Oct 28th, 9:00 AM - Oct 30th, 5:00 PM

Descriptive Analysis of Characteristics: A Case Study of a Phone Call Network Graph

Orgeta Gjermëni
University Ismail Qemali, o.gjermeni@gmail.com

Miftar Ramosaco
University Ismail Qemali, miftar.ramosaco@gmail.com

Descriptive Analysis of Characteristics: A Case Study of a Phone Call Network Graph

Orgeta Gjermëni[1], Miftar Ramosaco[1]

[1] Department of Mathematics, University Ismail Qemali, Str. Kosova, 9400 Vlore, Albania,
{ o.gjermeni, miftar.ramosaco}@gmail.com

Abstract. Nowadays, systematic collection of data has necessitated a detailed statistical analysis as a necessary tool to make a mathematical characterization of them with the purpose of gathering information about the present or the future. Our aim in this paper is to analyze a landline phone call network graph from the perspective of descriptive analysis. We explore the characteristics and structural properties of the network graph constructed using an anonymous collection of data gathered from a Call Data Records of a telecommunication operator center located in south of Albania. The R statistical computing platform is used for network graph analysis.

Keywords: Landline phone call, Network Graph, Descriptive analysis, R.

1. Introduction

A landline phone call network graph is a network graph which represent a system where the set of vertices is the set of active phone callers, and the set of edges is the set of phone call relations (communication relations) between them. The structure of such network graphs and the particular patterns of interactions inside them, can have a big effect on the behavior of the communication system. This network graphs can be seen both as technological and social network graphs.
Descriptive analysis of characteristics is an important task to explore the structural properties of the network graphs. This tasks range from the calculation of simple metrics, summarizing topological structure (global and local) to complex relational patterns. Understanding and analyzing network graphs is an area of science that stands between some disciplines (Mathematics, physics, the computer and information sciences, the social sciences, biology etc.). Various statistical and visual analyzing platform exist to study network graphs, such as: Pajek, Gephi, Python, and R [1]. In our study we have choose to use R statistical computing platform, as it offers great flexibility for network graph research.
Previous studies are conducted on phone call network graphs based on: the structure of the underlying network graph (cliques [2], degree distribution [3]). The majority of the studies are related to call network graphs is conducted on mobile data. The focus has been basically on the statistical properties of the social behavior of mobile network vertices [4, 5, 6, 7]. The purpose of this empiricale case study was to explore the characteristics and structural properties, in way that our findings b useful to provide business insights and help the operator to offer the right incentives to their customers.
We investigate about the candidate degree (strength) distribution. What is the dependency between the degrees (strength) of vertices? What is the cohesiveness of the graph? How transitive and dense is the network graph?

The outline of this paper is organized as follows: In Section 2, we describe material and methods that is applied to analyze data. Next, in Section 3, we see the results, and in Section 4, we discuss and conclude some remarks about the results.

2. Material and Methods

2.1. Data preparation

A phone call network graph $G = (V, E)$ is constructed using an anonymous collection of data, gathered from a Call Data Records of a telecommunication operator center, located in south of Albania. Data is related to phone calls only inside operator customers, not to foreign operators, during November 2014, and contains a total of 81591 phone calls. Of these, 41 phone calls which had no defined duration and 7442 phone calls which had a duration of less than 10 seconds, were excluded from consideration along the study. The motivation for this exclusion is that they may held incorrect results for being missed calls or wrong calls.
The phone call network graph G is constructed using only 90.83% of initial data set. The set of vertices in the network graph is denoted by V and represents the set of active landline operator customers, and E is the set of edges of G. Active customers are considered all them that have made at least a 10 seconds call duration. Each edge represents a communication relation between two customers. Specifically, if v_1 and v_2 are vertices of G, then an undirected edge (v_1, v_2) exists only if v_1 has called at least one time v_2 or the reverse. Multiple calls between any two vertices are given by a single edge, which is associated with a weight (to represent connectivity (0 or 1), the total number of calls, or the numbers of seconds of communication between two vertices during the interval of observation). In this way, during the statistical analysis, G is considered weighted in two ways:

1- edge weight shows the total number of calls;
2- edge weight shows the total duration of calls;

between the incident vertices during the interval of observation. The weighted matrices are $\boldsymbol{W'}$ and $\boldsymbol{W''}$, respectively.

