Mini Project III
Network Analysis

Naimur Rahman
Email: naimur.rahman@abo.fi

I. Introduction

In recent years, urban mobility has seen a significant shift towards sustainable and efficient transportation options. Helsinki City Bikes represents a pivotal initiative within this paradigm. Launched in 2016, this public bicycle-sharing system serves the Helsinki and Espoo metropolitan areas, aiming to alleviate the ”last-mile” challenge. The system has seen expansive growth due to its popularity, necessitating a comprehensive analysis to optimize its efficiency and accessibility.

However, the exploration and network analysis of such a dataset present multiple challenges. Firstly, the complexity of usage patterns over time and across different geographic locations requires advanced analytical techniques to understand user behavior. Secondly, the dynamic availability of bikes and docking stations introduces variability that must be accurately modeled to improve service reliability.

Conducting network analysis on this dataset is instrumental for several reasons - it can uncover the most critical nodes within the bike-sharing network, inform better placement of new stations, predict demand peaks, and ultimately enhance user satisfaction by reducing wait times and improving bike availability. Transferring the insights gained from this analysis to other cities like Turku could streamline the implementation of similar bike-sharing systems, thereby amplifying the benefits of sustainable urban mobility across Finland.

II. Tools Utilized

In this section, I delineate a range of different libraries that were instrumental in conducting my analysis of the Helsinki City Bikes dataset.

- **Pandas:** This library provided high-performance, easy-to-use data structures, and data analysis tools.
- **Matplotlib:** A plotting library for Python and its numerical mathematics extension, NumPy. Matplotlib was used to create static, interactive, and animated visualizations, offering a granular control over graph features.
- **Seaborn:** Built on top of Matplotlib, Seaborn specializes in statistical graphics and provided me with a high-level interface for drawing attractive and informative statistical plots, enhancing the interpretability of my findings.
- **NetworkX:** As a key tool for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks, NetworkX was integral to analyzing and visualizing the network graph of bike stations and their connections.
- **Plotly:** This interactive web-based tool enabled me to craft sophisticated plots that can be manipulated by end-users. Plotly's capability to handle dynamic and complex datasets made it an excellent tool for visualizing high-level summaries and patterns in my data.
- **Folium:** Leveraging the power of leaflet.js, Folium provided me with the means to visualize data that's been manipulated in Python on an interactive leaflet map. It was particularly useful for mapping the geographic distribution of city bikes stations across Helsinki and Espoo.

The combination of these tools provided a robust framework for my exploratory data analysis, allowing me to extract meaningful insights from the city bikes network and present them in an accessible manner.

III. Data Description

The provided dataset represents an extensive collection of over 10 million bicycle trips taken using the Helsinki City Bikes system from 2016 to 2020. Each entry details individual rides, capturing essential data such as departure and return times, station identifiers and names, distances traveled, trip durations, average speed, as well as geographic coordinates of the departure and return stations. The dataset also includes ambient temperature readings, adding a layer of environmental context to the trips.

This dataset is a testament to the program's reach and impact, documenting its expansion from a modest pilot of 46 stations to a comprehensive network of 350 stations and 3,510 bicycles. Riders have collectively spent over 280 years cycling, highlighting the system's role in urban mobility. This dataset provides a rich source for exploratory data analysis and city bike network analysis, offering insights into urban transportation patterns, seasonal usage trends, and the system's integration with the city's infrastructure.

The features given in the dataset with their corresponding types of data and potential applications are discussed below.

- **Departure:** Timestamp indicating when the bike ride started.
- **Return:** Timestamp indicating when the bike ride ended.
- **Departure_id:** A unique identifier for the departure station.
- **Departure_name:** The name of the departure station.
- **Return_id:** A unique identifier for the return station.
- **Return_name:** The name of the return station.
- **Distance:** The distance traveled during the trip, measured in meters.
- **Duration:** The duration of the trip, measured in seconds.
- **Speed:** The average speed of the trip in meters per second, calculated from the distance and duration.

