2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Are all Social Networks Structurally Similar?

Aneeq Hashmi
National University of Science and Technology
Karachi, Pakistan
Email: aneeqhashmi@yahoo.com

Faraz Zaidi
Karachi Institute of Economics and Technology
Karachi, Pakistan
Email: faraz@pafkiet.edu.pk

Arnaud Sallaberry
University of California,
Davis, USA
Email: asallaberry@ucdavis.edu

Tariq Mehmood
National University of Science and Technology
Karachi, Pakistan
Email: tariq.mehmood@nu.edu.pk

***Abstract*—The modern age has seen an exponential growth of social network data available on the web. Analysis of these networks reveal important structural information about these networks in particular and about our societies in general. More often than not, analysis of these networks is concerned in identifying similarities among social networks and how they are different from other networks such as protein interaction networks, computer networks and food web.**

In this paper, our objective is to perform a critical analysis of different social networks using structural metrics in an effort to highlight their similarities and differences. We use five different social network datasets which are contextually and semantically different from each other. We then analyze these networks using a number of different network statistics and metrics. Our results show that although these social networks have been constructed from different contexts, they are structurally similar.

I. Introduction

The web has provided a platform to build huge social networking webistes [16] and communication channels with hundreds and thousands of users. These networks provide challenging opportunities for researchers to analyze and explore how virtual societies exist in the cyberworld and how they impact our societies in the real world [1]. Moreover many useful applications for these online networks have been found both in business domain and in social relations. Business applications include information diffusion [13] and corporate communication [26], and social relations include searching individuals of similar interest, establishing discussion forums and exchanging information [2] with friends and family members distantly located.

Often these social networks are compared to other networks such as protein interaction networks [8] and computer networks [28]. For example, Newman studied the property of assortativity [22] only present in social networks where individuals of similar degree have the tendency to connect to each other. Another dimension is to study how these online social networks are similar to real world social networks [12]. Not much attention has been given to the differences and similarities of contextually and semantically different online social networks.

Semantics and *Context* refer to how social relations are created among individuals such as, direct communication through an email, personal liking of a photograph, or being part of a common group or community. These different forms of social networks [24] raises the question of whether different social networks have the same network structure or are they structurally different.

In this paper, we address this question and try to answer it empirically. We use five different social network datasets and compare them using different network statistics and metrics. Our results show high similarity among structural behavior of these networks with only slight differences.

The rest of the paper is organized as follows: In the next section, we review the literature where online social networks have been analyzed. Section III describes the data sets used for experimentation. In sections IV and V, we review a number of network statistics used for comparative study. Section VI describes how the samples were collected and we analyze different networks in section VII and finally draw conclusion and discuss future research prospects in section VII.

II. Related Work

A. Analysis of Online Social Networks

Jacob Moreno's [19] seminal work on runaways from the Hudson school for girls gave birth to sociometry. Since then, this field has grown steadily. Recent interest in this field was triggered by the work on small world [27] and scale free networks [4]. Further thrust to this field was given by the availability of large size social network data from online sites such as Facebook and Twitter. We briefly review some literature related directly to using online social network data.

Garton et al. [10] emphasized that earlier, research effort concentrated on studying how people use computers to communicate (computer mediated communication) rather than studying the social networks generated by this medium. They describe methods to identify sources to collect and analyze social network data focusing on how online communication systems provide a perfect platform to study virtual communities and interaction networks.

Kumar et al. [14] study the structural evolution of large online social networks using *Flickr* and *Yahoo! 360* data sets. The authors found that the network density followed similar

978-0-7695-4799-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ASONAM.2012.59

patterns concluding that both the graphs are qualitatively similar. They classified these networks in *singletons* who don't take part, a large *core* of connected users and a region of isolated communities forming a *star* structure.

Ahn et al [2] compare the structure of three online social networks: *Cyworld*, *MySpace*, and *Orkut*. They observe a multi-scaling behavior in Cyworld's degree distribution and that the scaling exponents of MySpace and Orkut are similar to those from different regions in the Cyworld data. They also validate the snowball sampling on Cyworld using degree distribution, clustering coefficient, degree correlation (also known as assortativity) [20] and average path length.

Mislove et al. [18] use *Flickr*, *LiveJournal*, *Orkut*, *YouTube* using degree, in-degree and out-degree, average path length, radius, diameter and assortativity metrics. Their analysis shows that social networks differ from other networks as they exhibit much higher clustering coefficient. They also show that social network have a higher fraction of symmetric links.

