Pragmatic Analysis of Diffusion in Computer Networks

Sajid Saleem, Abdul Samiah

***Abstract* – Social networks is a content transportation medium, that has become more popular than the conventional fax and email system in the current world even shadowing the online messengers. With social websites like Facebook, Twitter etc., one can spread any type of information whether it be text, picture or video. Information of even an enormous size can spread like an epidemic on social networks. The researchers are intrigued by the interesting behaviors such networks are exhibiting and are delving in these types of networks.**

Content diffusion in computer networks is among trusted networks, just like information sharing on social websites where you are only allowed to share / view / comment only among trusted members. Among trusted networks one can spread information to achieve a diversity of tasks like marketing products or diffusion of job vacancies through pop-ups. On the other hand there can also be a malware that can utilize the same network and infect the trusted network. With the help of study of diffusion one can predict how to benefit from spreading of information in a network. Our interest in this paper is to study influence which is amongst the origins of diffusion for information spreading in a network. We have analyzed three real computer networks and compared them with artificially generated complex networks of random and scale free equivalent to each of them. Our experiments are based simultaneously on the concepts of Linear Threshold and Independent Cascade Model. We have used five different methods / metrics of selecting initial seed nodes and then calculated influence for each of them. Our experiments also include comparing of the metrics on each of the real networks and its corresponding random and scale free networks. Our experiments not only show that small amount of initial seed can infect maximum network but also that some metrics have same effect on equivalent networks while some donot.

***Index Terms* – Network diffusion, Malware.**

I. INTRODUCTION

The globalization of computer and communication devices has led to ease of inter personal communication, where by, a person sitting in one part of the world can communicate with another person oceans apart. Initially there were emails for offline messages and messengers for communication between two persons if they are online. In recent years the advent of social media has gained much importance. With the advent of social networks like Facebook, Twitter etc. one can easily send messages to one another even if the person is not online and communicate live if they are online. People with common interests communicate through these social networks and are able to expand their social circle by finding, connecting and sharing with friends having common interest. Social media helps one transfer pictures, videos and text. Its importance is not due to the fact that one can spread information of any format easily at one platform, but because the diffusion process helps spread information very robustly. People tend to have more trust in word-of-mouth of known contacts rather than advertising campaigns [1] [2] [3]. In other words people are influenced more by known contacts rather than unknown ones. The core process of social networks is based on diffusion of information, and since social networks are based on humans who are influenced by their surroundings / friend circle, one can also say that social networks core process is diffusion of information based on influence. Internet consists of domains and networks and within the domains and networks a trust level is maintained. So one can say that internet, which is a network of networks, consists of networks that trust each other and information is spread based on that trust. When contents are transferred between trusted nodes, a user accepts contents assuming that the transferred content is the one required by it or some other useful information. That might be true in some of the cases but not in all of the cases. If some malware (virus) masquerades itself and starts infecting the network of trusted nodes, the nodes will assume the content to be the desired and hence also get infected [4] [5]. Malware will cause problems on the part of the user and the network because if one node is infected, the remaining trusted networks may also be influenced by it and the malware can diffuse freely within the network. Similarly for the transfer of useful information the same rule is applied and the trusted networks are able to benefit from it. This phenomena of influence intrigues the researchers into indulging themselves in the topic of information diffusion. How network diffuses information within is still an unanswerable question. The interconnection of people, that is, how they are connected, determines the information spread. This is also known as correlation of users. Diffusion cause in a network can either be full correlation or partial correlation. In this paper, we study one of the correlation source, influence, its effects and maximization on information diffusion. [6] Disagrees that diffusion of information is caused by influence only.

If a node x, replicates another node y action, it means that y has influence on x. Influence cause differs from one node to another. Influence of a node on interconnected nodes is an effect of external factor like trust [7] or maybe a popular (influential) node's action may influence others [8]. Recognition of influential nodes is an important task of influence/diffusion maximization (Chapter 19 of [9]). The behavior of network can be studied by discovering influential nodes and the quantity of information dissipated

Sajid Saleem and Abdul Samiah are with the Department of Electrical and Power Engineering, Pakistan Navy Engineering College, National University of Science and Technology, Karachi, 75350. E-mail: sajidsaleem@pnec.nust.edu.pk. Manuscript received June 30, 2015; revised on September 12 and December 15, 2015; accepted on December 30, 2015.

