Universidade Federal do ABC

Disciplina: BCM-0506 – Comunicação e Redes – 2021.1

Graph Modeling for Analysis of Traffic Load Distribution in Subway Lines

Felipe Nunes Melo, 11202130262
Henry Miyawaki, 11202131460
Igor Henrique da Costa Coelho, 11202022201

02/04/2023

1. Abstract

In the study, an analysis of the São Paulo metro system was conducted with the aim of identifying patterns and making projections. To do so, the stations that have the highest integration and connection with the network were explored. The analysis was based on graphs, which allowed the structure of the system and the interconnections between the stations to be visualized. This made it possible to identify the importance of some stations as transfer points and the relationship between the number of passengers and train flow. Central measures and trends were also used in the current and projected graph to better understand its behavior and that of the most important stations. Additionally, the study generated projections of how the graph will look with the future additions of lines.

2. Methods

When working with graphs, the adjacency matrix is a crucial tool for representing the graph mathematically. Therefore, to create our adjacency matrices, we needed to create three main datasets to set the connections of edges, vertices, considering their weights.

2.1 Data Gathering

Storing the data as JSON format, was created the first dataset, which is the station dataset. This dataset includes the actual lines and stations in the subway and train system of the state of São Paulo in 2022. The second dataset is the projection dataset, which contains planned projects to extend the existing rail transit system. The third dataset is the average number of passengers per business day.

For this study, the data was obtained from three main sources: the consortia contracted for the construction projects, the official websites of subway and train agencies. The first dataset, the station dataset, has two main keys that are not fixed and can be easily accessed and altered using a dictionary in the code. The first key is named "Linha" (Line). It represents the name of the line and includes all the stations associated with that line.

The second key is an array that contains objects representing the stations. Each object contains variables that provide information about the station and its connections to other stations, including integration points. These variables help identify the station and its properties within the transportation system.

One important rule of the station dataset is that the order of the elements in the array matters. If "Station 1" comes before "Station 2" in the array, the dataset indicates that "Station 1" comes before "Station 2" in the sequence of stations along the line. Therefore, the order of the stations in the array affects the formation of connections in the adjacency matrix and the overall representation of the graph.

Inside the object that represents a station in the array, there are typically three keys that provide important information about the station: the station name, its address, and what connections it has with other lines. An example of the used json:


```
[
 {
  "Linha": "Linha 1-Azul",
  "Estacoes": [
   {
    "Estação": "Estação Tucuruvi",
    "Endereço": "Av. Dr. Antonio Maria Laet, 100 – Tucuruvi",
    "Integração": ""
   },
   {
    "Estação": "Estação Parada Inglesa",
    "Endereço": "Av. Luiz Dumont Villares, 1721 – Parada Inglesa",
    "Integração": ""
   },
   {
    "Estação": "Estação Jardim São Paulo – Ayrton Senna",
    "Endereço": "Av. Leôncio de Magalhães, 1000 – Jardim São Paulo",
    "Integração": ""
   },
   {
    "Estação": "Estação Santana",
    "Endereço": "Avenida Cruzeiro do Sul, 3173",
    "Integração": ""
   },
   {
    "Estação": "Estação Carandiru",
    "Endereço": "Av. Cruzeiro do Sul, 2487 – Santana",
```

"Integração": ""
},
{
"Estação": "Estação Portuguesa – Tietê",
"Endereço": "Av. Cruzeiro do Sul, 1777 – Santana",
"Integração": ""
},
{
"Estação": "Estação Armênia",
"Endereço": "Rua Pedro Vicente, 47 – Bom Retiro",
"Integração": ""
},
{
"Estação": "Estação Tiradentes",
"Endereço": "Av. Tiradentes, 551 – Bom Retiro",
"Integração": ""
},
{
"Estação": "Estação Luz",
"Endereço": "Av. Prestes Maia, 925 – Luz",
"Integração": "Linha 4-Amarela,Linha 7-Rubi,Linha 11-Coral"
},
{
"Estação": "Estação São Bento",
"Endereço": "Largo São Bento,109",
"Integração": ""
},
{
"Estação": "Estação Sé",
"Endereço": "Praça da Sé , s/nº",
"Integração": "Linha 3- Vermelha"
},
{
"Estação": "Estação Liberdade",
"Endereço": "Praça da Liberdade, 133 – Liberdade",
"Integração": ""
},
{
"Estação": "Estação São Joaquim",
"Endereço": "Av. Liberdade, 1033, Liberdade",
"Integração": ""
},
{
"Estação": "Estação Vergueiro",

