Cross City Urban Transit Networks Analysis

Group 16
Mid Term update

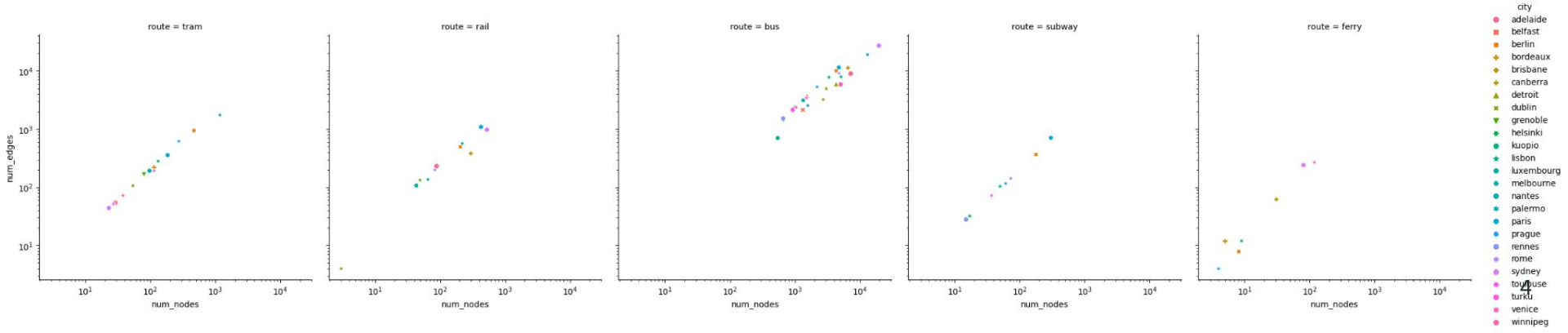
Aims

- How do the topological properties of public transport networks differ across different cities?
- Which public transport stops are the most important in each city based on their centrality measures?
- How do public transport networks compare across different cities (similarities and differences)?

Exploration

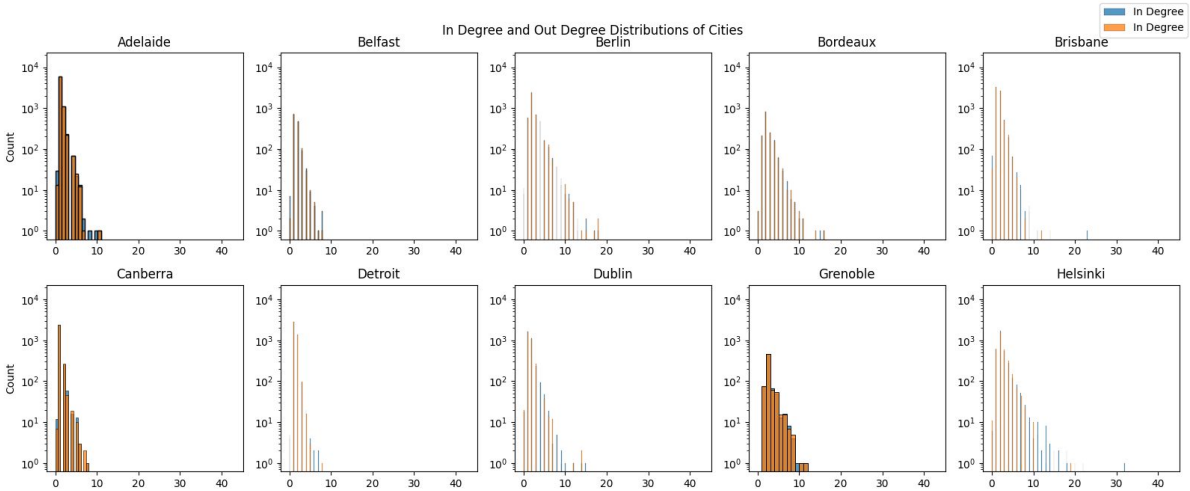
Preprocessing

- Two Main files for each city
  - Network_nodes.csv: For the nodes
  - Network_combined.csv: For the edges
- Merged nodes with same name and removed self loops between the nodes



