STM JOURNALS
Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720 (Online)
Volume 6, Issue 2
www.stmjournals.com

A Survey on Information Diffusion in OSNs based on Influence

Tabasum Sami[1,]*, Zahoor Ah Najar[2], Amjed Husain[2]
[1,2]Department of Information Technology, Central University of Kashmir, Jammu and Kashmir, India

Abstract
In social systems, information and ideas spread through the communication between different social entities. So it is important to detect the key factors in spreading the information. Finding influential entities in social network is an important issue so that they can influence largest population possible. At the same time, community structure plays a vital role in understanding the dynamics of a social system. Many researches have been done for identifying influential spreaders. In this study, we have divided information diffusion according to their influence. Three important distinctions have been done on the basis of influence and also an extensive survey on the influence due to a node, community detection and on influence maximization. We are trying to succinct various state-of-art methods and algorithms used by various researchers.

Keywords: *OSN (online social network), influence maximization, centrality, probabilistic model.*

****Author for Correspondence*** E-mail: tabasumsamibhat@gmail.com

INTRODUCTION

Online social network is a structure that is formed of social individual or organization having set of social entities, their interaction and the relationship among them. These social networks play a vital role in the propagation of information, ideas or influencing other individuals in the network. Information dissemination is often studied on a small world comparatively and on random networks. These random networks and topological differences are said to be the main reasons that lead to the spreading abilities of ties in the network. OSNs acts as hot-bed for content generated by users. Every user contributes to content spreading which then gets passed on to other members. User to user communication forms an idea for information spreading. The random networks can be formulated in graphs for better study and the individual contributes to the nodes and interaction between the users contributes to the edges of the graph. Then graph is manipulated according to the research process.

Social network research has received immense interest in different areas in recent years, including information diffusion models or study of the network or studying the influence of a particular node to the whole network. A social network forms different communities, where communities are group of people having some common interests. The individuals form different communities on the basis of their interest. It becomes a subgraph of the whole network in which people are densely connected to one another and share some common attributes. Although the communities keep on changing with every moment.

In this study, we are going to discuss influence regarding the information diffusion taken in account community structure and spreading of information from one node to the other node. Our research has been divided into three basic categories of information diffusion influence; these are:

1. Influence due to a node,
2. Influence due to a community, and
3. Influence maximization.

Finding influential nodes in the OSNs is one of the major challenges of information diffusion. It includes opinion leader related approach where opinion leader are those nodes in the network which form a bridge of information spreading. They have effective influence on other nodes in a network. Influence of these nodes is based on network structure, node attributes and on information sharing between them. Centrality measures are mainly used to measure the important nodes in a network. Besides centrality measures, PageRank algorithm can also be used to rank nodes according to certain criteria. Structure of network changes over time but communities are believed to be relatively stable. The main challenge in the community based influence is community detection. Many algorithms have been proposed for detection of communities in an OSN including Label propagation algorithm. The method assigned a property, known as activation value. This property is added to respective label, then pair (activation value and label) is propagated through activation spreading process. Moreover, LPA algorithm makes use of two weighted algorithms as well. Communities can include overlapping communities as well. Overlapping communities share a common node across two communities which is believed to be the influential node and helps in spreading information across various communities. These overlapping nodes form a bridge between two communities to communicate and spread information across.

Some application have discussed about the collaborative conditions among scientists and how it is going to benefit from the collaborative network within and across different communities [1]. They mainly focused on scientist collaboration so that they can be cohesive in nature in order to perform better decision making when they belong to same knowledge domain and the results reverse when they make decision by consulting different knowledge domain. It focuses on the interaction among scientists and how the information flows between them. It follows by taking the structural viewpoint and then discussing the collaborative network of scientists in a community by their ties which connects two or more scientists who co-write a paper.

Diffusion Model
We have two diffusion models as, independent cascade (ICM) and linear threshold (LT). Both models consider the nodes to be in either of two states (active/inactive state) where active state depicts that information is adopted by individual and inactive means that the individual has still not accepted the information. A node will change its form from active to inactive state. Spreading starts from seed set which consists of initial spreaders and in discrete time steps it unfolds. Both ICM and LT are probabilistic models in which each node is activated at certain probability.

ICM
In ICM seed set tries to activate its neighboring individual only once. At any time n, if a node is recently activated at time step n-1, then current node will try to activate its inactive neighbor at time n+1 only at once with a certain probability. At time n+1, if attempt is successful it will get activated otherwise it remains inactive. Every node independently tries to activate its neighbors and the spreading process gets completed when no inactive node changes to active state [2].

LTM
In LT model, every individual in the network is allotted a threshold random value between 0 and 1. This indicates the proneness of a node to get converted into active state. At time step n, each neighbor individual can get activated only if the threshold is less than the sum of all inactive neighbors. Spreading process gets completed until no activation is possible.