"Endereço": "Rua Vergueiro, 790",
"Integração": ""
},
{
"Estação": "Estação Paraíso",
"Endereço": "Rua Vergueiro, 1465 – Paraíso",
"Integração": "Linha 2-Verde"
},
{
"Estação": "Estação Ana Rosa",
"Endereço": "Rua Vergueiro, 505 – Vila Mariana",
"Integração": "Linha 2-Verde"
},
{
"Estação": "Estação Vila Mariana",
"Endereço": "Av. Profº. Noé Azevedo, 255",
"Integração": ""
},
{
"Estação": "Estação Santa Cruz",
"Endereço": "Rua Domingos de Morais, 2564, Vila Mariana",
"Integração": "Linha 5-Lilás"
},
{
"Estação": "Estação Praça da Árvore",
"Endereço": "Praça da Árvore, 39 – Praça da Árvore",
"Integração": ""
},
{
"Estação": "Estação Saúde",
"Endereço": "Avenida Jabaquara, 1634",
"Integração": ""
},
{
"Estação": "Estação São Judas",
"Endereço": "Avenida Jabaquara, 2438",
"Integração": ""
},
{
"Estação": "Estação Conceição",
"Endereço": "Avenida Engº. Armando de Arruda Pereira, 919",
"Integração": ""
},
{

```
        "Estação": "Estação Jabaquara",
        "Endereço": "Rua dos Jequitibás, 80 – Jabaquara",
        "Integração": ""
      }
    ]
  }
]
```


The projection dataset is comparable to the station dataset, as it consists of an array of lines and their corresponding stations stored within an object, allowing for the addition of new information for new stations and extensions containing the station names. This can be observed in the accompanying json format:


```
[
  {
    "Linha": "Linha Celeste",
    "Estacoes": [
      {
        "Estação": "Estação Vila Augusta"
      },
      {
        "Estação": "Estação São Bento"
      },
      {
        "Estação": "Estacionamento Belém"
      },
      {
        "Estação": "Estação Silva Teles"
      },
      {
        "Estação": "Estação Anhangabaú"
      },
      {
        "Estação": "Pátio Vila Medeiros"
      },
      {
        "Estação": "Estação Dutra"
      },
      {
        "Estação": "Estação Bosque Maia"
      },
      {
        "Estação": "Estação Pari"
      },
```

```
      {
        "Estação": "Estação Vila Maria"
      },
      {
        "Estação": "Estação Catumbi"
      },
      {
        "Estação": "Estação Jardim Japão"
      }
    ]
  }
]
```

The passenger boarding data is organized by daily average within a separate json file, which includes an object containing all stations and their corresponding daily average passenger counts (in thousands) for the Sao Paulo subway system. This data is organized by passengers per utility days of 2022 for each station.

To elaborate, the available data for the CPTM system only covers the months of January, February, April, May, September, and October. No data is available for the other months of the year. Therefore, to obtain daily averages for the entire year, we divided the monthly figures by 30 to obtain an approximate daily count. We then calculated the annual average, which is presented in the image below:

Image 1 - Table with passenger data entering by station for the available months, in the penultimate column the daily average for the year 2022, and finally the last column only has the number without the million to follow the pattern of the data provided by the subway official website.

Nome da estação/ linha	janeiro	fevereiro	abril	maio	setembro	outubro	média diária do ano de 2022	remove milha
linha 7 - Rubi								
Bras	234.185	281.033	297.213	362.268	327.718	337.126	10416	10
Luz	1.052.711	1.023.603	999.385	1.210.071	1.249.058	1.148.507	36687	37
Barra Funda E	1.090.714	2.762.470	1.112.088	1.925.505	3.656.580	3.373.727	78133	78
Agua Branca	151.885	149.542	141.980	196.984	203.392	186.033	5632	6
Lapa	611.154	573.970	640.580	663.986	681.664	670.902	21743	22
Piqueri	172.039	155.929	187.507	192.109	200.453	203.960	6327	6
Pirituba	334.778	316.480	335.925	395.006	413.402	397.814	12182	12
Vila Clarice	75.601	72.009	76.278	87.354	91.532	85.735	2700	3
Jaragua	365.090	334.421	369.365	409.178	423.383	401.086	12841	13
Vila Aurora	206.245	194.270	209.958	234.015	240.436	234.098	7400	7
Perus	487.115	462.935	492.328	544.786	558.920	536.831	17153	17
Caieiras	262.373	247.244	261.463	301.826	300.838	287.367	9162	9
Franco da Rocha	467.140	428.127	476.564	521.161	528.426	508.889	16424	16
Baltazar Fidelis	202.952	188.595	202.646	223.085	228.376	220.082	7051	7
Francisco Morato	750.748	687.665	752.682	808.968	813.455	783.007	25595	26
Botujuru	44.083	42.031	44.972	50.076	49.568	48.012	1550	2
Campo Limpo Paulista	85.327	81.054	88.368	107.528	103.098	100.519	3148	3
Varzea Paulista	52.420	47.038	54.400	57.709	58.163	57.618	1867	2
Jundiai	249.552	231.637	257.565	285.832	297.131	298.166	9057	9

For ViaMobilidade, we used a different approach as their data is organized by passengers per utility days monthly. We calculated the average for the entire year of 2022. It is worth noting that lines 8 and 9 were owned by CPTM, so their January data was in the CPTM pattern.

