Statistical Selection of Relevant Features to Classify Random, Scale Free and Exponential Networks[1]

Tania Turrubiates López[1,2], Claudia Gómez Santillán[1,3], Laura Cruz Reyes[2] Rogelio Ortega Izaguirre[3], Eustorgio Meza Conde[3]

[1] Instituto Tecnológico de Ciudad Madero (ITCM). 1ro. de Mayo y Sor Juana I. de la Cruz s/n CP. 89440, Tamaulipas, México.Teléfono: 01 833 3574820 Ext. 3024.
[2] Instituto Tecnológico Superior de Álamo Temapache (ITSAT). Carretera Potrero de Llano – Tuxpan Km. 6.5, Xoyotitla, Mpio. Alamo Temapache, Veracruz, México.Teléfono: 01 756 8440038
[3]Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada (CICATA). Carretera Tampico-Puerto Industrial Altamira, Km. 14.5. Altamira, Tamaulipas. Teléfono: 01 833 2600124.

tania_251179@hotmail.com, cggs71@hotmail.com, lcruzreyes@prodigy.net.mx, rortegai@ipn.mx, emezac@ipn.mx.

Abstract. In this paper a statistical selection of relevant features is presented. An experiment was designed to select relevant and not redundant features or characterization functions, which allow quantitatively discriminating among different types of complex networks. As well there exist researchers given to the task of classifying some networks of the real world through characterization functions inside a type of complex network, they do not give enough evidences of detailed analysis of the functions that allow to determine if all of them are necessary to carry out an efficient discrimination or which are better functions for discriminating. Our results show that with a reduced number of characterization functions such as the shortest path length, standard deviation of the degree, and local efficiency of the network can discriminate efficiently among the types of complex networks treated here.

1 Introduction

Any natural or artificial complex system of the real world as Internet, can be modeled as a complex network, where the nodes are elements of the system and the edges interactions between elements [1][2]. The term complex network refers to a graph with a non trivial topologic structure, this has motivated the study of topological characteristics of real networks [3][4] to identify characteristics that allow the discrimination among different types of complex networks and in this way optimize the performance of processes carried out in this networks as: search of distributed resources [1][5], traffic management and design of routing algorithms.

[1] This research was supported in part by CONACYT and DGEST

Advances in Artificial Intelligence and Applications
Research in Computer Science 32, 2007, pp. 185–192

Received 17/07/07
Accepted 31/08/07
Final version 23/09/07

Up to now the best way to identify the type of complex network has been observing the graphic of the degree distribution. This work presents a statistical selection of a set of characterization functions that allow quantitatively identified the type of complex network.

1.1 Characterization Functions of Complex Networks.

The characterization functions provide information about the topological characteristics of a complex network, when they are applied on a network, a vector of characteristic is obtained [6]. Through this vector, the behavior of the complex networks is characterized and analyzed. The characterization functions used in this work were: average degree of the network (Avg), standard deviation of the degree (Std), shortest path length (L), diameter (D), clustering coefficient (CG), global efficiency (EG) and local efficiency (EL), these characterization functions are described and detailed in [6].

One characterization function widely used to identify the type of complex network is the degree distribution which provides graphic information about the connectivity behavior of the nodes; through this function different types of networks as random networks, scale-free networks and exponential networks can be identified.

1.2 Three Types of Complex Networks.

Toward ends of the 50's and beginnings of 60's Erdös and Rényi focused in studying statistical properties in graphs where the connection among the nodes is established randomly, these graphs are called random networks and they are characterized for approaching a degree distribution binomial when the number of nodes n is small; when $n \to \infty$, the degree distribution approaches a Poisson distribution, in these graphs the nodes have approximately the same degree, nearby to the average degree of the network [7].

In the decade of the 90's diverse researchers [8][9], discovered networks of the real world as Internet exhibits a degree distribution following a power law distribution, these networks are called scale-free, because independently of the scale (number of nodes) the main characteristic of the network does not change, a reduced set of nodes have a very high degree and the rest of the nodes have a small degree [10][11].

Some natural networks exhibit an exponential distribution [12][13], where the majority of the nodes have a degree closer to the average degree and other nodes with a high degree can be observed, these networks are called exponential networks. To analyze the characteristics of these networks and to understand the phenomena carried out in them, generation models of complex networks have been created.