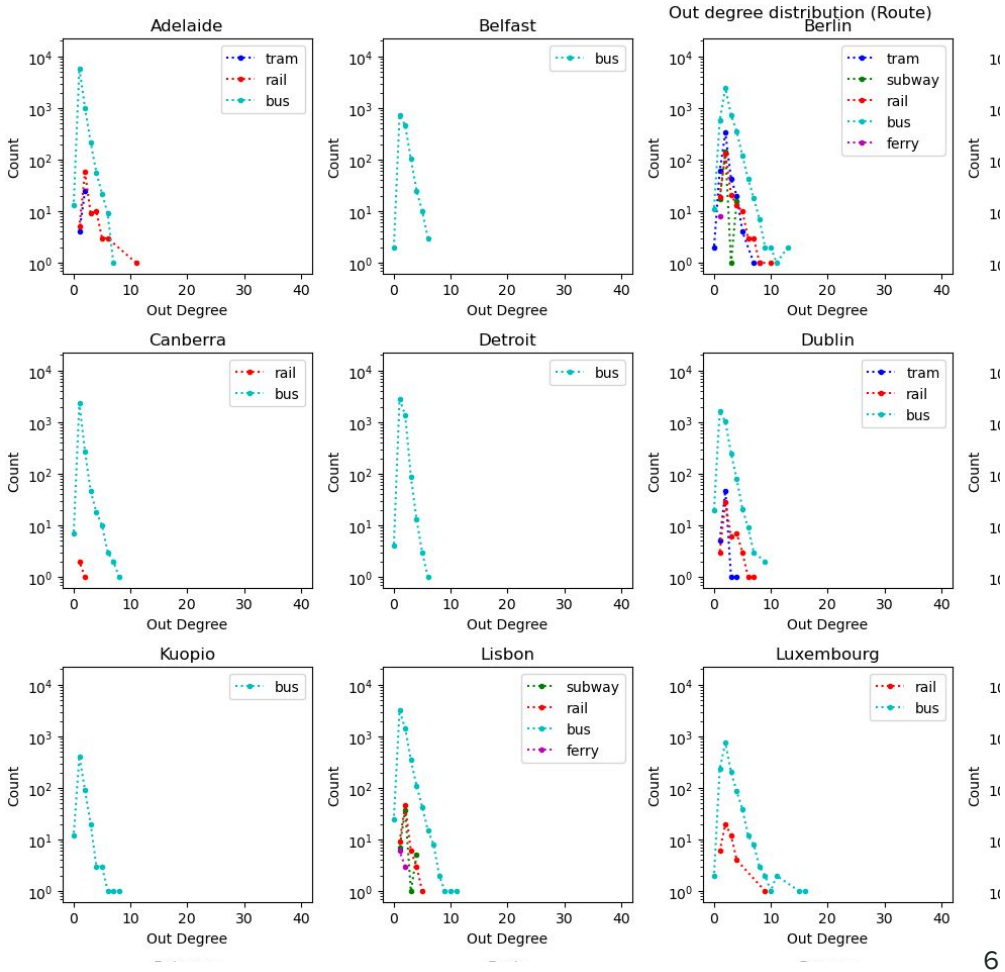
Tasks - Degree Distribution

- In-degree follows a very similar pattern to the out-degree distribution of graph nodes
- The average degree has a median of 2.03 and mean of 2.0 telling us that each station is connected with two other stations.



