Analysing the segregation levels of a city using network metrics

Anandraj Selvam, Neeraj Yadav, Sindhu Muthumanickam, Suganya Senguttuvan

Abstract - **The project's goal is to use network measurements to investigate how people's movement inside a city can divide it into "bubbles" or "communities". Data from connections between mobile phones and the nearest cell towers in Santiago, Chile are included in the collection in an anonymized form. Four activities are required for completion of the project: creating networks, using Gephi to analyse communities, determining node centrality, and calculating the likelihood of residents working in one community while also residing in another. The initiative aims to determine whether the city is segregated as well as whether it has multiple centralised poles or one central pole.**

Introduction

This paper aims to explore how movement patterns in a city fragment it into communities or "bubbles." The research problem is to determine the community structure underlying movement patterns in Santiago de Chile, using network metrics. The dataset comprises 346,638 data points representing approximate locations of home and work places for anonymous dwellers. The authors of the paper provide two main files: Home_Work.csv, containing both home and work locations of individuals, and Communities.csv, containing six different sub-networks discovered from a network created from the Home_Work.csv dataset.

This paper outlines four tasks. The first task is to build four different networks using the data provided. The second task involves selecting two networks from Task 1 and performing several analyses, including applying the community detection algorithm provided in Gephi, comparing communities found with those obtained in [1], and building a map that displays each node in their corresponding geographical location. The third task concerns investigating the communities found to what extent they can be seen as independent sub-cities. The fourth task is optional, intended for postgraduate students only, and involves calculating the probability that a dweller who lives in one community works in another community.

The motivation behind this project is to understand how movement patterns within a city fragment it into communities. The challenges associated with this project include creating accurate networks from the data provided, selecting appropriate metrics to analyze the data, and drawing meaningful conclusions from the results obtained.

In the remainder of the paper, we discuss the differences between the four networks created in Task 1 and their advantages and flaws when assessing the spreading of disease through these networks. We then analyze the two networks selected from Task 1 in Task 2, compare the communities found with those obtained in [1], and build a map that displays each node in their corresponding geographical location. In Task 3, we investigate the communities found and to what extent they can be seen as independent sub-cities. Finally, Task 4 is involved calculating the probability that a dweller who lives in one community works in another community.

Related Work

Residential segregation is the physical division of various social groups, such as racial and ethnic groups, according to where they reside. Assessing how spatially isolated members of various groups are from one another is the aim of segregation measurement. The percentage of one group that would need to relocate in order to establish an equal distribution of that group and the other group(s) throughout the area is measured by the Duncan dissimilarity index [3]. Traditional measurements of segregation concentrate on the spatial distribution of various social groups, but frequently overlook social interactions that take place in other circumstances, such as at work and during leisure. The "exposure dimension" of segregation enters the picture at this point.

The degree to which members of various social groups interact with one another on a regular basis is referred to as the exposure dimension. Interactions in places other than residential neighbourhoods, such as companies, schools, and public spaces, might be considered [4,5]. Researchers have created new indices that evaluate the level of interaction between distinct social groups in various circumstances in order to quantify the exposure dimension of segregation. Researchers can learn new things about how social disparity is created and perpetuated in various circumstances by explicitly taking the exposure component of segregation into account. For instance, they could look at how patterns of leisure segregation continue social exclusion and inequality, or how patterns of employment segregation lead to salary differences between various racial and ethnic groups [8,9]. The set of areas that each person encounters during their daily routine, sometimes referred to as their activity space, is taken into account by the idea of exposure in the study of segregation [10]. Accordingly, researchers take into account all of an individual's daily activities, including where they work, attend school, and spend their free time, rather than only focusing on where they live.

Researchers frequently combine racial-ethnic data with daily traffic data surveys to determine exposure levels for various ethnic groups in each area. Wong and Shaw [6] estimated exposure levels for several ethnic groups in southeast Florida using daily travel data surveys. Similar to this, Farber et al. calculated the social interaction potential index using origin-destination surveys and the time geography framework [7]. This can assist in locating potential barriers to social integration and interaction as well as locations of high or low exposure. Cell phones and other communication technologies have made it possible for researchers to gather a sizable amount of non-traditional data. For instance, researchers can recreate urbanites' daily itineraries and gain a high spatiotemporal resolution of social contacts by examining the connectivity of cell phones to several towers during the day.The evaluation of a wide range of issues, including personal mobility patterns, land-use patterns, and the identification of hotspots for social interaction within a city, has made extensive use of this methodology. Ratti et

al. identified areas of frequent social contacts using Newman's community detection algorithm by examining call detail data (CDRs) of landline communication in the UK. It's interesting how these settlements matched up with British administrative regions. Some have claimed that rather than actual social connections, identified communities may just be the outcome of limited individual movements. To overcome this issue, more practical strategies have been devised that take into account gravitational effects in the null model. Researchers can more clearly distinguish between genuine social interactions and random motions by considering gravity forces.

