International Journal of Scientific Research in Mathematical and Statistical Sciences **Research Paper**

Volume-5, Issue-4, pp.100-104 August (2018) **E-ISSN:** 2348-4519

An Analysis of Email-Eu-Core Network

A. Bharali

Dept. of Mathematics, Dibrugarh University, Dibrugarh, India

Corresponding Author: a.bharali@dibru.ac.in

Available online at: www.isroset.org

Accepted 14/Aug/2018, Online 30/Aug/2018

Abstract: In this paper a complex network approach is adopted to analyze the Email-Eu-Core Network. This communication network displays small-world (SW) properties with an average path length of 2.0834 and a clustering coefficient of 0.372, and it also exhibits assortative mixing on degree of nodes, that means the high-degree nodes in the network tend to have connections with high-degree nodes. Being an assortative network, it does percolate more easily but less robust to targeted node removal. A study of the robustness of Email-Eu-Core network is also carried out for random and targeted node failures.

Keywords: Network measures, Complex Network, Email-Eu Network, Small-world, Assortative mixing, Scale-free Network.
2010 MSC: 05C82, 90B06, 91B74.

I. INTRODUCTION

In recent times, Networks associated with various disciplines are studied alongwith their statistical properties, and based on these statistical properties those networks can be characterized. Moreover, several network models are proposed [2][14], which can be very helpful in understanding the statistical properties. Communication network is one of the four major classes of complex networks [12]. An example of communication network is the Internet, which has been studied as a real-world complex network.

The Email-Eu network was generated using e-mail data from a large European research institution. Information about all incoming and outgoing e-mails between members of the research institution were anonymized. The members represent the nodes of the network and existence of an edge between a pair of nodes indicates that at least one e-mail in exchanged between the pair of members. The e-mails only represent communication between institution members (the core), and the dataset does not contain incoming messages from or outgoing messages to the rest of the world [9]. An efficient communication network would have small characteristic path length, high clustering coefficient, which are the properties of small-world network (Watts and Strogatz, 1998) [14]. A small-world network is a graph in which most vertices are not neighbors of one another, but most vertices can be reached from every other by a small number of hops, attributing to its small characteristics path length. Another common property of many large real networks is that, the vertex distribution follows power-law (Barabasi and Albert, 1999) [2]. This kind of networks is known as scale-free (SC) networks.

Different Communication networks have been studied as complex network perspective to analyze their topological structures and functionalities [12][13][15]. Most of the networks are found to have small average path length (~log N, N is the number of nodes) and high clustering coefficients. Communication Networks are modeled, analyzed and characterized as complex networks using different network parameters. Links between some structural properties of networks, such as degree distribution, average path length, and betweenness, with communication network performance are also established. A detailed of such study of communication network as complex network with special reference to wired and fixed packet switching network may be found in [16].

Robustness is the ability of a network to perform correctly when it is subject to some failures or attacks [4]. In other words, robustness is all about back-up possibility or simply alternative paths (Ellen et al., 2013)[3]. Robustness is one of the most anticipated property of any communication network and the level of vulnerability of a network can be evaluated by studying the responses of it subjected to different nodes and edges failure scenarios e.g., random, most central nodes and link failures [1]. Attack Vulnerability of Complex Communication Networks is a very popular and demanding topic that has received considerable attention during the past decade [17]. Another important property that a communication network is expected to exhibit is assortative mixing on their degree of nodes. A network is said to exhibit assortative mixing if the nodes in the network that have many connections (high-degree) tend to be connected to other nodes with many connections (high-degree) [10][11].

In this paper, we analyze the structural and topological properties of Email-Eu-Core. We calculate different network measures like the diameter, average path length, clustering coefficient etc., to study the structure and different centrality measures, to find out key members in the network. The coefficient of mixing is also calculated to determine the nature of mixing exhibited by the network. A study on robustness of the network is also carried out for some key node failures.

The rest of the paper is organized as follows: in the next section, we discuss some measures used in complex network analysis. In section 3, we present the description and a visualization of Email-Eu-Core network and calculate the values of different metrics to study the network structure. In this section we also discuss the nature of mixing on degree of nodes of the network. In section 4, we present a preliminary study of robustness of the Email-Eu-Core network for random and targeted failure by removing key nodes from the network. Conclusions are presented in section 5.

