Social Network Analysis and Mining (2019) 9:40
https://doi.org/10.1007/s13278-019-0586-6

ORIGINAL ARTICLE

A node activity and connectivity-based model for influence maximization in social networks

Bhawna Saxena[1] · Padam Kumar[1]

Received: 29 September 2018 / Revised: 23 May 2019 / Accepted: 25 May 2019 / Published online: 1 August 2019

Abstract
Influence maximization deals with the problem of identifying *k*-size subset of nodes in a social network that can maximize the influence spread in the network. In this paper, the problem of influence maximization using two aspects, node connectivity and node activity level, has been studied. To measure node connectivity, the widely popular and intuitive measure of out-degree of node has been used, and for node activity, node's past interactions have been taken into consideration. For studying influence spread, two activity-based diffusion models, namely *Activity-based Independent Cascade* model and *Activity-based Linear Threshold* model, have been proposed in which influence propagation is driven by a node's activity that it has actually performed in the past. Activity-based models aim at studying influence spread by incorporating a more realistic aspect corresponding to user behavior. Motivated by the belief that activity is as important as connectivity, *UAC-Rank* algorithm for the identification of initial adopters has been proposed.

Keywords Online social networks · Influence maximization · Node activity · Node connectivity

1 Introduction

The emergence of so many online social networks (OSNs), like Facebook, Twitter, Instagram, Snapchat, has changed the way people connect and interact worldwide. OSNs have become an integral part of people's lives and have led to an explosion in the volume and variety of information available for research and analysis. Social network analysis is gaining momentum due to its use in varied applications, and the need to analyze the behavior of individuals in online environments.

A social network has three main components—(i) *users*—who connect and interact with each other, (ii) *network*—formed by the connecting links between the users, and (iii) *content*—the information that is being exchanged.

The members of a social network interact and exchange information with each other, thus diffusing information through the network. The propagation of information across the network through interactions among the members is known as *information diffusion*. A pervasive feature of information diffusion analysis is *social influence analysis*. Social influence refers to how the opinion of an individual impacts the opinion of another individual. According to Merriam-Webster dictionary, influence means "the act or power of producing an effect indirectly, or without apparent use of force or exercise of command." Influence displays certain properties, like it is dynamic, propagative, transitive, composable, measurable, subjective, asymmetric, etc. (Peng et al. 2018).

Various works being done on OSNs can be broadly divided into three categories, namely—*structural analysis*, *social data analysis*, and *social interactions analysis* (Kurka et al. 2016). Structural analysis deals with the structure and functionality of the network, like topology-based analysis, information diffusion models, etc. Social data analysis makes use of the data being generated and exchanged over the network and covers areas like sentiment analysis, emerging topic prediction, social recommendation systems, etc. Work in the category of social interactions analysis is based on user interactions. It covers areas like, cascade prediction, influence analysis, etc.

In today's world, OSNs are also emerging as one of the most impactful marketing and information spreading

✉ Bhawna Saxena
bhawna161@yahoo.co.in

[1] Department of Computer Science & Engineering and Information Technology, Jaypee Institute of Information Technology, A - 10, Sector - 62, Noida, Uttar Pradesh 201309, India

platforms. Consider a hypothetical scenario wherein a company wishes to promote their new product using OSN. It needs its online marketing campaign to reach out to a large number of people in a short span of time. Every campaign has a pre-decided budget, and so the company can give out samples to a limited number of people only. It would thus want to identify a few select people who, through their connections, can help in maximizing the spread of information about the product. So it needs to identify initial adopters who are influential and can influence their friends, especially those friends who can further influence their own friends and so on, till the influence spreads to a large number of people. Besides the identification of initial adopters, equally important aspect is the model for spread of influence. So in view of information diffusion within a specified time frame, when deciding upon which node to choose next, a node's activity level should be taken into consideration. Edges with higher activity frequency should be assigned higher propagation probability as compared to those with lower activity frequency. The problem of finding initial adopters who can eventually lead to a large number of people getting influenced is called *influence maximization*. It finds use in multiple domains like viral marketing, targeted marketing, political campaigns, search engines, recommendation systems, etc. Influence maximization aims at addressing tasks like "finding the most influential users in a network," "finding the influence spread of an influencer," or "who can influence whom" (Peng et al. 2017).

