From Small World Phenomenon to Correlation Analysis in a Temporal Landline Phone Call Network Graph Series

Orgeta Gjermëni*, Miftar Ramosaço
Department of Mathematics, University "Ismail Qemali", Str. Kosova, 9400 Vlore, Albania.

* Corresponding author. Email: o.gjermeni@gmail.com

Manuscript submitted July 14, 2017; accepted September 4, 2017.
doi: 10.17706/ijapm.2017.7.4.275-283

Abstract: Is a temporal landline phone call network graph series led by the presence of small world phenomenon? Are order and average vertex degree of the network graphs associated to small – world – ness? How are related size and order of the network graphs in this temporal series? A continuously graded notion of small – world – ness is used to study the presence of small world phenomenon. Spearman's and Kendall's correlation coefficients are used to perform a non – parametric correlation analysis between small – world – ness and order/average vertex degree. Linear regression on log – transformed quantities is used to analyse the relationship between size and order. It is achieved by the study that, the presence of small – world – ness is confirmed in each time step of the series, and there is no significant association between small – world – ness and graph order/average vertex degree. A significant positive power relationship between size and order is found.

Key words: Non–parametric correlation, small world, temporal network graph series.

1. Introduction

Different systems from real world are represented by network graphs with the purpose of topological studies [1]. A phone call network graph of a landline phone operator represents a system in which the vertex set is the set of the clients engaged in making or receiving calls, and the edge set is the set of communication relations between them. Scientific studies are concerned mainly in mobile network graphs [2]-[6] and those which analyze landline network graphs, considering dynamic evolution are rare. Network graph topology may vary over the observation time points. Temporal dynamics inferential studies in network graphs goes back to the 1970s and 1980s [7]-[11].

In most of the cases, topological relations in network graphs are supposed to be completely regular or completely random. Watts & Strogatz [12] showed that real world systems stand between these two topologies. Real systems are characterized by high level of clustering like regular lattices, and small geodesic distances like random graphs [13]. These network graphs are called by Watts & Strogatz as 'small world' in analogy with the phenomenon of 'small world' [14] known also as 'six degrees of separation' [15]. Association between different phenomena or processes has been a subject in many applied researches. Probabilistically speaking, it has to do with relationships between random variables. Practically, establishing the dependence or the independence, which is a property of the entire joint distribution of the random variables, is commonly replaced with establishing the correlation.

In this paper are tackled the following questions, and are aimed their answers: Is a temporal landline phone call network graph series led by the presence of small world phenomenon? Are order and average vertex degree of the network graphs associated to small – world – ness? How are related size and order of the network graphs in this temporal series? In this way 30 network graphs which were part of a temporal network graph series were constructed by splitting a phone call data records set for each day of a month. Time steps of the series are represented by each of the network graph. Small world phenomenon, in the case of a temporal landline phone call network graph series, is studied by applying a continuously graded notion of small – world – ness [16], [17] which led to a procedure for a general statistic test. Throughout, the association between small – world – ness versus the order and versus the average vertex degree of the network graph is studied by using a non – parametric correlation analysis. Linear regression on log – transformed quantities is used to analyse the relationship between size and order.

2. Material and Methods

2.1. Data Preparation

A landline phone call data records is provided by a local telecommunicating operator located in south of Albania. Client identities are substituted by numbers, with the purpose of conserving the privacy. The study is based only on phone calls inside the operator's clients' network, and not outside it. The reason of this restriction was that the information about phone numbers which are not operator clients would be incomplete. Phone calls belong to November 2014 and were in total 81591. From these, 41 phone calls which were without call durations and 7442 phone calls with a duration less than 10 seconds are excluded from the study. The reason of this exclusion is that these calls could be lost calls, or wrong calls and could affect the accuracy of the study. The total data set which was used to conduct the study was 90.83% of the initial data set. Active clients are considered only them that are engaged at least in one phone call (made or received) with a call duration at least 10 seconds.

