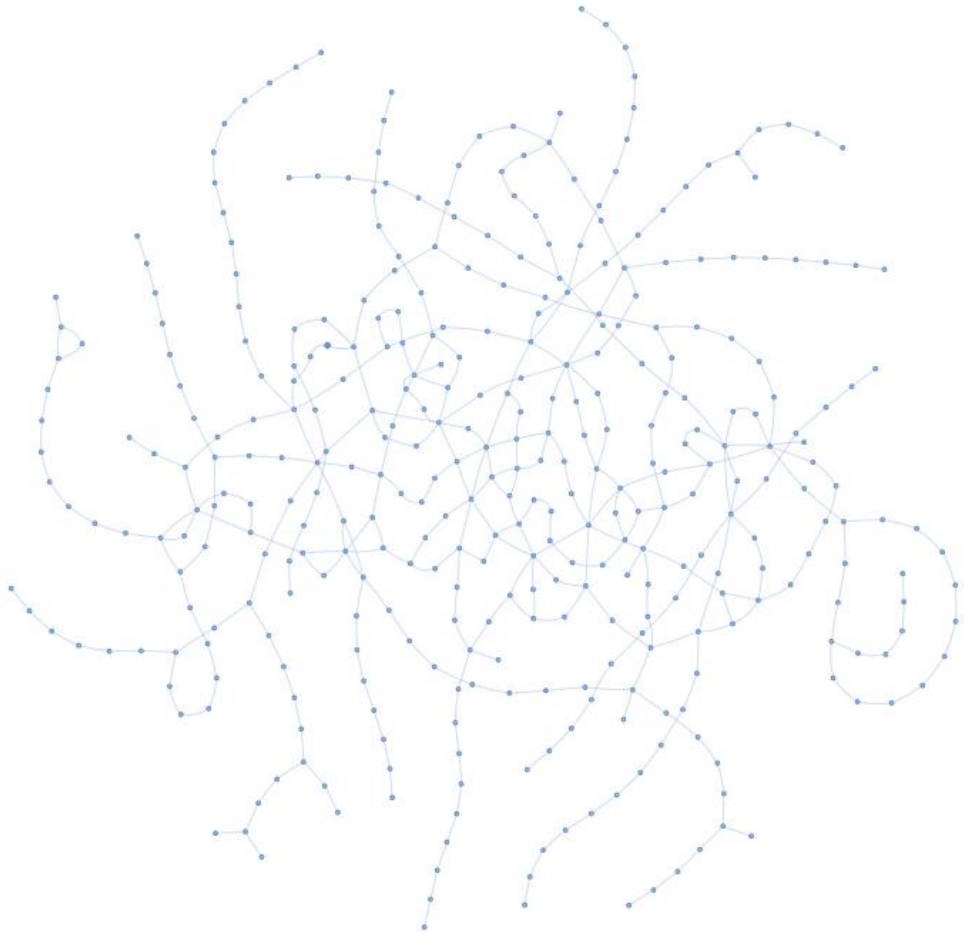
Is the network sparse? Please justify.

The network is quite sparse, given the sparsity factor of 0.006. This means that for every 184 possible link/edges between nodes, there is only 1 in the network.

Max Edges: 67896

Sparsity: 0.0063

QUESTION 5
Plot the network using the library pyvis and include it as an image in your document.



QUESTION 6
Looking at your visualisation from item (5), what do you think would be the typical node degree of this network and why?

Most nodes seem to have between two and four connections, with a few having more. Those with four could be where rail lines cross, while those with two connections can be stopping points on a longer route, some of which may run to the end of a given line.

Since most seem to have 2 connections, 2.5 should be a safe estimate.

QUESTION 7
What is the average degree of the network? How does it compare to your answer to item (6)?

Average Degree: 2.3306

The calculated average compares well to the estimate, the estimated proportion of nodes with two connections seem to have been a bit lower than reality, which would explain the average degree being closer to 2.

QUESTION 8
Does the network have a core-periphery structure or a structure in which hubs are situated at the centre of star-like components? Please justify.

Visually the network does seem to have a core-periphery structure, with the core having densely connected nodes with the sparse, tentacle-like structures stretching from the edges. The edge-structures also don't connect to new "hubs".

QUESTION 9
How many connected components does the network have?

The network consists of a single connected component. From the nodes specified in the dataset description, the node count retrieved in question 3 from the constructed network, and the manual node count, all totalling at 369, it can be confirmed that no nodes has been discarded or left out.

From the visualization it can also be confirmed that a single component exists.

QUESTION 10
What are the top 10 stations in terms of degree, closeness and betweenness centrality?

