Analyze the relationship network of long-distance trains in Portugal

Luís Marques
ISCTE, Avenida das Forças Armadas,
1649-026 Lisboa, Portugal
lclms@iscte-iul.pt

***Abstract*—The theory of graphs denotes a great employability, it is a resource that allows to creating knowledge through the analyzes that it provides. This article proposes a new approach in the analysis of a network of relationships between trains. Dataset represents a day of train production that allows us to apply a set of metrics and measurements to identify the characteristics of the network.**

Keywords—Railway, theory of graphs, networkx, python

I. INTRODUCTION

Several works were developed for analysis of rail networks [1]–[5], these works rely heavily on a network infrastructure perspective [6]. However, the graph theory, as described in [7]–[9] has several features that can be used in different perspectives. Thus, this article approaches the perspective of the network analysis of the relationships between railway circulations. To analyze the network of train relationships, and to see which trains are most important, we use graph theory. For this, we use a set of measures and mathematical metrics [10] that lead to the most important trains. And by way of implementing this analysis, we turn to networkx a library that allows us to implement analytical and graphical algorithms. Thus, we will notice the centrality of degrees, the centrality of proximity and centrality of interaction, these three metrics that will give us a view of the important nodes in the dataset under analysis. In a group strand, let's abort subgroups and clusters. This paper is organized in a contextualization of graph theory, description of the dataset to study and two parts of analysis, part analysis at the level of nodes and finally part of analysis a group level.

II. THEORY OF GRAPHS

In the year 1736 Leonhard Euler wrote the first article on graph theory. This author studied the problem of the bridges of Königsberg, the problem posed by Euler was that he went through the seven bridges of the city without repeating any. Euler discovered that the problem had no solution.

Graphs have been heavily implemented in several real-world representations. They can be used in social networks, website pages, neural networks among other applications.

Fig. 1 - Bridges of Königsberg

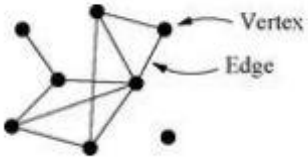

Fig. 2 - Illustration of a small network, a graph is usually denoted as $G = (V, E)$

III. DESCRIPTION OF DATASET

Railway movements are tasks which are planned and serve as a daily production objective for rail transport operators. However, there are several types of disturbances that can affect the production of trains. We can for example, describe some cases to frame the need to realize the importance of identifying the most important circulations.

- Crew: The crews follow a stopover, which may include one or more train tasks. In this sense, if a crew member fails or gets caught in a delayed or suppressed train, the task to be accomplished may be compromised.
- Circulating Material: Similar to crews, the rolling stock also complies with a planned, usually referred to as rolling stock rotation. The problem associated with this point is malfunctions that can compromise normal rotation.
- Infrastructure: Trains have their own very complex, highly security-oriented infrastructure. Therefore, when there are problems associated with infrastructure integrity, a number of constraints are being lifted, e.g., speed and cadence restrictions of traffic per section. In this way, the restrictions can affect the normal tracking

of the rolling stock rotations and the normal fulfillment of the scales of the crews.

The dataset that we are going to use in this article is the relations of the railroad circulations of a day.

IV. Dataset Information

Name	Train Graph Network
Type	Graph
Number of nodes:	56
Number of edges:	63
Average degree	2,2500