Leskovec et al. [15] studied Flickr, Delicious, Answers and LinkIn to develop a network evolution model. They also discuss how the number of connections drop off exponentially with individuals more than 2 hops away. Another interesting result from this study pointed the differences in the growth of new members where Flickr grows exponentially, LinkIn grows quadratically, Delicious grows superlinearly and Answer grows sublinearly.

Lewis et al. [16] investigate *Facebook* data emphasizing five distinct features. First, the correctness of data is ensured as it is downloaded from the internet. Second, the dataset is complete as it contains information about all the existing social ties in the network. Third, the data is collected over four years allowing temporal analysis of the social dynamics taking place in the network. Fourth, data on social ties is collected for multiple social relations: *Facebook Friends*, *Picture Friends* and *Housing Friends*. Finally, with users providing data for their favourite music, movies and books: the dataset is quite rich and provides new research opportunities.

Benevenuto et al. [6] use an entirely different approach to study and analyze social networks by studying the click streams generated when a user accesses a social network site. Four online social networks: *Orkut*, *MySpace*, *Hi5* and *LinkedIn* were used. The authors studies patters such as how frequently and for how long people connect to these networks, and how frequently they visit other people's pages. They also compared the click stream data and the topology of the friends social network of Orkut. Results reveal publicly *visible* social interactions such as commenting profiles as well as *silent* social interaction such as viewing profile and photos.

Rejaie et al [23] study *MySpace* and *Twitter* with the intent of finding the active population of these networks. They develop a measurement technique using the numerical user IDs assigned to each new user and the last login time of each user. This in turn helps to identify short lived users on the site and are termed as *tourists*. Results show that the number of active users in these networks is an order of magnitude smaller than the total population of the network.

Interesting observations about online social networks can be found in [12]. More comprehensive and recent review of literature on social networks can be found in [7], [25].

B. Network Statistics and Metrics

There are a number of network statistics and metrics in the literature. A detailed description of the metrics we have used is given in section IV and section V. We only consider node metrics that are widely used in the research community, or the most representatives ones as these basic metrics have been used to derive new variants. We limit our study to metrics applicable on undirected and unweighted graphs.

III. Data Sets

We have used a number of different data sets representing a variety of social networks used for analysis by the research community. The data sets are described below:

Twitter Friendship Network: Twitter is one of the most popular social networks in the world. A friendship network is extracted by crawling the twitter database using the api (api.twitter.com). Given a single user, the api returns a list of all the friends of the given user. We recursively applied this method to gather data of 2500 users starting from a single user. The complete network has 22002 edges.

Epinions Social Network: This is a *who-trust-whom* online social network of a customer analysis site Epinions.com (http://www.epinions.com/). Members of the site can either agree or disagree to trust each other. All the reliable contacts interact and form a of Trust which is then shared with users on the basis of review ratings. We have downloaded this data from the stanford website (http://snap.stanford.edu/data/) where it is publicly available in the form of a text file. The network contains 75879 nodes and 508837 edges.

Wikipedia Vote Network: Wikipedia is a free encyclopaedia which is written collectively by assistants around the world. A small number of people are designated as administrators. Using the complete dump of Wikipedia page edit history, we selected all administrator elections and vote history data. Users are represented by nodes in the network and a directed edge from node *i* to node *j* represents that user *i* voted on user *j*. Again, the data is available from stanford website with 7115 nodes and 103689 edges.

EU Email Communication Network: This network was generated by using email data from a huge European research institution. Information was collected about all emails (incoming and outgoing) for a period of Oct 2003 to May 2005. Nodes represent email addresess and an edge between nodes *i* and *j* represents that *i* sent at least one email to *j*. The network contains 265214 nodes and 420045 edges and available from stanford website.

Author Network: is a collaboration network of authors from the field of computational geometry. Two actors are connected to each other if they have co-authored an artifact together. The network was produced from the BibTeX bibliography obtained from the Computational Geometry Database 'geombib', version February 2002. The database is made available on Pajek datasets website (http://vlado.fmf.uni-lj.si/pub/networks/data/). We only consider the biggest connected component containing 3621 nodes and 9461 edges.

All these five datasets model contextually and semantically different social relations from each other. Twitter network is a friend network and represents mutual acceptance from both individuals. Epinions network is similar in the sense that it requires mutual acceptance but differs as it requires a certain degree of trust rather than friendship. Wikipedia network is a directed network which represents the voting behavior of users to select administrators and is completely different from the previous two contexts. The fourth dataset is the Email network which is also a directed network where users are related to each other if a user has communicated to the other through email. Finally the Author network is an affiliation network [21] which are based on bipartite graphs and are related to each other by having an affiliation to a common research artefact.