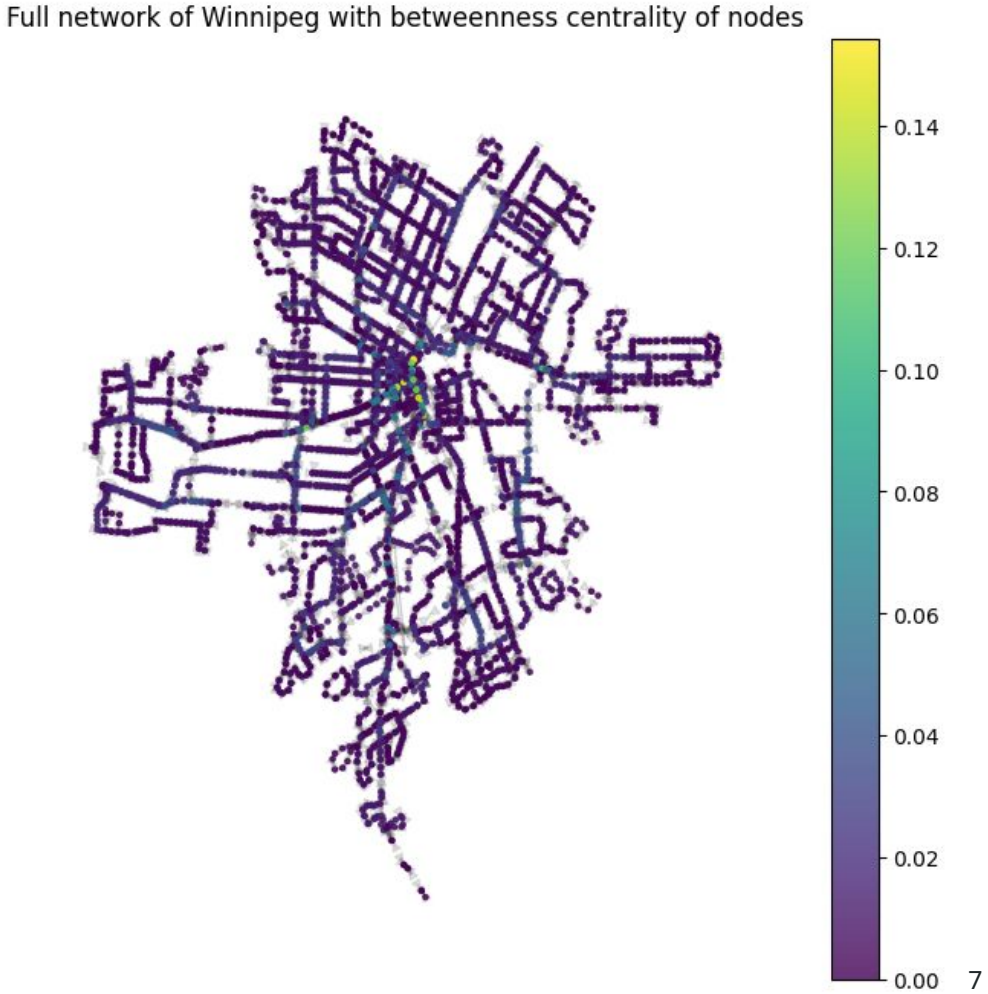
Tasks - Mode wise degree distribution

- The existence of a large number of nodes with low degrees and a smaller number of nodes with higher degrees is observed.
- This implies that the networks cannot be modelled by a random network model.



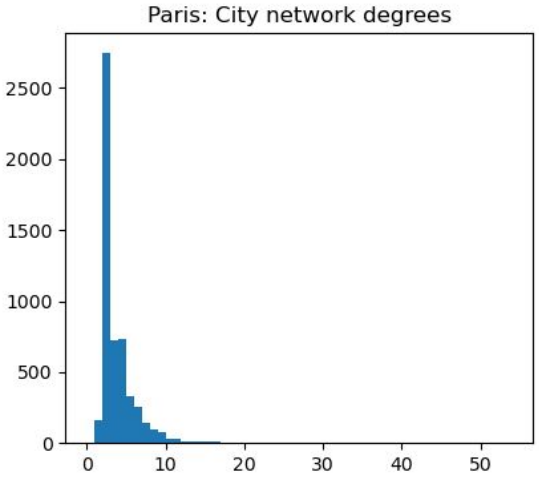
Task - Centrality Measures

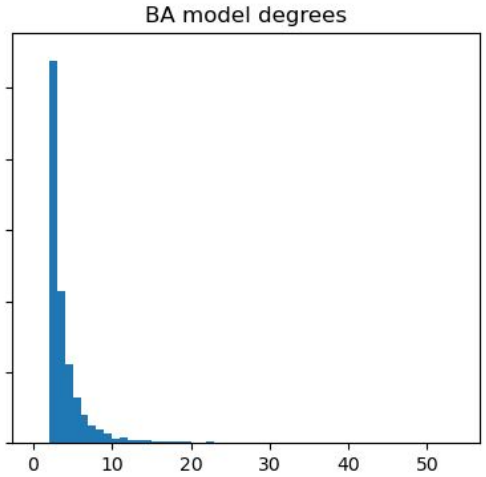
- Average clustering coefficient
- Betweenness centrality
- Central/Transit stations have high betweenness centrality