The small world phenomenon in the communication system is studied by observing 30 network graphs which are constructed by splitting the data set for each day of the month. The network graphs are denoted by $G_i = (V_i, E_i)$, where the vertex set (active phone clients) is V_i and the edge set is E_i. Each edge represent a communication relation between two phone clients. Thus, if $v_1$and v_2 are vertices, then an undirected edge (v_1, v_2) is between them only if v_1 has received at least one phone call from v_2 or the reverse. Multiple relations between two vertices were simplified in only one edge. The temporal network graph series is $G_1, G_2, \dots, G_{30}$.The network graph G_i is constructed based only on the data of the ith day.

The statistical computation analyses in this study is conducted based in these packages: igraphdata [18], igraph [19], Kendall [20], ggpbur [21] in R statistical computation platform [22].

2.2. Basic Definitions

The order of a network graph is the number of vertices and the size is the number of edges on it. Let $< l_{G_i} >$ be the average geodesic distance [1] between vertex pairs in G_i. We distinguish two type of clustering coefficient, based from the local and global perspective of a not weighted and undirected network graph. The clustering coefficient from the local perspective (vertices) was defined initially from Watts & Strogatz [12], [23] and is denoted by

$$cl(v) = \frac{\tau_\Delta(v)}{\tau_3(v)}, \tag{1}$$

where $v \in V_i$ and $\tau_\Delta(v)$ is the number of triangles in G_i in which is included the vertex $v \in V_i$, while $\tau_3(v)$ is the number of connected triples of vertices in G_i, such that two of the edges are simultaneously incident with the vertex v. The clustering coefficient of G_i is computed with the formula

$$cl(G_i) = \frac{1}{|V_i|}\sum_{v\in V_i} cl(v). \tag{2}$$

In case that vertices had degree equal to zero or one, the clustering coefficient is taken $cl(v) = 0$ [1]. Transitivity of a network graph G_i is denoted by $cl_T(G_i)$ [23], [1] and considers the network graph as whole, from global perspective:

$$cl_T(G_i) = \frac{3\tau_\Delta(G_i)}{\tau_3(G_i)}, \tag{3}$$

where $\tau_\Delta(G_i) = \frac{1}{3}\sum_{v\in V_i}\tau_\Delta(v)$ is the number of triangles in G_i, and $\tau_3(G_i) = \sum_{v\in V_i}\tau_3(v)$ is the number of connected triples. Although $cl(G_i)$ and $cl_T(G_i)$ are seen both form the global perspective, they are not equal but often have approximated values.

Let G_i be a network graph of order $|V_i| = n$ and size $|E_i| = m$. The equivalent random network graph Erdös – Rényi [13] with same order and size is denoted by $E{-}R$, which is constructed uniformly, and each edge has the same probability. A semi categorical definition of small – world – ness is [12]:

Definition 1: A network graph G_i is a small world network graph if $< l_{G_i} > \geq < l_{E-R_i} >$ and $cl_T(G_i\,) \gg cl_T(E - R_i)$.

Definition 1′: A network graph G_i is a small world network graph if $< l_{G_i} > \geq < l_{E-R_i} >$ and $cl(G_i\,) \gg cl(E - R_i)$.

A new categorical definition was proposed to measure quantitatively the small world phenomenon in a network graph [16], [17]. Let

$$\lambda_{G_i} = \frac{<l_{G_i}>}{<l_{E-R_i}>}, \tag{4}$$

and

$$\gamma_{G_i}^T = \frac{cl_T(G_i)}{cl_T(E-R_i)}, \qquad S_i^T = \frac{\gamma_{G_i}^T}{\lambda_{G_i}}. \tag{5}$$

Similarly,

$$\gamma_{G_i} = \frac{cl(G_i\,)}{cl(E-R_i)}, \qquad S_i = \frac{\gamma_{G_i}}{\lambda_{G_i}}. \tag{6}$$

Definition 1 and 1′ imply that $\lambda_{G_i} \geq 1$ and $\gamma_{G_i}^T \gg 1$ ($\gamma_{G_i} \gg 1$), from which we have $S_i^T > 1$ ($S_i > 1$).

Definition 2: A network graph G_i will be a small world network graph if $S_i^T > 1$ ($S_i > 1$).