Table 1 – Info Dataset

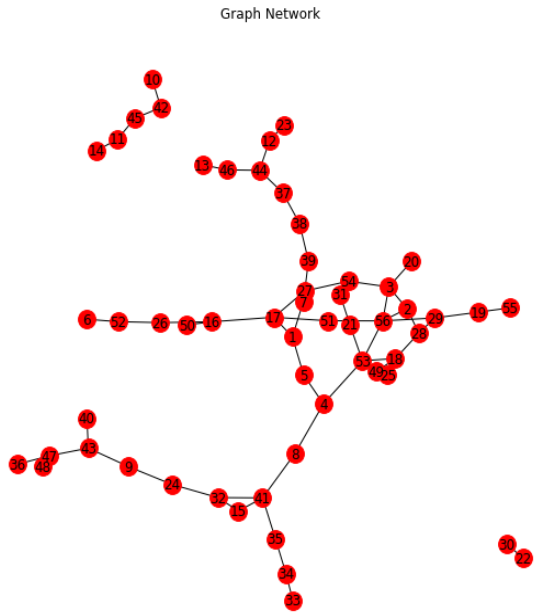

Fig. 3 - Graph Representation

V. Node-Level Analysis

A. *Degree Distribution*

It denotes the frequency of degrees of nodes, follows a power law distribution which means that few nodes have a high number of connections. Thus, nodes with higher degree have more impact on the network, in this way are considered more important.

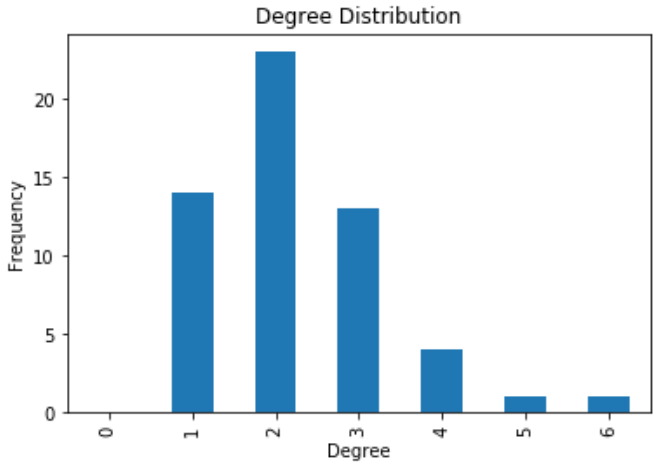

Fig. 4 - The degree distribution of network - The highest concentration is in the value 2 of Degree

B. *Betweenness Centrality*

Centrality can be described as the importance of a node. Demonstrates the number of times a node needs to pass through a node to reach another node. Thus, nodes with high centrality hold much importance because they form bridges between groups of nodes.

[('4', 0.3658810325476993), . . . ('42', 0.00202020202020202), ('11', 0.00202020202020202), ('10', 0.0), ('14', 0.0), ('15', 0.0), ('25', 0.0), ('20', 0.0), ('23', 0.0), ('49', 0.0), ('30', 0.0), ('22', 0.0), ('33', 0.0), ('36', 0.0), ('40', 0.0), ('13', 0.0), ('48', 0.0), ('50', 0.0), ('6', 0.0), ('55', 0.0)]

Node 4 presents the highest degree of centrality of intercession. At the other end, nodes 42 and 11 are located at the edge of the network and are not present in any of the shorter paths of the network. Still in this network 17 nodes appear without nodes of centrality.

C. *Closeness Centrality*

You can define how close a particular node is relative to the others. The centrality is given by the sum of the geodetic distances of one node, for all other nodes in the network. A given node may be important if it is relatively close to the remaining nodes in the network.

[('4', 0.2013986013986014), ('53', 0.1921601334445371), ('5', 0.19128268991282688), . . . 0.02909090909090909), ('14', 0.02909090909090909), ('30', 0.0181818181818181818), ('22', 0.0181818181818181818)]

Nodes 4, 53 and 5 reach the highest centrality degrees because they are present in the middle of the network and can reach other nodes with a few steps. However. On the contrary, nodes 14, 30 and 22 present values that indicate a lower degree of centrality.

D. *Degree Centrality*

In this degree centrality metric, it tells us the importance of a given node that is defined by the number of nodes it is connected to. Thus, the greater the number of adjacent nodes the greater importance is a given node. This is a local measure because its value is calculated taking into account the number of links that a node has to the other nodes directly adjacent to that node. Nodes with a high degree of centrality serve as a means of interconnecting network nodes.

[('53', 0.10909090909090909), ('56', 0.09090909090909091), . . . 0.01818181818181818), ('6', 0.01818181818181818), ('55', 0.01818181818181818)]

Note that node 53 and 56 show the highest levels of centrality, and it is the nodes that have the maximum number of connections. On the other hand, nodes 6 and 55 have the lowest degree of centrality.