However, this had a negligible impact on the calculated average. Furthermore, for ViaQuatro, which includes only Line 4, we calculated an average for the entire year of 2022, as we had data for all months. Once we normalized all the data, we created a JSON file in the following pattern:


```
[
 {
  "Linha": "Linha 1-Azul",
  "Estacoes": [
   {
    "Estação": "Estação Jabaquara",
    "Passageiros": {
     "Média Diária": 64
    }
   }
  ]
 },
 {
  "Linha": "Linha 5-Lilás",
  "Estacoes": [
   {
    "Estacao": "Estação Capão Redondo",
    "Passageiros": {
     "Média Diária": 84
    }
   }
  ]
 }
]
```

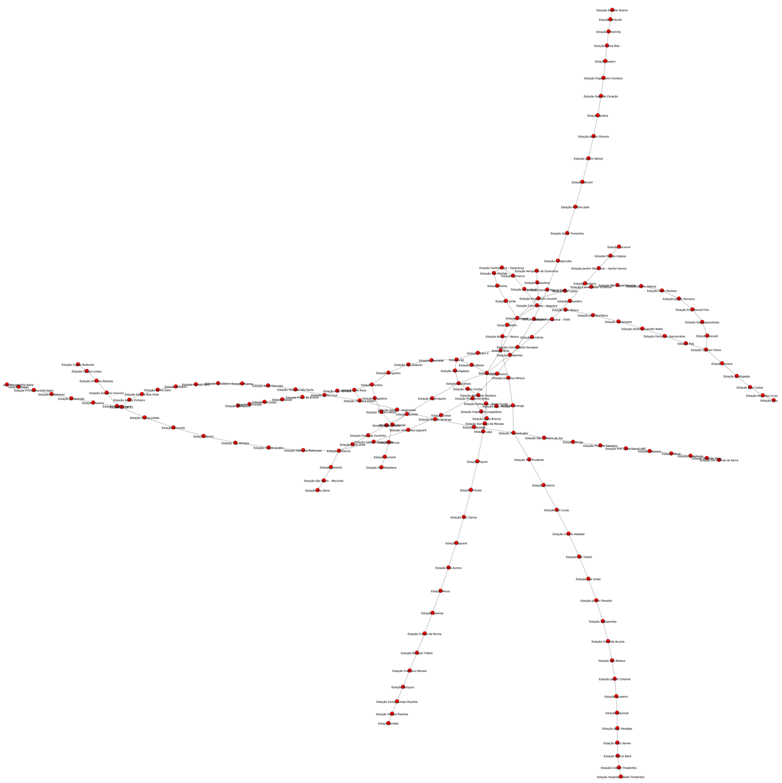

2.3 Data Process

After obtaining the data, an adjacency matrix was created in order to apply plotting methods with Igraph. This allows for analysis to be performed on the constructed networks. To accomplish this, a Python script was created that connects the nodes to their predecessors and successors within each object of the main array. Through this script, we can not only analyze the São Paulo subway network, but also any other dataset that has the same relationship.

Params as weights can be applied or not, depending on the dataset. In this study, daily averages were used as weights, but other data can also be used. These weights are processed and applied by another script at the time of graph plotting. Finally, scripts were created for each plot, with their parameters centralized in a file called "config".

3. Results

Image 3 - Graph created of the subway and train stations of São Paulo



The graph analyzed in this study is composed of 173 vertices, each representing a metro station in São Paulo, Brazil, while the edges represent the connections between them, including line integrations. Through the use of Python scripts and the Igraph module, data was collected to determine some important properties of the graph. The diameter of the graph was calculated as 38, the average degree was 9.68, and the average path length was 14.21. Additionally, the daily average of passengers was used to identify the five stations with the highest flow in the network and the five stations with the highest harmonic centrality.

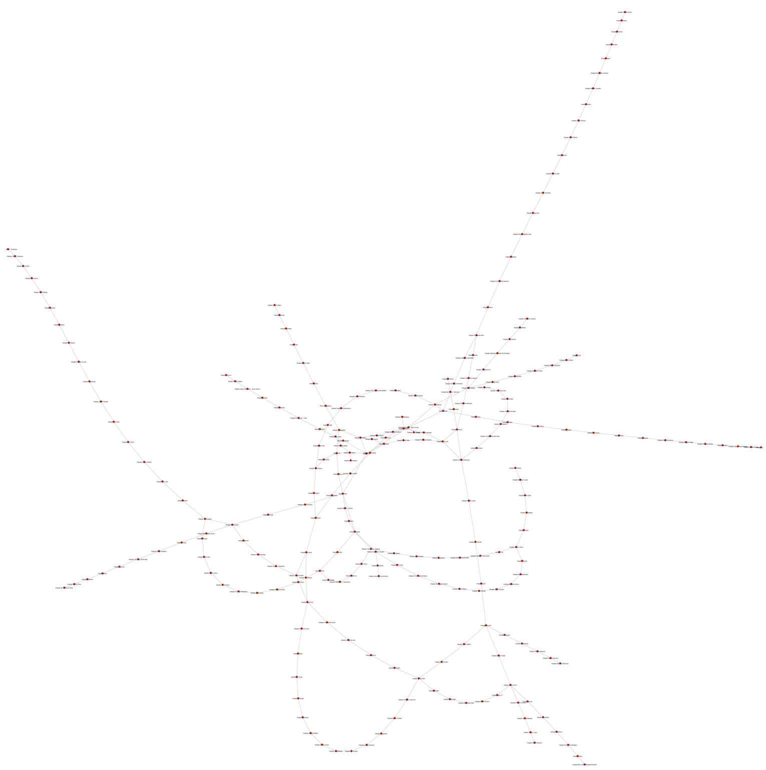
Table 1 - The highest passenger flow stations

