Public Transport Network Analysis

Final project for the Data Analytics course for the MSc in Computer Science at University of Milano-Bicocca.

Authors: Nassim Habbash (808292), Ricardo Matamoros (807450)

Introduction

This project aims to analyze the structure of a city network using averaged graphs of the different means of public transport.

Average graphs represent the connectivity between regions in a city. Using this method allows to overcome the disjointed nature of the different networks of transport and possibly noisy datasets without having to expend significant resouces into data quality processing. Given different average graphs representing different types of connections (neighboring regions, transport direct connections), it is possible to analyze and compare each to understand how optimal a public transport network is. Simulation of random failures and attacks can also shed insight on the robustness of the network.

Project structure

The project has been developed to allow easy reusability/recomputation of the transit feed data for different times or cities. As each public transport regulator may implement the GTFS format differently, slight modifications might have to be made. The code provided in the project's GitHub repository is tested and works on Milan's Azienda Mobilità Ambiente e Territorio GTFS feed, the README containing the instructions to reproduce this work.

Data management

All data has been obtained by OpenMobilityData. The data consists of Milan's GTFS feed from the 4th of March 2019 to the 4th of April 2019. As the data is quite extensive, a PostgreSQL Docker container has been setup for ease of use to contain Milan's GTFS database, with automatized database population.

The GTFS (General Transit Feed Specification) is a collection of at least 6 and up to 13 CSV files, describing a transit system's scheduled operations. The necessary data to reconstruct a network is contained in the following tables:

- **Stops**: Defines stops where vehicles pick up or drop off riders.
- **Stop times**: Provides the times when a vehicle arrives at and departs from individual stops for each trip.
- **Trips**: Defines trips for each route. A trip is a sequence of two or more stops that occur during a specific time period.
- **Route**: Defines transit routes. A route is a group of trips that are displayed to riders as a single service.

Through a set of different queries and some data processing, tables containing every stop for every route for metro lines, buses and trams have been produced.

Following are some statistics for the resulting graphs:

- **Metro**:
  - Density: 0.02
  - Nodes: 106
  - Edges: 109
- **Bus**:
  - Density: 0.001
  - Nodes: 2110
  - Edges: 2163
- **Tram**:
  - Density: 0.004
  - Nodes: 449
  - Edges: 445

Data analytics methodology

The graphs of metro lines, buses and trams are disjoint: although different stops refer to the same street, they are located at different coordinates and have a different unique IDs. As such, metro stations, bus stops and tram stops never intersecate, although there are some exceptions for buses and trams sharing the same stop (same unique ID).

It's not sensible to conduct a connectivity analysis and failure test on a disjointed network: although not directly connected, people can and will move from a metro station to its nearest bus or tram station,

to move from point A to point B. In case of failures between stops, spatially close stops work as a back-off to allow paths A and B, with C failed, to be reached by a working stop D.

Graphs derived from large transit systems networks and how to represent mobility has been extensively studied. An approach to represent trajectory data of city-wide traffic dynamics in spatial and temporal domains has been proposed by Kim et al. (2016). An approach building on the previous, by Yildirimoglu and Kim (2018), analyzes different traffic flows on a shared spatial grid, allowing for multiple traffic flows to be analyzed as a real world network.

This project takes inspiration from both these approaches to analyze how Milan's transport network behaves in the spatial domain.

Partitioning a network in cells

The partitioning of a network in regions (hereon cells) according to the spatial relation of its nodes (the stops in this case) is based on a method developed by Adrienko and Adrienko (2011).

Let S be a set of seed points, where a point $p(lat, lon)$ represent a stop. Let γ be the radius specifying the radius of a cell. We'll have:

$$\forall p_n \in S \begin{cases} p_n \in C_i.m, C_i.c = avg(C_i.m) & : dist(p, C_i.c) < \gamma \\ p_n \in C_{i+1}.m, C_{i+1}.c = p_n & : otherwise \end{cases}$$

Where C_i in the presentated formula is a cell of points each inside the radius of the cell's centroid, $C_i.m$ represents the cell's members and $C_i.c$ is the cell's centroid. The centroid of each cell is estimated iteratively by finding the mean of each point as they get assigned to each cell. After all the points p have been grouped into cells G_i, the cells are emptied and the points redistributed to the closest cell.

The set of cells C represents the partitions into which the stops are grouped. From this, it is possible to compute the Voronoi diagram of C, which partitions the space into geometrical cells with $C_i.centroid$ as a center, delineating the boundaries between each cell. It's complementary graph, computed through the Delaunay triangulation of the centroids, represents the *graph of neighboring regions*, while the Voronoi diagram represents the *spatial subdivision of the area.*

TODO: include voronoi diagram and Delanay diagram images

References

Adrienko, and Adrienko. 2011. "Spatial Generalization and Aggregation of Massive Movement Data."

Kim, Jiwon, Kai Zheng, Sanghyung Ahn, Marty Papamanolis, and Pingfu Chao. 2016. "Graph-Based Analysis of City-Wide Traffic Dynamics Using Time-Evolving Graphs of Trajectory Data." In.

Yildirimoglu, Mehmet, and Jiwon Kim. 2018. "Identification of Communities in Urban Mobility Networks Using Multi-Layer Graphs of Network Traffic."