COMPLEX NETWORK ANALYSIS: MEXICO’S CITY METRO SYSTEM

Olivia Sashiko Shirai Reyna(a), Idalia Flores de la Mota(b)

(a), (b) Posgrado en Ingeniería de Sistemas- Facultad de Ingeniería Edificio Bernardo Quintana 3er, Piso, Departamento de Sistemas, UNAM, Cd. Universitaria, Del. Coyoacán, C.P. 04510, México C.D.M.X  
(a)sashikosr@outlook.com, (b) idalia@unam.com

ABSTRACT

The metro system from Mexico City has previously been analyzed, but only by parts, specific case studies to some stations (transfer, transit or terminals) or metro lines (individually) and not to the entire system as such. This study is important since it will give us information about the system that is not yet known, it will help us to correctly identify risks to minimize them, as well as delays in the lines, make improvements to the system, have an adequate planning, establish different policies to improve and satisfy the system needs. Tools such as simulation will be used to create scenarios and search for alternatives for improvement in the system, as well as, where appropriate, other tools such as optimization will be used. This paper uses different techniques such as Complex Networks Methodology, Statistics, Simulation and Risk Analysis.

Keywords: transport system, Complex Networks

1. INTRODUCTION

Currently the Metro System has 12 lines which are distributed within Mexico City and part of the State of Mexico.

The Metro of Mexico City has a total of 384 convoys, of which 285 are in operation and 99 out of service for the following reasons: 33 for lack of spare parts, 20 for being in reserve, 17 for maintenance, 15 by general review, 7 by revision of breakdowns, 5 by work of modernization, and one more by special works and another by reprofiling of wheels.

2. PROBLEM

During a period of 14 months, from January 2017 to February 2018, the Metro system of México City presented 28,400 breakdowns during its operation. That is, an average of 2 366 failures per month or 77 per day, according to data obtained via transparency request.

The main problems that occurred in that time lapse were braking traction with 5 980; in the door system with 4 169; the automatic piloting with 4 043; the mechanical equipment with 3 144; and the generation of energy with 2 554.

While the lines that presented the highest number of breakdowns in all 2017 and in the first two months of 2018 were the line 3 that goes from Indios Verdes to Universidad with 4 459; line 1 that runs from Observatorio to Pantitlán with 3 631; the line A that travels from Pantitlán to La Paz with 3 394; followed by line 7 that goes from El Rosario to Barranca del Muerto with 2 985; and line 5 that goes from Politécnico to Pantitlán with 2 862 failures.

The subway system has its main problems due to factors such as the elements' wear of the gear change, which has been caused by natural wear, cracks or fractures in lines with greater age, as well as lack of lubrication in rails and settlements differentials caused by the settlements of the subsoil of the city.

While the electronic installations have normal deterioration in the equipment, which affects the Automation and Control systems, then they must operate in safety conditions, over electrical installations. In lines 8 and A were delays due to lack of power of that type due to the failure of a general switch that occurred due to the voltage variation of the CFE (Comisión Federal de Electricidad).
During 2017, the impairment rate in hours due to service faults on all Metro lines was 17.6, while the actual service was 7,454.3, with a service percentage of 99.76, according to the information presented by the Metro.

In January and February of 2018, the hours without service reached 6.8, while the active hours were 1,205. The striking thing about the numbers is that in just two months they have reached almost a third of the total hours of failure that were recorded in all last year, which indicates that the problem is going up.

Only last February there were 4 hours total losses due to faults that have occurred in the service of all Metro lines, which exceeds the highest months of 2017 that were June and December with 3.2 and 3.1, respectively.

The Metro reported different problems: Technical, Operational, Social and Financial problems. The technical problems are those related to the operation of the network such as the control system, braking system, door opening system, capacity of the wagons, lack of spare parts, among others, most of these problems are due to lack of maintenance. Other problems are found in the operation of the system, which are those related to the rules and policies with which the Metro operates, such as the number of trains operating per schedule, action policies within the platforms, such as safety measures, evacuation, action measures in case of mishaps such as earthquakes, fires, terrorist attacks, among others. Another type of problem is the social problems, which are associated with people such as the flow of passengers, crowds, violence inside and outside the wagons, street vendors, among others. Finally, we find the financial or budgetary problems, since the Metro does not have enough money to maintain, buy spare parts of trains, rehabilitation of trains that are out of circulation or put in circulation new trains.

Therefore, the next questions are made: which the most likely failure stations are? how faults will propagate to other lines? how the network connectivity is? which are the alternate routes in case of failures?

So, we will focus this analysis with Complex Networks to identify the stations that have the most important problems and its vulnerability, and we will create different scenarios from which we will have the simulation of the whole system and how it works with the different scenarios.

3. METHODOLOGY

For the methodology we will follow the next steps.

