MLNS Final Report

Bike Sharing Schemes Community Detection

Zachary Guenassia
ESSEC & CentraleSupélec
Paris, France
zachary.guenassia@student-cs.fr

Armand Picard
ESSEC & CentraleSupélec
Paris, France
armand.picard@student-cs.fr

Thomas Taylor
ESSEC & CentraleSupélec
Paris, France
thomas.taylor@student-cs.fr

Elie Trigano
ESSEC & CentraleSupélec
Paris, France
elie.trigano@student-cs.fr

Bicycle sharing schemes (BSS) have become an essential part of sustainable urban mobility, providing last-mile connectivity and complementing conventional modes of transportation. To improve BSS adoption and expand its reach, it is crucial to understand travel patterns and identify bottlenecks and inefficiencies. In this project, we propose a station-level characterization of the London BSS network using spatiotemporal features. Our methodology involves preprocessing the BSS rental data to extract relevant features, applying different algorithms and techniques to identify clusters of stations with similar travel patterns over two different periods. By visualising the results of our approach, we identify insights that can inform decision-making and improve the sustainability and effectiveness of urban transportation systems. Our analysis is based on a subset of 1,469,851 unique shared bicycle trips in June and July 2014, distributed across 750 stations in London and more than 2 millions of trips for the corresponding period but in 2022. The proposed approach offers a solution to understanding travel patterns in BSS systems, leveraging community detection algorithms to capture the network's underlying structure and improve the reliability and attractiveness of BSS systems.

MOTIVATION AND PROBLEM DEFINITION

The use of Bicycle Sharing Schemes (BSS) has become increasingly important in urban mobility as they provide a complementary effect to conventional modes of transportation and last-mile connectivity to transit systems. With over 600 BSS systems worldwide, including successful deployments in Paris, London, and Washington D.C., BSS offers sustainable solutions to urban transportation by contributing to the resolution of the problems of congestion and pollution. To encourage adoption and increase the expansion of BSS, it is essential to understand relevant spatial travel patterns and adjust design and management strategies, such as pricing, marketing, and expansions. Understanding trip patterns can lead to advanced bicycle relocation strategies and more reliable service provision, making the system more attractive to users. To address the network complexity and noise of shared mobility systems, we will propose a characterization of the London network based on spatiotemporal utilisation features. This framework extracts clusters, providing insight into mobility patterns and identifying bottlenecks and inefficiencies to help decision-makers plan operations and manage infrastructure. We will also try to assess the evolution of these clusters eight years later to infer some information about whether the city flows have changed.

Notation:

G = (V, E): London Bicycle Sharing Scheme (BSS) network, where V is the set of stations and E is the set of edges connecting stations.

T: set of observed trips in the BSS network.

C: set of clusters of stations in the BSS network.

S_i: set of stations in cluster $i \in C$.

n_i: number of stations in cluster $i \in C$.

$t_{ij}(t)$: number of trips between stations i and j at time t.

$s_i(t)$: number of trips originating from station i at time t.

$d_j(t)$: number of trips terminating at station j at time t.

B_{ij}: binary decision variable indicating whether there is a direct connection between stations i and j.

Formal Problem Definition:

Given a set of observed trips T in the London BSS network, the goal is to identify clusters of stations with similar travel patterns and to infer the underlying structure of the network. The problem can be formulated as follows:

Cluster Identification: Partition the set of stations V into clusters C such that the within-cluster trip volume is maximised and the between-cluster trip volume is minimised.

Network Inference: Determine the connections between stations in each cluster S_i by optimising a binary decision variable B_{ij} indicating whether there is a direct connection between stations i and j.

Hardness of the Problem:

The problem of clustering stations and inferring the network structure is known to be NP-hard due to its combinatorial nature and the large number of potential clusterings and network structures. Therefore, heuristic algorithms are typically used to find approximate solutions within reasonable time constraints.