Station Name	Daily Average
Estação Sé	337.000
Estação Luz	321.000
Estação República	198.000
Estação Belém	173.000
Estação Várzea Paulista	144.000

Table 2 - The highest harmonic centrality stations

Station Name	Centrality
Estação República	0.170
Estação Bresser – Mooca	0.150
Estação Santa Cecília	0.145
Estação Belém	0.142
Estação Chácara Klabin	0.132

Image 4 - Graph created of the projection subway and train stations of São Paulo



The projection graph of the subway and train lines in São Paulo exhibits two line extensions - the Green and Silver lines - and features four new lines: Brown, Sky Blue, Gold, and Orange. This new graph reveals a diameter value of 44, a mean degree of 2.15, and an average path length of 14.34.

In the current state of the analysis, twelve clusters have been identified in the first graph representing the current stations. These clusters represent distinct patterns of behavior that can be observed in the data. However, in the projection graph, a total of fifteen clusters have been identified. This suggests that there may be additional patterns or trends that are not immediately apparent in the current stations but can be observed in the projected data.

Image 5 - Cluster formed on the current train and metro network

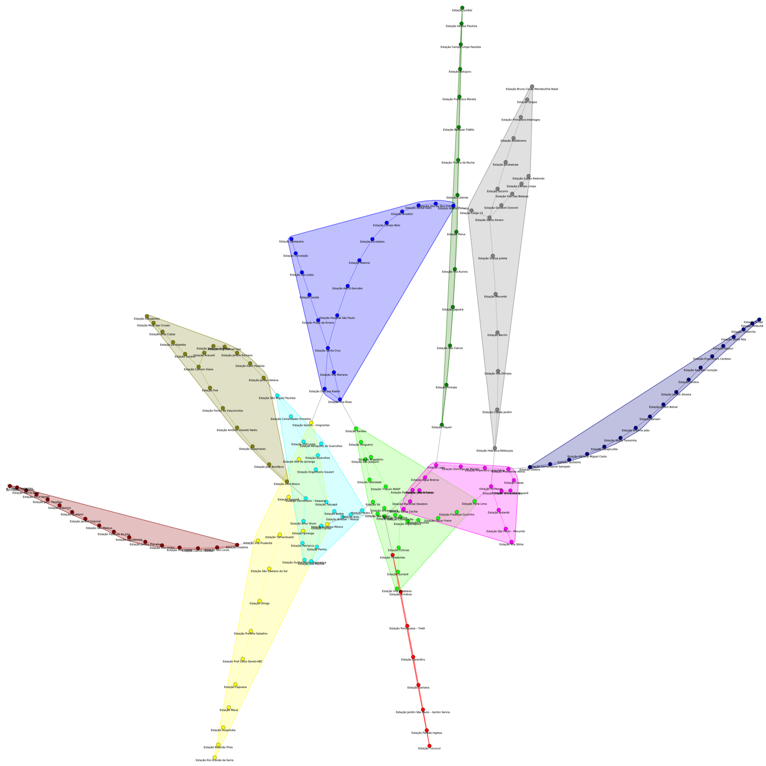


Image 6 - Cluster formed on the projection train and metro network

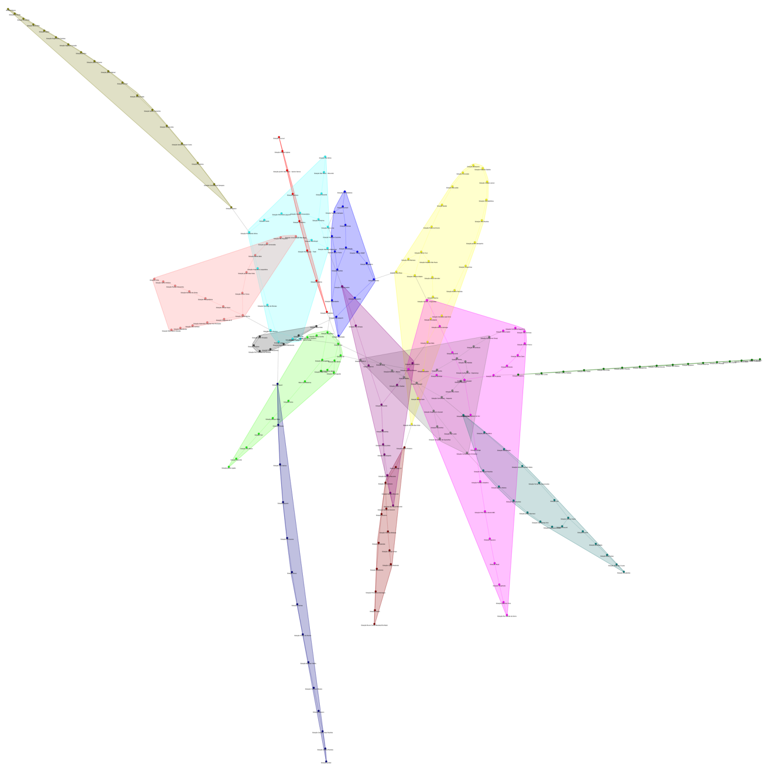


Table 3 - The highest number of integrations stations

Station Name	Connections
Estação Luz	5

Estação Brás	4
Estação Tatuapé	3
Estação Palmeiras Barra funda	3
Estação Sé	2

3.1 Discussion

The projection graph of a metro and train station can provide valuable insights into the network's structure and efficiency. In this study, we analyzed the changes in three variables of the projection graph: diameter, average path, and degree, to understand how the network has evolved by adding the new lines and stations.

Diameter: Our analysis revealed that the diameter of the projection graph has increased significantly compared to the current graph of the subway and train system. This indicates that the network has expanded and become more complex. However, the increase in diameter does not necessarily imply that the graph has grown in size. Instead, it suggests that there are more indirect connections between stations that were not present before. These indirect connections can lead to more efficient routes and better accessibility for passengers.

Average path: The projection graph's average path has also grown, though less dramatically than the diameter. The network's expansion with new stations is responsible for this rise. Longer average paths result from the fact that as stations are added, the average distance between them grows. The analysis also showed that the average path length increase is not as significant as anticipated. This is due to the integration of the new stations with the new lines, which has increased connectivity and decreased the distance between stations. Because of the new line integration, the average path has grown, but not as significantly as it might have.