II. SOME NETWORK MEASURES

In this section, we present the definitions of some measures used in network analysis.

Shortest path: There exists a set of paths between any given pair of nodes. The shortest path of a network is the path that has the lowest number of hops between the source and destination pair in the network.

Diameter: Diameter of a network is the longest shortest path between any pair of nodes in the network. If L_{ij} is the shortest path between nodes i and j, then Diameter,

$$d = \max_{i,j\in V} L_{ij}.$$

Average shortest path length: The Average shortest path length (L), also known as the characteristic path length, is defined as

$$L = \frac{1}{N(N-1)} \sum_{i,j=1,\ i\neq j}^{N} L_{ij},$$

where L_{ij} is the shortest path between the vertex i and j and N is the total number of vertices in the network. For a random network of size N and $< k >$, it is $\frac{\log(N)}{\log(<k>)}$.

Network clustering coefficient [3]: In a network, if node A is connected to node B and node B is connected to node P, then there is a intensify probability that node A will also be connected to node P. The clustering coefficient of a network is defined as:

$$\mathrm{C} = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}},$$

where a 'connected triple' means a node with edges running to an unordered pair of others. Clustering coefficient is also known as network transitivity. For a random network of size N and $< k >$, it is $\frac{<k>}{N}$.

Coefficient of mixing [10]: A network is said to be assortative if the high-degree nodes in the network tend to have connections with other high-degree nodes, otherwise the network is called disassortative. For example, if people prefer to link with others who are like them (such as similar social norms, language, culture etc. are to name a few), then the network shows assortative mixing or assortative matching and if they prefer to link with those who are different, it shows disassortative mixing.

Betweenness [3]: The Betweenness of a node is the number of shortest path going through the node. Similarly we can define edge betweenness.

Network Efficiency [8]: The efficiency in a network between any two vertices is inversely proportional to the shortest distance between the vertices.

Graph density [5]: The ratio of the actual number of edges M to the maximum number of possible edges is known as Graph density. It is usually denoted by *D.*

Reachability [7]: Reachability of a network is the probability of the connectivity between any pair of its nodes, (u, v), which is represented by R, and the reachability of node R_i is calculated as follows:

$$R_i = \frac{\textit{number of nodes reachable from i except itself}}{N-1}.$$

The Reachability (R) of the whole network is the average of all R_i. Clearly if the network is a fully reachable network, then R = 1and if it has an isolated component then it is always 0.

III. NETWORK DESCRIPTIONS AND ANALYSIS

The network under consideration represents the "Core" of the email-Eu network [9], which contains links between members of the institution and people outside of the institution. But in the Email-EU-Core network the e-mails only represent communication between institution members (the core), and the incoming messages from or outgoing messages to the rest of the world are not considered in the dataset. Each individual belongs to exactly one of 42 departments at the research institute [9] [18].

3.1 Structure and Visualization of Email-Eu-Core Network

We compute the average path length and network clustering coefficient using *MATLAB BGL* 4.0.1. The average path length of Email-Eu-Core Network is found to be 2.0834 and the network clustering coefficient is 0.372. The average path length of the network is less than that of the random network of same size, and similarly the clustering coefficient is much higher. So we can say that it shows small-world (SW) network properties. The diameter of Email-Eu-Core network is found to be 7, which implies that an e-mail can reach from any member to any other member in the network in not more than 7 hops.