2.3. Small World Hypothesis Testing

A quantitative definition of small – world – ness is adopted, which led us to a procedure for a general statistic test for the presence of small – world structure as defined by Watts & Strogatz [16], [17]. An equivalent random network graph $E - R_i$ [13] with same order and size, is created for each of the network graphs G_i, from the series $G_1, G_2, \ldots, G_{30}$ and after that is computed S_i^T and S_i. To ensure the robustness of the categorization, network graphs are tested for significance using Monte Carlo sampling. The null hypothesis [12] is that:

H0: The system G_i is an Erdös – Rényi [13] random network graph.

Thus, for each of the network graphs G_i, are constructed M equivalent random network graphs, computing S_i^T and S_i related to each of the ith E-R network graphs. The 99% confidence limits for the null hypothesis are defined for each of the 30 days. The procedure is described as follow [24]:

1) Let $\hat{F}$ be the empirical distribution of the M data S_i^T (or S_i), i.e. the probability which puts mas $1/M$ at each data.
2) A random generator is used to draw M new points S_i^{T*} (or $S_i{}^*$) independently and with replacement from $\hat{F}$, so that each new point is an independent random selection of one of the M 'original' data points. These new points, which are called 'bootstrap sample', are a subset of original data points.
3) The mean $\overline{S_i^T}$ (or $\overline{S_\iota}$) value is computed for the bootstrap sample.
4) Steps 2 and 3 are repeated 10000 times, each time using an independent set of new random numbers to generate the new bootstrap sample. The resulting sequence of bootstrap mean values is $\overline{S_\iota^T}^1, \overline{S_\iota^T}^2, \dots, \overline{S_\iota^T}^{10000}$
5) Let $[a^*, b^*]$ be the central 68% interval for the $\overline{S_\iota^T}$values such that $\frac{\#\{S_i^T < a^*\}}{10000} = 0.16$ and $\frac{\#\{S_i^T < b^*\}}{10000} = 0.84$. The bootstrap estimate of the standard deviation σ, is $\hat{\sigma}^{(B)} = \frac{b^* - a^*}{2}$, based upon the fact that a normal distribution puts 68% of its probability within one standard deviation of the mean. Half of the length of the interval from 16th percentile to 84th percentile is a reasonable definition of the normal – theory estimate of standard deviation.

The upper 99% confidence limit is $CL^{0.01} = 1 + 2.58\hat{\sigma}^{(B)}$, where by definition $S^T = 1$ for an E – R network graph. A network graph with $S_i^T > CL^{0.01}$ is considered to significantly differ from a random network graph [16, 17] and the hypothesis H_0 is rejected. In a similar way is done the same for S_i.

2.4. Correlation Analysis

Correlation is a bivariate analysis that measures the strength and the direction of the association between two variables. Correlation reflects only some parts of the joint distribution. Correlation coefficient is a scalar measure of association between paired observations. Different correlation coefficients, leads to different results of correlation analysis and different interpretation results. Below we will refer to two non – parametric correlation coefficients: Spearman's (ρ), and Kendall's (τ). Let (X_i, Y_i), $i = 1, \dots, n$ be a paired sample, and (R_i^X, R_i^Y) be the corresponding ranks for the sample.

Definition 3: The Spearman's correlation [25], [26] coefficient is defined as the Pearson's correlation of ranks

$$\rho(X, Y) = r(R_i^X, R_i^Y). \tag{7}$$

Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data. It does not make any assumption about the data distribution. To test the significance of this coefficient we use this hypothesis test:

H_0: $\rho = 0$ and Ha: $\rho \neq 0$.

The total number of possible pairing of X with Y is $\frac{n(n-1)}{2}$, where the sample size is n. The procedure is as follow:

- Begin by ordering the pairs by X values. If X and Y are correlated, then they would have the same relative rank orders.
- Now, for each Y_i, count the number of $Y_j > Y_i$ (concordant pairs n_c) and the number of $Y_j < Y_i$ (discordant pairs n_d).

Definition 4: The Kendall's correlation [25], [26] coefficient is defined by:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}. \tag{8}$$

The Kendall's correlation coefficient is the difference between the probability that the observed data are in the same order, versus the probability that the observed data are no in the same order. To test the significance of this coefficient we use this hypothesis test:

H_0: $\tau = 0$ and H_a: $\tau \neq 0$.

