A Network Tour of Data Science: Project proposal

Team 03
Dönz Jonathan, Esguerra Martin, Vojinovic Stefano

December 2019

Introduction

The recent swine flu (H1N1) epidemics outbreak of 2009, the ebolla crisis of 2014 and the one currently taking place in the region of the Democratic Republic of Congo (2019) are alarming illustrations of the large scale and deadly impact pathogens can have today. The deadliest epidemics that occurred in the last 200 years was the Spanish flu, that started in 1918 and caused between 50 and 100 millions casualties [1]. This represents 3 to 6 % of the global population living at that time [2].

This deadly pandemic took place in an era where airline transportation systems did not exist. Considering the much larger transportation network available today, and the staggering speed at which individuals can travel, it is relevant to ask oneself what would be the impact on our global population if a comparable pandemics started today. In this project, we seek to answer the following research question: **Would humanity survive a Spanish flu-like pandemic today?**

Plan

To answer this question, we will first build the graph of the global airline transportation system. The nodes of the graph will correspond to airports. The directed edges $(i, j) \in E$ will be the estimates of the flow of people from airport i to airport j per unit of time. We may also add node features such as the number of people living in the neighbourhood of the airport.

We will then proceed with the following steps:

1. Study the type of graph and its characteristics.
   We expect the graph to be scale free with a small γ parameter
2. Make a spectral clustering.
   We expect to see clusters based on geographical proximity
3. Simulate the Spanish flu pandemic using a Susceptible-Infected-Recovered (SIR) model. We will estimate the pathogen's spreading parameters (β, μ) from historical data
4. Study the spread from the spectral clustering perspective
5. Simulate different control strategies such as:
   - Random vaccination
   - Selective immunization
   - Isolating the biggest hubs
6. Compare the different control strategies and conclude

Appendix

The datasets we will be using are the following.

- `airports.csv`: contains records of over 10,000 airports around the world. There are 18 features per record, such as the standardized airport identification code (*IATA code*), latitude and longitude. It is available at `https://openflights.org/data.html`.
- `routes.csv`: contains records of over 60,000 flights between 3,321 airports on 548 airlines. The features of interest will be the IATA codes of the source and the destination airports allowing to join the flights with the airports from the `airports.csv` dataset. It is available at `https://openflights.org/data.html#route`.
- `regions.csv`: contains all countries' regions (provinces, states, etc.). It contains the same `local_code` variable as the `airports.csv` dataset allowing to join the regions with the airports. It is available at `https://openflights.org/data.html`.
- `world_cities.csv`: contains the population of more than 11,000 cities in the world. It may be used to join the airports with the population of their neighbourhood. It is available at `https://simplemaps.com/data/world-cities`.

References

[1] Niall Johnson and Juergen Mueller. Updating the accounts: Global mortality of the 1918-1920 "spanish" influenza pandemic. *Bulletin of the history of medicine*, 76:105–15, 02 2002.

[2] Jeffery K Taubenberger and David M Morens. 1918 influenza: the mother of all pandemics. *Revista Biomedica*, 17(1):69–79, 2006.