2.2. Basic definitions

Order of a network graph is referred to the number of vertices in it, while the *size* is the number of edges. Let l be the *mean geodesic distance* [8] between vertex pairs in an undirected network graph of order n. The geodesic distance between two vertices is the shortest distance between them. The *diameter* d, of a network graph is the length of the longest geodesic distance between any pair of vertices. The Breadth – first search algorithm is used in R to compute l and d.
Vertex degree [8] in a network graph is the number of edges incident on that vertex. Let denote with d_v the degree of the vertex v. We will define as p_x the fraction of vertices v that have $d_v = x$. This can be interpreted also as – the probability that a vertex chosen uniformly at random has a degree equal to x. The set of $\{p_x\}_{x\geq 0}$ defines the *degree distribution* of the network graph. *Vertex strength* [9] measures the strength of vertices in terms of the total weight of their connections in weighted network graphs. It is denoted by s_v. We will define as p_s the fraction of vertices v that have $s_v = s$.The set of $\{p_s\}$ defines the *strength distribution* of the network graph. *Average neighbor degree (strength)* of a vertex is the sum of neighbor vertex degree (strength).
In our paper, we describe network graph cohesion by looking at: maximal cliques, the clique number, density, and clustering (or transitivity). *Cliques* are network sub-graphs that are fully

cohesive. A *maximal clique* is a clique which is not subset of a larger one. The size of the largest clique is referred to as the *clique number*. The largest cliques are always maximal, but a maximal cliques is not necessarily the largest. The algorithm implemented in R for finding maximal cliques is given by Eppstein et. al. [10]. *Density* shows how close a network graph is to being complete [11]. It is measured as the frequency of realized edges relative to potential ones, and is denoted by $den(G)$.

We will distinguish two types of clustering based on local and global view perspective in undirected and not weighted network graphs. *Clustering coefficient* from the perspective of vertices was introduced by Watts and Strogatz [12, 13] and it is denoted by $cl(v)$ where $v \in V$. The corresponding clustering coefficient of the whole network graph takes the form, $cl(G) = \frac{1}{|V|}\sum_{v\in V} cl(v)$. In case of vertices with degree equal to zero or one we put $cl(v) = 0$ [8]. Network graph *transitivity* $cl_T(G)$ [13, 8] considers the network graph as a whole and it is referred to as the "fraction of transitive triples".

There is also an extension of the concept of clustering in local vertex level quantity in the weighted undirected simple network graphs defined by Barrat et al. [9] as: $cl_{\boldsymbol{W}}(v) = \frac{1}{s_v(d_v-1)}\sum_{u,t}\frac{(W_{vu}+W_{vt})}{2}A_{vu}A_{vt}A_{ut}$, where s_v is the strength and d_v the degree of vertex v, A_{vu} and W_{vu} are elements of adjacency and weighted matrix, respectively. The corresponding *weighted clustering coefficient* of the whole network graph takes the form, $cl_W(G) = \frac{1}{|V|}\sum_{v\in V} cl_W(v)$.

The tendency of vertices to be connecting to other vertices, according to a certain characteristics, is referred in the social network graph literature as *assortative mixing*, while measures that quantify the extent of assortative mixing in a given network graph have been referred to as *assortative coefficients* [13]. The assortativity coefficient used is attributed to Newman [14, 15].

2.3. Statistical Computation

Statistical computation analysis is conducted based on two packages, igraph [16] and igraphdata [17] in R [1] statistical computing platform.

3. Results

After simplification, our network graph $G = (V, E)$ was an undirected connected network graph. It had no edges for which both ends connect to a single vertex and no pair of vertices with more than one edge between them. The order of G was $|V| = 3287$ and its size was $|E| = 56259$. The network graph G had a diameter $d = 6$ and the mean geodesic distance was $l = 2.7226$.