The process of influence maximization can be carried out by making use of node connections (in a social network, a node represents a user), node behavior and/or the content being propagated. Centrality, a measure based on node connections, is one of the most popular measures used for identifying influential nodes in a network. There are many types of centralities, like degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, etc., which are based on node connections and where in the network the node is located. Centrality measures often find use in applications involving viral marketing. For applications such as trend analysis, topic-related predictive analysis, or targeted marketing, node behavior and content play an important role.

For studying the spread of influence, node activity-based models have been proposed, namely *Activity-based Independent Cascade* (AbIC) model and *Activity-based Linear Threshold* (AbLT) model. AbIC and AbLT models draw inspiration from Independent Cascade (IC) model and Linear Threshold (LT) model, respectively (Kempe et al. 2003). In the proposed activity-based models, the propagation probability over a connection (edge) has been computed based on the actual activity performed by the user in the past. Considering real-world scenarios (like the one mentioned above), where the goal is to spread information across the network within a given time frame, user activity frequency should be considered as a decisive factor. This thought forms the basis of computing propagation probability as per user activity frequency. It has been believed that the activity performed by a user gives a better picture of real user behavior and should be given adequate consideration when computing a user's influence potential. Driven by this thought, a heuristic-based algorithm *UAC-Rank* has been proposed, which addresses the issue of identifying initial influential users (seed set) for the process of influence maximization, by considering both user activity (node behavior) and connectivity (node connections).

The rest of the paper has been organized as follows. Section 2 presents a brief discussion about some of the existing algorithms and models developed for addressing the problem of influence maximization. In Sect. 3, the concepts of user connectivity and communication and why they should be brought together have been discussed. Section 4 briefly describes the problem definition. In Sect. 5, the proposed UAC-Rank algorithm aimed at the identification of initial users for maximizing the spread of influence has been presented. Section 6 presents details about the proposed activity-based models, AbIC, and AbLT. Section 7 presents the experiments performed and also an analysis of the results obtained. Section 7.1 elaborates on the details of the datasets that we have used to evaluate our work. Finally, Sect. 8 concludes the paper and discusses the future scope of work.

2 Related work

Influence maximization is one of the key research problems in social network analysis that aims at identifying a small set of highly influential nodes (initial adopters) that are able to maximize the influence spread across the network.

The process of influence maximization involves two major activities—identification of seed nodes and development of diffusion model. Domingos and Richardson (2001) were the first to work on the problem of influence maximization. Since then, a number of algorithms and models have been developed for influence maximization in social networks (Alshahrani et al. 2018; Chen et al. 2009; Deng et al. 2015; Goyal et al. 2011; Heidemann et al. 2010; Jianqiang et al. 2017; Kempe et al. 2003; Liu et al. 2017; Morone and Makse 2015; Sheng and Zhang 2018; Tong et al. 2017; Wang et al. 2014; Zhu et al. 2017). Based on the technique used, the algorithms and models can be broadly classified into two categories—greedy based and heuristic based.

2.1 Greedy-based techniques

- Kempe et al. (2003) proposed a greed-based solution for selection of influential nodes, wherein a node is chosen based on the concept of marginal gain, i.e., in every iteration that node is selected which contributes maximum gain toward the influence spread process. Additionally, they have proposed three diffusion models, namely—IC, LT, and Weighted cascade (WC).

- Inspired by the *Credit Distribution* (CD) model (Goyal et al. 2011) which studies diffusion of influence in a network based on information propagation traces, *Credit Distribution with Node Features* (CD-NF) has been proposed which incorporates user static influence as well as user dynamic influence (Deng et al. 2015). Edge probabilities are derived from past propagation traces which capture information pathways that get created right from the user who introduced a topic to the most recent user. Pathways consider the node as well as the content being propagated. Furthermore, *Greedy algorithm with Node Features* (GNF) for identifying initial seed set has also been proposed.
- Zhu et al. (2017) proposed a greedy algorithm for *Structure-Hole-based Influence Maximization* (SHIM) which is based on the belief that structure–hole nodes are more influential when aiming at spreading information between communities. Influence potential of structure holes having value above a given threshold is then quantified to generate a candidate set, from which seeds are selected.
- *LPIMA* algorithm is based on label propagation community detection and aims at influential node identification (Sheng and Zhang 2018). Candidate node set is generated by quantifying the influence potential of community nodes using LeaderRank centrality.
- Tong et al. (2017) have developed two adaptive seed user selection strategies—*A-Greedy* and *H-Greedy*. Under this approach, seeding strategy is adaptively constructed such that selection of nodes in current round is dependent on the outcome of the previous rounds. Additionally, they have also proposed *Dynamic Independent Cascade* (DIC) diffusion model wherein the propagation probability between nodes is not static and is randomly selected from a pre-defined distribution.