 ISSN – 1999-4974

through them also noting the time taken for the networks influence / diffusion / infection with information.

Using Independent Cascade Model (ICM) [10] and Linear Threshold Model (LTM) on the network datasets; real net- works (RN) and their equivalent random (RD) and scale free (SF) [11] networks, we have calculated influence on each of the network. ICM and LTM concurrently compute influence of k nodes (initial seed is taken as k percentage of total nodes) on the network. The selection of k nodes is based on five different methods: random, degree, betweenness, closeness and Eigen. The categorization of the remaining paper is as follows: Section II summarizes the related work, the details of datasets is given in Section III, Section IV discusses results and observation of the experiments performed followed by concluding remarks in Section V.

II. RELATED WORK

Valente et al. [12] delivered an inspiring work on creating threshold model of the diffusion of innovations for adopter categorization based on social network. His work gave a dual topology for adopter categorization that is, either based on entire social system or on individual personal network. His effort also show how diffusion of innovation is guided by external influence and opinion leadership.

Kempe *et al*. [3] through word-of-mouth referral gives natural and general model of influence propagation. Basing their idea on greedy algorithm, they propose a decreasing cascade model that initially searches for active nodes and spreads a particular action in the entire network eventually. In order to accommodate very fast spreading of influence a large no of active sets is initially chosen by the algorithm.
Kimura and Saito [1] propose a couple of ICM based natural scenario that computes effectively good estimation of influential node's diffusion quantity in an ICM social network of large scale. They further experimentally display that discovery of influential nodes can be achieved through better approximations in the course of small propagation probabilities through links.

Daly and Haahr [4] work on MANETs providing solution for information dissemination based on small world dynamics. Based on centrality characteristics some bridge nodes are selected and SimBet routing is proposed that uses between- ness centrality metrics and social similarity which is locally determined. They show that SimBet Routing is better than Epidemic routing and PRoPHET routing.

Apolloni et al. [13] based on synthetic population utilized actual social network in their work under realistic environment. They proposed a model that put its basis on agent's likeness amongst themselves and the time interval of agents contact. Their results display that agents strength of links amongst themselves and their interval of contact are the factors on which information diffusion depends.

Bakshy et al. [14] discover and model change in adoption rate affecting social influence such that friends play an important role in adoption of content, sharing among strangers is less rapid compared to among friends and that influencers in a network are different than early adopters.

Gomez et al. [15] proposed an algorithm that discovers near optimal networks based on the assumption static network and changes when it gets infected. The scalable algorithm, called NETINF, can be used to study the real networks properties.

Bakshy et al. [16] study social networks role in online social diffusion. They prove through experimentation that nodes exposed to information are most likely to spread it sooner than those who are not affected. They also give the concept of weak and strong ties that weak ties play a more dominant role as they are responsible for propagation of novel information. Reid and Hurley [17] study systematically diffusion in net- works with community structures by initially replicating and enhancing work on networks with non-overlapping community structures and then study diffusion on overlapping structures, Studying contagions using SIR diffusion model.

Myers et al. [18] propose a model for information diffusion in which information can reach a node via social network link or via external social influence.

Luu et al. [19] establish a probabilistic framework that can be utilized for macro level diffusion models like Bass Model (BM). They also establish other models using this framework like degree distribution coupled with linear influence by neighboring adopters and for variable degree distribution introduce multi-stage diffusion models.

III. DATA SETS

We have used 03 real network datasets that contain network mapping data consisting of paths to internet networks from a test host. 1)"Imp" 1 [20] is a collection of recording over 90,000 registered networks from a test host since 1998, 2) "Opte" 2 data serves a multitude of purposes like modeling the internet, analyzing IP space and IP space distribution and 3) "Oregon" 3 a symmetrized snapshot of the internet structure at autonomous systems level that is reconstructed from BGP tables posted by the University of Oregon Route Views Project. Mark Newman created this snapshot from data for July 22, 2006. For each of the network we constructed an equivalent scale free (SF) [11] and random (RD) network.