1. First Step
   This step is maybe one of the most difficult steps because we need to look up for all the data that is relevant for the study.

2. Second Step
   So, with this information the second step was to analyze the data with basic statistical techniques.

3. Third Step
   We can create the network in different ways, the most common is with the adjacency matrix. An adjacency matrix is a square matrix used to represent a finite graph or network.

4. Fourth Step
   Once the network is obtained, we used the methodology of complex networks, specially to analyze the topology or structure of the networks, for example, the clustering, the closeness, betweenness, assortativity, and more metrics of complex networks. Also, we can have the degree distribution of the networks and we can have a good approach of the network’s behavior. With all the metrics and the degree distribution we can classify the networks into one of the different networks model (Random Networks, Small World Networks and Scale-free Networks).
5. Fifth Step
The fifth step consists in translate the information to time series, so we did a decomposition of time series into the three components series (Seasonal, Trend and Random), we obtained the ACF (Auto Correlation Function), PACF (Partial Auto Correlation Function). In this step we also can create time series models like ARIMA (Autoregressive Integrated Moving Average) models to do some forecast of the data.

6. Sixth Step
For the simulation process we will build different scenarios of the network to analyze the different structures and the vulnerability so we can compare which network is better. We can have different scenarios for example what happen if we delete one node or an edge.

4. RESULTS
For the statistical analysis, we used the R software, which is an open source programming language and software environment for statistical computing and graphics. For this work, we used specific R software packages, such as, igraph, networks, tkplot, sand, sna, forecast, TimeSeries, TSA and others. Software allows us to generate graphs/networks, compute different network metrics like clustering or transitivity, different centrality metrics, plot networks, create mathematical models, forecast data and more functions. Also, we used a BI (Business Intelligence) software that allow us to have some data preparation just like an ETL (Extract, Transform, Load) process and to create reports and visualization of our data.

According to the methodology, at the first step we have the data of the number of passengers by station and trimester from the first trimester from 2011 to the first trimester of 2019.

Computing basic statistics: First, we analyze the number of passengers per line, to have the ranking of the lines with more passengers.

Table 1 Number of Passengers per Line

<table>
<thead>
<tr>
<th>Line</th>
<th>Passengers</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line 2</td>
<td>2,399,777,835</td>
<td>17.87%</td>
</tr>
<tr>
<td>Line 1</td>
<td>2,128,428,724</td>
<td>15.85%</td>
</tr>
<tr>
<td>Line 3</td>
<td>1,944,304,705</td>
<td>14.47%</td>
</tr>
<tr>
<td>Line B</td>
<td>1,313,948,094</td>
<td>9.78%</td>
</tr>
<tr>
<td>Line 8</td>
<td>1,129,922,146</td>
<td>8.41%</td>
</tr>
<tr>
<td>Line 9</td>
<td>952,297,672</td>
<td>7.09%</td>
</tr>
<tr>
<td>Line 5</td>
<td>882,986,511</td>
<td>6.57%</td>
</tr>
<tr>
<td>Line 7</td>
<td>828,596,887</td>
<td>6.17%</td>
</tr>
<tr>
<td>Line A</td>
<td>794,763,580</td>
<td>5.92%</td>
</tr>
<tr>
<td>Line 12</td>
<td>592,338,103</td>
<td>4.41%</td>
</tr>
<tr>
<td>Line 4</td>
<td>249,863,002</td>
<td>1.86%</td>
</tr>
<tr>
<td>Line 6</td>
<td>215,006,325</td>
<td>1.60%</td>
</tr>
<tr>
<td>Total</td>
<td>13,432,233,584</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Then, we analyze the number of passengers per station to also have a ranking of the station with the highest numbers of passengers.

Table 2 Top 10 Number of Passengers per Station

<table>
<thead>
<tr>
<th>Line</th>
<th>Station</th>
<th>Passengers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line 3</td>
<td>Indios Verdes</td>
<td>345,139,908</td>
</tr>
<tr>
<td>Line 2</td>
<td>Cuatro Caminos</td>
<td>344,277,759</td>
</tr>
<tr>
<td>Line A</td>
<td>Pantitlán A</td>
<td>300,460,361</td>
</tr>
<tr>
<td>Line 5</td>
<td>Pantitlán 5</td>
<td>267,067,692</td>
</tr>
<tr>
<td>Line 8</td>
<td>Constitución de 1917</td>
<td>259,450,656</td>
</tr>
<tr>
<td>Line 2</td>
<td>Tasqueña</td>
<td>259,221,934</td>
</tr>
<tr>
<td>Line 9</td>
<td>Pantitlán 9</td>
<td>254,180,021</td>
</tr>
<tr>
<td>Line 1</td>
<td>Observatorio</td>
<td>218,684,664</td>
</tr>
<tr>
<td>Line 3</td>
<td>Universidad</td>
<td>217,461,690</td>
</tr>
<tr>
<td>Line 2</td>
<td>Zócalo</td>
<td>204,023,284</td>
</tr>
</tbody>
</table>

From the table 2 we can see that the station Indios Verdes is the most crowded, but we also can notice that Pantitlán is a hub so, the cumulative number of passengers is higher than at Indios Verdes, this is important because this means that we need to have special attention in this station.