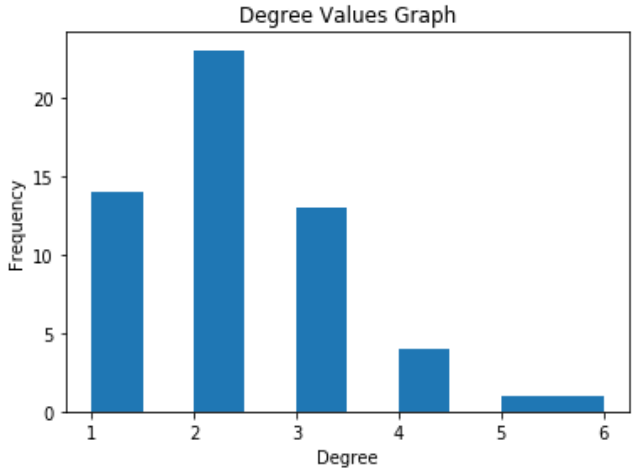

Fig. 5 - Degree Centrality graphic representation

E. *Density*

Density is defined as the degree that the nodes of a given network are connected to each other. We can also define the density as the number of edges of the network, divided by the maximum possible number of edges between the nodes in that same network. This measure is useful in exploring the dynamics of the network, such as the speed at which information diffuses between nodes. The density can assume values between 0 and 1. In our analysis, we obtained a value of 0,041 which means that we have a low density of the network.

Analyze for a particular node

Since node 53 presents a great level of centrality, its density presents 0,428 which we can verify that it is very close to the density of the network.

F. *Reciprocity*

It is a measure of the tendency for the creation of connections that are mutually directed between a pair of nodes. For a given node, reciprocity is the ratio of the number of nodes that have incoming and outgoing connections. Thus, reciprocity can be an indicator that reflects the importance of a relationship between two nodes. The reciprocity of our dataset results in 1 because the graph is not directed.

G. *Eco Net*

This analysis represents a common network taking into account a node in the center. By selecting a particular node, we can perceive its neighborhood. Thus we can perceive all the relationships of a given node. In our case, we selected node 53 because of its centrality and consequently importance in the network.

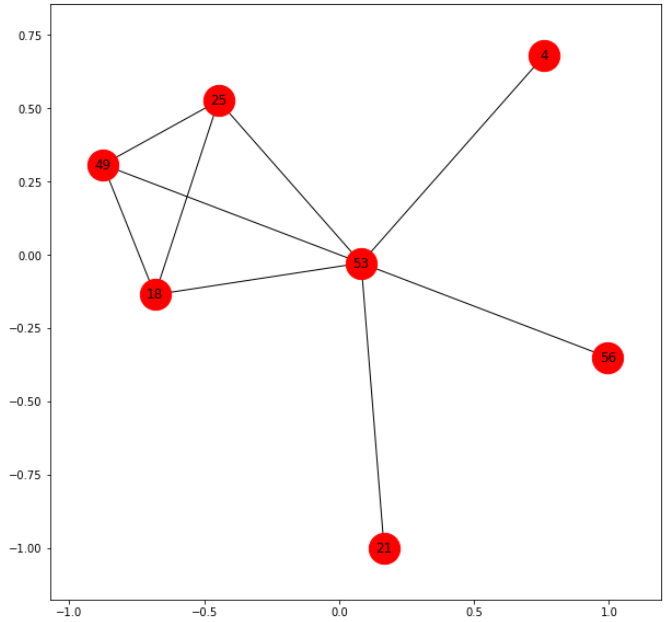

Fig. 6 - Ego network of a node 53 from the network graph - This node appears for 9 occasions in the network

H. *Eigenvector Centrality*

This metric is proportional to the sum of neighborhood results e.g. PageRank, it is a measure of related influence. Who is the most important node of the network. It may thus mean that a node with a small number of contacts may have a high level of importance depending on the level of importance of the adjacent nodes.