IV. Network Statistics

Table I shows some basic network statistics calculated on the above described data sets. We briefly define these statistics below:

Density refers to the Edge-Node ratio of a network representing the average degree of a node in the network. **Highest Degree (HD)** is the highest node degree a node has in the network. **Diameter** is the number of edges on the longest path between any two nodes in the network. **Girth** of a graph is the path length of the shortest cycle possible. **Clustering Coefficient Global (CCG)** is the measure of connected triples in the network. **Average Path Length (APL)** is the average number of edges traversed along the shortest paths for all possible pairs of network nodes. **Alpha(α)** is the constant obtained when a power-law distribution is fitted on the degree distribution of the network.

Density values for Epinions and Wikipedia networks are comparatively very high representing high number of connections for each node in the network. High density of networks can be one reason for having high clustering coefficient for a network but in the presented datasets, the networks with the lowest density have the highest CCG values which represents an important structural trait for these network as they have slightly higher number of transitive triples. For the author network, this is inherent due to the construction method of the network as research artefacts with three or more than three authors will all form triads. This observation is more interesting for the email network where people exchange emails forming triads whereas relatively low values for the twitter network suggest that friend of a friend phenomena is not quite common when compared to the email network. Girth values of 3 for all these networks represents the presences of smallest possible cycle in the network.

The APL and α of all the networks are quite close to each other again representing the similarity among the different networks. Low APL, High CCG and α values between 1.5 and 2 for twitter, email and author network represent the small world and scale free properties for these networks. The α value close to 1.2 for epinions and wikipedia network and cannot be classified as scale free networks. The histogram of degree distribution for all these networks is presented in Figure 1.

	Twitter	Epinions	Wikipedia	Email	Author
Nodes	500	500	500	500	500
Edges	3099	13739	11672	2396	2404
Density	6.18	27.47	23.34	4.79	4.80
HD	237	278	281	499	102
Diameter	11	7	12	7	10
Girth	3	3	3	3	3
CCG	0.19	0.43	0.35	0.54	0.60
APL	2.6	1.93	2.10	1.98	2.87
α	1.57	1.202	1.209	1.87	1.66

TABLE I
Basic Statistics for the Data Sets used in Experimentation. HD= Highest Node Degree, CCG= Clustering Coefficient Global, APL=Avg. Path Length, α=Power Law Fitting Constant

V. Network Metrics

In this section, we briefly describe a number of network metrics frequently used in network analysis. All the metrics considered are node level metrics or can be derived for nodes. Metrics are grouped together into Element Level Centrality, Group Level Cohesion and Network Level Centrality metrics. The metrics we have considered for experimentation are most widely used metrics in network analysis. An exhaustive study remains part of our future work.

A. *Element Level Centrality Metrics*

Element level metrics are calculated on individual elements of a graph. The term centrality refers to the idea where these elements are central in some sense in the graph.

Degree of node is an element level metric which refers to the number of connections a node has to other nodes. Degree distribution of nodes has been one of the most important metric of study for networks as the degree distribution of most real world networks follow power law [17].

B. *Group Level Cohesion Metrics*

Group Level Metrics are calculated for a small subset of nodes within the graph. The two metrics we consider here in our study are cohesion metrics that give a measure of how closely a group of nodes is connected to each other.

Local Clustering Coefficient [27] is a group level metric which counts the degree of connectedness among neighbors of a node.

Strength [3] is another group level metric which extends the notion of calculating triads in a network. This metric quantifies the neighborhood's cohesion of a given edge and thus identifies if an edge is an intra-community or an inter-community edge. The idea is to quantify whether the neighbors of a node connect well to each other or are loosely connected to each other. The values range between [0,1] such that low values indicate poor connection whereas high values indicate strong connections among the neighbors of a node.

C. Network Level Centrality Metrics

Network Level Metrics require the entire graph for calculation. Centrality in the context of network level metrics is a structure level metric which calculates how central a node is, in the entire network.

Betweenness Centrality [9] calculates how often a node lies on the shortest path between any two pair of nodes in the network. High betweenness centrality for many nodes suggest that the entire network has pockets of densely connected nodes or communities. Low values of betweenness centrality suggest that nodes of the entire network are well connected to each other representing the absence of well defined boundary structure for communities.

Eccentricity [11] also tries to capture the notion of how central a node is in the network. The eccentricity of a node is the maximum distance between v and any other node u of G. High values of eccentricity for many nodes in the network represent that there are people connected through long chains in the network which pushes these individuals far from the dense core as described by Kumar et al. [14].