- **Departure_latitude:** The latitude coordinate of the departure station.
- **Departure_longitude:** The longitude coordinate of the departure station.
- **Return_latitude:** The latitude coordinate of the return station.
- **Return_longitude:** The longitude coordinate of the return station.
- **Temperature:** Ambient temperature during the trip, measured in degrees Celsius.

This dataset is well-structured for geospatial and temporal analysis, providing opportunities to study patterns in bike usage across different times of the day and seasons, understand mobility within the network, and correlate usage with weather conditions due to the inclusion of temperature data. The geographic coordinates allow for mapping the trips and analyzing spatial patterns, which can be visualized using tools like folium for mapping, and Plotly or Seaborn for creating interactive graphs and statistical plots. The distance and duration fields offer direct measures of trip lengths, while the speed calculation allows for the analysis of travel times and congestion in the network. The unique identifiers for stations paired with their names enable the combination of network analysis with geographic information systems (GIS) to optimize the positioning of bike stations and improve urban transportation planning.

IV. Data Preprocessing for Helsinki City Bike Dataset

As a preliminary step in my data analysis, I have performed a series of preprocessing tasks on the Helsinki City Bike dataset. The objectives of these tasks were to facilitate more efficient data manipulation and to prepare the dataset for further analysis.

A. Conversion to datetime64 Objects

One of the initial preprocessing steps involved the transformation of the ’departure’ and ’return’ columns. Originally, these columns were formatted as object types, which in the context of a pandas DataFrame, are typically strings. To leverage pandas’ powerful time-series functionality, I converted these string representations of dates and times into datetime64 objects using the pd.to_datetime function. This conversion is crucial for any subsequent temporal analysis, as it allows for the utilization of datetime methods and properties inherent to pandas.

The specific conversion format I used was ’%Y-%m-%d %H:%M:%S.%f’, where each component represents a particular segment of the timestamp, including the year, month, day, hour, minute, second, and microsecond. This level of precision ensures that all temporal data is accurately represented and can be handled appropriately in later stages of the analysis.

B. Column Renaming

Following the datetime conversion, I streamlined the dataset by renaming several columns to names that are more succinct and more descriptive of the data they contain. This renaming process was achieved using the rename method of the pandas DataFrame, which provides a clear mapping from the original column names to the new ones.

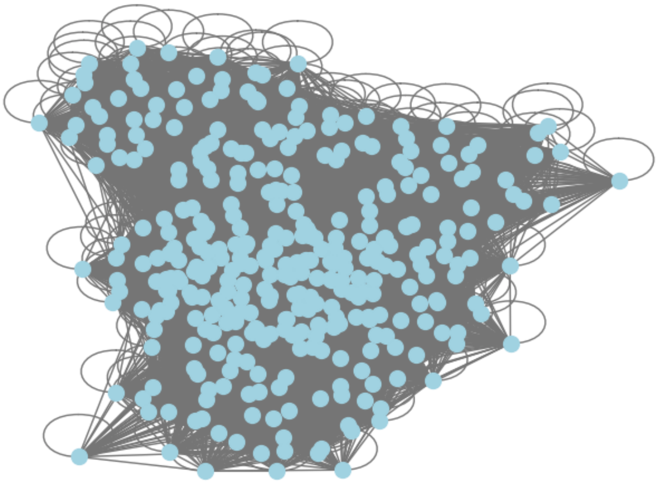


Fig. 1: Nodes in the network

These changes not only make the dataset cleaner but also enhance the accessibility and understandability of the data for analysts and stakeholders.

In summary, these preprocessing steps are fundamental in shaping the dataset into a more analysis-friendly form, setting a solid foundation for any forthcoming data exploration, and analytical tasks related to the Helsinki City Bike program.

V. Data Analysis

The network depicted in Fig. 1 is characteristic of a complex and highly connected system, with certain nodes acting as major hubs within the network. The level of connectivity indicates a well-utilized network but also points towards potential areas for targeted improvements and interventions.