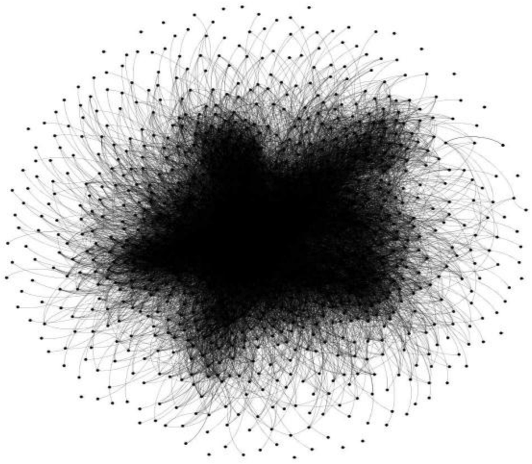

Figure 1: Email-Eu-Core Network

We also compute the network density, closeness and efficiency using *KONECT* tool box, a tool box for the *Matlab* programming language for analyzing large networks. The efficiency of the network is found to be 0.1726, which is much higher than its density. Since, the efficiency of a random network is almost equal to the rewiring probability of the network. So by comparing the network for its efficiency with a random network whose rewiring probability is equal to the density of the network, we can conclude that it is not a poorly connected network, as per the efficiency is concerned. All the calculated network measures are listed in Table 1. We start our analysis by providing a visualization of the network in Figure 1. There we generate the nodes of the network and the links between them. The visualizations are rendered via an implementation of the *Fruchterman-Reingold* (FR) force directed algorithm [6], which tends to create highly clustered “cores”.

Network Measure	Value
Number of nodes(Members)	1005
Links	25571
Connected Components(Weakly)	10
Diameter	7
Average degree	25.44
Average path length	2.0834
Network Clustering Coefficient	0.372
Average Closeness	0.0023
Network Efficiency	0.1726
Network Density	.025
Network Reachability	.7863

Table 1: Metrics of Email-Eu-Core Network

Rank	In-degree	Out-degree	Betweenness	Eigenvector	Closeness
1	#161	#161	#161	#161	#845
2	#63	#83	#87	#108	#996
3	#108	#122	#6	#63	#61
4	#122	#108	#122	#435	#83
5	#87	#87	#63	#122	#122
6	#435	#63	#108	#184	#108
7	#184	#14	#65	#129	#87
8	#130	#250	#83	#257	#63
9	#65	#184	#378	#130	#250
10	#129	#435	#130	#250	#435

Table 2: Most important nodes based on different centrality measures

3.2 Degree distribution and Mixing on degree of nodes

The cumulative degree distribution of Email-Eu-Core network follows power law-regime as shown in Figure 2 with a scaling exponent of 1.312, for a wide range of degrees, although it deviates from it for some degrees. It can be explained by the fact that a member who receives more e-mails needs to reply more and e-mails are most exchanged between friends (equi-important members).

The value of the coefficient of mixing of the network is found to be positive (0.0131), so the network is assortative in nature on degree of nodes i.e., members with more connections tend to connect with the members with more number of connections. The topological assortativity can be explained by the preferential attachment model and hence it percolates easily.

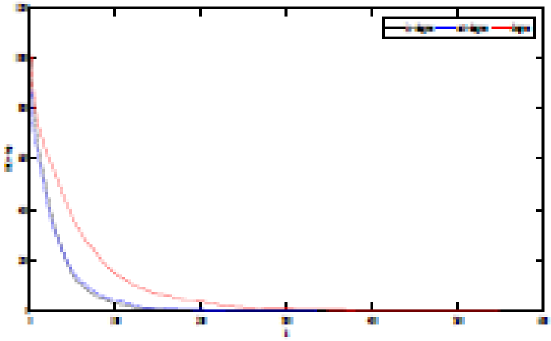

Figure 2: Scale free property of degree distribution of Email-Eu-Core network.

IV. ROBUSTNESS ANALYSIS OF EMAIL-EU-CORE NETWORK

The robustness of Email-Eu-Core network is studied for both targeted attacks and random failures. At the very outset we calculate different node centrality measures associated with the network to find out the important nodes, the top 10 nodes are listed in the Table 2. A high in-degree means the member receives plenty of e-mails, similarly a high out-degree corresponds to a member who sends plenty of e-mail. Betweenness centrality can capture those members that are in the center of the network but not necessarily those with higher authorities. Closeness centrality is a measure of the degree to which a member is near to all the members in the network. Eigenvector centrality measures the recursive influence of a node in a network. If a member has fewer connections but the connections are themselves high in degree then the member may have a higher value of eigenvector centrality. So from Table 2 we can conclude that the member #161 receives and sends highest number of e-mails lies in the center of the network and also is the most influential member in the network. But it is not close to all the members of network; in fact he or she is not in the top 10 list of members who is close to all the members. #161 is an ideal candidate for the chief of an institution who has the most influence in the institution but he or she also may not be accessible to all the members. #6 is a member who is ranked 3 based on betweenness but not there in any of the other lists. He or she is a member who acts as the middle man in all the communications.