Degree: The degree of a station in a projection graph represents the number of connections it has with other stations. According to our analysis, some stations' degrees have increased while others have stayed largely unchanged. This is explained by the network's expansion, which raises the number of potential connections between stations. However, due to their position or the lack of new lines in the area, several stations' degree has not greatly improved.

In addition to the changes in diameter, average path, and degree, our analysis also revealed that the number of clusters in the projection graph increased from 12 to 15 with the addition of new lines. Clusters are groups of vertices that are densely connected to each other, but not as densely connected to vertices in other clusters. The increase in the number of clusters suggests that the new lines have generated new integrations and groups of vertices in the network, leading to better connectivity and more efficient routes.

Additionally, the rise in clusters shows that the network has grown more structured and organized, which can improve the network's overall functionality. The network is now more accessible and flexible thanks to the new lines, which have increased the choices for passengers to move between clusters.

To better understand the influence of the projection graph adjustments on the network, we assessed the average daily passing through each station and identified the stations with the most links to other stations. It was noticed that some stations in the network were more connected and had substantially higher average daily passages than other stations, such as the central station.

These highly connected stations play a critical role in the network's functionality and are often referred to as "hubs". Hubs are essential for maintaining the network's integrity, and their shutdown can have a significant impact on the network's efficiency. Our analysis identified several stations as hubs, such as the central station, which had the highest average daily passage and the most connections with other stations.

Therefore, it is important to consider the potential impact of closing or disrupting these hub stations when planning for emergencies or making changes to the network. Our findings suggest that shutting down these hubs could lead to significant disruptions and delays in the network, which could have negative impacts on passengers and the city's transportation system as a whole.

From this analysis, we found that the top five hubs are respectively Luz, Brás, Sé, Tatuapé, and Palmeiras-Barra Funda. With this information, diagrams were created to visually demonstrate which lines were disconnected in the event of a failure at each of these stations, and through which stations they can be reconnected.

In the diagrams below, the following pattern was used: the square boxes represent the stations that connect the lines, the circles represent the station lines to which the station belongs and connects, the lines symbolize the connections, and the dashed lines marked with an "x" are the connections broken when removing the central station from the analysis, i.e., the hub.

image 7 -Diagram showing stations for reconnecting train lines after removing the station "Luz".