RELATED WORK

Smart data related to Bicycle Sharing Schemes (BSS) has led to significant research efforts aimed at improving our understanding of BSS, supporting evidence-based policymaking, and optimising bicycle relocation logistics. One study analysed Barcelona BSS data using spatio-temporal analyses, clustering techniques, and various machine learning algorithms to test performance. Another study used autoregressive predictive models on the same dataset to estimate station-level time-series. Vienna's BSS was analysed to obtain distinct clusters using partitioning algorithms on usage time-series data, and a predictive method was used to forecast ridership volume. In Paris, BSS stations were analysed based on usage counts using a novel Expectation Maximisation (EM) model, and identified clusters were related to their spatial relationships. These efforts aim to provide a characterization of the network based on the usage profiles of the stations and to improve the reliability and attractiveness of BSS systems.

METHODOLOGY

Our methodology addresses the problem of understanding travel patterns in the London Bicycle Sharing Scheme (BSS) system and identifying patterns in the network.
There are several parts that are included in our methodology such as:

1. Data collection & preprocessing
2. Cluster Identification & Visualisation
3. Network Inference

Data collection & preprocessing

We collected data from the Transport for London (TfL) Open Data API, which provides information on unique bicycle IDs, origin and destination station IDs and names, and start and end times for each rental.
We decided to focus on the period of June and July of 2014 for our analysis. This was the first time period where they started tracking all the

data. We wanted to choose this period, analyse it and perform our community detection and dynamics in order to compare with recent data of 2022. This would be quite insightful to see if the patterns matched in 2014 are still the same in 2022.

When checking the quality of our data we can see that the dataset is clean, there are no null values and we can directly start the preprocessing task.

For the preprocessing here are the major decisions we took:

- Renaming columns: This step is necessary to ensure that column names do not contain spaces, which can be problematic when working with the data in code.
- Converting date columns: Converting the date columns to datetime format allows for easier manipulation and analysis of the data based on dates and times.
- Dropping "unnamed_0" column: this column is not necessary for analysis, dropping it can help to reduce unnecessary clutter in the data.
- Calculating duration of trip: this is a useful feature to have, as trip duration can be an important variable in analysing bike usage patterns. We already had the "Duration" variable but it was a trip in seconds so not very interpretable quickly.
- Excluding weekends: we decided to exclude weekends in our analysis. Indeed, we decided to only focus on week days as for us it made more sense and patterns would have more chances of occurring as people could use bikes to go to work, university or other point A to point B trips.
- Calculating trip frequency: This step calculates the number of trips between each pair of stations, which can be a useful variable in network analysis. It would be used for community detection as a weight.
- Merging trip frequency with original dataframe: This step adds the trip frequency variable to the original dataframe, which allows for more comprehensive analysis.
- Calculating weighted edge: This step calculates a weighted edge variable based on trip duration and trip frequency. This can be useful in network analysis to account for the importance of different edges in the network. We have decided on a threshold of 80% importance on the trip frequency and 20% on the minute duration.

These preprocessing steps done we can have a clearer understanding of our dataset and to have better community detection analysis.

We also wanted to familiarise more with the dataset.

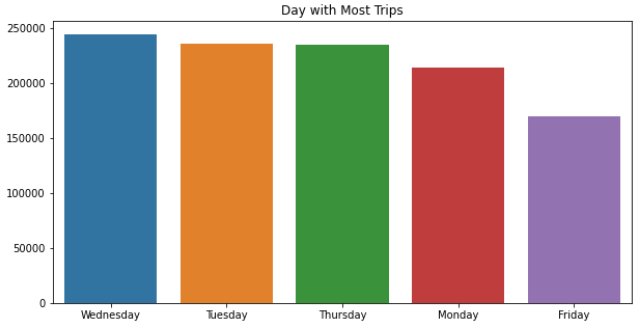


Figure 1 : usage per day of the week analysis

In the graph above we can see that the most used day is Wednesdays, tightly close to Tuesdays and Thursdays. We observe that on Mondays and Fridays there are less trips, perhaps as these are days that are near the weekend so people that go on big weekends do not work or do not go to classes.

