Outline Team 6
Can Network Analysis Pedal Hannover Towards Better Bike Sharing?

Akshita Sharma, Aleksanteri Kylmäaho, Emil Schallwig,
Jannis Wiederhöft, Yannick Oefler

University of Mannheim, 68131 Mannheim, Germany

Keywords: Directed Weighted Graphs, Edge Prediction, GCN, GAT, GCN-GRU, Degree Centrality, Betweenness Centrality, Closeness Centrality, K-Core Analysis, Louvain Methode, Girvan-Newman Method.

1 What is the problem you are solving?

Public transportation is becoming increasingly complex as more and more people concentrate in cities, making it progressively difficult to travel from point A to point B. At the same time, cities are confronted with the challenge of becoming more climate-friendly. Bike rental systems are a rapidly growing solution to this problem. Users can rent a bike at designated stations for a rental fee and return it to any station after their ride, all with just a few clicks on their phone. This provides a convenient, environmentally friendly, and flexible way to travel. To ensure these systems operate smoothly, careful, data-driven planning is essential. In our project, we focus on the bike rental network in Hannover, operated by Bike Citizens. Our primary goal is to predict the demand at individual bike stations for each month in the year 2024. By identifying which stations are likely to be used most frequently in the future, we can detect potential shortages in bike availability. In such cases, the company should consider manually reallocating bikes to better meet the expected demand. Beyond this, we aim to address additional challenges:

Q1: Where should the bike storage be located? It is in the company's best interest to position storage facilities centrally within the network, allowing for quick and efficient bike redistribution whenever demand arises.

Q2: How should the company divide the bike network into groups of stations to be serviced by different maintenance teams? This question focuses on structuring the network so that maintenance tasks can be managed efficiently and cost-effectively, with specific teams assigned to designated groups of stations.

Q3: Where should the company place bike repair kits? Strategically positioning do-it-yourself repair kits at key locations can not only reduce maintenance efforts but also enhance user comfort and satisfaction.

Q4: Where should the city build new bike lanes? The decision of the population to switch from car or tram to bike depends on various factors, including the time it takes to travel from A to B. Therefore, we aim to identify popular routes that currently have low average speeds. We suggest enhancing bike lanes along these routes.

2 What data will you use?

We will use data from the Geospatial Data Management of the Hannover Region. The GML files provide information on intensity, and average speeds of bike rides tracked by the Bike Citizens app, in aggregated form based on the street network graph for the Hannover region for the respective year (see title). Each quarter is represented as a separate layer, with a total of four GML files for each quarter from 2020 to 2024.

The GML files contain multiple `ogr:featureMember` entries. Each `featureMember` describes an individual bike ride. Each `gml:LineString` describes a route with a `gml:posList`, which provides the coordinate pairs (latitude and longitude) for the course of the bike ride. Each record has a unique `ID` that identifies the bike or station. The `YEAR` and `MONTH` attributes indicate the year and month of the ride. `TRACKS_FWD` and `TRACKS_BAC` refer to the number of rides in the forward and backward directions along the route. `SPEED_REL` is an attribute that provides the relative speed of the bike ride, representing the efficiency of the trip.

3 How to solve the problem?

Our core strategy to address the challenges outlined in Section 1 involves representing the Bike Citizens usage data as a time series of monthly weighted directed graphs. This graph-based representation provides a powerful framework for applying network analysis techniques to understand detailed bicycle usage patterns and make demand predictions. In these graphs, nodes are defined as the unique start and end points of road segments, identified using the `gml:posList` coordinates from the GML data. Each road segment is then modeled using directed edges that connect these nodes, explicitly capturing the two potential directions of travel. Edge weights directly quantify the usage frequency based on the provided `TRACKS_FWD` and `TRACKS_BAC` counts, and relative speed (`SPEED_REL`), `YEAR` and `MONTH` will be associated with each edge as an attribute.