Association between the small –world – ness and network graph order, and the association between the small – world – ness and the average vertex degree is studied by using Kendall's and Spearman's correlation coefficients. There will be a significantly correlation coefficient if $p \leq 0.05$ and if the value of $p \geq 0.20$, there is no correlation. More data are needed if $0.05 < p < 0.20$.

Furthermore, we are interested to see the relationship between the size and the order of the network graphs. For this, it is performed a linear regression on log – transformed quantities and is estimated the best fitting. It is assumed that hypothesis H_0 is true.

H_0: The residuals are normally distributed.

After that, it is controlled the hypothesis H_0 through the Shapiro –Wilk normality test [27]-[29]. If the p–value is less than the chosen alpha level 0.05, the hypotheses H_0 is rejected, and in this case there will be evidence that the data doesn't come from a normally distributed population.

3. Results

A data description of the temporal network graph series $G_1, G_2, \ldots, G_{30}$ related to the order $|V_i| = n_i$, size $|E_i| = m_i$, and average vertex degree $< d_i >$, is given in Table 1. Also the upper 99% confidence limit $CL^{0.01}$, is computed for each of the measures of small world phenomenon S_i^T and S_i.

Before applying the correlation analyses, with the purpose of studying the relationship between small – world – ness and network graph order and the average vertex degree, it is done a visual inspection of the data normality, which is given in Fig. 1. Q-Q plots draw the correlation between a given sample and the normal distribution.

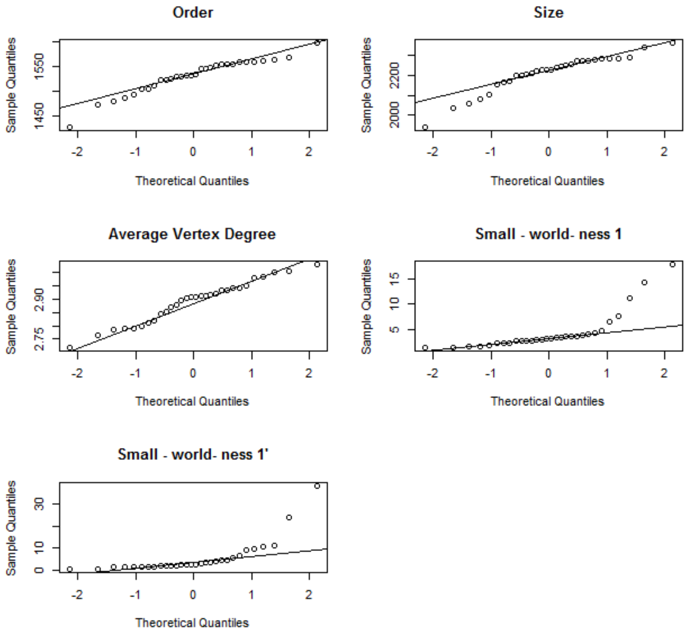

Fig. 1. Visual inspection of the data normality. "Small – world – ness 1" is referred to $S_i^T$and "Small – world – ness 1' " is referred to S_i.

From normality plots in Fig. 1 it is shown that there is no bivariate normal data distribution if we compare $S_T \sim n$, $S \sim n$, $S_T \sim < d_v >$, and S $\sim < d_v >$, and for this reason it is used a non – parametric correlation analysis. At first are constructed the scatter plots to visually see this relations, which are shown in Fig. 3, and after that are computed two type of non – parametric correlation coefficients, Kendall's τ and Spearman's ρ which are given in Table 2. The relationship between size and order of network graphs in the temporal series, the scatter plot and the best fitting is given in Fig. 2.