Closeness [5] is another network level metric which is the inverse sum of distances of a node to all other nodes. Closeness of a node represent on average, how close or how far it lies from all other nodes in the network. These nodes are good candidates to spread information as individuals with low values representing people that are closely connected to all other nodes in the network.

VI. Experimentation

As the first step to perform a comparative analysis of various networks using different metrics, we perform sampling on all these data sets to obtain small size networks. This is due to the calculation complexity of Network Level Metrics used in this study. We sampled equal size networks in terms of number of nodes.

We used random repeated sampling collecting 10 samples of size 500 nodes from each data set giving us a total of 50 graphs. Next we calculated different metrics on these samples. For each sample, we calculated the frequencies of the resulting values giving us a distribution of how these metric values occur in the network. For example, in case of the *degree*, we calculated the frequencies of the degree values obtained for the network. Next, for each data set we calculated the average of these frequencies and used these values to comparatively analyze different networks.

VII. Inferences and Observations

Figure 1 shows the frequency distribution calculated for the above described metrics. These metrics either return values between 0 and 1, or have been normalized in this range to facilitate comparative study. Furthermore we have applied binning to calculate frequencies where the values have been rounded off to 2 decimal places giving us bins in the range $[0.00, 0.01, 0.02, \cdots, 1.00]$. The values on the horizontal axis for the graphs in Figure 1 represent the bin number, i.e. bin 0 refers to the frequency of nodes for the value 0.00, bin 1 refers to the value 0.01 and so on. One final modification to these graphs is that we have cut the extreme bins for Degree distribution, Strength, Betweenness Centrality and Closeness as there was very less information available in these bins.

From the degree distribution of the five networks in Figure 1 the graphs for the author and the twitter network are quite similar. The most interesting observations are for the wikipedia and the epinions network where we can see a linear decay in the degree distribution of the two networks which shows a non-scale free behavior of the two networks. The email network has a very high peak for very low values showing that most of the individuals in this network have used email very rarely for communication purposes.

The clustering coefficient frequencies have a similar behavior as all the networks have peaks in their frequency values. For example, the twitter network has a peak at bin 11 which refers to a value of 0.11. This shows that around 30 nodes have a clustering coefficient of 0.11. Other networks have a peak which starts from bin 21 to 51. The lowest peak is for the email network although the global clustering coefficient of this network is higher than other networks as shown in Table I. This suggests that the triads in the email network are not concentrated around nodes part of the core of the network but are well spread out in the whole network.

A similar observation can be made about the frequencies for the strength metric as values gradually rise and fall off for every dataset. Wikipedia and epinions networks have frequencies quite close to each other, the email network has its peak shifted on the right and twitter's peak shifted on the left. This means that the email network has more dense components of size 4 as compared to twitter network which does not have many such nodes.

Betweenness centrality has the most perfect match for all these networks. This is due to a few nodes with very high degree present in all networks. These nodes in turn play a central role in connecting short paths among pairs of nodes. This finding can be reinforced by the low APL values for all networks and the HD values shown in Table I.

Eccentricity values of different networks follow each other very closely. This is again an implication of the presence a few very high degree nodes in the network as the maximum distance among any pair of nodes does not vary much, as all nodes use these high degree nodes which act as short cuts in these networks.

The most variation in the frequencies is for *closeness*. The email network has initially high values as opposed to other networks but remains very low for other values. This is because it has a node with exceptionally very high degree as it is connected to all other nodes. This reduces the average closeness of all pair of nodes. The twitter network has peaks around bin number 7, 27-28, 35 and 42 which is quite different from other networks. Wikipedia has also different peaks but they are shifted towards the right when compared to twitter network, which signifies higher frequencies for high closeness values. Epionions and Author networks have peaks at bin 24 and 46 respectively which gradually decrease for higher bins.