Here are some key points that can be inferred from Fig. 1.

- **Size of the Network:** The network comprises 346 nodes and 27,365 edges. This suggests a densely connected network, implying that each node (which represent a bike station) has, on average, many connections to other nodes. This indicates a high level of interconnectivity within the network, which is typical for urban transport systems where many routes or trips are possible between different locations.
- **Node Connectivity:** The ‘Node degree’ section lists specific nodes with their corresponding degree value, which measures the number of connections or edges a node has to other nodes within the network. For example, ‘Haukilahdenkatu’ has the highest degree listed with 299 connections. This indicates that Haukilahdenkatu is a major hub or a highly popular location within this network, perhaps due to its geographical position, importance as a transit point, or the presence of amenities that attract more trips.

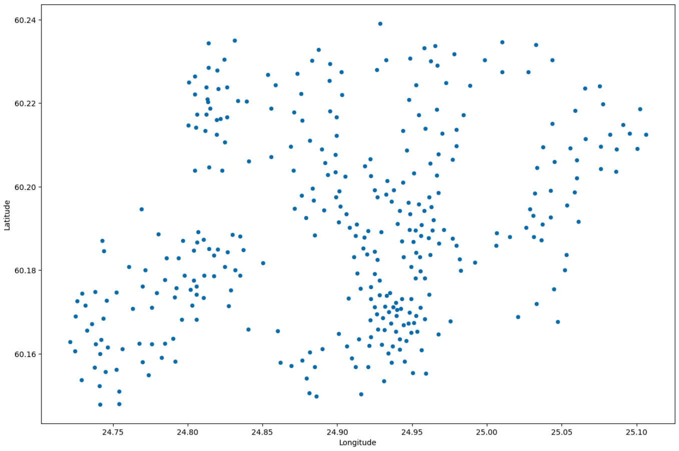
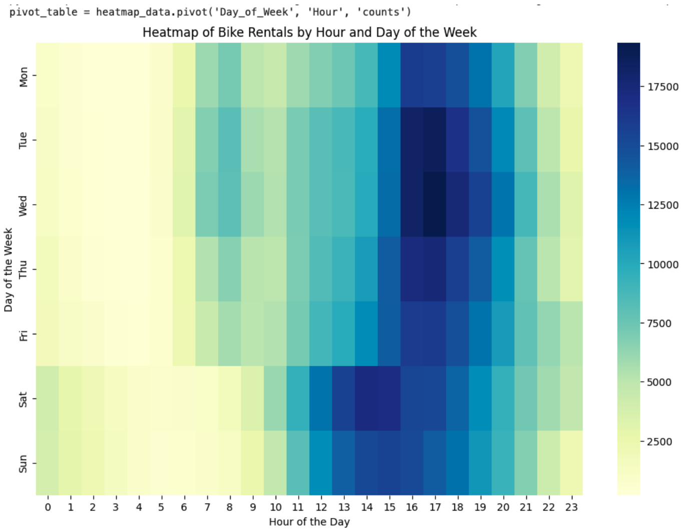


Fig. 2: Scatter plot to see the clusters

- **Network Layout:** The visualization shows a core-periphery structure, where many nodes are densely interconnected in the center, and some nodes are on the periphery with fewer connections. This pattern is common in transportation networks where central areas have higher connectivity.
- **Implications for Urban Planning:** For urban planners or city administrators, the nodes with the highest degrees might be areas that require additional resources or infrastructure development due to the high traffic. Additionally, the network analysis could inform decisions about where to place new bike stations, how to optimize bike paths, or where to focus maintenance efforts.

The scatter plot from Fig. 2 provides a visual representation of departure locations for Helsinki City Bikes, plotted according to latitude and longitude coordinates. The distribution of data points suggests a diverse range of starting points for bike trips, with noticeable clusters that may correspond to areas of high demand. The absence of duplicates, ensured by preprocessing the data, allows for an accurate interpretation of station distribution. Understanding these spatial patterns can assist in resource allocation and service optimization for the bike-sharing network.

