Empirical Analysis of Seed Selection Criterion in Influence Mining for Different Classes of Networks

Owais Hussain, Zainab Anwar, Sajid Saleem, Faraz Zaidi

Empirical Analysis of Seed Selection Criterion in Influence Mining for Different Classes of Networks

Owais A. Hussain
Muhammad Ali Jinnah University
Karachi, Pakistan
Email: owaishussain@outlook.com

Zainab Anwar
Karachi Institute of Economics and Technology
Karachi, Pakistan
Email: zainabanwar@pafkiet.edu.pk

Sajid Saleem
National University of Sciences and Technology
Karachi, Pakistan
Email: sajidsaleem@pnec.nust.edu.pk

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

***Abstract*—Recent years have seen social networks gain lot of popularity to share information, connecting millions of people from all over the world. Studying the spread of information, or *Information Diffusion* in these networks has shaped into a well known field of study with numerous applications in areas such as marketing, politics, and personality evaluation. Researchers have studied information diffusion under various models and opted centrality-based algorithms that offer better results over many other approaches. These algorithms try to select initial seed nodes effectively so as to maximize influence in a network in minimum time. However, since different networks follow different structural properties, motivating the need to study different diffusion strategies for networks with different structural properties. In this paper, we aim to empirically analyze four different measures of centrality to select seed vertices for influence mining on four classes of networks: small-World networks, scale-free networks, small world-scale free networks and random networks. These networks are generated equivalent in size to four semantically different real world social networks. We use two most frequently used diffusion models: Independent Cascade model and Linear Threshold model for analysis. Our results show interesting behavior of various centrality measures for the above said classes of networks.**

I. Introduction

With the recent development of social networks and online communication channels, communication and sharing of information has gained lot of popularity with billions of users for these communication channels. Online Social networks connect people having relationships, common interests, colleagues and friends irrespective of physical distances and geographical constraints with almost negligible communication cost. As communication in social networks has gained volume, it has also attracted researchers in studying information diffusion, i.e. how information spreads in a social network. On the other hand, popular online social networks including Facebook, Twitter, Flickr and LinkedIn provide rich support to researchers, to observer and analyse their network data, which they, in return; use to gain marketing advantages. For example, a travelling agency willing to advertise a family holiday trip to a resort would target a specific group of users in a network having travelling interests and more than average followers, or a political campaign runner will look for economists and journalists to publicize their policies. Viral Marketing and spread of innovations [1], virus propagation [2], rumors and everyday news are all being spread through social networks, either to achieve a desirable objective or just for entertainment. Its effects are similar to that of word-of-mouth phenomena well studied in sociology [3]. People tend to have more trust in word-of-mouth of known contacts rather than advertising campaigns [4], [5].

Given a network represented by a graph $G(V, E)$ with the set of vertices V representing people and the set of edges E representing links among people, if any node $v_1 \in V$, replicates the action of another node $v_2 \in V$, we may assume that v_2 has influence on v_1. The frequency of occurrence of this pattern helps us conclude whether our assumption is true or not. The causes for influence may vary from node to node. The influence of a node on other connected nodes can be due to some external factors like trust [6] or maybe the action of an influential node which is very popular, or they may be influenced by seeing the influential nodes performing that action [7]. When the influential node performs an action, because of its influence, nodes connected to it will do the same action i.e. replicate the action. In other words, it means that information diffusion starts whenever the action is replicated. A very important task for the maximization of diffusion due to influence is the identification of influential nodes (Chapter 19 of [8]). By determining influential nodes we can study the behaviour of a network and how much information is spread through the nodes and similarly how much time will it take for the network to get diffused or infected with the information. The role and structural position of individuals in a social network play an important part in determining how influential a person is, in a network. An important goal of researchers is to identify from a given network, a small set of such influential people which can, ideally influence the entire network in minimal time.

One of the most frequently used methods of picking up this set (called seed) of users to spread information is based

on centrality measures [9]. While there are various types of centrality, like degree centrality, closeness centrality, betweenness centrality, Eigen-vector centrality, they are applied to a network naively, without careful analysis of which of them will perform best on structurally different classes of networks.