Also, we were interested in analysing the sub boroughs of the start and finish stations of the different rides. Indeed, in our dataset, in the StartStation_Name/EndStation_Name we usually have STREET, SUBBOROUGH. They are not officially listed in the list of London boroughs, so we qualified them as "sub" boroughs.

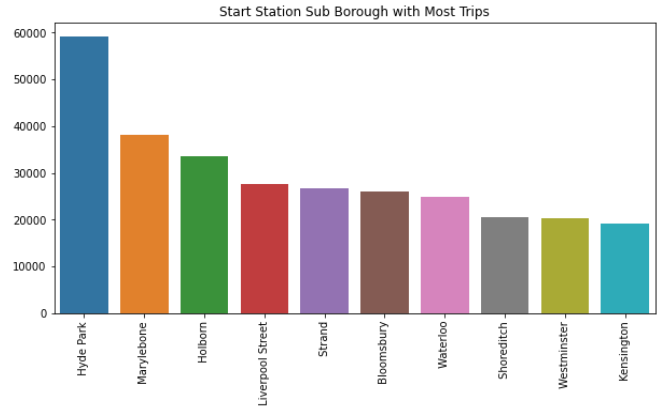


Figure 2: Start station sub borough with most trips

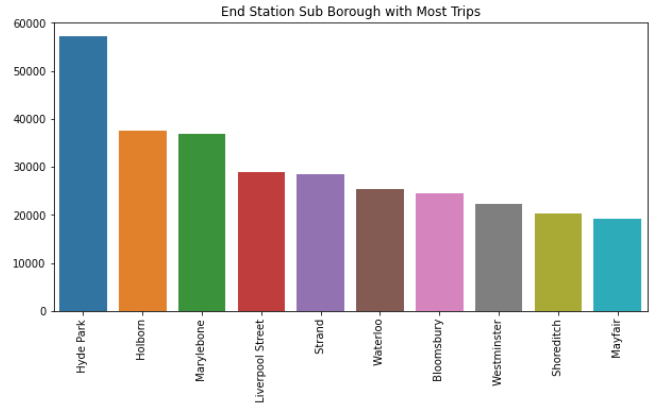
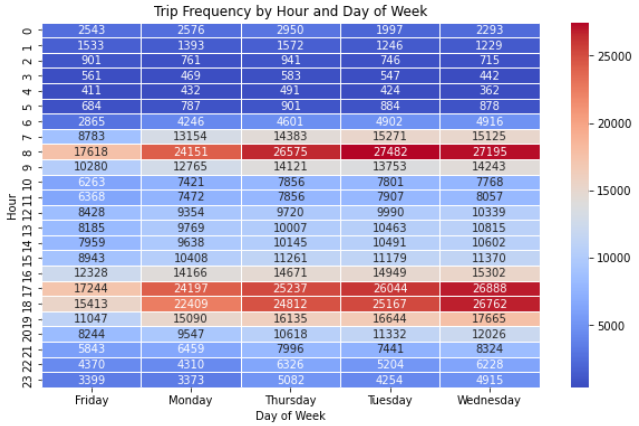


Figure 3: End station sub borough with most trips

We observe that the most frequent sub borough is Hyde Park, for start and end travel for our bike sharing scheme in London. The rest in our top 10 of most frequent stations are the same boroughs, alternating places between start and end stations.

Finally, we wanted to have an analysis of bike usage in terms of frequency per hour per day.



This was very interesting as we can see that the most frequent hours of usage during weekdays is 8 am, 5 pm and 6pm which would coincide with our hypothesis of patterns based on work, school, and other regular trips.

Our Data Cleaning, Preprocessing & Exploration part is finished. We are now going to go to our community detection part.