RELATED SURVEY
Community structure favors a lot to know the functionality and association of a social network, since they contribute to the real organization and associations. Some researches have been done which have displayed community structure having significant effect on spreading of an information. However most existing methods do not consider influence due to community structure in account which somehow restricts their usage.

Influence Based on Individual
Muhuri *et al*. proposed an innovative method for finding the influential nodes of a network [3]. The algorithm comprised of two important

factors which includes edge contribution factor and clustering coefficient. It can identify influential nodes from a dynamic and large online network. The edge contribution factor includes the degree centrality and connection with the neighboring edges, which is integrated with clustering coefficient in order to make it efficient in terms of time. Clustering coefficient tests the neighborhood links based on their connectivity.

Wang *et al*. proposed an algorithm which was based on clustering degree algorithm (CDA) for identification of the most influential nodes in weighted complex network [4]. CDA makes use of Kendall's tau coefficient and have high precision in finding influential spreaders. The degree of a weighted node has been defined according to the node strength and its degree. Thus, the clustering degree is calculated according to the topological network structure that is obtained to calculate propagation capability and contribution from its neighbors.

Amrita *et al*. proposed a weighted k shell degree neighborhood method to identify influential spreader that is efficient in spreading process, which does not need the completeness of network structure [5]. The effectiveness of the proposed method is verified with six real networks in comparison to Susceptible-Infected-Recovered (SIR) model and Kendall τ rank correlation coefficient as a reference.

Yang *et al*. considered the dynamic nature of networks which most of the existing systems compromise on by modelling dynamic networks as edge weight updates stream using two important models of diffusion process (linear threshold and cascading model) [6]. The paper deals with finding influential individuals in a dynamic network by an algorithm, which incrementally update sample RR set which consists of set of influencers that are found by each poll in polling based method irrespective of network change. In polling based method, a social network is given in which a poll is conducted by randomly chosen node and tries to find those nodes which are likely to get influenced by random node; then those nodes are taken as a set. Maintaining RR (Reversely reachable) sets and using polling based method are two key ideas for finding top K influential individuals in a dynamic network to approximate the influence with quality guarantee. Two diffusion models review polling based method for calculating information dissemination and formulate that influence in dynamic network. Based on a particular user defined threshold, nodes are tracked and then top K influential nodes are to be found after that. The proposed algorithm showed that ranking of the top nodes constantly has good quality as compared to the two heuristics and the incremental algorithm outperforms rerunning of sampling algorithm by some order of magnitude from the scratch in terms of scalability and memory usage with respect to the input size, and the number of RR sets needed to approximate the influential nodes exactly. Overall, the algorithm proved to be efficient and effective in both time and space.

The paper by He *et al*. deals with the finding of multiple spreaders by dividing the network [7]. A novel method has been described in which a network has been divided into various super nodes and then according to centrality index K spreaders are taken from it. If the super node includes a spreader, then the edges incident/connected to it will not be taken as spreader anymore. Normally, a community evolves from many sub-communities and the hierarchy can be found by Blondell method, communities are replaced by super node and these super nodes are then stored in Red black Binary tree. SIR method is used to test the method. The method outperforms other methods like degree centrality-core, and cluster Rank in most cases and further it is compared with k-media method. It slightly outperforms k-media as well. The proposed method can pick K spreaders more rapidly in large networks.

Influence Based on Community Structure

Sattari *et al*. proposed a modified label propagation algorithm (LPA). LPA being very simple fast tries to generate huge communities decreasing the accuracy of community detection [8]. LPA can detect only communities that are disjoint, while the real network communities overlap. To overcome the problem of over-propagation of more than one labels, improved method based on spreading activation used in semantic networks and centrality measurement, a little source nodes are taken randomly, where each of them has an activation value which represents the power of source nodes for spreading process to its neighbors. The method proposed is Spreading Activation-based LPA (SALPA), which tags every node with a pair of (activation value and name of label). Then, it applies two weighting algorithms for graphs; based on breadth-first search, each node directs activation value to other node. It is used to detect overlapping communities by making use of spreading activation process till the value goes below the defined threshold. Lastly, each node adds the pair (activation value and name of label) or updates based on its neighbors' activation value.

Ahajjam *et al*. proposed a leader community detection approach for detecting communities [9]. The earlier approaches used to detect the communities needed to know the size and the partition of the communities in advance, but this approach overcomes these by developing a new scalable approach as leader community detection. The method follows two important steps as detection of leaders and then finding the communities. Detection of leader nodes is done through eigenvector centrality and then the nodes are ranked according to the centrality measure; the nodes with the highest eigenvalue in the adjacency matrix are placed at the top. After finding the leader nodes, the second step includes detection of community by similarity function to find the neighbor of the leader nodes which are found in the earlier step. Based on certain threshold for similar nodes will form a community.

