Midterm report
Analysis of bus public transport in Malaga using graphs

Jesús Moncada Ramírez

Learning from Networks 2022-23

The current state

Up to now, we have divided the work into **three sections**, the first two are finished and the last one is just started. Here is a summary of what has been developed in each section:

1. Importing the datasets.

We have created the correspondent **Python data structures** for the CSV files of our dataset. For this, we have used the library *pandas*; in this way, the data structures are *pandas DataFrames*, which allow us to perform a lot of implemented operations. After importing, we renamed some column labels of the data, translating it from Spanish to English. Finally, we have provided a brief view of each dataset showing some samples. After all, this section oversees data preprocessing.

2. Creating the graph.

Here, we have created the graph using the library *NetworkX.* We have added the nodes (all the bus stops in the city), and the edges (all the connections between two stops belonging to the same line). Before adding the edges, we had to choose a measure of the quality of each edge, this is, its weight (which has been called ***velocity***). Finally, we have decided to make the *velocity* of an edge inversely proportional to the **distance** between the two stops and directly proportional to the **number of buses** in its line. It is also influenced by the product by a constant that has been set empirically. All the values we got have been normalized and applied to a threshold (see section *Problems found*).

The last thing in this section is the creation of an **image** of the generated graph to check if the elements were properly imported. Two images have been generated, one of them where the edge color depends on the bus line, and the other where the edge color depends on the *velocity* attribute.

3. Node centralities

In this section, which is not finished yet, we have started with node **centralities**. We calculated the closeness centrality of all the stops and analyzed the data.

The code has been uploaded to a GitHub repository called *malaga-bus-graph*. With this, we do a more specific control of versions and make the project open-source.

Problems found

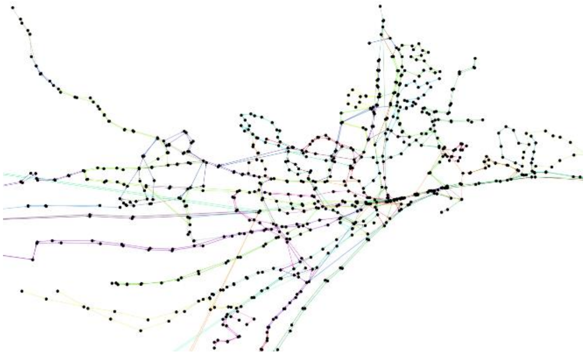
During the development of this first version, we have found some problems, a solution to all of them has been proposed, but it could change with the development of the project:

- **Inconsistencies in the dataset.** We have discovered that the data offered by the Malaga council are not perfectly correlated. An example of this is the dataset "*lines_and_stops.csv"* where you can find bus lines never referenced in "*routes.csv"* and stops never referenced in "*stops.csv"*. The solution to this has been creating the graph using only the data from *"lines_and_stops.csv"* and using the other files to get extra information. In some cases where the absent information was needed, we simply set a default value.

- **Values of the *velocity* attribute (weight)**. After setting the attribute *velocity* for all the nodes, we found that the data contained extreme values (especially very high *velocities*). This is the case of two stops of very important lines that are very close geographically. To avoid this, we have normalized the data (against the maximum) and applied thresholding with an empirical value.
- **Library *Mplleaflet*.** *Mplleaflet* is a very interesting library to display information on maps using Python. Unfortunately, it presents bugs, and the installation is quite complicated. Finally, we managed to create the map, but this section has been labeled as optional in the Jupyter notebook of the project.

Completed goals

Concerning the objectives established in the project proposal, we have completed the first one (*Create a graph representing all the bus lines and stops of the EMT…)* and we have started with the second one (*Compute some graph analytics at the node level…*).



Part of the drawing generated for our graph.

Some results

One of the results obtained up to now is the **drawing of the generated graph**, which perfectly represents the shape of the city of Malaga. Another is the identification of the **more central stop** (with the highest closeness centrality), which turned out to be one of the busiest in the city center, called *Paseo del Parque*.

Modifications

Based on what we have said, the biggest modification we would do to the project proposal would be the reduction of the initial objectives, as we think we are not going to have time to develop them. The objective we would discard would be the one related to **random graphs** (*Do some experiments with random graphs*…). This topic will be retaken only in case we have time.

Some questions

At this point, any comments on the following questions would be very helpful:

- Between two stops there are usually a lot of edges (one for every bus line that connects them), each one with its *velocity*. Would it be a good idea to put together all those edges and create a unique edge whose *velocity* is the sum of all the previous ones, for example? We have been thinking about it and we don't know if it will bring advantages or disadvantages. Up to now, the graph has multiple edges between nodes.
- Has it been a good idea to normalize and apply thresholding over the values of the *velocity* attribute? The motivation was to have more uniform values for the graphical representation because otherwise, the colors weren't distinguishable at all.
- Regarding the next steps, we think about finishing the node centralities (betweenness, current flow closeness, current flow betweenness, harmonic… and more that the library allows computing). What should be our next steps? Computing the graph clustering coefficient? Or maybe try some algorithms about motifs?