Table 1. Network Graph Topological Properties, Small – World – Ness Values and the Upper 99% Confidence Limit $\mathrm{CL}^{0.01}$. M is the Number of Simulated Equivalent E-R. Values of Measures Associated with (*) Are Rounded to Hundredths

i	$\|V_i\|$	$\|E_i\|$	$< d_v^i >*$	S_i^T*	$CL_T^{0.01}*$	S_i*	$CL^{0.01}*$	M
1	1597	2226	2.79	2.78	1.27	2.96	1.65	966
2	1428	1940	2.72	1.70	1.18	2.25	1.38	972
3	1561	2344	3.00	2.74	1.26	2.57	1.32	990
4	1534	2286	2.98	3.46	1.32	2.77	1.43	990
5	1522	2172	2.85	2.25	1.21	2.66	1.41	977
6	1530	2231	2.92	14.16	1.24	23.66	1.26	982
7	1531	2283	2.98	1.49	1.19	2.22	1.47	986
8	1560	2363	3.03	2.80	1.26	4.12	1.49	991
9	1487	2080	2.80	6.48	1.26	11.18	1.64	977
10	1564	2273	2.91	2.92	1.20	1.67	1.28	989
11	1545	2245	2.91	3.87	1.29	1.68	1.30	983
12	1531	2221	2.90	3.32	1.18	1.75	1.32	980
13	1547	2199	2.84	2.82	1.26	3.58	1.70	986
14	1552	2290	2.95	4.41	1.27	4.67	1.45	989
15	1555	2239	2.88	3.57	1.20	4.86	1.30	988
16	1554	2168	2.79	1.39	1.14	0.87	1.18	968
17	1555	2281	2.93	3.12	1.22	1.58	1.25	985
18	1560	2272	2.91	2.22	1.32	1.54	1.43	984
19	1494	2101	2.81	7.63	1.29	9.04	1.45	972
20	1568	2250	2.87	2.99	1.34	2.18	1.57	981
21	1523	2284	3.00	17.78	1.27	38.21	1.32	992
22	1505	2207	2.93	1.82	1.25	2.70	1.46	985
23	1479	2059	2.78	3.60	1.22	1.76	1.24	976
24	1512	2202	2.91	11.09	1.41	9.66	1.52	985
25	1559	2256	2.89	4.07	1.29	2.30	1.38	978
26	1545	2272	2.94	2.38	1.35	3.79	1.62	984
27	1530	2157	2.82	3.41	1.39	6.53	1.87	981
28	1505	2213	2.94	3.74	1.24	5.53	1.44	986
29	1524	2226	2.92	4.67	1.34	10.97	1.67	991
30	1472	2036	2.77	1.72	1.14	0.66	1.11	961

Table 2. Results of Non–parametric Correlation Tests

	Model	**Kendall**		**Spearman**	
		Tau	**p - value**	**Rho**	**p - value**
I	$S_T \sim n$	-0.1319476	0.3087	-0.1892253	0.3166
II	$S \sim n$	-0.196764	0.129	-0.2793856	0.1349
III	$S_T \sim < d_v >$	0.1034483	0.4358	0.1372636	0.4679
IV	S $\sim < d_v >$	0.1218391	0.3567	0.2378198	0.2049

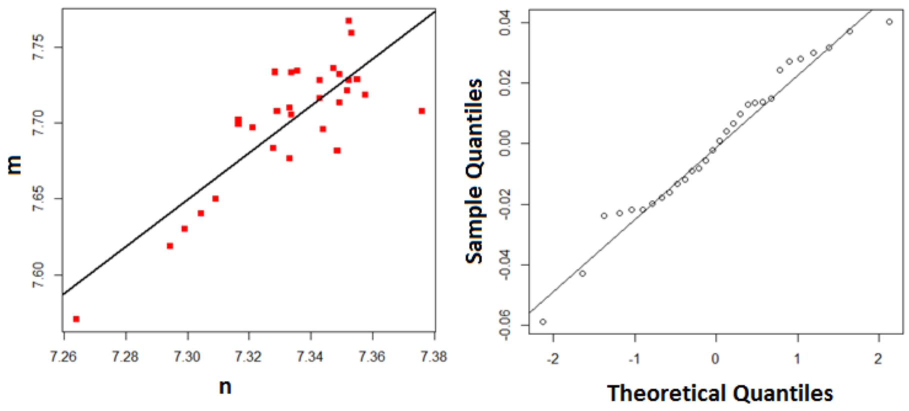

Fig. 2. The best fitting for the correlation between size and order of network graphs in log – log scale is $\mathrm{m} = \mathrm{e}^{-3.66}\,\mathrm{n}^{1.55}$ ($\mathrm{R}^2 = 0.69$). For the model $\log \mathrm{m} \sim \log \mathrm{n}$, the value of the Shapiro – Wilk Statistic on the residuals is W=0.9704 and the p- value is 0.5501. Normal Q-Q plots of the model $\log \mathrm{m} \sim \log \mathrm{n}$.