The heatmap from Fig. 3 provides a visual representation of bike rental patterns by hour and day of the week. The data suggests that bike rental activity peaks during the midday to early evening hours, particularly from around 3 PM to 6 PM, with Friday being the most active day, indicating a possible correlation with the end of the workweek and leisure activities. The darkest shades occurring during these times suggest the highest concentration of bike rentals. Conversely, the early morning hours from midnight to 5 AM show the least activity across all days, as indicated by the lighter shades. The relatively consistent rental patterns during the weekdays, with increased activity in the afternoons, may suggest commuting behavior, while the more varied and less intense patterns on the weekend imply leisure or non-routine travel. This information could be crucial for planning bike fleet logistics, maintenance schedules, and promotional activities targeting increased usage during lower-activity periods.



```
pivot_table = heatmap_data.pivot('Day_of_Week', 'Hour', 'counts')
```

Fig. 3: Heatmap to compare weekdays and hours together

The histogram in Fig. 4 illustrates the frequency of bike rentals in Helsinki against varying temperature ranges. The observations can be summarized as follows:

- Bike rental frequency is relatively low when temperatures are below 0°C, suggesting that colder conditions likely discourage biking activities.
- The highest frequency of bike rentals occurs within the moderate temperature range of 5°C to 20°C, with pronounced peaks around 7-8°C and 14-15°C. This indicates that these temperatures are more comfortable and preferable for cyclists.
- As temperatures move towards the extremes, especially below freezing and above 20°C, the frequency of bike rentals drops off, which might be attributed to the discomfort or additional challenges posed by these conditions, such as icy roads or excessive heat.
- The roughly symmetrical distribution around the central temperature ranges implies that there's a clear preference for biking in mild and moderately warm conditions.

This data could be useful for urban planners and transportation departments to optimize the availability of bikes and maintenance schedules, ensuring adequate service during peak usage temperatures. Additionally, it can help in designing targeted campaigns to promote biking during less popular temperature ranges, if deemed necessary.

The findings from the analysis of the bike-sharing network in Helsinki, as shown in Fig. 5, Fig. 6, and Fig. 7, could be summarized as follows.

- **Network Composition and Activity:**
  - The bike-sharing network comprises 346 nodes and 27364 edges, which indicates a high level of complexity and interconnectivity among different stations in the network.
  - The average degree of 158.173 indicates that, on average, each node is connected to about 158 other nodes, suggesting a high level of direct routes available between stations.

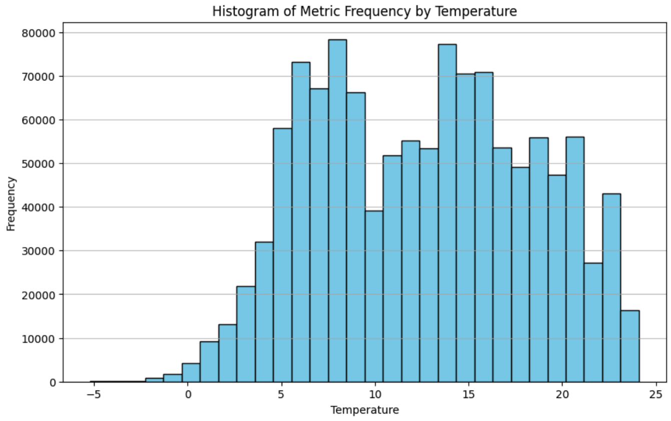
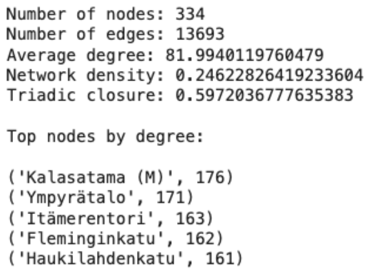