Nodes are preferentially attached in SF model and number of edges can be tuned for each new node generating a SF network.

RD network is generated with n nodes and m edges and on Poisson distribution $p_k = e^{-\lambda}\frac{\lambda^k}{k!}$ based degree distribution p_k . All nodes in our generated network had $\lambda > 1$ specifying that mean degree of the network is what most nodes degree is close to.

[1]www.cheswick.com/ches/map

[2]www.opte.org

[3]http://www-personal.umich.edu/˜mejn/netdata/

All the networks are treated as undirected. Table I shows the number of nodes and edges and the edge-node ratio of these networks. The comparison of real data with artificially generated SF and RD networks help us verify the properties of different models.

TABLE I. NETWORK STATISTICS FOR DIFFERENT REAL NETWORKS

Network	Nodes	Edges	Edge-Node Ratio
Imp	190383	228354	1.2
Opte	35836	42387	1.2
Oregon	22963	48436	2.1

ISSN – 1999-4974

IV. EXPERIMENTATION

For experimental computation and analysis, we have generated a simulated network scale free and random network equivalent to the 03 different real networks giving us 09 networks in total. The software used is R. The results were exported in csv format and imported in excel to present graphs. The Table II shows highest degree node values, clustering coefficients, maximum clique, and girth and average path lengths for the generated networks in comparison to real networks.

TABLE II

RD=RANDOM NETWORK AND SF=SCALE FREE. TABLE SHOWS DIFFERENT METRICS CALCULATED FOR THE REAL AND SIMULATED NETWORKS FOR COMPARISON.

	Highest Degree of a Node		
Data Set	Real Network	RD	SF
Imp	994	30	400
Opte	259	11	372
Oregon	2390	15	378
	Clustering Coefficient		
Imp	2.65e-3	0.41e-3	1.81e-3
Opte	5.49e-3	0.08e-3	0.96e-3
Oregon	11.15e-3	0.19e-3	1.82e-3
	Maximum Clique		
Imp	07	03	04
Opte	05	03	03
Oregon	17	03	04
	Girth		
Imp	03	03	03
Opte	03	03	03
Oregon	03	03	03
	Average Path Length		
Imp	15.4609	4.2320	5.5062
Opte	16.7494	10.6898	5.9042
Oregon	3.8424	7.1048	4.8947

As seen from Table II, there is a huge difference between the highest degree nodes of the 09 networks with Imp real network having the highest degree of a node and Opte random network having the least high degree of a node.

Nodes in a graph have a tendency to cluster together. The measurement of this degree of clustering is clustering coefficient (CC). The largest value of CC is of Opte real network and the smallest is of Opte random network.

Set of nodes forming a complete graph / subgraph within a network is a clique. The maximum clique is the largest complete graph / subgraph within a network. Oregon has the largest maximum clique value whereas the lowest clique value is among 04 networks.

Length of shortest cycle within a graph is called Girth. As shown in Table II all the networks have girth value of 03. Shortest distance between node pairs within a network is Average Path Length (APL). The largest value of APL is of Opte real network and the lowest if of Oregon real network. We have used Linear Threshold Model (LTM) and Independent Cascade Model (ICM) [10] for our experiments. Our experimentation utilizes five methods for initial seed calculation. These methods are degree, betweenness, closeness, random and Eigen. Because of the large no of nodes in Imp we take the initial seed as 0.15% of the original sample for all the 09 networks. This helps us create a uniformity and standardization for experimentation. The LTM and ICM computes the total influence exerted by 0.15% of initial seed and results are also compared in the form of percentage.

V. RESULTS AND DISCUSSION

Figure 1 shows the results of influence caused by 0.15% of initial seed values chosen using random, degree, betweenness, closeness and Eigen methods and applied on all the 03 real networks in comparison to their corresponding random (RD) and scale free (SF) networks. 03 networks used for experimentation are of variable sizes in no of nodes and edges.

In real networks, degree based influence is providing maximum diffusion in almost all except betweenness method in Oregon real network.