Node removed	Number of edges	$\Delta L\%$	$\Delta R\%$	$\Delta E\%$
#161	25026	1.66	-0.242	-1.275
#122	25193	-0.24	-0.331	-0.695
#108	25199	-0.74	-0.458	-0.637
#63	25203	0.42	-0.025	-0.290
#87	25216	1.31	-0.025	-0.579
#83	25224	0.10	-0.127	-0.348
#435	25264	0.29	-0.025	-0.174
#184	25270	0.32	-0.025	-0.174
#6	25292	-1.29	-0.814	-0.869
#130	25297	0.01	-0.140	-0.174
#65	25323	0.50	-0.025	-0.232
#378	25353	-1.33	-0.661	-0.985

Table 3: Change of network measures after removal of a key node based on degree & betweenness.

Then we analyze the effect of targeted attack on the network by removing ten most important nodes (members) from the network and observe the changes in different centrality measures like the average path length, network efficiency and reachability for the remaining network. The values are noted in the Table 3. We also calculate these measures after removal of the top 5 nodes orderly, based on degree and betweenness. A positive sign indicates an increment in that parameter whereas a negative sign suggests a decrement. We also analyze the robustness of the network for random failures. In this case we remove nodes from the network in an increasing order from 5%-99% of the total nodes of the network. A graphical representation of the change of values of different network measures is shown in Figure 3. In Figure 3 we observe that there is very slight change in the values of the network parameter even after removal of 40-50% of the total nodes. And as expected we observe a sudden drop in the values of the network parameters after removal of 90% of its nodes. From this observation we may conclude that the network is robust towards random failures. We choose the nodes for removal randomly and an average of 5 observations is considered to minimize the effect of selection of nodes on the values of the network parameters. But in Table 3 we observe that removal of key nodes can affect the network parameters. The removal of the highest node (#161) based on degree and betweenness results in significant change of network parameters. There is a clear drop in the values of the parameters like network efficiency and reachability whereas we observe an increment in the values of the average path length. So we can conclude that in absence of the node (#161) the performance of the communication network decreases. It is also true in case of the failure of the other important nodes. From Table 4 we can have an idea about the cumulative effect of the failure of these important nodes together. We observe that the removal of the top five nodes reduces the efficiency of the network by more than 4% and reachability is dropped by almost 2%. The average degree of the nodes is also dropped by more than 7%.

Node removed	Number of edges	$\Delta L\%$	$\Delta R\%$	$\Delta E\%$
1	25026	1.66	-0.242	-1.275
2	24671	0.031	-0.279	-1.912
3	24392	1.04	-1.335	-2.955
4	24017	0.889	-1.679	-3.708
5	23651	1.53	-1.704	-4.056

Table 4: Change of network measures after removal of top 5 nodes orderly based on betweenness.

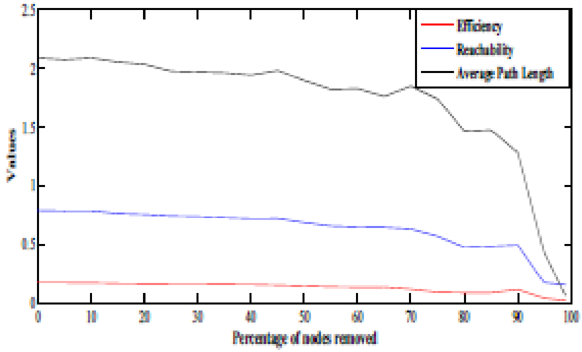

Figure 3: Changes in network measures with the removal of nodes.

V. CONCLUSIONS

In this paper we analyze Email-Eu-Core network as a complex network. The network is constructed by considering each member as a node and a link is considered for each pair of directly connecting members via e-mails. The results show that the network has scale-free distribution that indicates there are some high-degree nodes (called hubs) in the network. The traffic (e-mails) of the network is found to be clustered on an interconnected group of high degree nodes.