[('53', 0.492437945678543), ('18', 0.3959492221182913), ('25', 0.3482572059284773), ('49', 0.3482572059284773), ('56', 5.1865563981718094e-17), . . . ('22', 5.1865563981718094e-17)]

The node 53 reaches the centrality of the highest eigenvector, while node 22 reaches the lowest value. In the Degree Centrality test, this node appears first. However, the same does not happen with node 56 which in this test lies

behind nodes 18, 25 and 49. This means that it is in a border position of centrality.

I. PageRank

PageRank is a variant of eigenvector centrality measure. Calculates a classification based on connections in a given network, taking into account the conformity with the relative importance of each node within a set of nodes. It can assume values between 0 and 1.

Node	Value
53	0,033
56	0.029
47	0,029
41	0,028
17	0,027

Table 2 – PageRank results

We returned node 53 first to consolidate its importance in the node, although in the Eigenvector Centrality test node 56 did not reach the first places here is second which denotes its centrality and equal importance. The node in this test 47 is assumed, which assumes the third position, which thus places it in a border zone of centrality.

VI. GROUP ANALYSIS

A. Cliques

It is a graph in which each node is connected to any other node. When we find the maximum of Cliques on a network we can analyze the relationships between the nodes, since we have the ability to obtain the group dynamics. To find Cliques in a graph in the computational sense we face an NP-Complete problem, which means that the problem grows exponentially as the data size grows.

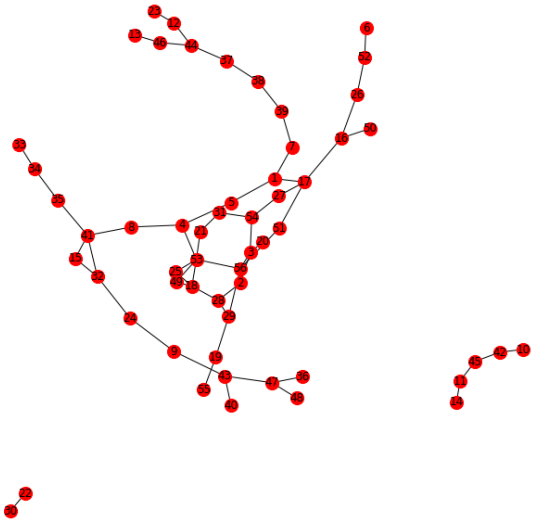

Fig. 7 - Cliques Graph

Number of Network Cliques

54

Cliques (repeat) sizes on the network

[2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 4, 2, 2]

Maximum clique size condition which was 4

[[53, 49, 25, 18]]

And also their number:

1

This means that we have 1 different maximum Cliques in the graph, each of which consists of 4 nodes. This means that this 4 knot variation binds.

We see the average query size:

2,074

[{49, 18, 53, 25}]

B. Clustering Coefficient

It is a measure that describes how many nodes tend to form subgroups, also called Cliques as we saw. This metric ranges from 0 to 1. Values near 1 indicate that the network denotes the small world phenomenon.

The *nx.clustering(g)* function calculates the collation coefficient for each node, higher values mean higher Cliques value. In this way, these values represent segments that are connected.

{'1': 0,

'10': 0,

'11': 0,

'12': 0,

'13': 0,

'14': 0,

'15': 1.0,

'16': 0,

'17': 0,

'18': 0.5,

'19': 0,

'2': 0.3333333333333333,

'20': 0,

'21': 0,

'22': 0,

'23': 0,

'24': 0,

'25': 1.0,

'26': 0,

'27': 0,

'28': 0,

'29': 0,

'3': 0.16666666666666666,

'30': 0,

'31': 0,

'32': 0.3333333333333333,

'33': 0,

'34': 0,

'35': 0,

'36': 0,

'37': 0,

'38': 0,

'39': 0,

'4': 0,

'40': 0,

'41': 0.16666666666666666,

'42': 0,

'43': 0,

'44': 0,

'45': 0,

'46': 0,

'47': 0,

'48': 0,

'49': 1.0,

'5': 0,

'50': 0,

'51': 0,

'52': 0,

'53': 0.2,

'54': 0,

'55': 0,

'56': 0.1,

'6': 0,

'7': 0,

'8': 0,

'9': 0}

Function to calculate the average coefficient of the network, we denote a low value that represents few clusters.