Fig. 4: Histogram to analyze bike counts vs temperature

- **Network Density and Cohesion:**
  - A network density of 0.5847 reflects a tightly knit network, with a greater proportion of possible connections being realized, beneficial for a bike-sharing system.
  - Triadic closure, sitting at 0.7127, hints at a high level of local interconnectedness, meaning that stations tend to create tightly connected triads, which is indicative of strong local user communities or frequent short trip loops.
- **Node Centrality Analysis:**
  - The nodes with the highest degrees are 'Haukilahdenkatu', 'Paciuksenkaari', 'Huopalahdentie', 'Itämerentori', and 'Laajalahden aukio', indicating that these stations are the most connected within the network, and hence, likely to be the busiest or most critical for maintaining the network's integrity.
  - When considering betweenness centrality, 'Haukilahdenkatu' has the highest score, suggesting that it serves as a significant transfer hub within the network, connecting different parts of the network and facilitating a large number of shortest paths.
  - In terms of eigenvector centrality, 'Haukilahdenkatu' again has the highest value, indicating not only that it is well-connected but also that it is connected to other well-connected nodes, reinforcing its importance in the network's structure.
- **Degree Distribution:**
  - The degree distribution histogram exhibits a right-skewed distribution, with most nodes having a degree less than 200, but with a tail extending towards 300, implying that while most stations have a moderate number of connections, a few nodes act as hubs with significantly more connections.



```
Number of nodes: 334
Number of edges: 13693
Average degree: 81.9940119760479
Network density: 0.24622826419233604
Triadic closure: 0.5972036777635383

Top nodes by degree:

('Kalasatama (M)', 176)
('Ympyrätalo', 171)
('Itämerentori', 163)
('Fleminginkatu', 162)
('Haukilahdenkatu', 161)
```

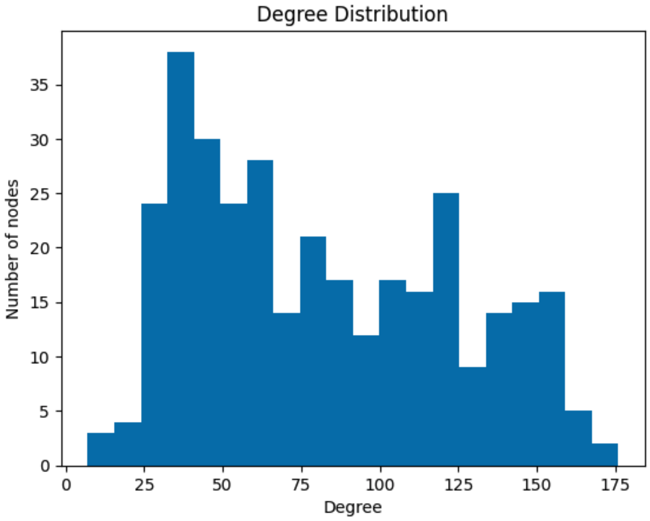
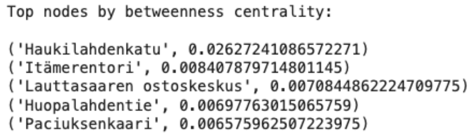


Fig. 5: Degree distribution among the nodes



```
Top nodes by betweenness centrality:

('Haukilahdenkatu', 0.02627241086572271)
('Itämerentori', 0.008407879714801145)
('Lauttasaaren ostoskeskus', 0.0070844862224709775)
('Huopalahdentie', 0.00697763015065759)
('Paciuksenkaari', 0.006575962507223975)
```