2.2 Heuristic-based techniques

- Chen et al. (2009) have proposed two degree discount heuristic algorithms for identification of initial seed set, namely—*SingleDiscount* and *DegreeDiscountIC*. Seed nodes are selected based on their degree. For each seed node, degree of their neighbor nodes is discounted. SingleDiscount discounts the degree by 1, and DegreeDiscountIC discounts the degree based on degree of neighbors and number of neighboring nodes already selected as seed nodes. Performance of *DegreeDiscountIC* has been found to be comparable to greedy algorithm for IC model.
- *PRDiscount* is a heuristic scheme for initial seed selection (Wang et al. 2014). *PRDiscount* assigns an influential power to each node, and when a node gets selected into the initial seed set, the degree of its neighbors is discounted accordingly so as to lessen the neighborhood overlapping effect. *PRDiscount* draws inspiration from *PageRank* algorithm (Page et al. 1998) wherein, the PageRank of a web page is computed on the basis of its predecessors' PageRank values. The PageRank of a web page gets divided equally among all its outgoing paths. Performance of *PRDiscount* has been found to be better than *DegreeDiscountIC* and comparable with greedy algorithms.
- *Local index rank* (LIR) is a novel topology-based algorithm for mining top *k*-nodes (Liu et al. 2017) and is based on the "rich-club phenomena" which suggest that nodes with high degree are connected with other high-degree nodes. LI score is assessed for each node by computing the difference between node's degree and its neighbor's degrees. Nodes having LI = 0 are believed to be leaders in their local neighborhoods and are selected as seeds.
- Alshahrani et al. (2018) proposed *PrKatz* algorithm which aims at identifying top *k* influential users on the basis of Katz centrality measure and propagation probability computed using node degree.
- *UIRank* is a user influence ranking algorithm that aims at identifying influential users in micro-blog networks based on influence of their tweets and importance of their location in the network (Jianqiang et al. 2017). Influence of tweets has been measured by computing retweet to read ratio and comment to read ratio. User contribution measures the importance of user location based on out-degree centrality, betweenness centrality, and out-degree closeness centrality values for a user.
- *Linear Threshold Rank* (LTR) proposes a new centrality measure which can be used to identify initial seed users (Riquelme et al. 2018). LTR measures the influence of a node based on its capacity to influence other nodes and its own resistance to getting influenced.
- A *novel PageRank-based algorithm* that uses the concepts of users' connectivity and communication activity has been developed to identify the top *k* users in a network who are unlikely to quit using the network (Heidemann et al. 2010). A weighted activity graph is first derived, in which the weight of an edge has been computed based on number of interactions that take place over it. Users' centrality scores have then been determined using the weighted activity graph, and the approach has been inspired by PageRank algorithm.
- Agarwal and Mehta (2018) proposed algorithm for seed set identification using genetic algorithm with dynamic

edge probabilities. Dynamic probabilities have been computed using topical affinity propagation method.

A summary of the current state-of-art algorithms is given in Table 1.

On the basis of the aforementioned works, it has been observed that the researchers have adopted many approaches to tackle the problem of influence maximization. Under greedy-based approach, techniques of structure–hole-based influence maximization, label propagation community detection, adaptive greedy-based user selection are among the state-of-art techniques. Under heuristic-based approach, state-of-the-art techniques include degree discount-based algorithms, PageRank inspired algorithms, rich-club phenomena-based method, centrality-driven measures, genetic algorithm inspired techniques. It has been found that most of the initial seed selection methods make use of node's degree or connections to measure its influence potential. In some works, it has been suggested that user (node) activity should also be considered when measuring the influence of a node. All these techniques make use of the node's connections, but have not considered the real behavior of the node when quantifying its influence potential. When computing the influence potential of a node, the actual activity performed by the node over a period of time should also be considered as an important aspect, as it gives a realistic measure of how much the node is making use of its connections. Frequency of interacting with neighbors is as important as having neighbors. A node might be having a lot of connections and might be placed quite centrally in the network, but if it is not using those connections, then it is more like a dormant node and its influence potential should be adjusted accordingly. Thus, having a large number of connections should not be the only criteria under consideration for quantifying the influence potential of a node. When dealing with time-critical activities, techniques for influence maximization should also give adequate weightage to the actual past behavior of the node. Hence, in the proposed work, a novel method for addressing the problem of influence maximization has been proposed that makes use of both user connectivity and user activity.