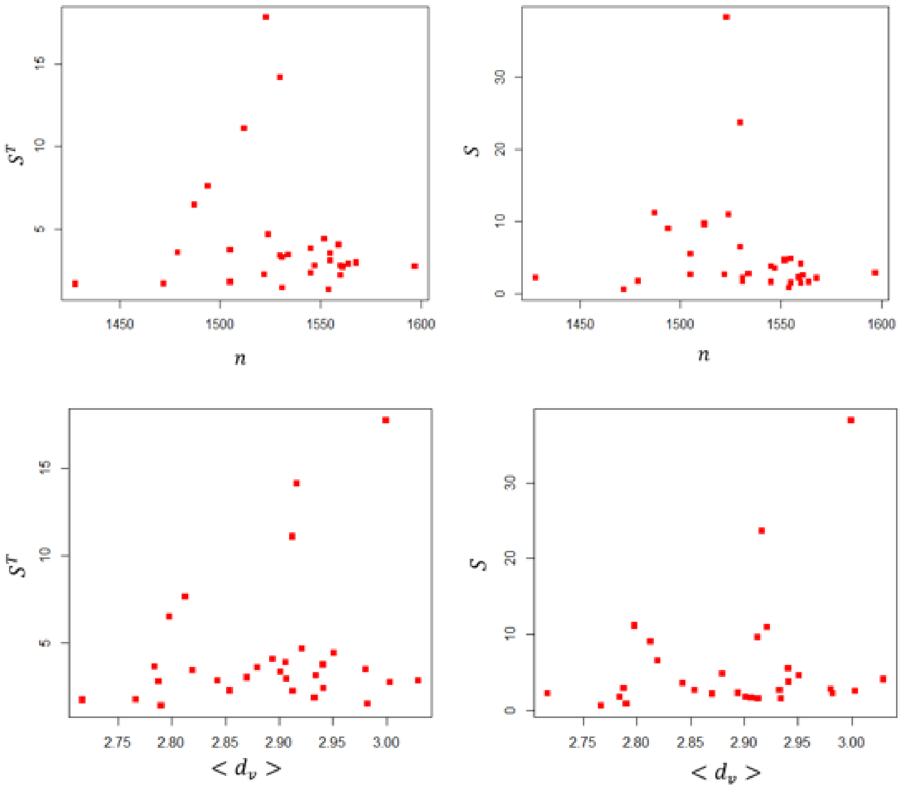

Fig. 3. Scatter plots of small – world – ness in relationship with the order of the network graph and in relationship with the average vertex degree of the network graph.

4. Conclusion

As a conclusion, 30 network graphs which were part of a temporal network graph series were constructed by splitting a phone call data records set for each day of a month. By applying a continuously graded notion of small – world – ness, the presence of small – world – ness is confirmed in each time step of the series. After computing Spearman's and Kendall's correlation coefficients, as part of a non – parametric correlation analysis between small – world – ness versus order, and between small – world – ness versus average vertex degree, is found that there is no significant association between them. Linear regression on log – transformed quantities is used to analyse the relationship between size and order, and a significant positive power relationship between them is found.

Acknowledgment

The first author thanks University “Ismail Qemali” of Vlora, Albania.

References

[1] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167-256.

[2] Dong, Z.-B., *et al.* (2009). An experimental study of large–scale mobile social network. *WWW2009*. Madrid, Spain.

[3] Seshadri, M., *et al.* (2008). Mobile call graphs: Beyond power-law and lognormal distributions. *KDD'08*. Las Vegas, Nevada, USA.

[4] Onnela, J.-P., *et al.* (2007). Structure and tie strengths in mobile communication networks. *PNAS*, *104(18)*, 7332-7336.

[5] Nanavati, A.-A., *et al.* (2006). On the structural properties of massive telecom call graphs. *CIKM'06*. Alington, Virginia, USA.