This approach does not need to know the size and without preceding knowledge to know the number of communities to detect. It uses a new version of leader rank algorithm to detect the communities by using similarity function to detect the leader from similar nodes [7]. The algorithm is tested on real life social networks and is suitable for undirected and unweighted graphs with the complexity of O (n^2). It represents a significant improvement over the previous version of leader Rank algorithm.

Bhat *et al*. proposed an HOC tracker which is used for tracing the evolution of hierarchy and overlapping communities in a network [10]. Multiple smaller communities form a bigger community and these communities have been dynamic in nature with each time adding and deleting of nodes. The proposed method deals with the dynamic nature of the network by addressing most of issues of DENGRAPH–IO. Hence, it aims to trace the evolution applicable to undirected/directed and weighted/unweighted networks.

It uses a density based approach for identification of communities for density based clustering. Cluster is obtained by perceiving in underlying database the neighborhood of every object. If the neighborhood of a given radius contains more than μ objects, where μ is minimum cluster size, a fresh cluster with that object is taken as a core node forming density based community. Moreover it uses the distance function to know whether a pair of node belongs to the same community or not which in turns includes Jaccard coefficient, topological overlap and cosine similarity. It works on mutual core connected maximal subgraphs which consist of core nodes. After finding the initial structure of community, evolution of communities is taken in consideration. The results show that the given method is faster than some of benchmark methods like CFinder and OSLOM and is used to identify community evolution of real-world communities.

Zhang *et al.* proposed a method that does not rely on the structure of network by using a probabilistic model known as COSINE [11]. It makes use of users' high order proximity which is believed to exist occur naturally in network consisting of multiple communities. It exploits the first, second and high order proximity. The proposed model assumes that the diffusion process starts with an unobserved network at time zero. The nodes that have direct connection to the source node (start node) get infected first. Once the user is infected, it continues to infect its direct connections. Each spreading process takes random time and once a user gets infected it cannot be infected again. The data which is inputted to the algorithm is cascade data. Each cascaded data contains tuple set of infected users and the time to get infected; it uses Euclidean squared distances for timestamp.

It models temporal dynamics as well as community preserving methods by using Gaussian mixture model (GMM). Experimental studies are carried on both real and synthetic real world networks. It helps better in understanding social interactions, effectiveness and scalability by investigating performance with different training cascades. It also helps in prediction of diffusion by checking through test cascades whether a user gets infected in a given time or not.

Bourigault *et al*. proposed a model which is an embedded version of independent cascade model of diffusion to obtain more robust probabilities of diffusion than any other classical model [12]. The model was based upon some hypothesis that the infection is a binary process, diffusion networks are unknown, influence relationship are independent of propagated content and infection probabilities do not vary with time between users. Based upon this assumption, they formulated an alternate approach to IC model which is based on representative learning. The model assumes the diffusion space as continuous space, relative position of an individual is going to define content transmission likelihood. Moreover, it also developed a stochastic gradient method for large datasets. Considering modified version of IC model, they make use of partial orders of infection rather infection timestamp exactly. The infection depends on all the previously infected nodes and calculates the infected probability. It makes use of the distance to calculate probability of transmission. The experiments are performed on large collection of datasets. The results showed that the performance remains same with using less parameter learning.

Influence Maximization

Influence maximization forms a recent focus in analysis of social network analysis. Influence maximization is finding the initial users, which will influence the widest range possible and to largest number in a network. It can be somewhat related in determining influential spreaders in a network.

Wilder *et al*. formulated an algorithm which queries individual nodes to learn their links [13]. The algorithm starts by locating a seed set which will be influential as the global optimum by using certain queries. The proposed ARISEN algorithm uses querying leveraging community structure by finding seed set of influential nodes. It was applied for undirected graph. Seed set of k nodes are selected with aim of increasing expected size of resulting cascade influence. The diffusion model here used is Independent Cascade model. Initially the seed nodes are only active nodes in network; then each node tries to activate connected neighbor with a probability which is assumed to be same for all edges. For each neighbor to become active, Jump-Crawl model is adopted and when neighbor becomes active its all connected neighbors are revealed. The graph is obtained by stochastic model. The experiment is done on a real world network of households of a village in India. The results showed that the proposed system outperforms several baselines and is global optimum.

Zhao proposed an algorithm based on community structure that is used to find the influential nodes in a network [14]. The method uses IM algorithm which is based on label propagation; it includes two steps namely finding seed set and then seed set is manipulated further in the second step namely label phase [7]. In this step, each node of seed set is labelled by a unique label and rest nodes are kept with empty label sets. Then each node is updated by its label at time step n synchronously by:

$L'_v(t)=\text{argmax}\sum\delta(l, L_u(n-1))$

$L'_v(t)$ is a set of label at time n for node v, δ is kronecker delta function and $L_u(n-1)$ is set of label at time n-1 for node v. Labels tries to expand by attempting to acquire more neighbors when label reaches a particular node, they start competing to grab the node and neighboring node accepts the most frequent label. The labels which defeat other labels will be the most influential node and will gradually occupy other nodes. New label set is derived and then by calculating centrality, we get the top k influential active seed node set.