1.3 Generation Models of Complex Networks

The generation models of networks are an important tool to reproduce graphs sharing topological characteristics of networks of the real world making possible its study [11]. Some generation models are the Erdös-Rényi (ER) model [7] that reproduces random networks, the Barabási-Albert (BA) model [14] that reproduces scale free networks and the Liu model [15] that reproduces scale free and exponential networks.

2 Related Work

A recent application in the field of complex networks, is used information obtained through the characterization functions with the objective to discriminate among different types of complex networks, this application is related with the area of pattern recognition also known as classification. The classification of naturals and artificial structures modeled as complex networks implicate an important question: what characterization functions to select in order to discriminate among different types of complex networks [6].

In the classification area the selection of features (in this case, characterization functions) has great benefits: to improve the performance of classification procedure, and to construct simple and comprehensible classification models, these are achieved leaving aside, irrelevant and redundant characteristics that can introduce noise [16][17].

The work reported in [6], used a classification procedure to identify the type of a network with unknown nature, the results show that the type of network assigned to the networks, vary according to the characterization functions selected and an excessive number of characterization functions can compromise the quality of the classification.

An experimentation to evaluate the stability and separability of different types of networks is presented in [18], taking into account 6 topologies of networks to discriminate, 47 characterization functions are used as inputs to the classifiers. This work does not present evidences of a selection of relevant and not redundant characterization functions.

In [19], the objective is to determine from a set of models those describing more adequately networks that represent biological systems, a technique described in [20] was utilized for extracting characteristics serving as inputs to the classifier permitting to relate an instance generated by a model with a real instance. Though the results of classification obtained are good, they do not show if the characteristics extracted by the technique improve the performance of the classifier regarding the information provided by characterization functions extensively used in the field of the physics and the social sciences. In [21] networks in 3 types of topology are classified by means of a neural network; the eigenvalues of the adjacency matrix are utilized as inputs of the neural network.

3 Experimentation

Following a general scheme of procedures described in [22] an experiment was designed, to determinate, given a set of characterization functions, the functions permitting quantitatively discriminate among three types of complex networks: random networks, scale free networks and exponential networks.

In other words, 3 different populations are defined for each type of networks, from which topological characteristics are extracted, if the averages of those characteristics are significantly different and the populations do not overlap then those characteristics can help to distinguish networks of those populations. Three possible results of classifying networks can be observed:

- Case 1: There are significant differences among the 3 types of networks according to a set of characteristics, in this way through these characteristics a new network can be classified inside a type of network.
- Case 2: There exist significant differences among some of the types of networks according to a set of characteristics.
- Case 3: There does not exist significant differences among the 3 types of networks according to a set of characteristics, these two last cases lead to wrong classifications of new networks that need to be identified.

Instances of complex networks were generated to carry out the experimentation, through the models of E-R (random networks), BA (scale free networks) and Liu (exponential networks), with 200, 512 and 1024 nodes by each type of network; subsequently topologic characteristics were extracted such as: average degree (Avg), standard deviation of the degree (Std), clustering coefficient (CG), global efficiency (EG), local efficiency (EL), the shortest path length (L), diameter (D). The statistical packages MINITAB and SAS were utilized to study these characteristics.

In the related works, the networks have the same number of nodes and edges, so the average degree is equal for each type of network; in this experimentation the number of edges with the purpose of introducing variability to the experiment and carry to correct conclusions is not determined. In [18] it was observed that determining the number of edges aid to understand phenomena of interest but introduces dangers in statistical tests.

In this way, two factors that can influence in the process of characterization were identified; these factors can be controlled without affecting data normality: the type of network and the number of nodes. The factor, type of network, is composed of 3 levels, represented by the 3 types of networks being analyzed. The factor, number of nodes, is also composed of 3 levels, 200, 512 and 1024 nodes. Giving the characteristic of the problem a two-factor factorial design was chosen. The significance level α was set in 0.05.

The operation characteristics curves was used to determine the number of instances of networks n, appropriate for detecting significant differences. Through this procedure it was determined that with n = 4 instances a probability of 98% is obtained to detect differences up to 0.01 this procedure can be found in [22]. Due that this experiment was extended to the classification, n = 4 was taken as the minimum number of instances to generate; n = 30 instances was set for the experimental design.

According to the fixed effects model describing to a two-factor factorial design [22] a hypothesis were formulated to detect significant differences. A Multivariate Analysis of Variance General (GML) was carried out to obtain results permitting to reject or to accept the hypothesis, these results are discussed subsequently.