Two well known classes of networks studied extensively are: the small world networks [10] with low average path lengths and high clustering coefficients and scale free networks [11] with degree distribution following power-law. Many real networks works posses both the properties of small world and scale free networks thus giving us a hybrid classification, small world-scale free networks. We also use random networks [12] for comparative analysis.

This paper tries to establish a relationship among different centrality based seed selection methods and different classes of networks in an attempt to analyse these effects in the case of social networks. We propose that centrality-based methods behave differently on structurally different types of networks. In order to prove our postulate, we use two most fundamental and widely used models to study diffusion and influence, linear threshold (LT) and independent cascade (IC) model [13]. The LT model determines how an inactive node is influenced by its active neighbours, each inactive node v chooses a minimum influence threshold. If the incoming influence exceeds this threshold, v becomes active. On the other hand, IC model defines how active nodes will attempt to influence an inactive node in their neighbour. Each active node u attempts to influence all of its inactive neighbours v only once, independent of past propagation and other active neighbours of v.

The LT and IC model work simultaneously and calculate influence of k nodes (k is a percentage of total nodes in the network to be used as initial seeds) on the network. We study this model on four different classes of networks, including small world (SW) [10], scale free (SF) [14], small world-scale free (HK) [15] and random (RD) networks [12]. We also compare these artificially generated networks with four equivalent size real social networks which are a Blog network, Twitter social network, Epinions who-trust-who network and co-atuhor network of researchers. The use of real networks not only allow us to select realistic density for the artificially generated networks, but also to compare the behavior of LT and IC models on real networks as well. The results reveal quite interesting behavior and can be outlined below:

- We study the effects of different seed selection strategies for different classes of networks and find similarities based on the degree distribution of these networks. Networks with power law degree distribution behave similarly and networks with poisson distribution behave similarly for our experiments.
- High degree nodes play an important role in influencing other nodes in the network.
- High density of networks and very high connectivity of high degree nodes play an important role in reducing the time required to influence other nodes.

Rest of the paper is organized as follows: in section II we have summarized the related work in which diffusion in social networks via influence is studied; section III explains the diffusion models we used to carry out our experiments and the different centrality metrics used for seed selection; section IV provides the details of datasets used for experimentation and section V describes the experimental setup. In section VI, results and observations are discussed and finally, we have given concluding remarks in section VII.

II. Related Work

Valente *et al.* [16] designed a threshold model for diffusion in social networks assuming that behaviour of an individual is either to engage or not to engage in an activity, in which some of the people are engaged depending on their threshold value. Concluding that threshold lags occur in this model, whose magnitude indicates the degree of delay in threshold activation. Kempe *et al.* [4] give natural and general model of influence propagation through word-of-mouth referral. They propose a decreasing cascade model, based on greedy algorithm, that initially searches for active nodes eventually spreading a particular behaviour through the entire network. The algorithm chooses a large no of active sets initially so that the spreading can commence very swiftly. Kimura and Saito [17] propose two natural special cases of Independent Cascade model, which efficiently calculate good estimate of quantity for influential nodes in large scale IC based social networks for information diffusion. They propose better models than IC for extracting influential nodes and experimentally demonstrate small propagation probabilities through links can give good approximations for discovering influential node sets. Jackson and Yariv [18] study a diffusion model on social networks which are connected through undirected graphs and each node can either adopt or or decline new changes. The authors randomly select initial nodes and then diffuse information observing a threshold point called "tipping", a point after which majority of the population adopt changes. It is based on the theory that if a substantial amount of population adopts a behaviour then the behaviour/change spreads to majority of population, or otherwise it collapses. Apolloni *et al.* [19] used a realistic social network which is based on synthetic population under realistic conditions. They presented an interaction model based on the similarity of agents linked with each other and the duration of contact of agents. They found that information spreading depends on the duration of contact and strength of links between agents. Bakshy *et al.* [20] examined the interaction of social influence and social networks while adopting online content. They applied different models of social contagion which captures the rate at which a user adopts an asset following the adoption of one more of their friends. They also found a slight correlation between number of assets transferred and strength of tie between two friends. Gomez *et al.* [21] developed a scalable algorithm NETINF that finds provably near-optimal networks, assuming the network is static and observing the times when it gets infected only. The algorithm is evaluated on very large datasets of information spreading between news and blogs sites. Using this algorithm properties