3.1 What preprocessing steps will be required?

The primary preprocessing task is to convert the raw GML data into a sequence of monthly weighted directed graphs. This process begins by parsing the GML files from 2020 to 2024 using the Python library `NetworkX`, leveraging the `YEAR` and `MONTH` attributes present in the data. Data from the years 2020 to 2023 will be used as the training set to develop our predictive models, while the 2024 data will be reserved for evaluating model performance on unseen data.

Standard data cleaning procedures, including coordinate validation, will be applied. To account for the evolving structure of the network over time, we will define a fixed set of nodes comprising all unique locations observed throughout the 2020–2023 training period. This consistent node set ensures stable input dimensions for temporal models. In each monthly snapshot, nodes from this set that are inactive (i.e., not involved in any recorded rides) will still be included in the graph but will have no associated edges for that month. Crucially, if a specific edge (i.e., a route segment) is not present in a given month, its weight (i.e., the ride count) will be set to zero—accurately representing the lack of usage during that period.

The result of this preprocessing phase is a cleaned and structured time series of monthly graph snapshots, which are ready for network analysis and can be directly used as input for graph neural network models.

3.2 Which algorithms you plan to use?

To estimate the demand for bikes at a specific station in a given month in 2024, we first aim to predict the directed, weighted edges (i.e., `TRACKS_FWD` and `TRACKS_BAC`) between two nodes (i.e., bike stations). For the prediction task, we plan to test three models: Graph Convolutional Network (GCN), Graph Attention Network (GAT), and a hybrid GCN-GRU model. As a baseline, we will also include the historical monthly averages from the years 2020 to 2023 for comparison.

In a subsequent step, we propose analyzing the in-degree and out-degree of each node using the model's predicted values. We hypothesize that if a node's out-degree significantly exceeds its in-degree, it may indicate a shortage of bikes at that location. In such cases, the company should consider manually reallocating bikes to better meet the demand.

We will answer the questions *Q1-Q4* using the model's predictions—if they are reliable. Otherwise, we plan to use the actual data from 2024. To answer the question *"Q1: Where should the bike storage be placed?"*, we plan to use closeness centrality as a metric. The goal is to place storage facilities in central positions within the network, enabling the company to redistribute bikes quickly and efficiently whenever demand arises.

To address the question *"Q2: How should the company divide the bike network into groups of stations to be serviced by different maintenance teams?"*, we will apply k-core analysis. Stations that belong to high k-cores are more densely connected within the network, making them ideal candidates for bundled maintenance alongside nearby stations. We also consider incorporating more advanced community detection methods such as Louvain or Girvan-Newman to further refine the grouping of stations.

Additionally, we aim to address the question, *"Q3: Where should the company place bike repair kits?"* To answer this, we plan to measure betweenness centrality.

Stations with high betweenness centrality serve as critical hubs in the network, facilitating connections between different parts of the system. These locations are ideal candidates for placing repair kits, as they are likely to be visited by many users.

Finally, we aim to address the question *"Q4: Where should the city build new bike lanes?"*. To answer this, we will analyze the average relative speed (`SPEED_REL`) in conjunction with the most heavily used edges in the network. Our hypothesis is that routes with high traffic but low average speeds may require infrastructure improvements—such as enhanced or extended bike lanes—to better meet demand and improve overall efficiency.

We reserve the option to apply more advanced methods and metrics to address all questions, should our further research uncover appropriate techniques.

4 How will you evaluate, measure success?

To evaluate the weighted edge predictions of the models GCN, GAT, and GCN-GRU, we compare their performance against our baseline—the historical monthly averages from previous years. For this comparison, we use the metrics MAE, MSE, and RMSE to assess and quantify the prediction errors. For questions *Q1–Q4*, we acknowledge that no established benchmarks exist within our specific context, which makes evaluation inherently more challenging. As such, we treat these questions in a more exploratory manner. The insights gained from them should be considered an additional layer to our project—intended to guide our graph analysis based on realistic, scenario-driven considerations.