**Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**

# Study of links between people in urban areas based on mobility data for the city of São Paulo

Matheus de Moraes Gonçalves Correia[1]
INPE, São José dos Campos, SP
Jéssica Domingues Lamosa[2]
UNIFESP, São José dos Campos, SP
Vander Luis de Souza Freitas[3]
Department of Computing - UFOP, Ouro Preto, MG
Lívia Rodrigues Tomás[4] Leonardo Bacelar Lima Santos[5]
CEMADEN, São José dos Campos, SP

**Abstract**

Our study explores the average degree and clustering of a complex mobility network designed to model and simulate the COVID-19 pandemic. To construct this network, we utilized mobility data collected in São Paulo, creating a network in which each node represents an individual, and each edge weight denotes the duration of contact between individuals during a typical day. By analyzing data from an Origin-Destination Research, we calculated the average degree and weighted clustering coefficient of the network for various minimum contact duration. We aimed to understand the effect of increasing minimum contact duration on network structure. Our findings indicate that networks with different minimum contact duration remained sparse, as the average degree of the generated graphs decreased.

**Keywords**. Complex Networks, COVID-19, Brazil, São Paulo

## 1 Introduction

Since the emergence of the SARS-CoV-2 virus in early 2020, it has rapidly spreaded worldwide, resulting in a devastating pandemic. In Brazil, the virus has infected more than 35 million people and caused over 689,000 deaths since the first case was recorded in São Paulo on February 25, 2020 [3].

Computational models have become increasingly crucial in dealing with the complex problems presented by the pandemic. In particular, mathematical models focused on epidemiology have gained traction as a valuable tool to forecast, and highlight the importance of actions to reduce the number of cases [1]. One promising approach in this area is the use of mobile network modeling based on actual data from the region under study [5].

A mobility network consists of locations connected by the flow of people [6], which is essential for understanding virus transmission on a large scale, particularly in a vast country like Brazil [4]. Such networks provide insight into complex systems by treating regions as nodes and movements between them as edges in a graph [9].

[1] matheusmgc@id.uff.br
[2] jd.lamosa@unifesp.br
[3] vander.freitas@ufop.edu.br
[4] liviatomas@gmail.com
[5] santoslbl@gmail.com

DOI: 10.5540/03.2023.010.01.0104

Against this backdrop, this work uses mobility data from São Paulo as the basis for our study. We build a contact network relating people to the time they spend in contact with others, using graph theory to analyze how this mobility network is structured. Specifically, we investigate how different values of a minimum contact duration threshold affect the construction of the network.

In Section 2, we describe the metrics we use and how we manipulate the mobility data. We present our results and analyses in Section 3, and conclude in Section 4. This study adds to the growing body of literature using computational models to understand the epidemiological dynamics of the COVID-19 pandemic.

# 2 Methodology

Graph theory has emerged as a powerful tool in the computational analysis of mobility data. In our study, we utilized this framework to construct a graph where each vertex represents an individual, and the edges represent the relationships between them. The weights of these edges vary depending on the duration of time that the individuals spent in contact with each other. This approach allows us to model the mobility patterns of individuals and provides a way to study the network structure of the resulting graph.

## 2.1 Graphs Theory

A network is formally defined as a graph $G(V, E)$, where $V$ represents a non-empty set of vertices (nodes), and $E$ represents a set of non-ordered pairs of vertices, which correspond to the edges (links) of $G$ [7]. The adjacency matrix, $A = a_{ij}$ for $i, j = 1, ..., N$, is used to represent the links between vertices in the graph. The value of $a_{ij}$ is 1 if there is an edge between vertices $i$ and $j$, and 0 otherwise, where $N$ is the total number of nodes in the graph. Edge weights are assigned to the edges in a matrix $W = w_{ij}$ for $i, j = 1, ..., N$, where $w_{ij}$ represents the weight of the edge connecting vertices $i$ and $j$. In the case of our study, the weight is the duration of contact between individuals. Notably, the adjacency and weight matrices are constructed with a zero value in their main diagonal, which disallows self-loops (i.e., $a_{11}, a_{22}, ..., a_{NN} = 0$ and $w_{11}, w_{22}, ..., w_{NN} = 0$).