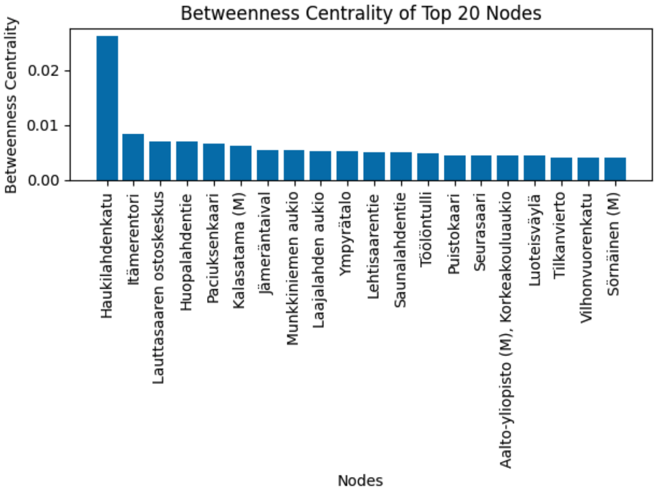
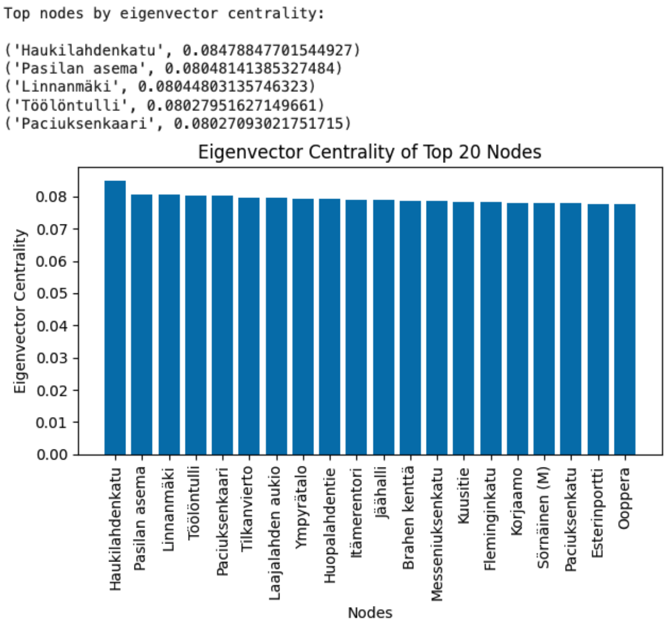


Fig. 6: Top betweenness centrality among the nodes

These findings suggest a well-utilized and highly integrated bike-sharing network in Helsinki, with specific nodes playing crucial roles in the network's connectivity and efficiency. This information can be used to guide operational decisions, such as where to allocate resources for maintenance, where to station additional bikes, and where to focus efforts for network expansion.

The essential insights we can extract from Fig. 9 regarding the betweenness centrality presented in clusters are outlined as follows.

- **Central Hub Identification:** The node with the highest



```
Top nodes by eigenvector centrality:

('Haukilahdenkatu', 0.08478847701544927)
('Pasilan asema', 0.08048141385327484)
('Linnanmäki', 0.08044803135746323)
('Töölöntulli', 0.08027951627149661)
('Paciuksenkaari', 0.08027093021751715)
```

Fig. 7: Top Eigenvector centrality among the nodes

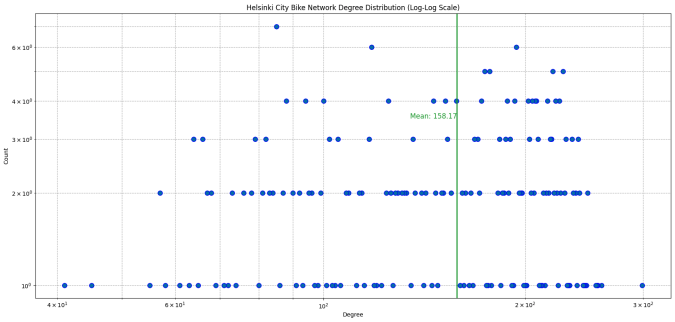


Fig. 8: Network degree distribution

betweenness centrality within the network is identified as 'Itämerentori'. This node acts as a critical hub within the network, suggesting it has the highest number of shortest paths passing through it compared to other nodes.