We continue with the methodology and we need to create a network. So, we have the structure of the metro system, characterize these data as an adjacency matrix, in this case, the nodes represent the stations and the edges represent the connections through the line. The next figure shows the structure of the system as a complex network.
Now, we have the system characterized as a complex network so now we can compute the different complex network metrics to study the topological structure of our network.

<table>
<thead>
<tr>
<th>Results</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>195</td>
</tr>
<tr>
<td>Edges</td>
<td>220</td>
</tr>
<tr>
<td>Max. Degree</td>
<td>4</td>
</tr>
<tr>
<td>Min. Degree</td>
<td>1</td>
</tr>
<tr>
<td>Mean Degree</td>
<td>2.25641</td>
</tr>
<tr>
<td>Diameter</td>
<td>39</td>
</tr>
<tr>
<td>Mean Distance</td>
<td>12.94618</td>
</tr>
<tr>
<td>Cliques</td>
<td>4</td>
</tr>
<tr>
<td>Density</td>
<td>0.011631</td>
</tr>
<tr>
<td>Assortativity</td>
<td>0.245905</td>
</tr>
<tr>
<td>Global Clustering</td>
<td>0.056962</td>
</tr>
<tr>
<td>Mean Local Clustering</td>
<td>0.017304</td>
</tr>
<tr>
<td>Closeness Centrality</td>
<td>0.059484</td>
</tr>
<tr>
<td>Degree Centrality</td>
<td>0.008988</td>
</tr>
<tr>
<td>Betweenness Centrality</td>
<td>0.144816</td>
</tr>
</tbody>
</table>

The minimum degree corresponds to 1 and it makes sense because they are the terminal stations, the maximum degree is 4 that corresponds to stations like Pantitlán, meanwhile the average grade is 2.25, which tells us that there are very few stations that are transfer.

On the other hand, the density is important, it tells how connected the network is, the real systems modeled with networks, in general, are not very dense, due to the cost of the links. The network has a density of 0.011 which indicates that the connectivity within the network is very low and poor.

Another of the metrics that we use in this analysis is the mean distance, which is the average of the distances between all pair of nodes, so we expect that the networks have a low average distance, which has to do with the small world property, but in this case we have a mean distance of 12.94 that is quit high in comparison with the number of nodes and edges.

In addition to the metrics that are listed above, we are interested in studying the topology of the network so clustering is important and, we start with the global clustering, it means what the tendency of the network is to form triangles or to be transitive, so, the global clustering is very low it is 0.056, so it has a low tendency to form triangles. While, if we look at the mean local clustering, it is very similar to the global clustering but in this case, it is lower (0.017) so we can say that there is no tendency to form small groups, that is, they remain in the whole group.

On the other hand, the betweenness centrality, helps us to identify how important a node is within a network, computing how many short paths pass through the node in question, so we compute the average of the intermediate centrality of each case and we obtained a value of 0.144, the network has a very low betweenness centrality. The closeness centrality focuses on computing the shortest paths of each node to all other nodes in the network, we have that the closeness is relatively high. If we talk about the correlation of nodes, we have the coefficient of assortativity that gives us values between -1 and 1, therefore we can say if a network is assortative or dissortative, so, our network has a value of 0.24 with this we can say that it is assortative.
We can observe that the distribution of degree seems to be binomial.

With all these results we are able to analyze and compare the behavior of the different stations and lines, in addition, we could analyze the topology of the whole system, which concludes the type of network model is and what specific characteristics and properties they share.

The next step is to perform the time series analysis, so first we organize and sort our data by the date (the most recent date and the end). Then we plot our time series just as the example of the figure 5, where we plot the 12 lines as a time series.

We use the time series decomposition that is a mathematical procedure that transforms a time series into a multiple different time series. The original time series is often split into 3 component series:

- Seasonal: patterns that repeat with a fixed period.
- Trend: The underlaying trend of the metrics.
- Random: also call “noise”, “irregular” or “remainder”, this is the residuals of the original time series after the seasonal and trend series are removed.

Figure 7 Decomposition Total Passengers Metro

To continue with our analysis, we use the ACF (Auto-Correlation Function) that gives values of autocorrelation of any series with its lagged values. We plot these values along with the confidence. We have an ACF plot. In simple terms, it describes how well the present value of the series is related with its past values. A time series can have components like trend, seasonality, cyclic and residual. ACF considers all these components while finding correlations hence it’s a complete auto-correlation plot.