The average path length of Email-Eu-Core network is found to be smaller and the clustering coefficient is found to be much higher than that of a random network of same size. This indicates that the network exhibits SW property and is characterized by power law-regime, with a decrement in the scaling exponent with time. The value of the coefficient of mixing is positive, which suggests that network is assortative in nature on degree of nodes. The topological assortativity is balanced by the traffic dynamics as the traffic is concentrated between high-degree nodes and hence it does percolate easily.

One of the important goals for analyzing these communication networks is to understand the organizational behavior of the institutions. Since a significant drop in reachability is observed only when a very high-degree node (hub) is removed, one can expect that the impact of the spread on the network is higher when the hubs are infected first than when random nodes are infected. Finally, we can summarize the findings as follows: Email-Eu-Core network is a Small-World network that exhibits power law-regime with an assortative mixing pattern on the degree of nodes. The network is robust against random failures but vulnerable to targeted attacks.

ACKNOWLEDGMENT

The author is thankful to the anonymous reviewers for their valuable suggestions in improving the manuscript.

REFERENCES

[1] Albert, R., Jeong, H., & Barabasi, A., (2000) Error and attack tolerance of complex networks. Nature, 406, 378-382.

[2] Barabasi A., Albert R., (1999) Emergence of Scaling in Random Networks. Science, 286, 509-512.

[3] Ellens W., et. al., (2013) Graph Measures and Network Robustness. arXiv preprint arXiv:1311.5064.

[4] Gribble, S., (2001), Robustness in complex systems. Proceedings of the 8th Workshop on Hot Topics in Operation Systems (HotOS-VIII).

[5] Faust, K., (2006) Comparing Social Networks: Size, Density, and Local Structure. Metodolo˜ski zvezki, 3(2), 185-216.

[6] Fruchterman, ThomasM. J., Reingold, EdwardM., (1991) Graph Drawing by Force-Directed

Placement. Journal of Software-Practice and Experience, 21 (11), 1129-1164.

[7] Hossain M., et. al., (2013) Australian Airport Network Robustness Analysis: A Complex Network Approach. Proceedings of Australasian Transport Research Forum, 1-21.

[8] Latora, V., et. al., (2001) Efficient Behavior of Small-World Networks. Phys. Rev. Lett., 87: 198701.

[9] Leskovec, J., Krevil, A., SNAP Datasets: Large Network Dataset Collection. http://snap.stanford.edu/data.

[10] Newman, M. E. J., (2002) Assortative mixing in networks. Phys. Rev. E, 89(20): 208701.

[11] Newman, M. E. J., (2003) Mixing patterns in networks. Phys. Rev. E, 67(2): 026126.

[12] Newman, M. E. J., (2003) The structure and function of complex networks. SIAM Rev., 45(2), 167-256.

[13] Wang, X., et. al., (2003) Complex networks: small-world, scale-free and beyond. IEEE Circuits and Systems Magazine, 3(1), 6-20.

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

[15] Wu, J., et. al., (2013) Analysis of Communication Network Performance From a Complex Network Perspective. IEEE Transactions on Circuits and Systems I: Regular Papers, 60(12), 3303-3316.

[16] Wu, J., (2014) Study of Communication Network Performance From A Complex Network Perspective. A Doctoral Thesis (The Hong Kong Polytechnic University), http://ira.lib.polyu.edu.hk/handle/10397/7429.

[17] Xia, J., et. al., (2008) Attack Vulnerability of Complex Communication Networks. IEEE Transactions on Circuits and Systems II: Express Briefs, 55(1), 65-69.

[18] Yin, H., et. al., (2017) Local Higher-order Graph Clustering. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

AUTHORS PROFILE

Dr A. Bharali obtained his master degree in Maths and Computing from IIT Guwahati and his PhD in Mathematics from Dibrugarh University. He is currently working in the Dept. of Mathematics, Dibrugarh University since 2009. He has published many research papers in reputed international journals and also authored a book. His research interest includes Graph Theory and Complex Networks. He has almost 10 years of teaching and research experience.