- **Betweenness Centrality Value:** The highest betweenness centrality value recorded is approximately 0.0097. This quantifies 'Itämerentori's' level of influence within the network as a conduit for travel routes.
- **Network Visualization:** The visualization of the network presents a color-coded representation of betweenness centrality values across different nodes. The gradient color scale from dark purple to bright green indicates the increasing level of betweenness centrality, with 'Itämerentori' likely represented by the greenest node, signifying its pivotal role.
- **Node Influence on Traffic Flow:** Nodes with higher betweenness centrality, evidenced by greener colors on the map, are influential in controlling the flow of bike traffic throughout the network. Their strategic location and connectivity make them important for the efficient functioning of the entire system.

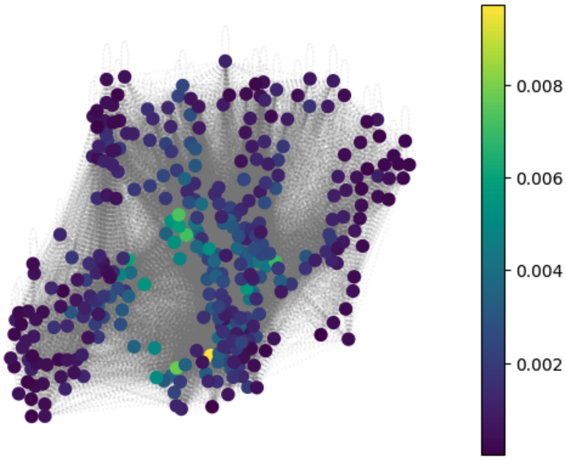


Fig. 9: Betweenness centrality in cluster forms

- **Network Robustness and Vulnerability:** Understanding the betweenness centrality helps in assessing the network's robustness and potential vulnerabilities. Nodes like 'Itämerentori' are vital for connectivity; however, they also represent points of failure that could disproportionately affect the network if disrupted.
- **Implications for Network Management:** The insights into the betweenness centrality offer crucial information for network management, maintenance prioritization, and potential improvements in the bike-sharing infrastructure. It suggests that bolstering the capacity and reliability of high-centrality nodes like 'Itämerentori' could significantly enhance overall network performance.

This analysis underscores the importance of 'Itämerentori' within the Helsinki City Bike Network and highlights the significance of betweenness centrality in understanding the network dynamics and in formulating strategic operational decisions.

VI. Conclusion

This report culminates in a comprehensive understanding of the Helsinki City Bike Network, offering critical insights into how strategic enhancements can be made to its infrastructure. The proposed approach aimed at addressing the core aspects of centrality analysis, betweenness centrality exploration, and community detection, providing a multifaceted view of the network's functionality and areas for improvement.

In my centrality analysis, I identified stations with the highest degrees of connectivity, such as Haukilahdenkatu and Paciuksenkaari, and sought to understand the underlying factors contributing to their popularity and pivotal role within the network. This analysis was challenging due to the complex interplay of various factors affecting station use, such as demographics, spatial distribution, and multimodal transport integration.

The betweenness centrality exploration focused on nodes like Itämerentori, which serve as vital connectors within the network. Understanding the significance of these nodes was occasionally difficult, as it required not just a static view of current usage, but also a predictive analysis of traffic flow patterns and potential future changes in user behavior.

Community detection emerged as a complex task, necessitating the segmentation of the network into distinct clusters to better understand localized demands and connectivity issues. The challenge here was in fine-tuning the detection algorithms to accurately reflect the real-world community boundaries and usage patterns without oversimplifying the network's intricate structure.

Despite these challenges, this report contributes significantly to the body of knowledge surrounding urban bike-sharing systems. By dissecting the network through various centrality measures and community detection, I have paved the way for targeted improvements that can lead to a more efficient, user-friendly, and sustainable bike-sharing system in Helsinki. This report not only serves as a guide for infrastructural enhancements but also as a strategic tool for city planners to optimize resource allocation, improve user satisfaction, and promote healthier, more eco-friendly urban transportation alternatives.