[6] Onnela, J.-P., *et al.* (2007). Analysis of a large-scale weighted network of one-to-one human communication.

[7] Holland, P.-W., & Leinhardt, S., *et al.* (1977). A dynamic model for social networks. *Journal of Mathematical Sociology*, *5(1)*, 5-20.

[8] Runger, G., & Wasserman, S. (1980). Longitudinal analysis of friendship networks. *Social Networks*, *2(2)*, 143-154,

[9] Wasserman, S., & Iacobucci, D. (1986). Statistical analysis of discrete relational data. *British Journal of Mathematical and Statistical Psychology*, *39(1)*, 41-46.

[10] Wasserman, S., & Iacobucci, D. (1988). Sequential social network data. *Psychometrica*, *53(2)*, 261-282.

[11] Iacobucci, D., & Wasserman, S. (1988). A general framework for statistical analysis of sequential dyadic interaction data. *Psychological Bulletin*, *103(3)*, 379.

[12] Watts, D.-J., & Strogatz, S.-H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440-442.

[13] Bollöbas, B. (2001). *Random Graphs*. Cambridge: Cambridge University Press.

[14] Milgram, S. The small-world problem. *Psychology Today*, *1(1)*, 61-67, 1967.

[15] Guare, J. (1990). *Six Degrees of Separation: A Play*. New York: Vintage Books.

[16] Humphries, M.-D., & Gurney, K. (2008). Network 'small-world-ness': A quantitative method for determining canonical network equivalence. *PLos ONE*, *3(4),* e0002051.

[17] Humphries, M.-D., Gurney, K., & Prescott, T.-J. (2006). The brain reticular formation is a small-world, not scale-free, network. *Proc. R. Soc. B*, *273*, 503–511.

[18] Csardi, G. (2015). Igraphdata: A collection of network data sets for the 'igraph' package.

[19] Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*.

[20] McLeod, A.-I. (2011). Kendall: Kendall rank correlation and Mann-Kendall trend test. *R Package Version 2.2,*

[21] Kassambara, A. (2017). Ggpubr: 'Ggplot2' based publication ready plots. *R Package Version 0.1.2*.

[22] Team, R. C. (2016). R: A Language and environment for statistical computing. Vienna, Austria.

[23] Kolaczyk, E. D. (2009). Descriptive analysis of network graph characteristics. *Statistical Analysis of Network Data*. Springer Science+Business Media.

[24] Efron, B. (1979). Computers and theory of statistics: Thinking the unthinkable. *SIAM Review*, *21*, 460-480.

[25] Sprent, P., & Smeeton, N.-C. (2001). Correlation and concordance. *Applied Nonparametric Statistical Methods*, 245-278.

[26] Hollander, M., & Wolfe, D.-A. (1973). Kendall and Spearman tests. *Nonparametric Statistical Methods*. New York, John Wiley & Sons.

[27] Royston, P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, *31(2)*, 115-124.

[28] Royston, P. (1982). Algorithm AS 181: The W test for normality. *Applied Statistics*, *31(2)*, 176–180.

[29] Royston, P. (1995). Remark AS R94: A remark on algorithm AS 181: The W test for normality. *Applied Statistics*, *44(4)*, 547–551.

Orgeta Gjermëni was born in Albania, 1986. She is graduated in mathematics from Faculty of Natural Sciences, University of Tirana Albania in 2009. She was part of the special group of mathematical students. Since 2013, she is a PhD candidate in mathematical engineering at the Faculty of Mathematical Engineering and Physical Engineering, Polytechnic University of Tirana, Albania. Her current work and studies are focused on real network graphs viewed from probability and statistical perspectives. Ms. Gjermëni is part of Department of Mathematics, Faculty of Technical Sciences – University "Ismail Qemali" of Vlora, Albania as a teaching assistant.

Miftar Ramosaço was born in Albania, 1967. He received his PhD in mathematics from University "Ismail Qemali" of Vlora Albania in 2015. He is currently a lecturer in the Department of Mathematics, Faculty of Technical Sciences, University "Ismail Qemali" of Vlora, Albania. His research interests are on applied mathematics and statistics. Mr. Ramosaço is author of some books.