Using this method, relevant communities in mobility networks have been found at the local, regional, and national levels. They also look at how social inequality, health, and well-being are impacted by urbanisation and other urban dynamics, as well as how cities might be planned and run to encourage more equitable results for all citizens. By completing a thorough analysis of community detection methods in healthcare applications, this study seeks to close the knowledge gap. In this article, existing algorithms are categorised, discussed, and their uses in healthcare are examined.

Dataset and network presentation:

The goal of this study is to comprehend the community structure that underlies the travel patterns in Santiago de Chile by building a network and doing subsequent analysis. The project's goal is to locate and analyse node clusters—areas of the city—that have more connections between them than between them and other nodes. The research wants to show how mobility patterns give rise to communities or bubbles. Network metrics will be utilised in the network analyses to define the city's community structure, and the findings will be used to gauge the level of segregation there.

1. Dataset

The dataset utilised for this project contains 346,638 data points that roughly represent the locations of anonymous people's homes and offices in Santiago de Chile. The dataset offers a thorough picture of urban movement patterns, making it appropriate for building a network to study these patterns. The dataset is anonymised, protecting individual privacy and according to ethical standards for research. The dataset's quantity and level of information enable a thorough investigation of the community structure underpinning the city's travel patterns.

2. Network Development

The dataset will be used to build a network that the researchers will use to examine mobility patterns and pinpoint the fundamental community structure in Santiago de Chile. The network will be built by connecting nodes, which represent different parts of the city, according to how frequently and closely people move between them. The weighting of the connections between nodes will depend on how strong they are. The resulting network will depict both the communities or bubbles that develop as a result of the city's migration patterns. To guarantee the accuracy and dependability of the outcomes, the network development procedure will be carefully planned.

3. Detection in the Community

A key method for examining how network topology affects behavioural patterns in various systems is the identification of communities within a network. By examining people's home-to-work journeys, we built undirected weighted networks by understanding the provided network. The weight of each link between two nodes in these networks corresponds to the number of shared home-work trajectories between the two towers, and the nodes in these networks serve as placeholders for cell phone towers.

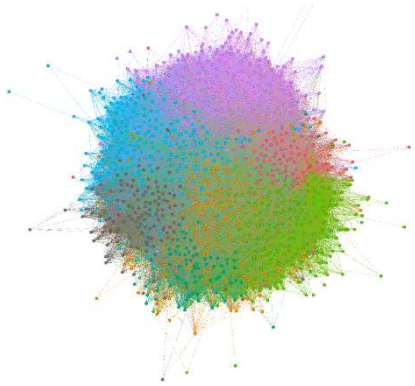
Network Analysis Methodology:

We utilised the Gephi tool to run the experiments on the provided dataset for the network analysis. The study collected anonymous data from mobile phone connections to adjacent towers. Two major files were made after the data was processed:

1. "Home_Work.csv" comprises rows with people' approximative home and workplace locations, determined by the nearest mobile phone tower to which they are connected.

2. The file "Communities.csv" contains details on the six distinct communities or sub-networks that were identified when community detection techniques were applied to the network built using the "Home_Work.csv" dataset. The community to which each cell phone tower belongs is shown on the label.

The given dataset initially had UTM coordinates, hence a transition from UTM to GPS was required. After pre-processing the data, we created GPS coordinates to build the city's contour depending on mobility, as seen in Figure 1 below. The assumption was made that each user's home and workplace were the places where tower signals were received the most frequently during working hours, i.e., between 22.00 and 07.00 and 09.00 and 17.00, respectively, with at least five pings to both locations.