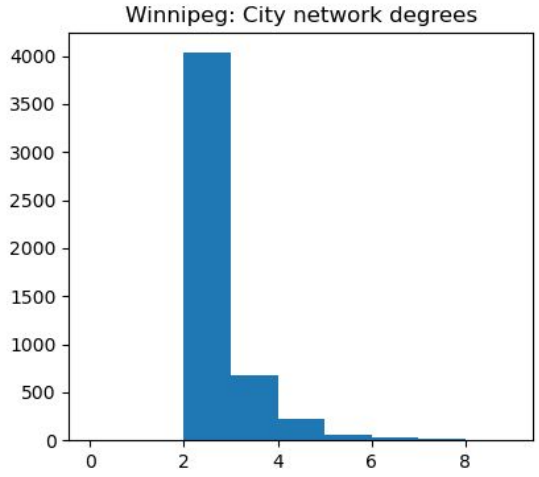


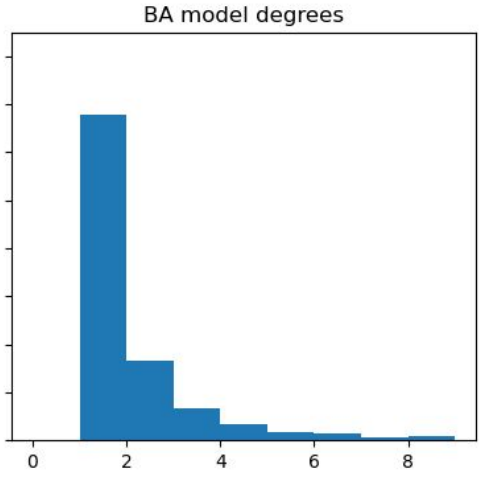
Task - Modelling networks

- Looking at the degree distribution above, we concluded that it cannot be modelled by a random network model.
- Comparing it with the Barabasi Albert generation model with an appropriate choice of q (from the degree distribution of network nodes).

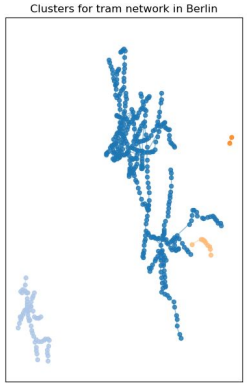


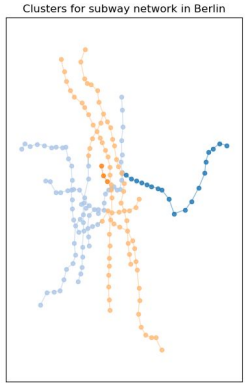


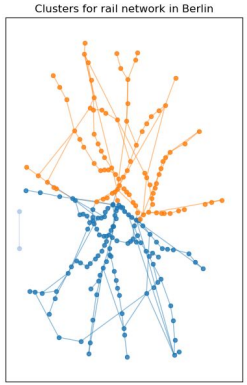


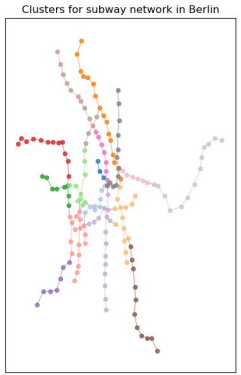


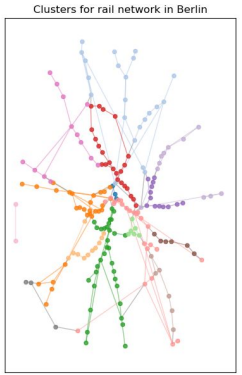
Spectral clustering & Community detection