of real networks can be studied. Bonchi [22] provided a survey on social influence and its propagation in networks. He discussed many models for influence maximization in viral marketing, emphasizing that available past propagation's details should be used in the models. He also highlighted the importance of using algorithms that can minimize the number of scans of propagation log. Bakshy *et al.* [23] experimented on Facebook dataset, generalizing that it proves the strength of weak ties study of Mark Granovetter [24]. Further, in the context of their study they explain phenomena of diffusion by these mechanisms (1) A link is shared by an individual and exposure to that link causes a friend to reshare that link (2) A web page is visited by friends and that link of the web-page is shared by them independent of each other (3) A link is shared by an individual external or within Facebook and a friend visit that link externally and share it on Facebook. Lewis *et al.* [25] study a dataset of Facebook activity of a cohort of college students(their friendships, tastes in music, movies and books). Finally suggesting that friends do share some tastes but not because they influence each other but because this similarity is the part reason of their becoming and remaining friends in the first place.

All these different studies target either real networks or synthetic data without focussing on well known classification of networks, namely small world, scale free, small world-scale free and random networks. Our objectives are thus clearly different from the earlier studies where we attempt to develop a generic understanding of how the properties such as average path length, clustering coefficient and power law degree distribution affects influence mining. Furthermore we consider five different methods of initial seed selection namely, random, degree centrality, closeness, betweenness and eigen to develop a better understanding of the interplay between structural differences of networks and centrality measures.

III. Diffusion model and Seed Selection

Starting with an undirected graph G consisting of nodes V and edges E, (u, v) represents that nodes u and v have a direct link with each other. A node u performs certain action a in t time and becomes active. u cannot become inactive once it activates. We use independent cascade model to exert influence from active nodes on their inactive neighbours. Each active node u has 0.5 probability to attempt to activate each of its inactive neighbours v in the next time step, regardless of the past propagation and independent of other active nodes. It may be that nodes are exerting influence on the same inactive node. If the attempt goes successful, then the inactive neighbour v becomes active and will further contribute in activating more nodes. If the attempt fails, the same active node will never have another chance of activating the same inactive node. Then, according to the linear threshold model, an inactive node does not always get activated on the basis of one successful attempt of one active neighbour only. Each inactive node chooses an activation threshold at random (because of lack of awareness of the influentiability of nodes) and if the sum of all incoming influence exceeds this threshold, only then can it become active; the LT model normalizes the influence weightage such that total weight less than or equal to 1.

In order to initiate the influence process, we choose a set of k nodes that have performed action a already. We do this by choosing k nodes called seed, from the data with highest centrality [26], where we used four different centrality measures, degree centrality, closeness, betweenness and eigen-vector along with random selection of seeds for comparative analysis. The four centrality measures are described below.

Degree centrality is the most simple centrality defined as the number of connections (degree) of a vertex in a network. These vertices are good candidates to influence other people as they have many social contacts in the society. *Closeness* [27] is a network level metric which is the inverse sum of distances of a node to all other nodes in the network. Closeness of a vertex or an individual in case of social networks, represents 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 represent people that are closely connected to all other nodes in the network. *Betweenness Centrality* [28] calculates how often a vertex lies on the shortest path between any two pair of vertices in the network. High betweenness centrality for vertices with a clear difference from others betweenness values suggest that the network has pockets of densely connected vertices or communities. Low values of betweenness centrality suggest that vertices of the entire network are well connected to each other representing the absence of well defined boundary structure for communities. Vertices with high betweenness centrality values are the ones which play the role of bridges between different communities and thus are able to influence people from different groups increasing chances to influence more people. Finally *Eigen-vector* [29] centrality is a measure of the importance of a vertex in a network. It is a network level metric calculated iteratively on vertices and assigns a relative score based on the idea that connections to high-scoring vertices contribute more to the score of the vertex in question than equal connections to low-scoring vertices. A vertex is considered important if it is connected to other important vertices implying that a node with high eigen-vector centrality might not itself have high connections but relies on its neighbors to influence other vertices.