We also used the PACF (Partial Auto-Correlation Function). Basically, instead of finding correlations of
present with lags like ACF, it finds correlation of the residuals (which remains after removing the effects which are already explained by the earlier lag(s)) with the next lag value hence ‘partial’ and not ‘complete’ as we remove already found variations before we find the next correlation. So, if there is any hidden information in the residual which can be modeled by the next lag, we might get a good correlation and we will keep that next lag as a feature while modeling. Remember while modeling we do not want to keep too many features which are correlated as that can create multicollinearity issues. Hence, we need to retain only the relevant features.

![PACF Total Passengers Metro](image)

Figure 9 Total Passengers Metro

The ACF and PACF plots are more common used to obtain the values of p and q to feed into the ARIMA model.

All these analyses are important because it show us which are the patterns, seasonality and trend that the passengers follow throughout the time.

For the simulation scenarios we made 2 even more scenarios can be constructed depending on the problem that the network has to face. The first scenario is when we delete Pantitlán (the four stations of the lines 1,5,9, A and it’s connections) station that is one of the most important because of the number of passengers and the connections. The second scenario is when we delete the station with the lowest number of passengers that in this case is Tláhuac.

![Scenario 1](image)

Figure 10 Scenario 1

![Scenario 2](image)

Figure 11 Scenario 2

On the scenario 1 there is a community (all the line A) that is completely disconnected from the whole system and in the case of the scenario 2 we only delete one station and this station is the last one of the line 12 so the only problem here is that the terminal station Tláhuac is completely disconnected from the system.

According with the methodology, we compute the different complex network metrics for both scenarios and then we analyze and compare the results with the original network.
Comparing the three-network metrics, we find that the maximum degree is the same but the minimum degree on the scenario 2 is 0 because we delete the node that is the only connection with the terminal Tláhuac, so Tláhuac had 1 degree and when we delete Tlaltenco, Tláhuac remain alone. The mean degree is almost the same, in the case of the diameter on the scenario 2 there is a difference of 2 nodes so is a smaller size, but where we find the greatest difference is on the mean distance because on the scenario 1 it increases a lot so this tell us that Pantitlán station is important in our system and if we delete this station our connectivity decrease so we cannot remove or change this station. On the other hand, the centrality metrics does not change so much, so the scenarios remain with almost the same characteristics of the original network.

We also plot the degree distribution of the scenarios 1 and 2.

Both plots follow a binomial distribution just like the original network.

5. CONCLUSIONS
We conclude that the degree distribution of the network follows a Binomial Distribution, and in this case the network follows a Random Network Model because of the binomial distribution on the degree, the mean distance is high (tends to $p \sim \log(N)$), the clustering is low (tends to $k/N$), where $k$ is the average degree of the nodes.

In random networks, the neighbors of a certain node are chosen at random, so there is no correlation between the degree of neighboring nodes. Finally, these networks are more robust to targeted attacks, but at the same time they are vulnerable to internal errors.

After the time series analysis, we concluded that there is no evidence of a growing trend in the number of passengers and we could find some patterns in the seasonal cycles.

It is difficult to find the behavior patterns in a macro level so for the next steps we will do the same analysis but in a medium and micro levels. We will use the same methodology on the stations and lines that are more crowded to find and implement real solutions for this complex system. Also, the simulation will help us to create different scenarios to improve the way the metro works.
ACKNOWLEDGMENT
To the PAPIIT DGAPA project IT102117, UNAM for its support to this research.

REFERENCES


Figueras J., Modelos de Simulación usando simio y redes de Petri, Universidad Nacional Autónoma de México. pág. 2. (2013).


Lara, F., Teoría, métodos y modelos de la complejidad social I. Seminario de Investigación, CCADET.


**ACKNOWLEDGMENTS**

To CONACYT scholarship and UNAM scholarship for the first author.

**AUTHORS BIOGRAPHY**

Sashiko Shirai received a Master in Operation Research in the Faculty of Engineering of the UNAM and she also received her bachelor’s in actuarial science in the Faculty of Science of the UNAM. She is studying her PhD in the Instituto de Investigación en Matemáticas Aplicadas y Sistemas (IIMAS) of the UNAM. She is an assistant professor at the Faculty of Engineering and the Faculty of Science. Her research interests lie in complex networks, statistics and operational research.

Idalia Flores received a Master with honors, being awarded with the Gabino Barreda Medal for the best average of her generation, in the Faculty of Engineering of the UNAM, where she also obtained her Ph.D. in Operations Research. Dr. Flores is a referee and a member of various Academic Committees at CONACYT as well as being a referee for journals such as Journal of Applied Research and Technology, the Center of Applied Sciences and Technological Development, UNAM and the Transactions of the Society for Modeling and Simulation International. She is a full-time professor at the Postgraduate Program at UNAM and her