Dong proposed an efficient algorithm based on ranking method as semi-local centrality to find the influence of a node more precisely [2]. This method starts with random walk to grab influential surround nodes. It is based upon an idea that if a node is surrounded by many influential nodes then the node has high probability to be influential itself. By evaluating its surrounding nodes starting from the source node v, it performs a random walk multiple times to all the paths in its neighborhood collecting all the nodes in the path. The cardinality of set obtained is used to find the centrality of source node v. Random walk starts from the source node up to length of path not larger than l. With probabilities, nodes are added to path. Hence it forms an evaluation for adding new nodes to a random path. Probabilities depend on three things namely node similarity, degree centrality and attenuation coefficient. By integrating three factors, probability is calculated.

CONCLUSIONS

In this survey study we have presented a brief survey on the information diffusion influence due to many entities. We have also discussed some challenges for almost each, and briefed about diffusion models. Influence can be studied in different ways and we can combine any two ways to get more precise information about the network. By manipulating community structure or the ties among many individuals can help in predicting future diffusion or even in prevention of spreading of information or time and probability of an information reaching from one node to another node.

REFERENCES

1. Lambiotte R, Panzarasa P. Communities, Knowledge Creation, and Information Diffusion. *J Informetr.* Jul 1, 2009; 3(3): 180–90p.
2. Dong J, Ye F, Chen W, *et al.* Influential Nodes in Complex Networks via Semi-Local Centrality. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. May 27, 2018; 1–5p.
3. Muhuri S, Chakraborty S, Setua SK. An Edge Contribution-Based Approach to Identify Influential Nodes from Online Social Networks. In *2016 IEEE International Symposium on Nanoelectronic and Information Systems (iNIS)*. Dec 19, 2016; 155–160p.
4. Wang Q, Ren J, Wang Y, *et al.* CDA: A Clustering Degree Based Influential Spreader Identification Algorithm in Weighted Complex Network. *IEEE Access.* 2018; 6: 19550–9p.
5. Namtirtha A, Dutta A, Dutta B. Weighted k Shell Degree Neighborhood Method: An Approach Independent of Completeness of Global Network Structure for Identifying the Influential Spreaders. In *IEEE 10th International Conference on Communication Systems & Networks (COMSNETS)*. Jan 3, 2018; 81–88p.
6. Yang Y, Wang Z, Pei J, *et al.* Tracking Influential Individuals in Dynamic Networks. *IEEE Trans Knowl Data Eng.* Nov 1, 2017; 29(11): 2615–28p.
7. He JL, Fu Y, Chen DB. A Novel Top-k Strategy for Influence Maximization in Complex Networks with Community Structure. *PloS One*. Dec 18, 2015; 10(12): e0145283, 1-10p.
8. Sattari M, Zamanifar K. A Spreading Activation-Based Label Propagation Algorithm for Overlapping Community Detection in Dynamic Social Networks. *Data Knowl Eng*. Jan 1, 2018; 113: 155–70p.
9. Ahajjam S, El Haddad M, Badir H. A New Scalable Leader-Community Detection Approach for Community Detection in Social Networks. *Soc Networks.* Jul 1, 2018; 54: 41–9p.
10. Bhat SY, Abulaish M. HOC Tracker: Tracking the Evolution of Hierarchical and Overlapping Communities in Dynamic Social Networks. *IEEE Trans Knowl Data Eng*. Apr 1, 2015; 27(4): 1019–3p.
11. Zhang Y, Lyu T, Zhang Y. COSINE: Community-Preserving Social Network Embedding from Information Diffusion Cascades. In *32nd AAAI Conference on Artificial Intelligence*. Apr 26, 2018.
12. Bourigault S, Lamprier S, Gallinari P. Representation Learning for Information

Diffusion through Social Networks: An Embedded Cascade Model. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining.* Feb 8, 2016; 573–582p.
13. Wilder B, Immorlica N, Rice E, *et al*. Influence Maximization with an Unknown Network by Exploiting Community Structure. *InSocInf@ IJCAI*. 2017; 2–7p.
14. Zhao Y, Li S, Jin F. Identification of Influential Nodes in Social Networks with Community Structure Based on Label Propagation. *Neurocomputing*. Oct 19, 2016; 210: 34–44p.

Cite this Article
Tabasum Sami, Zahoor Ah Najar, Amjed Husain. A Survey on Information Diffusion in OSNs based on Influence. *Journal of Artificial Intelligence Research & Advances*. 2019; 6(2): 25–31p.