Cluster Identification & Visualisation

In this section, we will perform cluster identification on our bike sharing data to identify groups of stations that are similar in terms of trip patterns. To achieve this, we will use several community detection algorithms including Louvain, Infomap, Girvan-Newman, and Walktrap. These algorithms will identify groups of stations that have higher connectivity within the group and lower connectivity between groups. By identifying such groups, we can gain insights into the underlying structure of the data and potentially use this information to improve bike sharing services. We will compare the results of these algorithms and select the one that provides the best insights.

Louvain

The Louvain algorithm is a popular community detection algorithm that aims to optimise modularity, a measure of how densely connected the nodes within communities are compared to how sparsely connected they are between communities. The algorithm works by initially assigning each node to its own community and then iteratively moving nodes between communities to improve the overall modularity score. The Louvain algorithm is computationally efficient and can quickly identify communities in large networks

The Louvain algorithm is a community detection algorithm that aims to optimise modularity. The algorithm starts with each node in its own community and iteratively merges communities to maximise the modularity of the resulting network partition.

The modularity Q of a network partition is defined as:

$$Q = \frac{1}{2m} \sum_{i,j} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

Where Q is the modularity score, m is the total number of edges in the network, A_{ij} is the weight of the edge between nodes i and j ,k_i and k_j are the degrees of nodes i and j, respectively, c_i and c_j are the communities to which nodes i and j belong, respectively, $\delta(c_i, c_j)$ is the Kronecker delta function, which equals 1 if nodes i and j belong to the same community and 0 otherwise.

For instance, our Louvain clustering has a modularity score of 0.2129.

Infomap

The Infomap algorithm is a community detection algorithm that uses information theory to identify communities in networks. The algorithm works by assigning each node a unique random walk path through the network and then clustering nodes based on the similarity of their paths. The algorithm seeks to minimise the amount of information required to transmit a message within and between communities, with the goal of finding the most compact and informationally efficient representation of the network.

Girvan-Newman

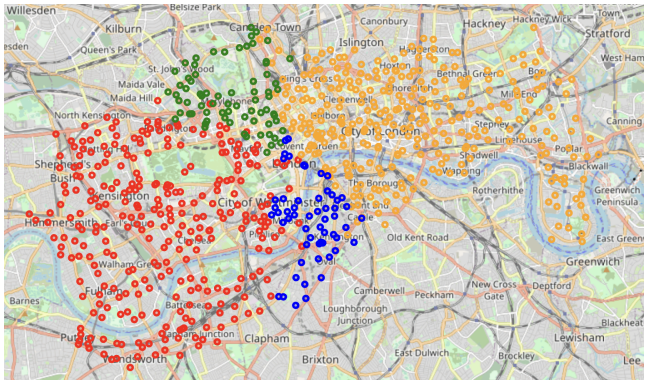
The Girvan-Newman algorithm is a hierarchical clustering algorithm that iteratively removes edges with the highest betweenness centrality until the network is split into its individual communities. Betweenness centrality measures the number of shortest paths between all pairs of nodes in a network that pass through a particular node or edge, with higher values indicating greater importance in facilitating communication between nodes. The Girvan-Newman algorithm is computationally expensive, but can be useful for networks with a clear community structure and well-defined bottlenecks.

Walktrap

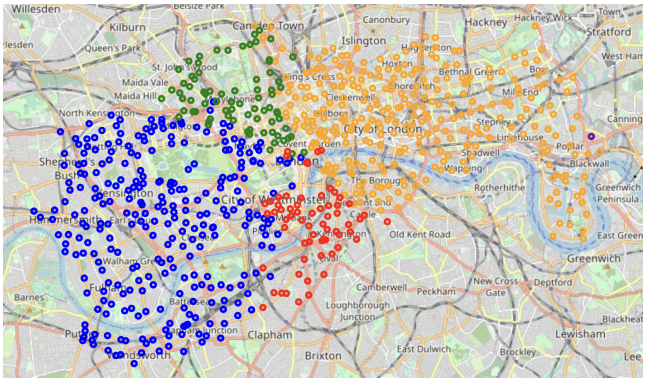
The Walktrap algorithm is a community detection algorithm that works by simulating random walks through the network and then clustering nodes based on their similarity in terms of the probability of transitioning between nodes in the same or different communities. The algorithm starts by treating each node as its own community and then iteratively merges communities that are most similar based on their random walk trajectories. The Walktrap algorithm is computationally efficient and can identify hierarchical community structures in networks.