In SF networks, the general trend is degree method influencing maximum closely followed by betweenness method which is closely tailed by Eigen method.

Random and closeness methods provide the least influence in all 09 networks.

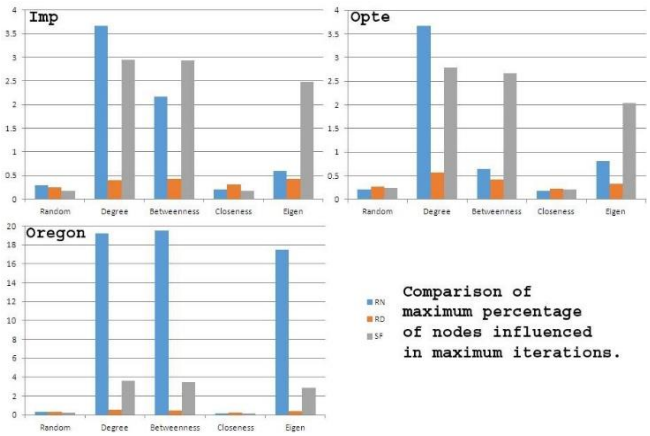

Fig. 1. Max Influenced Nodes: Figure showing graphs for the 03 real world networks, their equivalent simulated networks and the number of nodes influenced (in %) using different methods. RN= Real RD=Random, SF=Scale free.

Figure 2 shows that maximum 09 iterations are required for maximum influence under current parameters. Eigen generally is

ISSN – 1999-4974

taking the minimum iterations for maximum influence in real networks. In general 06 is the no of iterations required to influence / infect maximum of network in current settings.

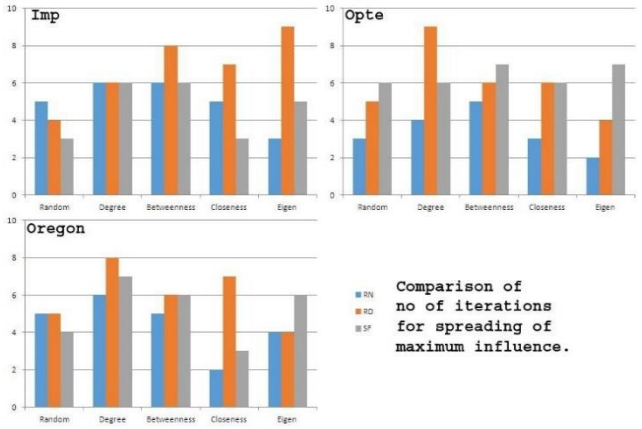

Fig. 2. Max Iterations for Influence: Figure showing graphs for the 03 real world networks, their equivalent simulated networks and the number of iterations the seeds influenced using different methods. RN= Real RD=Random, SF=Scale free.

Figure 3 shows that for all the 09 datasets / networks, betweenness method is taking the maximum amount of time to calculate the influence in the network. The next is closeness method followed by degree method. In some experiments Eigen takes least time while in others random takes less time. But the general trend from taking most time to compute influence is betweenness, closeness and degree respectively in descending order.

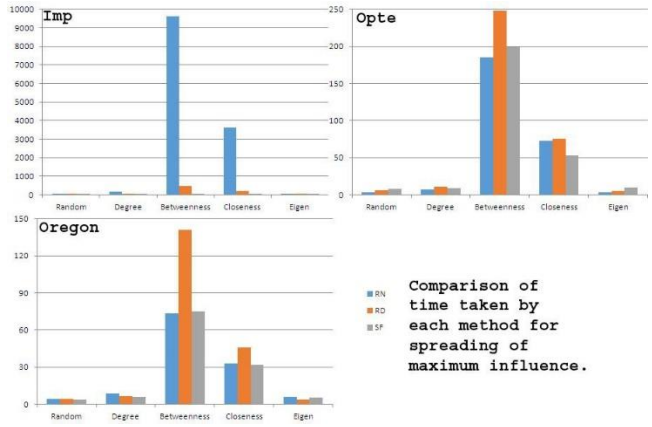

Fig. 3. Max Time for Influence: Figure showing graphs for the 03 real world networks, their equivalent simulated networks and the time taken by each method to influence the nodes. RN= Real RD=Random, SF=Scale free.