We use linear threshold (LT) and independent cascade (IC) [13] for our experiments. These diffusion models require as input, initial seed nodes which are considered to be influenced at the start of the experiment. We have used the above four centrality metrics to determine intial seeds for our experiments.

Algorithm 1 defines the implemented algorithm. In algorithm 1, G is a graph representing a social network, consisting of V nodes and E edges, where c is some centrality of a node. First, we initiate seed set S and its influence I_S. The *while* loop stores a set of nodes with highest centrality in S; this set is used to start influence spread process. Next, we execute the influence propagation process using LT and IC diffusion models. We run the algorithm until no untried nodes are left. In each iteration, active nodes in S are removed from V; the inner

Algorithm 1 Influence mining using various seed selection methods

Require: $G : (V, E, c)$
  $S \leftarrow NULL$
  $I_S \leftarrow NULL$
  while $|S| < k$ **do**
    $S \leftarrow S \cup (v|c_v = max(c); v \in V - S)$
  end while
  while $V \neq NULL$ **do**
    $A \leftarrow NULL$
    $V \leftarrow V - S$
    for all $v \in V$ **do**
      $N \leftarrow u|u \in S; (u, v) = 1$
      $\theta \leftarrow rand(0, |N|)$
      $sum \leftarrow \Sigma_{i=|u|} rand(0, 1)$
      if $sum \geq \theta$ **then**
        $A \leftarrow A \cup v$
      end if
      $V = V - v$
    end for
    $S \leftarrow S \cup A$
    $I_S \leftarrow I_S + |A|$
  end while
  return I_S

loop runs for each node v in set of inactive nodes V, where we choose a random threshold for v and sum up successful trials from each active neighbour of v in N; if the sum is greater than or equal to the threshold θ, then v is appended to the set of active nodes. At the end of the loop, v is removed from the inactive set, as it has already had its chance to get activated. Then in the outer loop, all the recently activated nodes join S to activate more nodes and the influence is recalculated by adding up the number of newly activated nodes in A. Finally, the total influence from S, i.e. I_S is returned.

IV. Data Sets

We have generated different artificial networks; small world network (SW) using the model of [10], scale free (SF) using the model of [14], small world-scale free network (HK) using the model of [15] and random network (RD) using the model of [12] to represent different types of networks. We have also used four real networks which are semantically different from each other and they represent different forms of social communication and are described below:

Political Blog hyperlinks network between weblogs on US politics, which was recorded in 2005 by Adamic and Glance [30]. *Twitter* is among the most popular social networks for communication between online users. We have used the dataset extracted by [31]. *Epinions* is a who-trust-whom online network at Epinions.com, and the data is available at (http://snap.stanford.edu/data). For experimentation purposes we have considered a small subset of the original network. The *Author* a co-authorship network is, generated from the BibTeX bibliography developed from the Computational Geometry Database and made at Pajek datasets (http://vlado.fmf.uni-lj.si/pub/networks/data/). We treat all these networks as simple and undirected and only consider the biggest connected component. Table I shows the number of nodes and edges and the density (edge-node ratio) of these networks. We have generated equivalent size networks for each of these real networks, using the four network generation models referred above. By using real data we not only select realistic density, but can also compare these models with real data. The rationale behind selecting multiple data sets for experimentation is that since the diffusion models use randomization, it would be more accurate to not base the analysis on single data set.

Network	Nodes	Edges	Density
Blog	1222	16714	13.6
Twitter	2492	17658	7.0
Epinions	2000	48720	24.3
Author	3621	9461	2.6

TABLE I
NETWORK STATISTICS FOR THE FOUR DIFFERENT SOCIAL NETWORKS USED FOR EXPERIMENTATION.