3 User connectivity and activity

One of the most popular and intuitive measures used for evaluating the importance of a node is its degree, which is a measure based on the number of connections a node has. It is a general opinion that more the number of connections a node has, the higher are its chances of spreading influence. Edge propagation probability in popular information diffusion models, namely—IC, LT, and WC, has either been kept constant for all edges (like in IC) or has been computed based on the degree of nodes (as in LT and WC) (Kempe et al. 2003). So, a node's connections seem to play a decisive role in these models used for studying the spread of influence across a network. User connectivity is a pivotal parameter for key user identification as well. Users having more connections, direct and indirect, are considered to be more important as they are likely to be more influential. Therefore, topological characteristics of nodes have been widely analyzed for the identification of influence spreaders. Depending upon the direction of an edge, high connection nodes can be classified into *Hubs* and *Authorities* (Kleinberg 1999). Nodes with high number of outgoing links are considered as *Hubs*, and nodes with high number of incoming links are considered as *Authorities*. So, nodes in a hub position seem to be more apt for the role of influence spreaders owing to a higher number of communication and interaction channels.

Studies have shown that user activity also plays an important part in key user identification. Consider a hypothetical scenario wherein an organization has only 15 days for running its online campaign. It would want to spread the information to more and more people in these 15 days. So, in this situation, which node should be chosen by the organization, a node with higher degree but lower activity level or a node with lower degree but higher activity level? Marketing campaigns are usually time-critical, and so as much as the number of connections is important, so is the frequency of communication between users. Say there are two nodes *A* and *B* in the network, such that *A* has five friends and it communicates with them daily and *B* has ten friends, but it communicates with them once in 15 days. Under this scenario, *A* seems to be a better influencer than *B*, even though *A* has fewer connections than *B*. This is because *A* has a higher probability of spreading the information, as it is interacting daily with its neighbors, whereas *B* is more like a dormant user who carries out interactions at a much lower frequency. Inspired by this thought, activity-based diffusion models have been proposed, wherein the edge propagation probability has been computed purely on the basis of a node's actual activity that it has carried out in the past, and influence spread has been studied from the perspective of user activity instead of user connectivity.

Using activity, an activity network can be derived from a social network, which represents the actual communication channel between nodes instead of static social connections (Heidemann et al. 2010). Figure 1a shows an undirected connectivity network based on the social links between users, and Fig. 1b shows its corresponding activity network based

Table 1 Summary of current state-of-art algorithms and models

Approach	Author reference	Activity (algorithm/model)	Features	Diffusion model used	Advantages	Disadvantages

Greedy	Kempe et al. (2003)	Seed identification (*Greedy*)
		Diffusion models (*IC, LT, WC*)
			Seed nodes are chosen based on marginal gain	IC, LT, WC	High accuracy;
					Higher influence spread of seed set
						High execution time;
						Low efficiency;
						Not suitable for very large networks

	Deng et al. (2015)	Seed identification (*GNF*) Diffusion model (*CD-NF*)	Edge probabilities derived from past propagation pathways; Consider node features as well as the content; Model incorporates both user static influence and user dynamic influence	CD-NF		
	Zhu et al. (2017)	Seed identification (*SHIM*)	Structure holes identified using adjacency matrix and degree matrix	IC		
	Sheng and Zhang (2018)	Seed identification (*LPIMA*)	Influence of community nodes quantified using LeaderRank centrality	IC		
	Tong et al. (2017)	Seed identification (*A-Greedy, H-Greedy*) Diffusion model (*DIC*)	Selection is adaptive; Dependent on past rounds; Propagation probability selected from pre-defined distribution	IC, DIC		