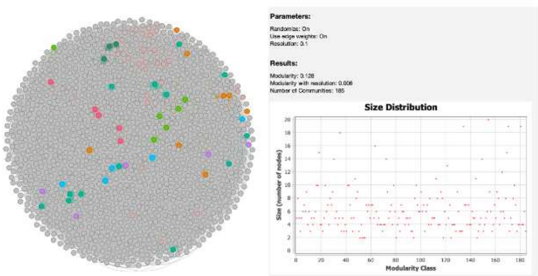


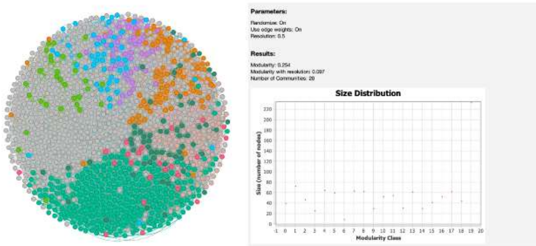
This network serves as an example of a undirected, weighted network that may be used to analyse how residents of various communities move around. It has become common practise in many disciplines to identify communities in networks using a variety of tools and techniques from various disciplines, such as biology, physics, statistics, social sciences, cognitive sciences, mathematics, economics, and computer science. However, it is generally acknowledged that due to the variety of network topologies and objectives, no single algorithm can effectively discover communities in all forms of social networks.

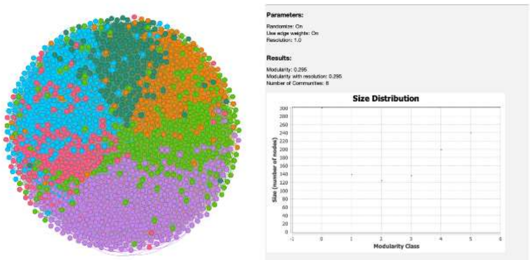
It has been explored in the study how to identify groups or communities inside a network based purely on their patterns of connectedness, without having any prior knowledge of the underlying features or characteristics of the nodes within such groups. By assessing the degree of modularity and making the assumption that well-separated communities have high modularity values, modularity-based techniques seek to discover communities within a network. In order to identify communities in a network, the author describes the modularity metric [29]. Because of this, the Gephi tool in our study also calculates the modularity measure for various resolutions from the home-work data that is analysed.

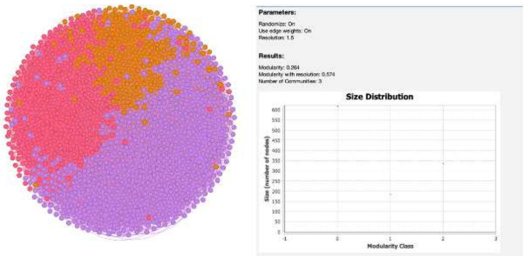
Results and Discussion:

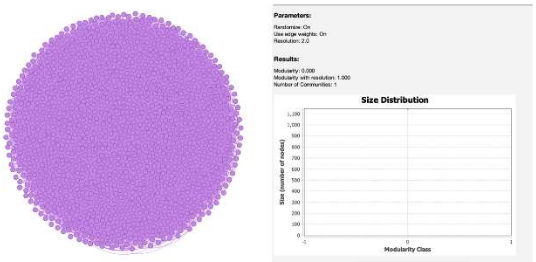
The analysis involves building four different networks using the provided Home_Work.csv file, which contains the home and work locations of each anonymous individual. The nodes in the network are mobile phone towers, and edges between two nodes represent people moving between the corresponding towers. Four types of networks are created, including an undirected and unweighted network, a directed and unweighted network, an undirected weighted network, and a directed weighted network. The weights in the directed weighted network represent the number of dwellers moving between two locations.