### 2.1.1 Network Degree

In an undirected network, the degree of a node $k_i$ represents the number of links that node $i$ has with other nodes. As there are no differences between incoming or outgoing edges in an undirected network, $k_i$ can be interpreted as the number of neighbors that node $i$ has. The degree $k_i$ for node $i$ can be calculated as the sum of the elements in the $i$-th row (or column, since the matrix is symmetric) of the adjacency matrix $A$, that is:

$$k_i = \sum_{j}^{N} a_{ij}, \tag{1}$$

where $N$ is the total number of nodes in the network and $a_{ij}$ is the element in the $i$-th row and $j$-th column of $A$ [2].

To calculate the average degree of the graph, denoted by $< k >$, we can use the formula:

$$< k >= \frac{1}{N} \sum_{i=1}^{N} k_i.$$

DOI: 10.5540/03.2023.010.01.0104

which gives an idea of the connectivity of the network as a whole. A higher value of $< k >$ indicates a denser network with more connections, while a lower value indicates a sparser network with fewer connections.

### 2.1.2 Clustering Coefficient

The clustering coefficient, denoted by $C$, is a metric commonly used in network analysis to measure the tendency of nodes in a network to form clusters or tightly interconnected groups. It provides insight into the local structure of a network by quantifying the extent to which a node's neighbors are interconnected. Specifically, for a node $i$ with degree $k_i$, the clustering coefficient is defined as the ratio of the number of links between neighbors of node $i$ to the total number of possible links between them [10].

It should be noted that for nodes with degree $k_i = 0$ or $k_i = 1$, the clustering coefficient is undefined since there are no neighbors to form connections between. Thus, by convention, $C_i = 0$ for such nodes [2]. For this study, a weighted clustering coefficient, $C^w$, was calculated, which takes into account the weights of the edges connected to each vertex:

$$C_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2} a_{ij} a_{ih} a_{jh}, \tag{2}$$

in which $s_i$ is obtained by summing the weights of adjacent edges for each vertex.

To obtain the average clustering coefficient of a graph, denoted as $< C >$, one needs to divide the sum of the weighted clustering coefficients $C_i^w$ for all nodes $i$ in the network by the total number of nodes $N$:

$$< C >= \frac{1}{N} \sum_i C_i^w. \tag{3}$$

## 2.2 Data

The contact network between individuals was constructed using real-world data from the city of São Paulo, Brazil. The data was obtained from the most recent Origin-Destination Survey of the city, which was completed in 2017 and is available in an open format on the Transparency Portal of the State of São Paulo Metro website [8].

The dataset covers a 24-hour period, starting at 00:00h and ending at 23:59h of the same day. Each individual in the dataset is identified by an identity number (ID PESS), and the dataset records the total number of trips made by each individual during that period (TOT VIAG).

The Origin-Destination (OD) zones are defined based on their urban homogeneity and socioeconomic characteristics. Each zone represents the smallest geographic unit which ensures the statistical representativeness of the data.

Thus, every trip made by an individual is numbered (N VIAG) and includes information on the origin zone (ZONE O) and destination zone (ZONE D) of the trip. The dataset also provides the exact departure time from the origin zone (H SAIDA and MIN SAIDA) and the arrival time at the destination zone (H CHEG and MIN CHEG).

DOI: 10.5540/03.2023.010.01.0104

Table 1: Profile of the first 5 lines of the dataset of the city of São Paulo.

| ID<br>PESS | N<br>VIAG | TOT<br>VIAG | ZONA<br>O | ZONA<br>D | H<br>SAIDA | MIN<br>SAIDA | H<br>CHEG | MIN<br>CHEG |
|---|---|---|---|---|---|---|---|---|
| 10001101 | 1.0 | 2 | 1.0 | 3.0 | 5.0 | 45.0 | 5.0 | 55.0 |
| 10001101 | 2.0 | 2 | 3.0 | 1.0 | 15.0 | 45.0 | 15.0 | 55.0 |
| 10001102 | 1.0 | 3 | 1.0 | 82.0 | 9.0 | 0.0 | 9.0 | 50.0 |
| 10001102 | 2.0 | 3 | 82.0 | 84.0 | 17.0 | 0.0 | 18.0 | 0.0 |
| 10001102 | 3.0 | 3 | 84.0 | 1.0 | 22.0 | 50.0 | 23.0 | 30.0 |