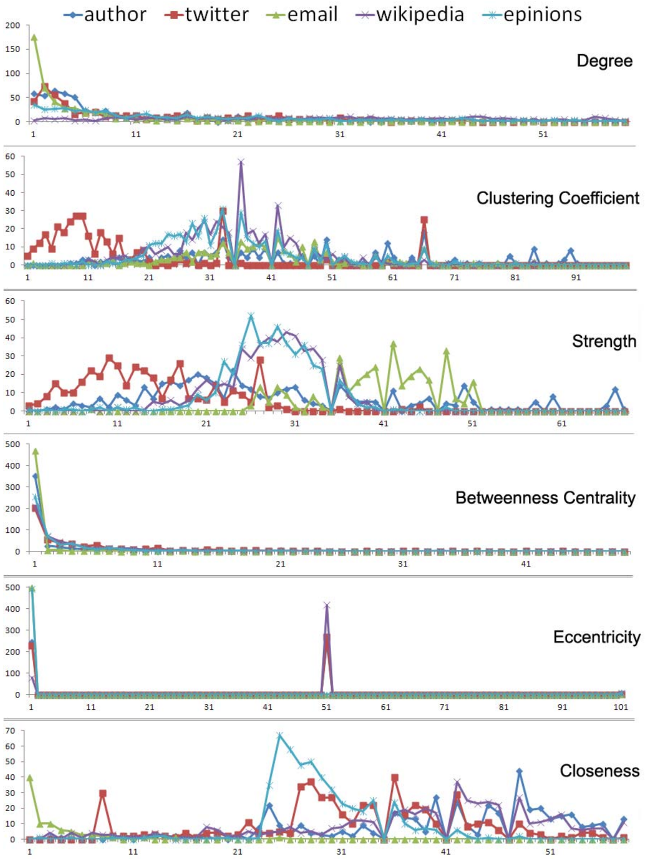

Fig. 1. Calculating different Network Metrics on the Five datasets. Horizontal axis represents bins and vertical axis represents the frequency with which nodes appear in that particular bin.

In general, the behavior of all these networks is similar when evaluated with the discussed metrics. Two findings can be quoted, one for the non-scale free behavior of two social networks, epinions and wikipedia. Second is the variations in frequencies for the closeness metric. Both these results highlight the slight structural dissimilarity among different forms of social networks.

VIII. Conclusion

In this paper, we have performed a comparative study to analyze contextually and semantically different social networks using different network statistics and metrics. Our results show that these network are structurally similar to each other in most of the cases. As part of future work, we intend to incorporate more data sets and more network metrics to perform a comprehensive comparative analysis of different social networks. We also intend to explore the possibilities of proposing a new sampling method which is robust against different structural metrics.

References

[1] A. Acar. Antecedents and consequences of online social networking behavior: The case of facebook. *J. of Website Promotion*, 3(1/2), 2008.

[2] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 835–844, New York, NY, USA, 2007. ACM.

[3] D. Auber, Y. Chiricota, F. Jourdan, and G. Melancon. Multiscale visualization of small world networks. In *INFOVIS '03: Proceedings of the IEEE Symposium on Information Visualization*, pages 75–81, 2003.

[4] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[5] M. A. Beauchamp. An improved index of centrality. *Behavioral Science*, 10:161–163, 1965.

[6] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, IMC '09, pages 49–62, New York, NY, USA, 2009. ACM.

[7] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, Feb. 2009.

[8] M. Cannataro, P. H. Guzzi, and P. Veltri. Protein-to-protein interactions: Technologies, databases, and algorithms. *ACM Comput. Surv.*, 43:1:1–1:36, December 2010.

[9] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.

[10] L. Garton, C. Haythornthwaite, and B. Wellman. Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1):0–0, 1997.

[11] P. Hage and F. Harary. Eccentricity and centrality in networks. *Social Networks*, 1:57–63, 1995.

[12] B. Howard. Analyzing online social networks. *Commun. ACM*, 51:14–16, November 2008.

[13] J. L. Iribarren and E. Moro. Affinity paths and information diffusion in social networks. *Social Networks*, In Press, Corrected Proof:–, 2011.

[14] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM.

[15] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 462–470, New York, NY, USA, 2008. ACM.

[16] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties, and time: A new social network dataset using facebook.com. *Social Networks*, 30(4):330–342, Oct. 2008.

[17] L. Li, D. Alderson, J. C. Doyle, and W. Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2:4, 2005.

[18] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, pages 29–42, New York, NY, USA, 2007. ACM.

[19] J. Moreno. Who shall survive? *Nervous and Mental Disease Publishing Company, Washington*, 1934.

[20] M. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, 2003.

[21] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[22] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122+, Sept. 2003.

[23] R. Rejaie, M. Torkjazi, M. Valafar, and W. Willinger. Sizing up online social networks. *IEEE Network*, 24(5):32–37, Sept. 2010.

[24] D. Rosen, G. A. Barnett, and J.-H. Kim. Social networks and online environments: when science and practice co-evolve. *Social Netw. Analys. Mining*, 1(1):27–42, 2011.

[25] J. Scott. *The SAGE Handbook of Social Network Analysis*. SAGE, 2011.

[26] P. B. Scott. Knowledge workers: social, task and semantic network analysis. *Corporate Communications: An International Journal*, 10(3):257–277, 2005.

[27] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998.

[28] B. Wellman. Computer networks as social networks. *Science*, 293(5537):2031–2034, 2001.