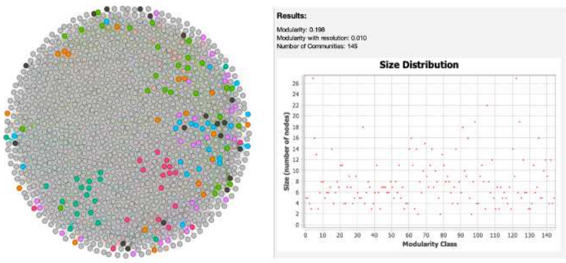


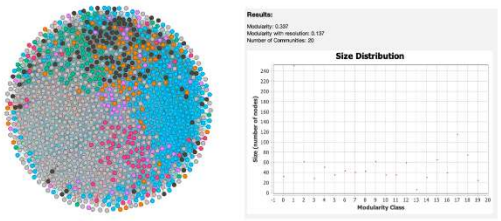


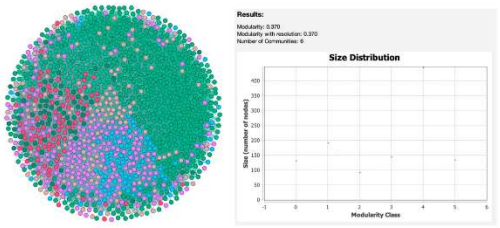


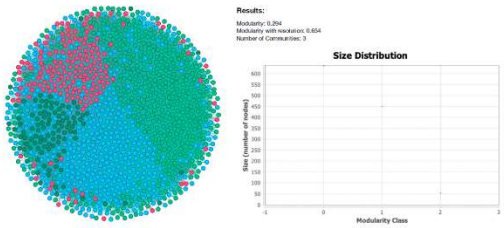


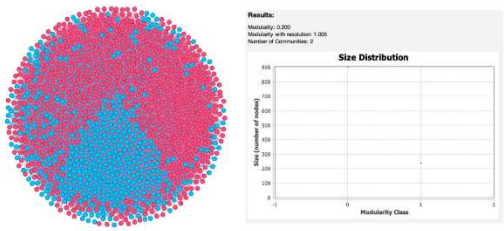






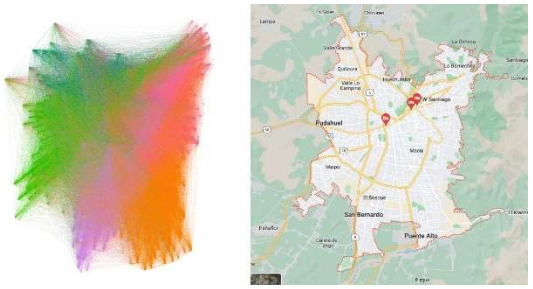




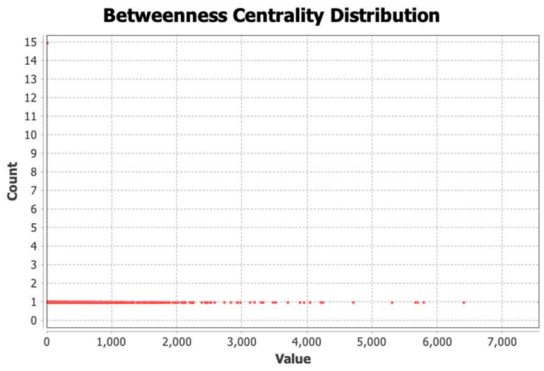


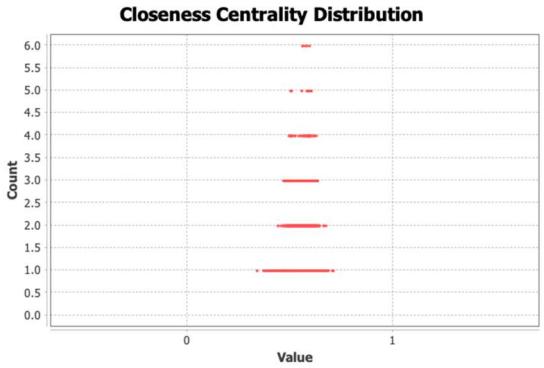
Resolution	Avg. Community Size (undirected & unweighted	Number of Communities (undirected & unweighted)	Std. Dev. of Community Size (undirected & unweighted)	Avg. Community Size (directed & weighted)	Number of Communities (directed & weighted)	Std. Dev. of Community Size (directed & weighted)
2.0	0.180	2	0.0	0.00	2	0.00
1.5	0.415	3	0.786	210.51	3	0.789
1.0	0.376	6	1.562	190.65	6	1.568
0.5	0.345	20	6.57	0.25	20	6.59
0.1	0.210	145	51.97	0.125	145	51.99

Two networks (undirected and unweighted, and directed and weighted) were selected to count the number of communities and to investigate the effects of each resolution value on the community structure. The findings demonstrate that, depending on the resolution value, a different number of communities are found by the Modularity algorithm. Different numbers of communities in the network were discovered using a range of resolution modularity values (2.0, 1.5, 1.0, 0.5, and 0.1). Particularly, just one community was found at a resolution modularity of 2.0, three at 1.5, six at 1.0, twenty at 0.5, and one hundred and eighty at 0.1. These results imply that the resolution modularity value utilised has a significant impact on the number of communities found in the network, and that a lower resolution modularity value may show a more fine-grained community structure within the network. The optimum resolution value out of these five options is 1.0 because it accurately represents the number of communities in Communities.csv. It is confirmed by mapping the nodes that correspond geographically to Santiago, Chile as given below.