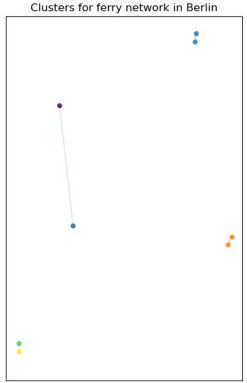


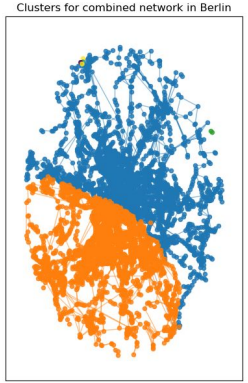


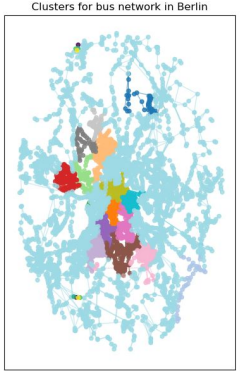


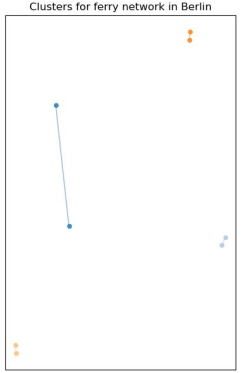


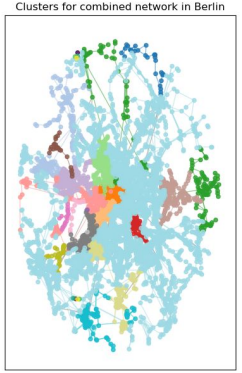










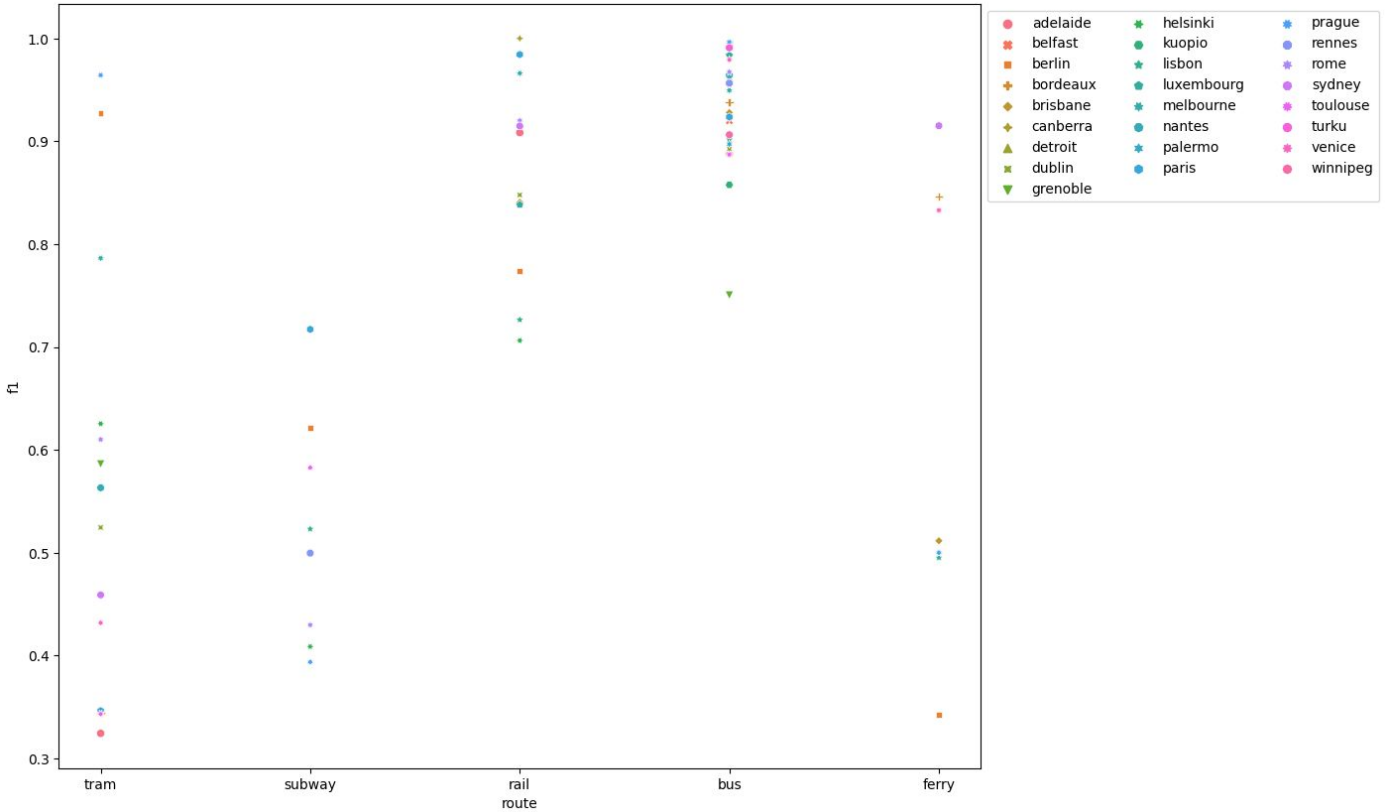


Exploitation

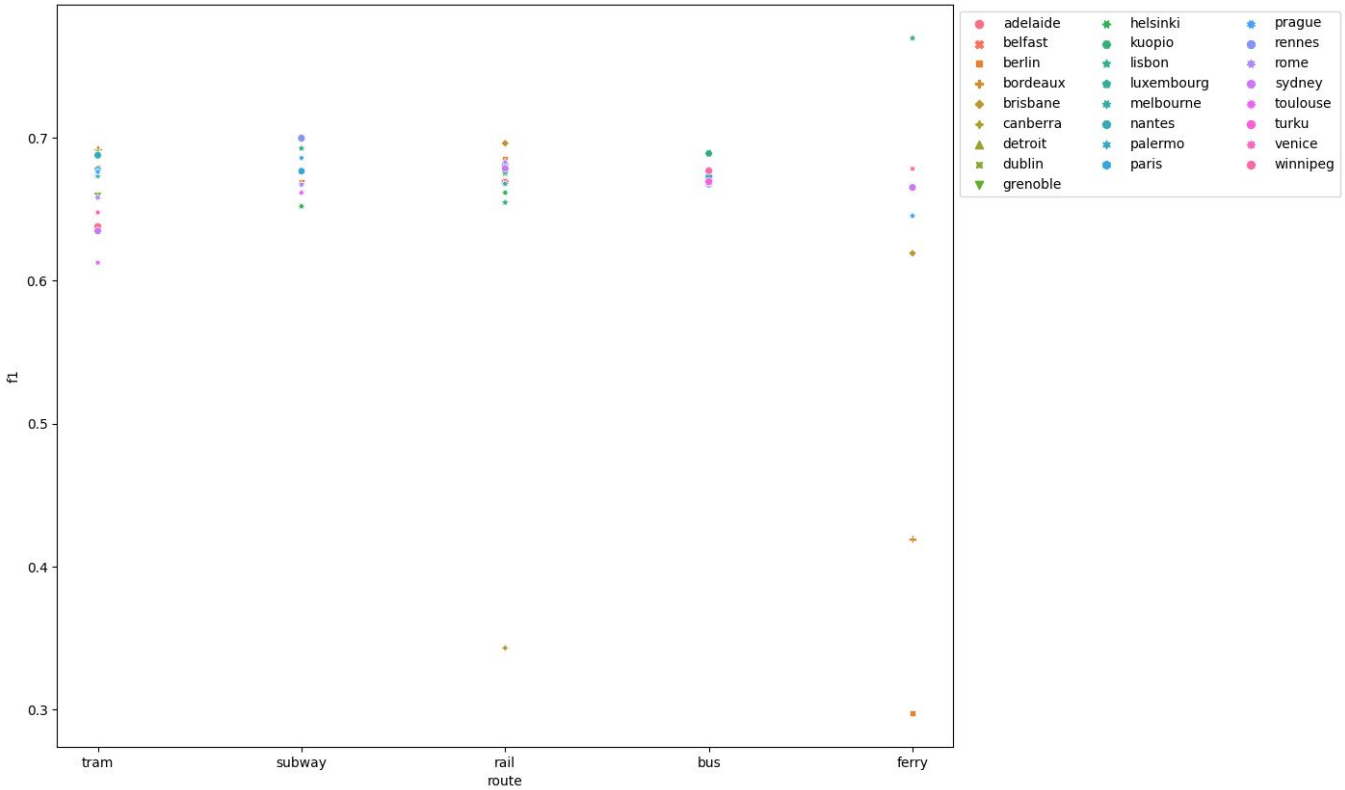
Task - Edge Prediction

- Given two nodes, predict if an edge exists from node 1 to node 2 (directed).
- Models
  - Logistic Regression
  - Node 2 Vec
  - Deep Learning (may be)
- We do transport mode wise analysis
- Only finished LR with 50:50 train/test split.
  - Hand crafted features
    - In degrees centrality
    - Out degree centrality
    - Betweenness centrality
    - Katz centrality
    - Number of Common predecessors
    - Number of common successors

Logistic Regression - Results



Node2Vec (3 features) - Results



Next steps & Issues

Task - Edge Prediction

- Try with Node2Vec models
- Edge prediction for whole network
- Use Deep Learning models for edge presence and category prediction for full network

Issues

- Storing the edge-level features is very expensive in terms of memory required.
  - Is it possible to access any computing clusters?
  - This will also be needed for using GNNs.
- Should we look at alternative exploitation tasks/methods to mitigate memory problem?

Thank you