Table 1 gives a summary of some basic statistics. 'Call strength' is referred to G when edge weight shows the number of calls between the incident vertices and 'Duration strength' is referred to G when edge weight shows the total duration of calls between. In the fourth row is given a summary of the call durations present in the data.

Table 1. Basic summary statistics related duration of calls, degrees and strength of vertices. Strength of vertices is related to the total number of calls or total duration of calls.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Degree d_v	1	5	16	34.23	45	844
Call strength $s_v^{W'}$	1	6	19	45.05	55	2713
Duration strength $s_v^{W''}$	10	604.5	2569	6859	7962	416800
Duration	10	34	80	200.4	209	17980

Non – cumulative degree and strength distribution in log – log scale are given in Fig. 1.

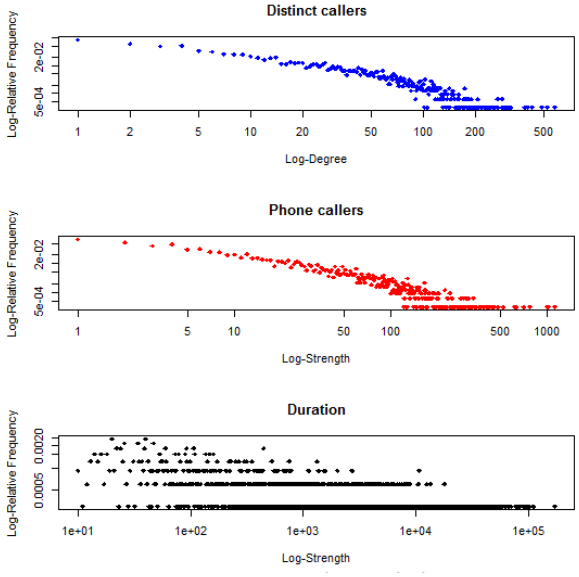

Fig. 1. *'Distinct callers'* gives the non – cumulative degree distribution of $\boldsymbol{G}$, while *'Phone callers'* and *'Duration'* are non – cumulative strength distribution related to the total number of calls ($\boldsymbol{W'}$) and the total duration of calls ($\boldsymbol{W''}$) of each vertex during the interval of observation.

Average neighbor degree (strength) versus vertex degree (strength) are given in Fig. 2 in log – log scale.

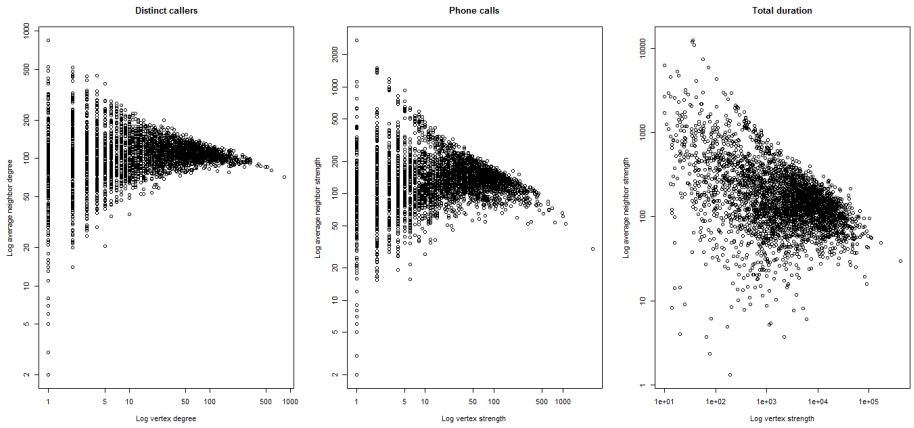

Fig. 2. In the first figure is given the average neighbor degree versus vertex degree, in the second figure is given the average neighbor strength versus vertex strength when edge weight is referred to total number of phone calls, and in the third one is given the average neighbor strength versus vertex strength when edge weight is referred to total duration of phone calls.

Values of assortativity coefficients are given in Table. 2.

Table 2. Assortativity mixing coefficients.