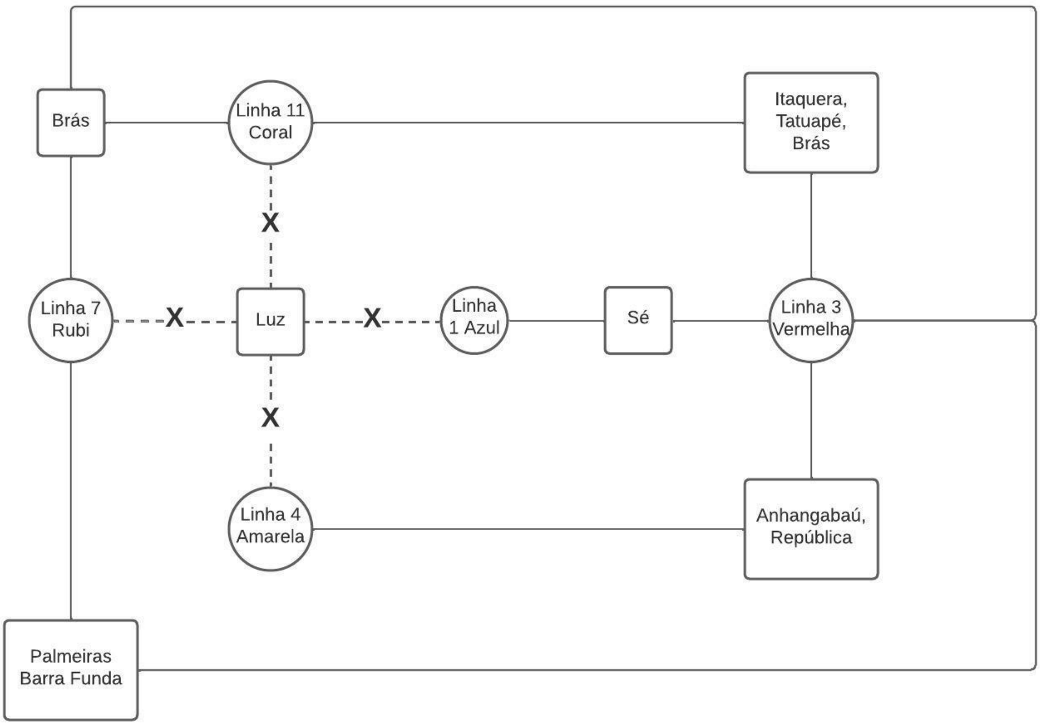


image 8 - Diagram showing stations for reconnecting train lines after removing the station "Brás".

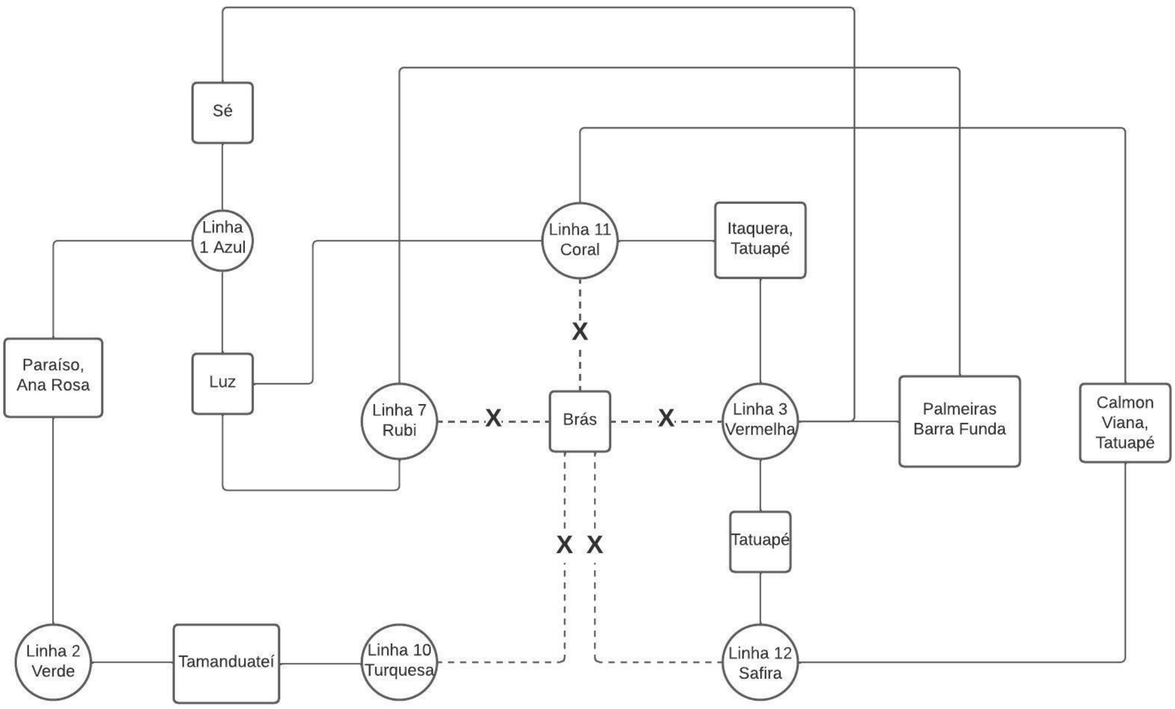


image 9 - Diagram showing stations for reconnecting train lines after removing the station "Sé".

