Using network metrics to explore how segregated a city is

Context and project description:

City dwellers usually move in the city according to predictable patterns, within well-defined areas and using quite repetitive routes (everyday home to work travels, for instance). This brings the question of what gives a city its unity if people are moving only within small fragmented parts of it. In particular, can we unveil the existence of sub-cities that concentrate movements within their boundaries?

In this project, we are particularly interested in exploring how movement patterns within the city fragments a city into "bubbles" or communities. We will use a dataset that contains 346,638 data points representing the approximate locations of the home and work places for specific anonymised dwellers from the city of Santiago de Chile.

Through the building of a network and subsequent analyses, the aim is to describe the community structure that underlies movement patterns in the city.

Dataset:

We will use the dataset provided in [1]. The data was obtained from anonymised mobile phone connections to the closest mobile phone towers[1]. The authors processed this data to create two main files:

1. **Home_Work.csv:** Each row contains both home and work locations of a particular person. Home and work locations are just approximate, as they correspond to the location of the closest cell phone tower to which the person connects.

2. **Communities.csv:** The authors of [1] created a network from the Home_Work.csv dataset and then applied community detection algorithms and discovered that the network can be divided into six different communities or sub-networks. This file contains the communities, where each cell phone tower contains a label of the community it belongs to.

[1] https://en.wikipedia.org/wiki/Cell_site

Tasks:

Task 1:

Starting from *Home_Work.csv,* build a network where nodes are mobile phone towers, and they are connected through people moving between them. You are asked to build the following networks:

a) Undirected and unweighted network
b) Directed and unweighted network
c) Undirected weighted network
d) Directed and weighted network

For creating the directed networks, you can simply create a directed edge in the home → work direction. The weights should correspond to the amount of dwellers that move from one location to the other.

Discuss the differences between each of the above networks. What are the advantages or flaws of each network if we want to assess the spreading of a disease through these networks?

Task 2:

Choose two networks that you have built in Task 1. For each one, perform the analyses described below.

a) Use the community detection algorithm provided in Gephi [2,3] to obtain communities from the network. Use 5 different "Resolution" values and show how each value is influencing the resulting community structure (number of communities, average and standard deviation on the size of communities, among others).

b) Compare the communities you have found with the ones obtained in [1] (available in *Communities.csv*). Discuss any influence that can be attributed to the particular parameter value or the network used (directed/undirected weighted/unweighted).

c) Build a map that displays each node (the towers) at their corresponding geographical location, and in which each community is denoted by a different colour (something similar to Figure1b in [1]).

Note: you have to mention which two networks of Task 1 you have chosen for this Task and *explain why* you chose these two over the other networks of Task 1

Task 3:

Finally, we are concerned about investigating to what extent the communities you have found can be seen as independent sub-cities. Is the city of Santiago a single centralized pole, or is it rather a multi-pole city?

a) For this part, choose only one network from Task 1 and one of the five partitions found in Task 2. Use two different centrality measures of your choice (you have to justify your choosing; use the lecture slides and bibliography, and refer to chapter 7 of [5] for a good review or also to [6] for a critical comparison of centrality measures) to explore how important each node (tower) is within the city. Contrast your findings with the communities found in Task 2. Are communities equally central regarding the nodes they contain? Compare the centrality distribution of different communities for discussing the above question.

b) Compare the distribution of centrality to the geographical location of the nodes. Are nodes that are more distant to the geographical centre of the city less central? Make use of plots that can help to illustrate your point (e.g. geographical distance to some central tower vs centrality).

Task 4:

*This Task should be conducted ONLY by PG students (and **not** by UG students).*

a) Similar to what was done in [1], and for the same partition used in Task 3, calculate $P_X(Y)$, the probability that a dweller that lives in community X works in community Y. Calculate this probability for each pair of communities (including the case when $X=Y$).

b) What would be the expected values of $P_X(Y)$ if the city wasn't segregated at all (e.g. if each dweller chose their workplace randomly between all the available locations within the city)?

c) Compare the values of $P_X(Y)$ obtained on the real data with that obtained for the random case described above. Can you tell from your results whether the city is segregated?

References:

[1] Dannemann, T., Sotomayor-Gómez, B., & Samaniego, H. (2018). The time geography of segregation during working hours. *Royal Society open science*, *5*(10), 180749.

[2] https://gephi.org/users/download/

[3] Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical review E*, *80*(5), 056117.

[4] https://networkx.org/documentation/stable/reference/algorithms/community.html#module-networkx.algorithms.community

[5] Newman, M. (2018). *Networks*. Oxford university press

[6] Landherr, A., Friedl, B., & Heidemann, J. (2010). A critical review of centrality measures in social networks. *Wirtschaftsinformatik*, *52*, 367-382.