Table 1. F_0 values calculated by GLM for each variable.

	Avg	Std	CG	EG	EL	L	D
Type of network	178.47	124.39	220.18	507.82	55.21	566.56	523.72
Number of nodes	32.80	50.45	1.70	10.50	1.07	39.11	21.07
Interaction	32.70	20.87	0.56	3.57	0.18	13.90	10.01

The values showed in Table 1, were compared with the criteria for rejection associate to the hypothesis formulated, it can be observed that according to the type of network, the characterization functions satisfy the criteria for rejection, therefore, it can be concluded that all the characterization functions differ significantly according to the type of network.

For the effect of the number of nodes, and the effect of the interaction between the type of network and number of nodes, the criteria for rejection are satisfied for Avg, Std, EG, L and D, therefore, it can be concluded that these characterization functions differ significantly according to the number of nodes presented in the network, and the interaction between the type of network and the amount of nodes.

The analysis of the residuals of Avg and CG, showed abnormalities, this is because in spite of the fact that the number of edges was not determined, Avg and CG for scale free and exponential networks, were defined by initial parameters employed by the models of generation, on the other hand the plots of Std, EL, EG, L, D showed normality.

MANOVA tests shown in Table 2, for the type of network, the quantity of nodes and the interaction, are statistically significant, by which the significant differences detected in the analysis of a variable at one time are reaffirmed, being real and not false positive. Due that the values of F0 are greater for factor of type of network, it is concluded that this factor has greater influence in the values obtained by the characterization functions.

Table 2. F_0-values calculated for the MANOVA test and values of F distribution.

MANOVA Test	Type of network		Number of nodes		Interaction	
	F_0	F	F_0	F	F_0	F
Wilks'	515.91	1.23	88.62	1.23	32.97	1.17
Pilliai's	391.57	1.23	34.52	1.23	17.04	1.17
Lawley–Hotelling	676.26	1.23	183.88	1.23	66.26	1.17
Roy's	1137.97	1.29	367.89	1.29	250.28	1.29

Once it was detected that the characterization functions differ significantly according to the type of network and this factor has greater effect, using the test of

Tukey, multiple comparisons were carried out, with the objective of detecting which characterization functions differ significantly in each level of the factor of type of network.

Tukey's tests carried out showed the means of the characterization functions Avg, EL, EG and CG are significantly equal for scale free and exponential networks, and for the random networks the mean is significantly greater. For the function Std, the means of the scale free and random networks are significantly equal and for the exponential networks it is significantly smaller. The characterizations functions where significant differences in the means of the random, scale free and exponential networks were observed are L and D. Then it can be conclude, according with the cases mentioned above, that the Avg, Std, EL, EG, and CG functions are weak relevant characteristics; the L and D functions are very relevant characteristics.

The characterization functions selected to discriminate among different types of networks, were L, Std and EL, in this way a characterization function of each group representing the cases 1 and 2 were taken into account. The Avg and CG were not taken into account because of the abnormalities presented in the plots of the residuals, thus D and EG also were excluded because of being highly correlated with L, which indicates redundancy.

To obtain an efficient classifier of networks, a quadratic discriminant analysis was carried out with different combinations of the characterization functions selected, using in both cases the cross-validation procedure. In Table 3, the percentages of networks classified correctly inside each type of network, are shown and the total percentage of correct classifications.

Table 3. Results of the Quadratic Discriminant Analysis.

Quadratic Discriminant Analysis				
Combination of variables	Random Networks	Scale Free Networks	Exponential Networks	Total
L	97.8% (88 / 90)	78.9% (71 / 90)	50.0% (45 / 90)	75.6%
L, Std	97.8% (88 / 90)	88.9% (79 / 90)	95.6% (84 / 90)	94.1%
L, EL	97.8% (88 / 90)	72.2% (63 / 90)	66.7% (60 / 90)	78.9%
L, Std, EL	98.9% (89 / 90)	100% (90 / 90)	100% (90 / 90)	99.6%

4 Conclusions and Future Work

An experimental design was presented for the purpose of selecting statistically characterization functions of complex networks relevant and not redundant, the results obtained in the experimentation show that the shortest path length, the standard deviation and the local efficiency, permit in a quantitative way to discriminate among random networks, scale free networks and exponential networks. By means of the quadratic discriminant analysis an accuracy of classification of the 99.6% was obtained, the size of the sample was justified statistically. As it can be appreciated in this work the number of characterization functions is very small in relation to the 47 functions utilized in [17], this reduces dramatically the time of computation required for the classification of complex networks.