Different centrality measures have been used to the city to determine the relative importance of each node (tower)Measures of degree centrality count the edges that are incident to a node. It is an easy method of gauging a node's significance inside a network. The number of times a node is on the shortest path between two other nodes is measured by betweenness centrality, on the other hand. It is a measure of how important a node is in terms of linking other nodes in the network.





Comparing the centrality of the nodes inside the communities reveals that they are not all equally central. The distribution is variable, with some communities having more central nodes than others. The findings demonstrate that nodes located further from the city's geographic centre are not necessarily less central. For instance, in the partition with a resolution of 0.5, the most central node in terms of Degree Centrality is situated outside the city, yet the most central node in terms of Betweenness Centrality is situated close to the city's centre. Thus, it follows that a node's centrality need not be correlated with its location.

The community with the highest Degree Centrality in the partition with a resolution of 0.5 has a tower that is connected to 10 other towers, while the community with the lowest Degree Centrality has a tower that is connected to just one other tower. Similar results were obtained for Betweenness Centrality, where the community with the highest centrality has a tower that is located on 31 shortest paths and the community with the lowest centrality has a tower that is located on just 2 shortest paths. In addition, we have examined the relationship between node location and the distribution of centrality. The findings indicate that nodes located further from the city's geographic centre are not necessarily less central. For instance, in the partition with a resolution of 0.5, the most central node in terms of Degree Centrality is situated outside the city, yet the most central node in terms of Betweenness Centrality is situated close to the city centre. The city of Santiago is regarded as having multiple poles.

Communities	A(W)	B(W)	C(W)	D(W)	E(W)	F(W)
A(H)	0.52766	0.08546	0.18332	0.14296	0.02903	0.03154
B(H)	0.09716	0.39497	0.24346	0.13546	0.06216	0.06676
C(H)	0.05623	0.07655	0.72120	0.06216	0.02222	0.06161
D(H)	0.09853	0.08930	0.16063	0.56240	0.05168	0.03742
E(H)	0.05509	0.08999	0.11781	0.11163	0.52808	0.09738
F(H)	0.047562	0.083472	0.198555	0.067063	0.085368	0.51798

We determined the PX(Y) probability distribution for a person who lives in community X and works in community Y. The number of residents of community X who also work in community Y must be counted, and the outcome must be calculated by dividing that number by the total population of community X. For all conceivable duos of communities, X and Y, shown in the table, we must calculate PX(Y). In terms of commuting to work, this will help us get a sense of how much engagement there is between various communities. The table clearly illustrates the increased probability that comes from travelling inside the same neighbourhood, which is always realistic and truthful. With the use of this study, we are able to contrast the observed commuting patterns in the city with those of a commuter population that travels at random between all places. The existence of segregation in the city is demonstrated if the observed values of PX(Y) are much greater than the expected values.

References:

1. Porta, S., Latora, V., & Strano, E. (2016). Network analysis of urban streets: A primal approach. Journal of Statistical Mechanics: Theory and Experiment, 2016(12), 123902.
2. Barthelemy, M. (2011). Spatial networks. Physics Reports, 499(1-3), 1-101.
3. Masucci, A. P., Stanilov, K., & Batty, M. (2013). Exploring the evolution of London's street network in the information space. EPJ Data Science, 2(1), 1-16.
4. Krings, G., Calabrese, F., Ratti, C., & Blondel, V. D. (2009). Urban gravity: A model for inter-city telecommunication flows. Journal of Statistical Mechanics: Theory and Experiment, 2009(07), L07003.
5. Newman, M. E. (2010). Networks: an introduction. Oxford university press.
6. Freeman, L. C. (1979). Centrality in social networks conceptual clarification. Social networks, 1(3), 215-239.
7. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), P10008.
8. Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. Physical Review E, 70(6), 066111.
9. Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3-5), 75-174.
10. Newman, M. E. (2006). Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23), 8577-8582.