Data Set	Real Network	RD	SW	SF	HK
	Highest Degree of a Node				
Blog	351	46	47	211	321
Twitter	237	27	27	253	319
Epinions	1192	77	72	373	560
Author	102	15	16	201	183
	Average Path Length				
Blog	2.7	2.5	3.2	2.4	2.2
Twitter	3.4	3.2	4.2	2.9	2.8
Epinions	2.2	2.2	3.0	2.2	2.0
Author	5.31	5.07	6.41	3.4	4.0

TABLE II
RD=RANDOM NETWORK, SW=SMALL WORLD, SF=SCALE FREE, HK=HOLME AND KIM MODEL(SMALL WORLD-SCALE FREE NETWORKS). TABLE SHOWS DIFFERENT METRICS CALCULATED FOR THE REAL AND ARTIFICIALLY GENERATED NETWORKS FOR COMPARISON.

V. Experimentation

For experimentation and analysis, we have generated five simulated networks each for small world, scale free, small world-scale free and random networks equivalent to the 4 different social networks giving us 80 networks in total. We use the arithmetic mean calculated over the five samples each to tabulate the results of the experiments. Table I shows some basic statistics for the four real networks used for experimentation. Table II shows highest degree node values, clustering coefficients and average path lengths for the generated networks in comparison to real networks.

We use linear threshold (LT) and independent cascade (IC) [13] for our experiments. These diffusion models require as input, initial seed nodes which are considered to be influenced at the start of the experiment. We have used four centrality based methods degree, betweenness, closeness, eigen-vector and random selection of nodes as initial seed. The LT and IC model calculates the total influence exerted by the seeds on the network. We measure two parameters for comparative analysis of these networks. The percentage of vertices influenced after

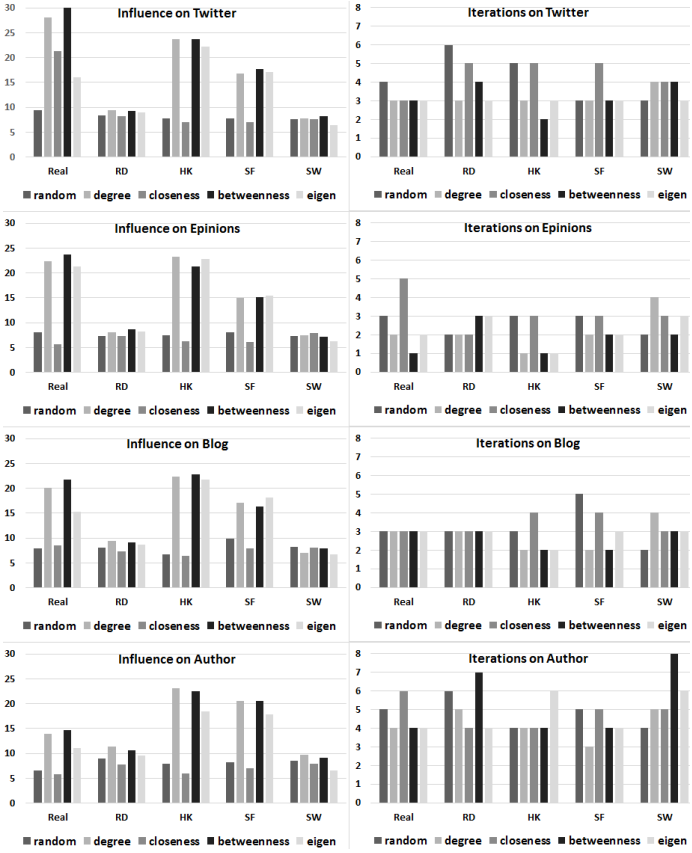

Fig. 1. Influenced Nodes: Figure showing graphs for the data sets and their simulated networks. On y-axis of left column, we have the percentage of total nodes influenced by diffusion with seed nodes $k = 5\%$ of total nodes, where on y-axis of right column, we have number of total iterations that the algorithm took to try to spread the influence to whole network. X-axes contain blocks of data sets: HK=Small world-scale free, RD=Random, SF=Scale free and SW=Small world networks, and within each block, we have seed selection methods used: random=Random seed selection method, degree=Degree centrality, closeness=Closeness centrality, betweenness=Betweenness centrality and eigen=Eigen-vector centrality.

the execution of the algorithm and the number of iterations required to achieve this percentage. The number of iterations can be considered as an indicator of the time required to influence the entire network as they represent how many iterations it takes to try to influence every node in a network. The ideal seed selection would be with maximum influence and minimum iterations for a network.