Type of Assortativity	**Coefficient value r**
Degree	-0.0638
Call strength	-0.0459
Duration strength	-0.0368

In Fig. 3 are given relationships between degrees, call strength and call duration per vertex.

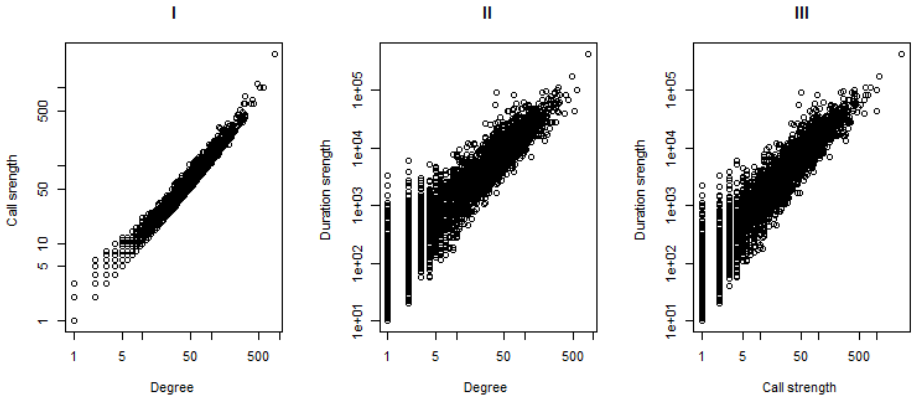

Fig. 3. The relationships between: degree and call strength (I), degree and duration strength (II), and call strength and duration strength (III) per vertex in log – log scale.

Related to network graph cohesion, Fig. 4 gives a histogram of the distribution of maximal cliques. The clique number of G is 11.

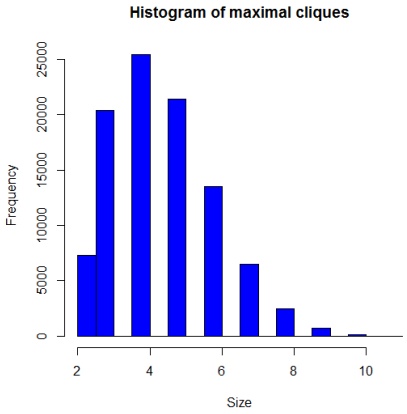

Fig. 4. The histogram of maximal cliques.

We see that $den(G) = 0.01, cl_T(G) = 0.083$ and $cl(G) = 0.093$. When G is considered as weighted, $cl_{W'}(G) = 0.098$ and $cl_{W''}(G) = 0.0938$.

Discussion and Conclusions

The diameter of our network graph respects the Stenley Milgram theory of "Six degree of separation". The mean geodesic distance can be considered quite small because $l \sim \log(|V|)$. From Fig. 1 we find that there is a somewhat linear decay in log - relative frequency as function of log – degree or log – strength. The degree (strength) distribution is suspected to be a power – law, but in this case it would be impossible to fit all the data. Double Pareto log – normal (DPLN) [18]distribution is a suspected distribution too. Degree (strength) distribution results are consistent with mobile phone calls empirical studies [4, 5, 6], although our network graph is with landline phone call data and undirected. Vertices with lower degree tend to have connection more with vertices of higher degree, while vertices with higher degree tend to have connection more with vertices of similar degree (Fig. 2).

The number of vertices of higher degree is quite low compared to that of lower degree. The same situation is when we take into account the strength of vertices – total number of calls and the strength of vertices – total call durations. All the assortativity coefficients in Table 2 are negative. This suggests that our network graph is disassortative mixing. Technological network graphs are disassortative mixing [14, 15].

Call strength as a function of degree, duration strength as a function of degree and duration strength as function of call strength showed a positive linear relationship in log – log scale axes (Fig. 3). This means that: ${s_v}^{W'} \sim {d_v}^{\alpha}$ for some $\alpha > 0$; ${s_v}^{W''} \sim {d_v}^{\beta}$ for some $\beta > 0$; and ${s_v}^{W''} \sim [{s_v}^{W'}]^{\gamma}$ for some $\gamma > 0$. Same conclusions are given also in [9].