nx.average_clustering(g)

Out: 0.08571428571428572

Let's compute the collation coefficient for a specific node. We select the node 53 being the node with more centrality.

ego_net = nx.ego_graph(g, "53")

len(ego_net)

Out: 7

nx.average_clustering(ego_net)

Out]: 0.4571

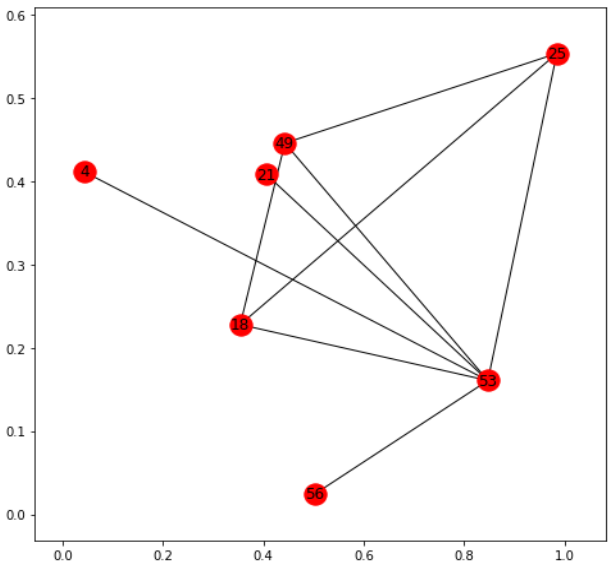

Fig. 8 - Coefficient of grouping for a node 53

Node Local Network 53

nx.clustering(ego_net)

Out: {'18': 1.0, '21': 0, '25': 1.0, '4': 0, '49': 1.0, '53': 0.2, '56': 0}

This function allows us to find the clustering coefficient for each node in the ego network. Node 53 is surrounded by a set of nodes that each have a Cliques coefficient of 1.

VII. CONCLUSION

The network that represents the daily production of trains, presents a set although reduced to an important node in its composition. However, we realize that the level of grouping is not high, which leads to a resilience of precedence. It is important to note that this study did not take into account two important factors, the spatial and temporal reference. Thus, in the sense, there is an opportunity to associate these factors to improve future research.

VIII. REFERENCES

[1] R. M. P. Goverde, "A delay propagation algorithm for large-scale railway traffic networks," Transp. Res. Part C Emerg. Technol., vol. 18, no. 3, pp. 269–287, 2010.

[2] F. Yang, J. Feng, and H. Zhang, "Power flow and efficiency analysis of multi-flow planetary gear trains," Mech. Mach. Theory, vol. 92, pp. 86–99, 2015.

[3] H. Bast et al., "Route Planning in Transportation Networks," pp. 1–65, 2015.

[4] D. Tönissen, J. Arts, and Z. J. M. Shen, "Maintenance location routing for rolling stock: Robust and Stochastic Programming Formulations," vol. 521, no. January, 2017.

[5] J. Wagenaar and L. Kroon, "Maintenance in Railway Rolling Stock Rescheduling for Passenger Railways," pp. 1–38, 2015.

[6] S. A. Khan, N. A. Zafar, F. Ahmad, and S. Islam, "Extending Petri net to reduce control strategies of railway interlocking system," Appl. Math. Model., vol. 38, no. 2, pp. 413–424, 2014.

[7] D. Jungnickel, Graphs, Networks and Algorithms, vol. 53, no. 9. 2008.

[8] Y. Zhu and H. Fan, "Vital Nodes Evolution Study on Railway Network of Silk Road Economic Belt," Sci. Res. Publ. Inc, no. August, pp. 115–123, 2016.

[9] M. Z. Al-Taie and S. Kadry, Python for Graph and Network Analysis. 2017.

[10] M. E. J. Newman, "The mathematics of networks," New Palgrave Encycl. Econ., vol. 2, pp. 1–12, 2007.