Heuristic	Chen et al. (2009)	Seed identification (*SingleDiscount, DegreeDiscountIC*)	Degree of neighbor nodes of chosen seed is discounted	IC, LT	Low execution time;
					High efficiency;
					Suitable for very large networks
						Seed set influence spread cannot be guaranteed

	Wang et al. (2014)	Seed identification (*PRDiscount*)	Influence potential of node takes into account influence potential of neighbors; Degree of neighboring nodes is discounted	WC		
	Liu et al. (2017)	Seed identification (*LIR*)	Seed nodes have highest degree in local neighborhood	IC, LT		
	Alshahrani et al. (2018)	Seed identification (*PrKatz*)	Katz centrality score; Propagation probability	IC		
	Jianqiang et al. (2017)	Seed identification (*UIRank*)	Influence of tweets; Importance of user location	Not specified		
	Riquelme et al. (2018)	Centrality measure for seed identification (*LTR*)	Potential to influence others; Resistance to getting influenced	LT		
	Heidemann et al. (2010)	Identification of top k users retained in network (*Novel PageRank based algorithm*)	User connectivity; Communication activity of users in the network	Not specified		
	Agarwal and Mehta (2018)	Seed identification (*Genetic Algorithm with Dynamic probabilities*)	Genetic algorithm; Topical affinity propagation	IC		

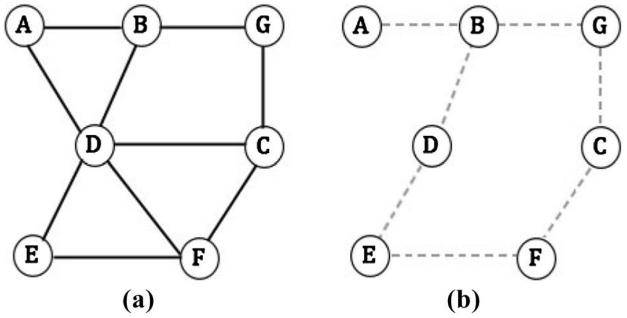


Fig. 1 **a** Connectivity network. **b** Activity network

on how users are communicating. As can be seen, node D has maximum connections, and so as per connectivity-based models/algorithms, node D would be considered as most influential. But even though node D has more connections than node B, it is not communicating along all its connections (as can be seen from the user activity network), and although node B has fewer connections than node D, it is utilizing all its communication channels. So the probable influence of node B is more than that of node D in such a scenario.

The proposed UAC-Rank algorithm for the identification of initial influential adopters leading to influence maximization in social networks is based on the aforementioned concept. UAC-Rank is based on the premise that both user connectivity and user activity frequency are important when measuring a node's influencing potential, and thus should be used together for identifying key adopters for influence maximization process.

4 Problem definition

A social network can be represented as a graph $G(V,E)$, where the set of vertices V represents the users and the set of edges E represents the connections between the users. Social networks could be both directed, like Twitter, and undirected, like Facebook. Every edge in a directed network has a source and target. In case of undirected networks, edges have no direction associated with them and there is only one edge between a pair of nodes. So an edge between nodes u and v would be considered as a two-way communication channel between the two nodes.

For successful viral marketing, the first step is to select a model to be used for information diffusion, then assess the influential capabilities of users, and finally select an initial set of adopters who can help maximize the spread. For a given network G, the aim is to find a k-node set S (such that $S \subset V$) that is likely to provide maximum spread of information over the network. The set S is known as the *seed set*, and k is the number of initial adopters. Let $\delta(S)$ denote the total number of individuals expected to adopt the information, then the goal of influence maximization is to find a k-node set S for which $\delta(S)$ is maximum.

When using a social network for viral spreading of information, the fact that how fast a node can spread information is as important as the number of connections it has. The work presented in this paper supports the premise that users with high activity levels, but fewer connections should be considered as equally good, if not better candidates as compared to dormant users with more connections. Most of the existing models address the issue of influence maximization based on either user connections or user activity. Degree seems to be the most obvious choice for tackling the influence maximization problem, as higher the degree of a node, more the paths to spread information. In some works, it has been suggested that node features like user activity and user's temporal behavior should also be considered when measuring the influence of a node.