Degree Centrality (most connected stations):

- Baker Street
- Stratford
- King's Cross St. Pancras
- Paddington
- Oxford Circus
- Waterloo
- Bank
- Earl's Court
- West Ham
- Green Park

Closeness Centrality (stations most 'central' in terms of shortest paths to all others):

- Green Park
- Westminster
- Bond Street
- King's Cross St. Pancras
- Oxford Circus
- Bank
- Waterloo
- Baker Street
- Euston
- Victoria

Betweenness Centrality (stations most frequently on shortest paths between other stations):

- Bank
- Waterloo
- King's Cross St. Pancras
- Green Park
- Baker Street
- Euston
- Stratford
- Westminster
- Finchley Road
- Liverpool Street

QUESTION 11
Which stations are central according to all three centrality measures?

When considering the stations present on the three lists resulting from question 10, the list is as follows:

- Baker Street
- King's Cross St. Pancras
- Green Park
- Waterloo
- Bank

These 5 are also present in the top 5, though in different orders, when calculating the "most central" nodes in other ways including:

- Rank Aggregation
- Standardizing and Summing Scores
- Multiplying Centrality Values
- Principal Component Analysis

QUESTION 12
Using the stations identified in item (10), which stations with high degree centrality are not betweenness-central? How do you interpret the role of such stations in the network?

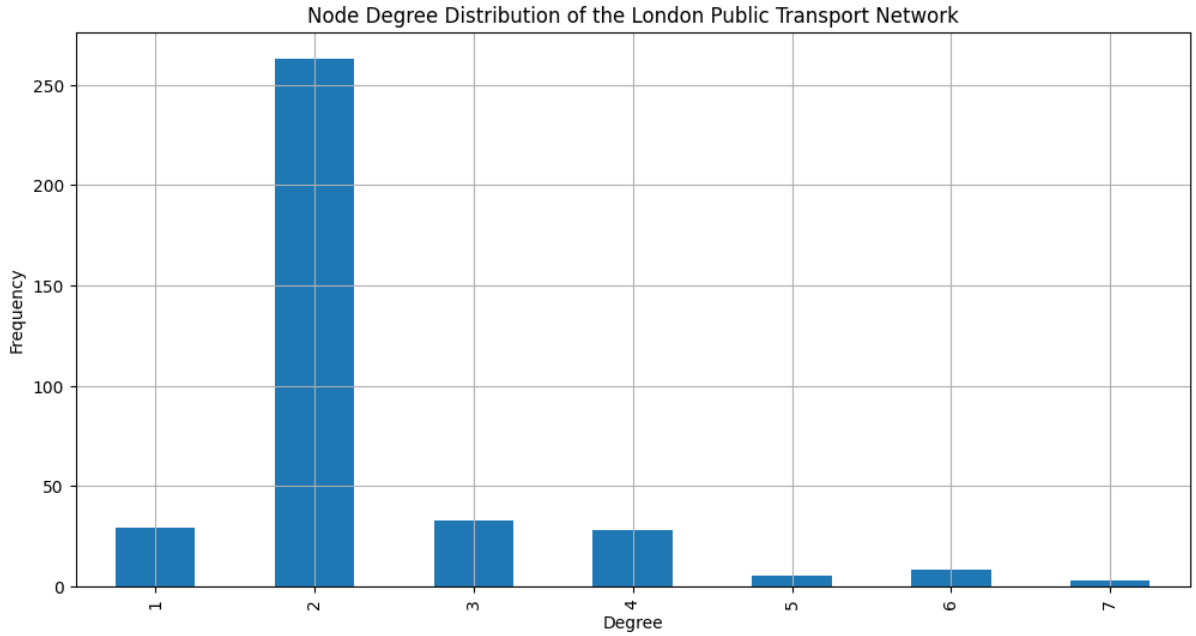
- Earl's Court
- Oxford Circus
- Paddington
- West Ham

High degree centrality with lower betweenness centrality indicates that these stations connect to many other stations directly, but they may not be the only or fastest route to get from one side of the network to another i.e. there may be faster alternatives (routes with less stations), or direct connections with stations with higher betweenness.

QUESTION 13
What is the proportion of stations where, at least, two different (train) lines intersect?

27.64%

QUESTION 14
Build and plot the node degree distribution for this network. How does it relate to your answer to item (6)?



From the graph it is clear that the overwhelming majority of nodes has two connections, which would then support the estimate in question 6 and the findings and elaboration in question 7.

With the degree distribution favoring 2, it would follow that the average should be near 2.

QUESTION 15
How many steps does it take, on average, to go from one station to another using shortest paths?

13.73 steps

QUESTION 16
What is the shortest path to go from the first to the second most central station by degree? What is the path length?

"Baker Street", "Stratford", and “King's Cross St. Pancras” tied for the first three places, the combination of each two will be used

Baker Street to Stratford:

Baker Street → Bond Street → Green Park → Westminster → Waterloo → Bank → Liverpool Street → Bethnal Green → Mile End → Stratford

9 steps.