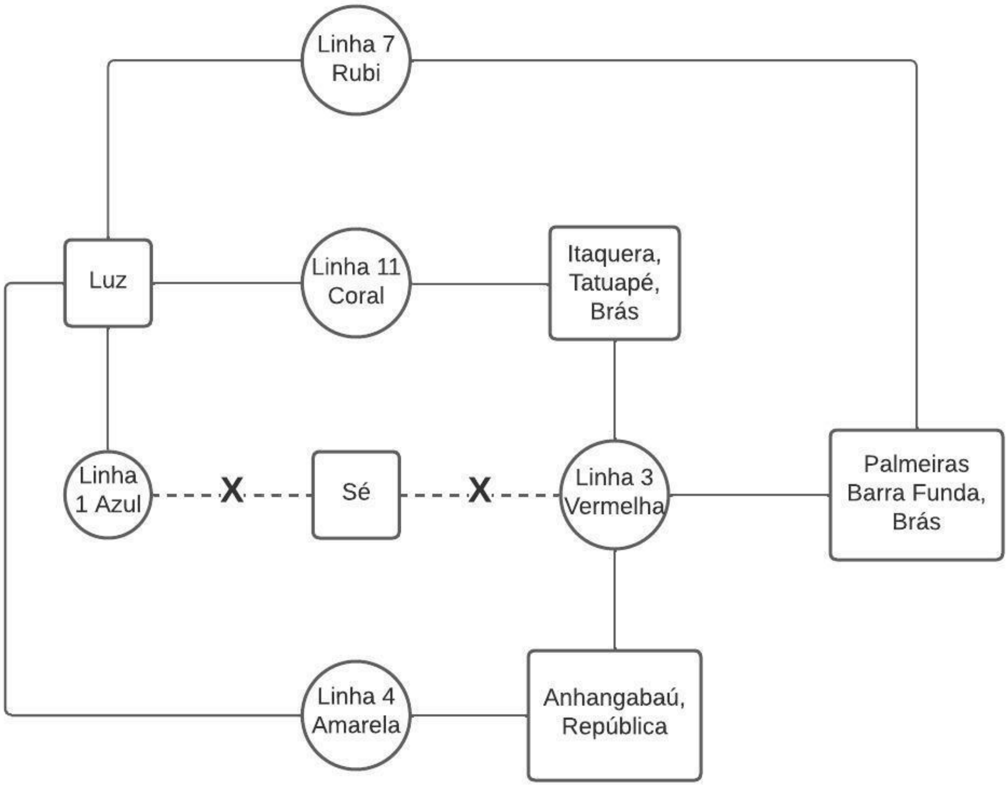


image 10 - Diagram showing stations for reconnecting train lines after removing the station "Tatuapé".

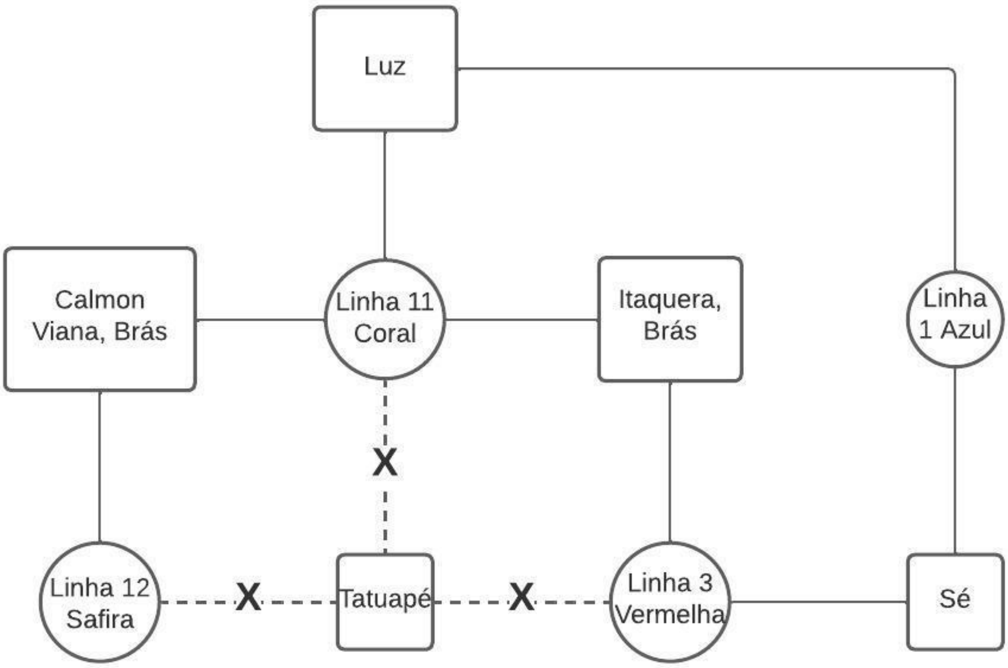
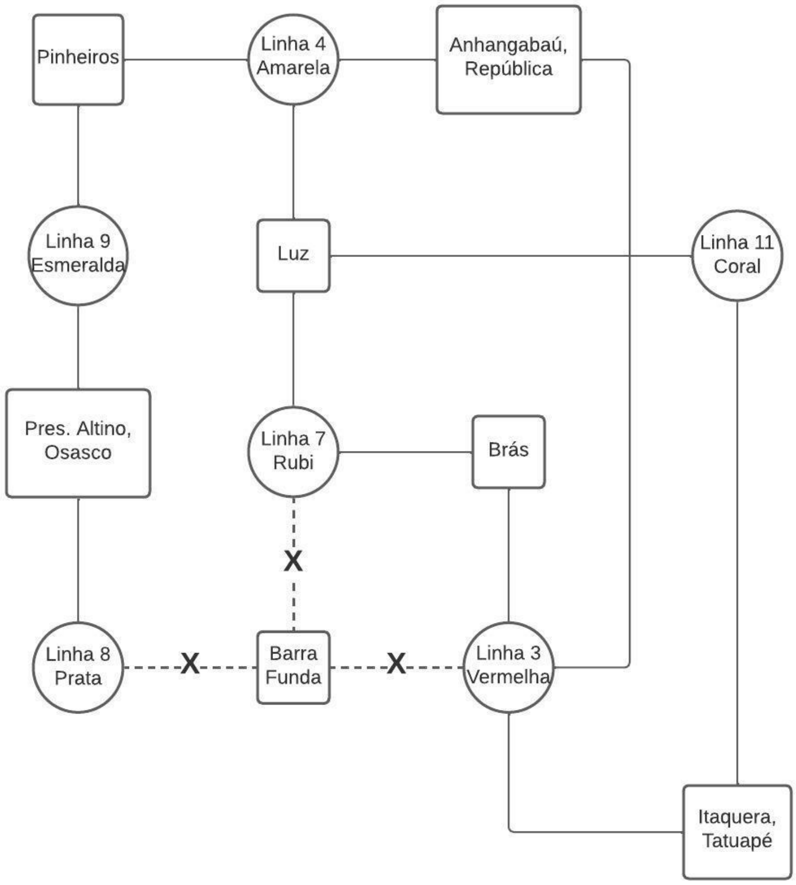