As stated earlier, in this paper, the concepts of user connectivity and user activity have been brought together and user activity-based information diffusion models have been presented. An algorithm for the identification of initial seed set based on these two aspects has also been developed. The motivation for the work presented in this paper comes from the belief that not just the number of connections, but user's activity level should also be considered when measuring user influence potential as well as influence spread.

5 UAC-Rank algorithm

Unlike existing works, the proposed algorithm UAC-Rank addresses the initial seed selection problem by incorporating both user's activity and connectivity with other members. The approach consists of the following four steps:

1. Transform the given network into a weighted activity network, such that each node (user) and each edge (connection) have a weight associated with it. Assign *Node Activity Weights* (NAWs) to the nodes based on the number of communications they have initiated in a given time span (as in (1)), and *Edge Activity Weights* (EAWs) to the edges based on the number of interactions carried out over them by their corresponding source node (as in (2)).

   (*The time span is a pre-decided duration over which the interactions of the nodes are being recorded and* *analyzed. If different networks are being compared, then same time span should be taken. But if networks are being analyzed independent of each other, then time span can be different for different networks*).

2. In the second step, incorporate the node connectivity aspect. The aim here is to identify the hub nodes on the basis of node connections. For this, determine the nodes' out-degree scores.
3. Rank the nodes by combining their NAW and out-degree values.
4. Thereafter, k-node seed set of initial adopters is identified for the given network.

5.1 Generating weighted activity network

The given directed social network has been transformed into an activity network by assigning nonzero NAWs to those nodes which have initiated some activity in a given time period and zero value NAW to all other nodes. The NAW assigned to a node is directly proportional to the number of communications it has initiated in the given time span.

Say, the set $N(v)$ depicts the neighbors of node v, and $E(u, v)$ depicts an edge from node u to node v, then Node Activity Weight(u) is computed as per (1), and Edge Activity Weight corresponding to each edge is computed as per (2).

Once a weight is assigned to all nodes and edges, the activity level associated with each node and edge is known.

$$\text{Node Activity Weight}(u) = \sum_{v \in N(u)} \text{Interactions from node } u \text{ to } v \tag{1}$$

$$\begin{aligned}&\text{Edge Activity Weight } (u, v)\\&= \sum \text{Total interactions from node } u \text{ to } v \text{ over edge } E(u, v)\end{aligned} \tag{2}$$

5.2 Identifying hub nodes

The initial seed set is a k-node set, which means that we need to pick up k number of seeds from among all the nodes as initial adopters. Out-degree of a node is computed on the basis of the number of direct edges going out from a node. Hub nodes have high out-degree.

5.3 Node ranking criteria

The node ranking criteria takes into consideration both user activity and user connectivity, giving equal weightage to both of these aspects. A rank is given to all the nodes in the network, as per the ranking criteria, specified in (3).

$$\text{UAC-Rank}(v) = (\alpha * \text{Node Activity Weight}(v)) + (\beta * \text{Outdegree}(v)) \tag{3}$$

where α and β are the weights given to the two aspects of ranking node activity and connectivity. In present work, equal weightage has been given to both factors, and hence, $\alpha = \beta = 0.5$.

5.4 Identifying influential initial adopters

After ranking the nodes, the node with highest rank value is selected and added to the initial empty seed set. To alleviate the overlapping effect (Chen et al. 2009; Wang et al. 2014), inactive (yet uninfluenced) neighbors of the selected node are identified (nodes with edges directed toward the selected node) and their out-degree is reduced by one. This is done because once a node gets influenced and becomes active, the probability that one of its inactive neighbors will influence it loses significance. Thus, the out-degree of all inactive neighbors of selected node is reduced by one.

Next, the NAW for all inactive neighbor nodes is also adjusted. Since the edge between the selected node and its inactive neighbor is no longer useful in the influencing process, activity performed by the neighbor node over this edge (toward the selected node) has lost its purpose and so is removed from the computation of NAW of the neighbor.

After adjusting the activity levels and out-degree of all the inactive predecessor nodes of a selected node, fresh rankings are computed for all the nodes in the network. The node with the highest rank is then selected into the seed set, and the process is repeated. Repetition is done k times to get k-node seed set.