### 2.2.1 Network Construction

To construct the network, we connected individuals who were present in the same zones during the same time periods. The number of individuals considered in this analysis is equal to the number of nodes, denoted as $N$, in the resulting graph. To achieve this, we divided the day into one-minute periods, and our code, implemented in Python, efficiently identified the zones each individual was in at each minute, disregarding travel time between zones. This information was then used to construct the weight matrix $W$, which represents the contact duration between individuals.

The weight matrix $W$ was employed to create the graph, and its analysis is presented in Section 3. See Table 1 for an example of the dataset used to construct the weight matrix.

Despite the large number of nodes and edges in the graph, we optimized the Python code to such an extent that it enabled us to calculate the relevant metrics within approximately 12 minutes.

# 3 Results

The weights of $W$ represent the duration of time two individuals maintained contact with each other if both were in the same zone. To determine whether an edge will connect two different people in the network, a threshold called the minimum contact duration ($mcd$) is defined. For example, if $mcd = 1$, it is enough that both individuals are present in a zone for 1 minute for an edge to be formed between them.

To further clarify the thresholds used in the present metrics, we started with a base threshold of 1 minute, $mcd = 1$, and varied it with in increments of 60, which corresponds to 1 hour in a day, until we reached $mcd = 1440$, which corresponds to 24 hours. For instance, if $mcd = 60$, two individuals needed to be in the same zone for at least 1 hour for an edge to be formed between them. These different thresholds allowed us to analyze the network structure and behavior for different levels of contact duration.

## 3.1 Average Degree

The degree distribution of the graphs generated by the different values of $mcd$ provides a visualization of how the connections between individuals behave. As seen in Figure 1, increasing the value of $mcd$ leads to a decay in the average degree. This is expected since a higher minimum contact duration means that fewer edges will be present in the network.

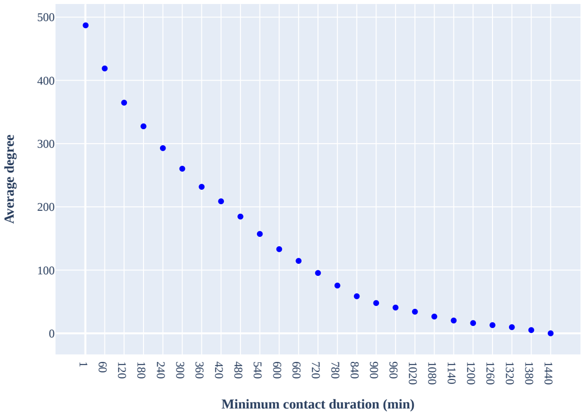

Figure 1: Average degree for a series of threshold values ranging from $mcd = 1$ to $mcd = 1440$, with increments of 60. Each point represents the average degree $< k >$ of a graph generated with the corresponding threshold value.

### 3.2 Average Clustering Coefficient

Figure 2 illustrates the mean clustering coefficient $< C >$ for varying $mcd$. It is observed that as the threshold $mcd$ increases, the mean clustering coefficient $< C >$ tends to increase to a certain extent, indicating that the network has a tendency to form clusters or groups.

This trend reaches its peak at $mcd = 660$, where the average clustering coefficient of the graph reaches its highest value. However, after this point, the mean clustering coefficient starts to decrease and eventually reaches a value of zero at $mcd = 1440$, which corresponds to 24 hours.

Furthermore, the interval between $mcd = 600$ and $mcd = 720$ indicates that the network contains many clusters or tightly interconnected groups. This range could be of particular interest for future research, as it provides a basis for more detailed investigation into the nature and characteristics of these clusters.

## 4 Conclusions

Working with mobility networks can be challenging due to their complexity and the high computational costs associated with generating and analyzing results. Moreover, obtaining and manipulating data for these networks can pose a significant obstacle to replicating the study in other regions.