In future works it is recommended to extend this experimental design for networks with greater amount of nodes and other types of networks, besides it is recommended to work with procedures and algorithms for selection of characteristics utilized in the area of machine learning in order to compare the results obtained with this experimental design, and work with real instances of graphs as Internet.

5 References

[1] V. Latora, and M. Marchiori, "Efficient Behavior of Small World Networks", Physical Review Letters, 2001, pp. 198701-1 – 198701-4

[2] A. L. Barabási, R. Albert and H. Jeong, "Mean-Field Theory for Scale-free Random Networks", Physica A, 1999, pp. 173-189.

[3] S. Maslov, K. Sneppen, A. Zaliznyak, "Detection of Topological Patterns in Complex Networks: Correlation Profile of the Internet". Physica A, 2004, pp. 529-540.

[4] R. Albert and A. L. Barabási, "Statistical Mechanics of Complex Networks", Reviews of Modern Physics, 2002, pp. 47-97.

[5] L. A. Adamic, R. M. Lukose, A. R. Puniyani and B. A. Huberman, "Search in power law network", Physical Review E, 2001, pp. 046135-1 - 046135-8.

[6] L. Costa, F. A. Rodrigues, G. Travieso and P. R. Villas, "Characterization of Complex Networks: A survey of measurements." arXiv:cond-mat/0505185v5, 2007.

[7] B. Bollobás and O. M. Riordan, "Mathematical results on scale-free random graphs", Handbook of Graphs and Networks, Wiley-VCH, Berlin, 2002, pp. 1-32.

[8] R. Albert, H. Jeong, and A. L Barabási, "Error and attack tolerance of complex networks", Nature, 2000, pp. 5234-5237.

[9] M. Faloutsos, P. Faloutsos and C. Faloutsos, "On power-law relationship on the internet topology", ACM SIGCOMM Computer Communication Review, 1999, pp. 251-262.

[10] A. L. Barabási, "Emergence of Scaling in Complex Networks", Handbook of Graphs and Networks, Wiley-VCH, Berlin, 2003, pp. 69-82.

[11] M. E. J. Newman, "The structure and function of complex networks", SIAM Review, 2003, pp. 167-256.

[12] L. A. N. Amaral, A. Scala, M. Barthelemy and H.E. Stanley, "Classes of small world networks", PNAS, 2000, pp. 11149-11152.

[13] P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna, "Small-world properties of the Indian Railway network", Physical Review E, 2003.

[14] A. L. Barabási and R. Albert, "Emergence of Scaling in Random Networks", Science, 1999, pp. 509-512.

[15] Z. Liu, Y. Lai, N. Ye and P. Dasgupta, "Connectivity distribution and attack tolerance of general networks with both preferential and random attachments", Physics Letters A, 2003, pp. 337- 344.

[16] L. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research, 2003, pp. 1157 - 1182.

[17] S. K. Singhi and H. Liu, "Feature Subset Selection Bias for Classification Learning", Proceedings of the 23rd ICML, 2006, pp. 849 – 856.

[18] E. M. Airoldi and K. M. Carley, "Sampling algorithms for pure network topologies: a study on the stability and the separability of metric embeddings" ACM SIGKDD Explorations Newsletter, 2005, pp. 13 - 22.

[19] M. Middendorf, E. Ziv, C. Adams, J. Hom, R. Koytcheff, C. Levovitz, G. Woods, L. Chen, and C. Wiggins, "Discriminative Topological Features Reveal Biological Network Mechanisms", BMC Bioinformatics 2004, 2004.

[20] E. Ziv, R. Koytcheff, M. Middendrof, and C. Wiggins, "Systematic Identification of statistically significant network measures. http://arxiv.org/PS_cache/cond-mat/pdf/0306/030 6610v3.pdf, 2005.
[21] W. Ali, R. J. Mondragón, and F. Alavi, "Extraction of topological features from communication network topological patterns using self-organizing feature maps", http://arxiv.org/abs/cs.NE/0404042, 2004.
[22] Montgomery, D.C. Diseño y Análisis de Experimentos. Limusa Wiley, 2004.