We tried using Girvan-Newman clustering, but the runtime was very big. Indeed, after more than two hours we did not succeed to have any output of our function. We also used information from TP5 viewed in the Lab as we did a Girvan-Newman, without any success. One of the main reasons that it takes a very long time is that the algorithm can be slow for large graphs, as it requires computing betweenness centrality for all edges, which is an expensive operation. In our case, our graph has 750 nodes and 121149 edges.

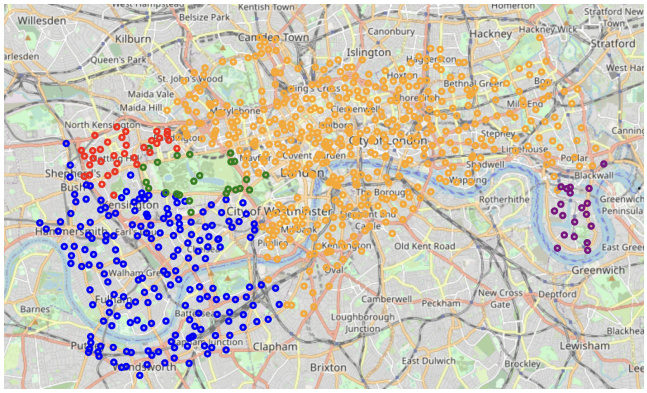
RESULTS



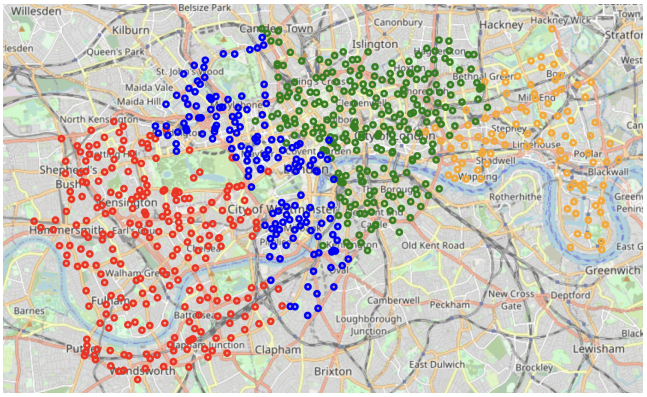
Louvain



Greedy



Infomap



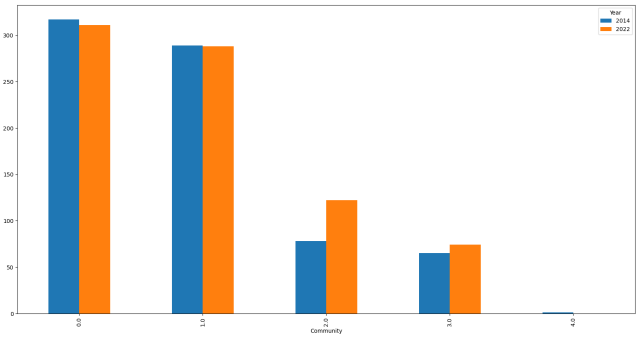
Walktrap

Community detection representation:

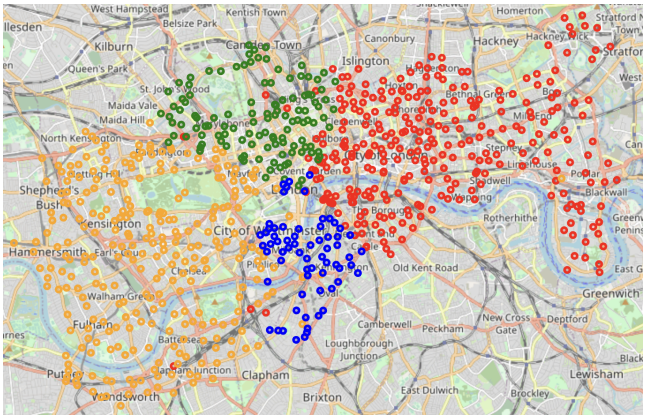
The four figures above represent the comparison of the different community detection algorithms where stations are coloured according to their assigned community.

Previous research has extensively discussed the technical distinctions among the four approaches, which are reflected in their respective output communities. In particular, Infomap is the only method capable of identifying well-known physical structures in London, such as Hyde Park and Canary Wharf. While the other three methods yield four clusters, the optimal solution provided by Infomap results in six modules. The largest community, accounting for over half of the trips in the network, is Central and East London and Regent's Park (orange). It is adjacent to South-West London (blue), and Hyde Park (green) clusters to the west, which are the second and third largest clusters in terms of flow. These three clusters are also adjacent to the fifth largest community, Notting Hill (red). Finally, Canary Wharf (purple) is located in the remote South-East and has the lowest flow of any cluster, bordering only the Central cluster.

Community Evolution



Here, we plotted the evolution of the distribution of communities among the London bike stations in 2014 (blue) vs 2022 (orange). We did the same comparisons with 2016, 2018 and 2020 and we obtained approximately the same results. What we can see is that the changes in communities come mainly from the construction of new bike stations rather than modification of behaviours.



On this plot with the greedy algorithm in 2022, we can see that some new stations were opened in North East London for instance. However the clusters are still more or less the same.

CONCLUSION

In our project, we tried to provide important information and implications for users, providers, and authorities involved in Bikeshare systems. We compared different methods for detecting communities in Bikeshare systems that can identify both system features such as directed links and exogenous network formation, as well as urban environments like Hyde Park. These methods can help us understand user behaviour and the geographic boundaries of Bikeshare trip-making. By using these techniques, we can encourage Bikeshare adoption based on detected communities, and connect communities by incentivizing users or expanding infrastructure.

Our study highlights the spatio-temporal complexity of Bikeshare systems, and suggests that future research should explore the driving factors behind our observations. It also suggests that more holistic approaches are needed to draw meaningful operational and political conclusions, which could include contextualising patterns in shared mobility systems with urban amenities or weather data, or in-depth analyses of the dynamics throughout the day. Our methodological contribution could be expanded to address time-varying networks of bike stations and communities by providing deeper insights.

REFERENCES

Some useful references that have treated the same subject:

Gao, S., Zheng, Y., & Li, D. (2017). Exploring spatio-temporal bike sharing patterns: A case study of the bike sharing system in Melbourne. Transportation Research Part A: Policy and Practice, 101, 125-135.

Wang, X., Zhou, X., & Ye, J. (2020). Revealing urban travel patterns with bike-sharing data: A review. Transportation Research Part C: Emerging Technologies, 116, 102598.

Lou, Y., & Yin, Y. (2018). A network flow-based method for improving the station-level performance of bike-sharing systems. Transportation Research Part C: Emerging Technologies, 89, 117-132.

Guo, Y., Wang, J., & Liu, Y. (2019). A data-driven model for predicting bike-sharing demand in stations. Transportation Research Part C: Emerging Technologies, 107, 14-25.

Vu, H. L., Lee, M., Kim, K. W., & Kim, K. (2020). Bike-sharing system: A comprehensive review of its existing literature. Transport Reviews, 40(1), 62-91.

Shi, X., Wang, Y., Lv, F. et al. Finding communities in bicycle sharing system. J Vis 22, 1177–1192 (2019). https://doi.org/10.1007/s12650-019-00587-0