VI. Results and Discussion

Figure 1 shows the results of all the four simulated networks as well as real networks. The first observation from the graphs for the influence spread is that the behavior of small world networks and random networks is the same for all the different seed selection strategies. An important similarity in both small world networks and random networks is the small average path length but since very high degree nodes are missing from these two classes of networks, the algorithm behaves similarly for all the different seed selection methods. Furthermore, it is important to note that small world networks have high clustering coefficient, i.e. presence of triads. Thus, small world networks should exhibit high influence spread, but to our surprise, no significant impact of clustering coefficient was evident during our experiments; rather, high-degree nodes contributed more in the influence spread.

For the case of real networks, scale free networks (SF) and small world-scale free networks (HK), betweenness centrality leads the different metrics in spreading maximum influence followed closely by degree centrality and then eigen-vector centrality. Closeness performs well for the Twitter network whereas for the other three datasets, the closeness and random selection have only minor differences. Eigen-vector performs well for small world-scale free and scale free networks, even better than betweenness centrality for the Epinions network as an exception, but for other networks its performance is behind betweenness centrality and degree centrality. It is interesting to note the behavior of eigen-vector for the four real world networks. The behavoir is quit variable as it performs poorly for Twitter network, worse than closeness, performs very close to betweenness and degree centrality for Epinions and has a notable difference for Blog and Author networks.

The second observation from the graph in figure 1 is the high similarity of real networks and the HK model for small world-scale free networks. This reaffirms that most real world networks are both small world and scale free in nature. A slight exception is the Author network which is generally less influenced using different attack strategies. Although in this case, its equivalent small world-scale free (HK) and scale free (SF) networks behave similarly. This is due to the difference in the highest degree node (see Table II) where Author network has a highest degree node of only 102 as compared to SF with 201 and HK with 183 as the degree of their highest degree nodes. This clearly suggests the high impact of these very high degree nodes in influencing a network.

In terms of number of iterations required to influence the maximum number of vertices in the network, betweenness centrality performs well along with degree and eigen-vector centrality except for the case of small world and random networks generated equivalent to the Author network. Even for the real Author network, more iterations are required. This is due to the low density of the overall network and absence of high degree vertices in small world and random networks, and the highest degree vertex in the Author network has not a very high value. As a result, its takes more iterations in an attempt to influence the entire network. An interesting observation about the Epinions networks also proves the above conjecture, as the highest degree node in the real Epinions network is with degree 1192, which is more than 50% of the nodes are connected to a single node. As a result the number of iterations required to make attempts to influence vertices is very low as compared to other real and simulated networks. The density of the network also plays its part as Epinions has a very high density of 24.3 as both scale free network and small world-scale free network of same density require minimal number of iterations when compared to other datasets.

Finally, as expected, the average path length attribute exhib-

ited uniform behaviour in almost all the experiments. In figure 1, the number of iterations on all of the Epinions and Blog data sets, having low APL remained between 2 and 5. Twitter data set took 6 iterations, once for RD in random influence method and 5 iterations, in 4 more occasions. However, on Author data set with highest APL, the number of iterations begin from 4 and hike to 8. This is because it requires less number of steps to traverse from one node to another in the network. Summarizing our findings below:

- The effects of different seed selection strategies are nullified if the degree distribution of networks follows poisson distribution.
- High degree nodes have a high impact in influencing other nodes. This is evident from the analysis of real networks, scale free networks and small world-scale free networks.
- High density of a network and high connectivity of very high degree nodes results in less iterations required, which in turn means less time is require to influence nodes in the entire network.
- Number of iterations is proportional to the Average path length. In low-APL data sets, influence spread requires less traversing as compared to in high-APL data sets.