Baker Street to King's Cross St. Pancras:

Baker Street → Great Portland Street → Euston Square → King's Cross St. Pancras

3 steps.

Stratford to King's Cross St. Pancras:

Stratford → Mile End → Bethnal Green → Liverpool Street → Moorgate → Barbican → Farringdon → King's Cross St. Pancras

7 steps.

QUESTION 17
Amongst the top 5 stations by betweenness centrality, which are directly connected or have at most one intermediate station between them?

The top 5 stations by betweenness centrality being:

Bank, Waterloo, King's Cross St. Pancras, Green Park, Baker Street

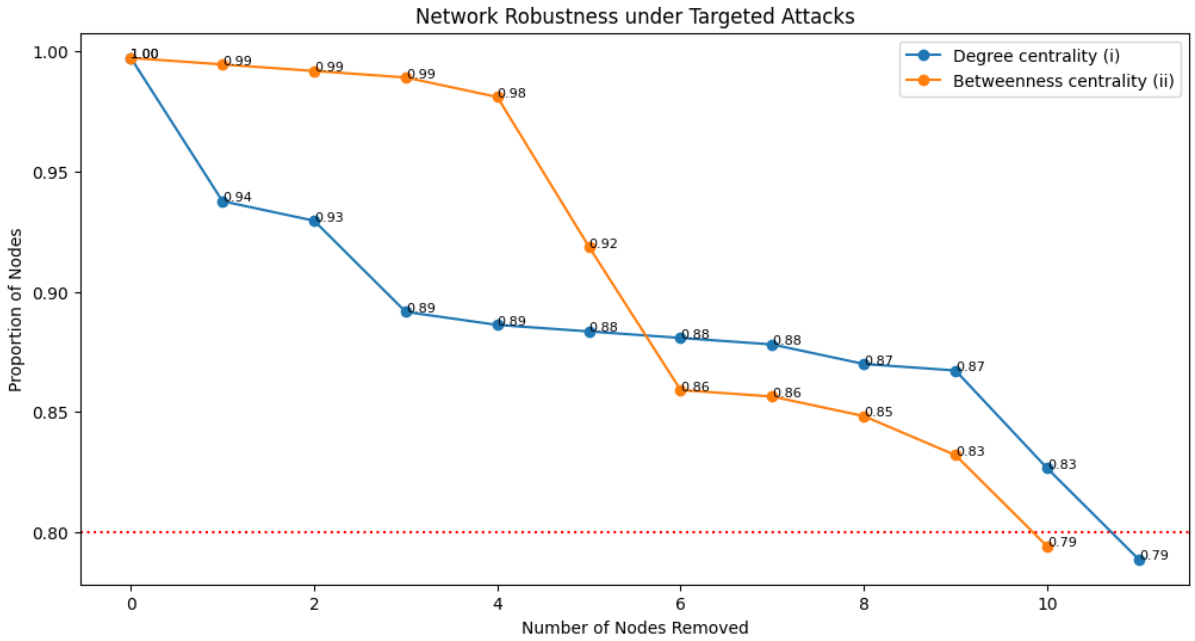
- Bank and Waterloo has a direct connection.
- Waterloo and Green Park has a connection via Westminster.
- Green Park and Baker Street has a connection via Bond Street.

QUESTION 18
You want to perform some robustness tests on this transport network by simulating attacks which disrupt one station at a time. In particular, you want to compare the vulnerability of the network when nodes are removed one by one in decreasing order of (i) their degree, as opposed to (ii) their betweenness centrality. Which strategy – (i) or (ii) – will be faster in reducing to less than 80% the proportion of nodes in the (largest) connected component? Please justify.

As displayed in the graph, targeting the network by betweenness centrality reduces the nodes connected to less than 80% the quickest within 10 disruptions. Targeting by degree centrality follows closely with 11 disruptions to reach 80%.

The betweenness centrality measure reflects the number of shortest paths that pass-through a given node. Nodes with high betweenness centrality often act as bridges between different parts of the network. When such nodes are removed, it can significantly disrupt the flow of the network, as many paths are severed at once.

Nodes with high degree centrality have many connections, but these connections may not be as critical for maintaining the network's integrity. Removing these nodes affects the network, but not to the same extent as removing betweenness-central nodes. While the network does become less connected, it may not fragment as quickly because alternative routes that do not pass through these highly connected nodes may still exist.



It is however notable that degree centrality targeting dropped below 90% a lot quicker than that of betweenness centrality. The graph also continues to change between which strategy leads in disrupting the network (see below).

This may however be attributed to the nature of the specific network, as both strategies share many nodes in common in terms of which nodes are generally higher ranked. (see question 10)