Our research aims to gain a deeper understanding of the structure of complex networks that represent the relationships between people based on their contact over time. To achieve this, we used metrics such as average degrees and clustering coefficient, and examined graphs with varying minimum contact duration.

As we increased the threshold $mcd$, we observed a decrease in the average degree of the network. However, analyzing the behavior of the average clustering coefficient is crucial to understanding the variance and peak observed in the graph, taking into account the weighted calculation thereof. Further investigation within this interval can shed light on the factors that influence this variation.

The peak observed in the graph may indicate the presence of tightly interconnected groups or clusters in the network. Therefore, we plan to conduct more extensive research within this interval

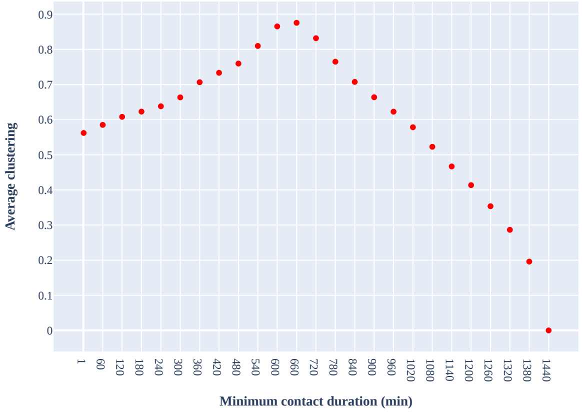

Figure 2: Average clustering coefficient for a series of threshold values ranging from $mcd = 1$ to $mcd = 1440$, with increments of 60. Each point represents the average clustering coefficient $< C >$ of a graph generated with the corresponding threshold value.

to better understand the structure and characteristics of these clusters, and how they relate to the behavior of the average clustering coefficient.

## Acknowledgements

This research was supported by scholarship 88887.668467/2022-00 from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and grant FAPESP 15/50122-0

## References

[1] Linda JS Allen et al. **Mathematical epidemiology**. Vol. 1945. Springer, 2008. DOI: `10.1007/978-3-540-78911-6`.

[2] Albert-László Barabási and Márton Pósfai. **Network science**. Cambridge: Cambridge University Press, 2016. ISBN: 9781107076266 1107076269.

[3] Wesley Cota et al. “Monitoring the number of COVID-19 cases and deaths in Brazil at municipal and federative units level”. In: (2020).

[4] Vander LS Freitas, Gladston JP Moreira, and Leonardo BL Santos. “Robustness analysis in an inter-cities mobility network: modeling municipal, state and federal initiatives as failures and attacks toward SARS-CoV-2 containment”. In: **PeerJ** 8 (2020), e10287. ISSN: 2167-8359. DOI: `10.7717/peerj.10287`.

[5] Vander Luis de Souza Freitas et al. “The correspondence between the structure of the terrestrial mobility network and the spreading of COVID-19 in Brazil”. In: **Cadernos de Saúde Pública** 36 (2020), e00184820. ISSN: 1678-4464. DOI: `10.1590/0102-311X00184820`.

[6] J. D. Lamosa et al. “Topological indexes and community structure for urban mobility networks: Variations in a business day”. In: **PLoS ONE** 16.3 (3 March 2021), e0248126. ISSN: 19326203. DOI: `10.1371/journal.pone.0248126`.

[7] P. O. B. Netto and S. Jurkiewicz. **Grafos: introdução e prática**. Vol. 2. Blucher, 2017. ISBN: 9788521211334.

[8] Relatório Sintese (Pesquisa Origem-Destino). https://transparencia.metrosp.com.br/dataset/pesquisa-origem-e-destino/resource/b3d93105-f91e-43c6-b4c0-8d9c617a27fc. Online; accessed 04-December-2022. 2017.

[9] Eduardo R Pinto, Erivelton G Nepomuceno, and Andriana SLO Campanharo. "Impact of network topology on the spread of infectious diseases". In: **TEMA (São Carlos)** 21 (2020), pp. 95–115. DOI: 10.5540/tema.2020.021.01.0095.

[10] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world'networks". In: **nature** 393.6684 (1998), pp. 440–442. DOI: 10.1038/30918.