VII. Conclusion

We have performed an empirical study of social influences and its effect in the diffusion of information in social networks using LT and IC model on four complex networks and real world networks. Our results show that for networks with degree distribution following poisson distribution, different seed selection methods have no effect whatsoever on the performance of the influence algorithm. For the case where the degree distribution follows power-law, the real networks, the scale free networks and small world-scale free networks, betweenness centrality performs well to select initial seed nodes followed closely by degree centrality. Furthermore high degree nodes play an important role in maximizing influence and reducing the time period required to spread the iterations in a network. These preliminary results are intended to help us create a better understanding of the performance of different seed selection methods for different networks. We intend to use this study to develop new metrics that can be used to determine better performing seed selection methods. We also intend to generalize these results by including other social networks. All the networks used have a limited size and we plan to include larger datasets to generalize our results.

References

[1] F. Alkemade and C. Castaldi, "Strategies for the diffusion of innovations on social networks," *Comput. Economics*, vol. 25, no. 1-2, pp. 3–23, 2005.

[2] W. Fan and K. Yeung, "Online social networksparadise of computer viruses," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 2, pp. 189 – 197, 2011.

[3] J. J. Brown and P. H. Reingen, "Social ties and word-of-mouth referral behavior," *Journal of Consumer Research*, pp. 350–362, 1987.

[4] D. Kempe, J. Kleinberg, and É. Tardos, "Influential nodes in a diffusion model for social networks," in *Automata, languages and programming*. Springer, 2005, pp. 1127–1138.

[5] P. Domingos, "Mining social networks for viral marketing," *IEEE Intelligent Systems*, vol. 20, no. 1, pp. 80–82, 2005.

[6] 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.

[7] N. E. Friedkin, *A structural theory of social influence*. Cambridge University Press, 2006, vol. 13.

[8] D. Easley and J. Kleinberg, *Networks, crowds, and markets*. Cambridge Univ Press, 2010, vol. 8.

[9] P. Bonacich, "Power and centrality: A family of measures," *American journal of sociology*, pp. 1170–1182, 1987.

[10] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, Jun. 1998.

[11] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[12] P. Erdös and A. Rényi, "On random graphs, i," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.

[13] D. Kempe, J. Kleinberg, and É. 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.

[14] A.-L. Barabási, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: the topology of the world-wide web," *Physica A: Stat. Mechanics and its Applications*, vol. 281, no. 1, pp. 69–77, 2000.

[15] P. Holme and B. J. Kim, "Growing scale-free networks with tunable clustering," *Physical Review E*, vol. 65, p. 026107, 2002.

[16] T. W. Valente, "Social network thresholds in the diffusion of innovations," *Social networks*, vol. 18, no. 1, pp. 69–89, 1996.

[17] 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.

[18] M. O. Jackson and L. Yariv, "Diffusion on social networks," *Economie publique/Public economics*, no. 16, 2006.

[19] A. Apolloni, K. Channakeshava, L. Durbeck, M. Khan, C. Kuhlman, B. Lewis, and S. Swarup, "A study of information diffusion over a realistic social network model," in *Computational Science and Engineering, 2009. CSE'09. International Conf. on*, vol. 4. IEEE, 2009, pp. 675–682.

[20] 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.

[21] 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.

[22] F. Bonchi, "Influence propagation in social networks: A data mining perspective," *IEEE Intelligent Informatics Bulletin*, vol. 12, no. 1, pp. 8–16, 2011.

[23] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proc. of the 21st international conference on World Wide Web*. ACM, 2012, pp. 519–528.

[24] M. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, May 1973.

[25] K. Lewis, M. Gonzalez, and J. Kaufman, "Social selection and peer influence in an online social network," *Proceedings of the National Academy of Sciences*, vol. 109, no. 1, pp. 68–72, 2012.

[26] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1979.

[27] M. A. Beauchamp, "An improved index of centrality," *Behavioral Science*, vol. 10, pp. 161–163, 1965.

[28] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.

[29] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[30] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 u.s. election: divided they blog," in *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*. New York, NY, USA: ACM Press, 2005, pp. 36–43.

[31] A. Hashmi, F. Zaidi, A. Sallaberry, and T. Mehmood, "Are all social networks structurally similar?" in *IEEE/ACM International Conf. on Advances in Social Networks Analysis and Mining*, 2012, pp. 310–314.