Our network graph, as a real network graph, showed that large cliques were relatively rare and the clique number was very small compared to the network graph order. No maximal cliques of size larger than 11 was observed, and it results consistent with [7]. This happens because real network graphs are often sparse. Based on density value, only 1% of possible undirected connections was active.

According to the values of transitivity and clustering coefficient, nearly 10% of connected triples close to form triangles. In order to investigate about the small world property, transitivity – clustering coefficient values and mean geodesic distance should be compared to random graphs samples. Also, the values about all the variables defined in this paper are all based in empirical

data. We cannot say how significant they are in this paper, but these are subjects of our future work.

References

1. R. C. Team, "R: A Language and Environment for Statistical Computing". Vienna, Austria 21 June 2016.
2. Abello J., Pardalos P. M., Resende M. G. C., "On Maximum Clique Problems In Very Large Graphs," in *DIMACS Series: External Memory Algorithms*, vol. 50, V. J. S. Abello J. M., Ed., American Mathematical Society, 1999, pp. 119--130.
3. Aiello W., Chung F., Lu L., "A Random Graph Model for Massive Graphs," in *32nd Annual ACM Symposium on Theory of Computing*, New York, 2000.
4. Dong Zh.-B., Song G.-J., Xie K.-Q., "An Experimental Study of Large – Scale Mobile Social Network," in *WWW2009*, Madrid, Spain, 2009.
5. Seshadri M., Machiraju S., Sridharan A., Bolot J.,Faloutsos Ch., Leskovec J., "Mobile Call Graphs: Beyond Power-Law and Lognormal Distributions," in *KDD'08*, Las Vegas, Nevada, USA, 2008.
6. Onnela J. – P., Saramaäki J., Hyvöven J., Szabó G., Lazer D., Kaski K., Kertész J., Barabási A. – L., "Structure and Tie Strengths in Mobile Communication Networks," *PNAS*, vol. 104, no. 18, pp. 7332-7336, 2007.
7. Nanavati A. A., Gurumurthy S., Das G., Chakraborty D., Dsgupta K., Mukherjea S., Joshi A., "On the Structural Properties of Massive Telecom Call Graphs," in *CIKM'06*, Alington, Virginia, USA, 2006.
8. M. E. J. Newman, "The Structure and Function of Complex Networks," *SIAM Review*, vol. 45, pp. 167-256, 2003.
9. Barrat A., Barthelemy M., Pastor - Satorras R., Vespignani A., "The Architecture of Complex Weighted Networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3747-3752, 2004.
10. Eppstein D., Löffler M., Strash D., "Listing all Maximal Cliques in Sparse Graphs in Near-Optimal Time," arXiv, 2010.
11. Wasserman S., Faust K., Social Network Analysis: Methods and Applications, Cambridge Univeristy Press, 1994.
12. Watts D. J., Strogatz S. H., "Collective Dynamics of 'Small-World' Networks," *Nature*, vol. 393, pp. 440-442, 1998.
13. E. D. Kolaczyk, "Descriptive Analysis of Network Graph Characteristics," in *Statistical Analysis of Network Data*, Springer Science+Business Media, LLC, 2009, pp. 79-122.
14. M. E. J. Newman, "Assortative Mixing in Networks," *Phys. Rev. Lett.*, vol. 89, no. 208701, 2002.

15. M. E. J. Newman, "Mixing Patterns in Networks," *Phys. Rev. E,* vol. 67, no. 026126, 2003.
16. Csardi G., Nepusz T., "The igraph software package for complex network research," *InterJournal,* vol. Complex Systems, p. 1695, 2006.
17. G. Csardi, "igraphdata: A Collection of Network Data Sets for the 'igraph' Package," 2015. [Online]. Available: CRAN.R-project.org/package=igraphdata.
18. Fang Zh., Wang J., Liu B., Gong W., "Double Pareto Lognormal Distribution in Complex Network," in *Handbook of Optimization in Complex Networks: Theory and Applications*, vol. 57, P. P. Thai M. T., Ed., Springer Science+Business Media, LLC, 2012, pp. 55-80.