VI. CONCLUSION

We have successfully performed a comparative study on 03 real networks of variable sizes and recorded in different environments from different universities, with their corresponding random and scale free networks to study the effect of diffusion by influence in a network just like a social network. Our findings include not only the comparison of different networks but also the comparison of the 05 different methods (random, degree, betweenness, closeness, Eigen) applied on them and their results and discussion are elaborated in the above sections. We can conclude that the information dissimenation in social and computer networks have similar properties.

In future we would like to expand our experiments and study the effects of even more parameters / methods on not only the above networks but also other datasets / networks.

REFERENCES

[1] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," in Knowledge Discovery in Databases: PKDD 2006. Springer, 2006, pp. 259–271.
[2] M. O. Jackson and L. Yariv, "Diffusion on social networks," Economie publique/Public economics, no. 16, 2006.
[3] D. Kempe, J. Kleinberg, and E´. Tardos, "Influential nodes in a diffusion model for social networks," in Automata, languages and programming. Springer, 2005, pp. 1127–1138.
[4] E. M. Daly and M. Haahr, "Social network analysis for routing in disconnected delay-tolerant manets," in Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing. ACM, 2007, pp. 32–40.
[5] A. Khelil, C. Becker, J. Tian, and K. Rothermel, "An epidemic model for information diffusion in manets," in Proceedings of the 5th ACM international workshop on Modeling analysis and simulation of wireless and mobile systems. ACM, 2002, pp. 54–60.
[6] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and corre- lation in social networks," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008, pp. 7–15.
[7] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of trust and distrust," in Proceedings of the 13th international conference on World Wide Web. ACM, 2004, pp. 403–412.
[8] N. E. Friedkin, A structural theory of social influence. Cambridge University Press, 2006, vol. 13.
[9] D. Easley and J. Kleinberg, Networks, crowds, and markets. Cambridge, Univ Press, 2010, vol. 8.
[10] D. Kempe, J. Kleinberg, and E´. Tardos, "Maximizing the spread of influence through a social network," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003, pp. 137–146.
[11] A.-L. Baraba´si, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: the topology of the world-wide web," Physica A: Statistical Mechanics and its Applications, vol. 281, no. 1, pp. 69–77, 2000.
[12] T. W. Valente, "Social network thresholds in the diffusion of innova-tions," Social networks, vol. 18, no. 1, pp. 69–89, 1996.
[13] A. Apolloni, K. Channakeshava, L. Durbeck, M. Khan, C. Kuhlman, B. Lewis, and S. Swarup, "A study of information diffusion over a real- istic social network model," in Computational Science and Engineering, 2009. CSE'09. International Conference on, vol. 4. IEEE, 2009, pp. 675–682.
[14] E. Bakshy, B. Karrer, and L. A. Adamic, "Social influence and the diffusion of user-created content," in Proceedings of the 10th ACM conference on Electronic commerce. ACM, 2009, pp. 325–334.
[15] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010, pp. 1019–1028.
[16] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in Proceedings of the 21st international conference on World Wide Web. ACM, 2012, pp. 519–528.
[17] F. Reid and N. Hurley, "Diffusion in networks with overlapping commu- nity structure," in Data Mining Workshops (ICDMW), 2011

ISSN – 1999-4974

IEEE 11th International Conference on, 2011, pp. 969–978.

[18] S. A. Myers, C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks," in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012, pp. 33–41.

[19] D. M. Luu, E.-P. Lim, T.-A. Hoang, and F. C. T. Chua, "Modeling diffusion in social networks using network properties," in Proceedings of the 6th International Workshop on Weblogs and Social Media (ICWSM2012), 2012, p. 2.

[20] B. Cheswick, H. Burch, and S. Branigan, "Mapping and visualizing the internet," in ATEC '00: Proceedings of the annual conference on USENIX Annual Technical Conference. Berkeley, CA, USA: USENIX Association, 2000, pp. 1–1.

 ISSN – 1999-4974