image 11 - Diagram showing stations for reconnecting train lines after removing the station "Palmeiras-Barra Funda" which is abbreviated as “Barra Funda”.



Upon examining the diagrams, it becomes evident that in the event of one of the primary hubs being removed, the other hubs consistently present themselves as viable options for re-establishing the connections. This is indicative of the exceptional connectivity rates displayed by these vertices. Consequently, it can be inferred that the elimination of any of these hubs may result in overburdening the flow to other vertices that have a significant passenger rate at the station.

4. Conclusion

This study analyzed the subway and train stations of São Paulo, Brazil, through a graph theory approach. The results showed that the diameter and average path length of the projection graph have increased compared to the current subway and train system, indicating an expansion in the network's indirect connections. The degree of a station in the projection graph represents the number of connections it has with other stations, and our analysis revealed that the degree of some stations has increased, while others have remained relatively constant. Furthermore, the number of clusters in the projection graph increased with the addition of new lines, indicating that the network has become more organized and structured, leading to better connectivity and more efficient routes. The study also identified the hubs in the network, which are essential for maintaining the network's integrity, and their shutdown can have a significant impact on the network's efficiency.

Our methodology has resulted in scripts that can be used to analyze other transportation networks that exhibit similar behavior. Therefore, this study provides valuable insights for urban planners and policymakers in making informed decisions about transportation infrastructure. The findings of this study can be used to improve the functionality of the transportation system in the city and ensure better connectivity and accessibility for commuters.

References

Portal da transparência do metrô de São Paulo (Transparency portal of the São Paulo subway). (2022) Report on expansion of construction and modernization. https://transparencia.metrosp.com.br/dataset/relatório-de-expansão-obras-e-modernização

Viaquatro. (2022) Database of transported passengers. https://www.viaquatro.com.br/linha-4-amarela/passageiros-transportados

ViaMobilidade. (2022) Database of transported passengers. https://www.viamobilidade.com.br/nos/passageiros-transportados

CPTM (São Paulo Metropolitan Trains Company). (2022) Database of transported passengers. https://www.cptm.sp.gov.br/negocios/Pages/Movimentacao-de-Passageiros.aspx?RootFolder=%2Fnegocios%2FMovimentao%20de%20Passageiros%2F2022&FolderCTID=0x0120001B071DE8D9072049B0F9B0E2E77B902E&View=%7B23EF9C34-E382-4D73-84B5-A63047037BE4%7D

MELLO, Celso F. de; RIBEIRO, João P. Topological analysis of public transportation networks in São Paulo using graphs. Caderno de Informática, v. 9, n. 2, p. 52-62 (2018). https://github.com/IgorRozani/Public-Transport-SP-Graph-Database/blob/master/database.cypher