Algorithm 1 describes the UAC-Rank algorithm for identification of seed nodes.

```
Algorithm 1: UAC-Rank(G,k)

Input parameters: Graph G = (V, E), size of initial seed set k

Output parameters: initial_seeds[ ] :  k-size set of selected nodes

Algorithm:

1: For each node, compute Node Activity Weight (NAW) using (1), out-degree (OD) value, and UAC-Rank using (3);
2: Compute Edge Activity Weight (EAW) for each edge in the network using (2);
3: Initialize seed_counter  = 1;
4: WHILE seed_counter <= k DO
5:      selected_node = MAX(UAC-Rank(vi));
6:      initial_seeds[seed_counter] ← selected_node;
7:      Find predecessors of selected_node and store them as neighbors_of_selected_node;
8:      FOR neighbor in  neighbors_of_selected_node
9:          IF neighbor not in initial_seeds
10:             NAW(neighbor)  = NAW(neighbor) – EAW(neighbor, selected_node);
11:             OD(neighbor) = OD(neighbor) - 1
12:         END IF
13:         Recompute UAC-Rank for each node as per (3);
14:     END FOR
15:     seed_counter = seed_counter + 1;
16: END WHILE
```

Computation of NAW, OD, UAC-Rank for each node, and EAW for each edge (Steps 1,2) takes O(n) time, where n is the number of nodes in the network. Identifying node with highest UAC-Rank from n nodes and finding the predecessor nodes of that node takes O(n) time (Steps 5, 7). Steps 9–12 perform some constant operations which take O(1) time. Recomputation of UAC-Rank for each node will take O(n) time (Step 13). Since steps 9–13 are being repeated for all nodes, the total time taken would be O(n^2). Further, the computational time required by steps 4–16 is O(kn^2), as the algorithm runs for k iterations (k being the size of the initial seed set).

So, the algorithm takes O(kn^2) time to complete where k is the size of the initial seed set and n is the total number of nodes (users).

6 Activity-based model for influence maximization

Kempe et al. have proposed two widely popular diffusion models, IC model and LT model (Kempe et al. 2003). In IC and LT models, a node is either in the active state (when node has adopted the information) or inactive state. Both of these models are progressive in nature, i.e., once a node switches to active state, it remains in that state and cannot switch back to inactive state. The basic idea behind the working of these two models is that, as more and more neighbors of an inactive node u become active, chances of u switching its state from inactive to active increase. Diffusion in both of these models progresses in discrete time steps.

Under IC model of information diffusion, we start with an initial set of active nodes, say S at time t. At time $t+1$, each active node u has a chance of activating its neighbor node v with a probability $p_{u,v}$. But, u gets only one chance to activate v, i.e., if u succeeds in its attempt, then v becomes active, but if u is unsuccessful at time $t+1$, then it will not get another chance to activate v during the course of the diffusion cycle. If multiple newly active neighbors of v attempt to activate it at the same time, then their attempts will follow an arbitrary sequence. The diffusion process continues until no more activations can be done. In the classic IC model, the propagation probability is the same between any two nodes.

Under LT model of information diffusion, each node in the network has a threshold value θ associated with it, which depicts the minimum amount of influence needed to activate that node. Each edge has a weight associated with it which represents the influence being exerted over that edge. Each neighbor u of node v influences it with a weight $b_{v,u}$, such that $\sum_u b_{v,u} \leq 1$. So node v gets influenced when the sum of influence exerted by its active neighbors exceeds its threshold value. So this means that if each active neighbor w of node v influences it with weight $b_{v,w}$, then v will get influenced when $\sum_w b_{v,w} \geq \theta_v$. In this model, also we start with an initial set of active nodes S and assign a random threshold to all the nodes. The threshold value for all nodes is kept same. At time t, nodes in S are active. At time $t+1$, we activate all those nodes for which the sum of influence exerted by their active neighbors exceeds their corresponding threshold value. This process keeps repeating till no more activations can be carried out.

Driven by the thought that vitality of a node is as important as its degree, in this work, user activity-based model for studying influence maximization has been proposed. Two models have been presented, namely—*AbIC* and *AbLT*.

AbIC is a user activity-incorporated variation of the IC model. In IC model, the propagation probability between any two nodes is considered same, which is not the case in real-world networks, where the influence propagating between any two nodes is not the same always. Based on this premise, in AbIC model of diffusion, the propagation probability between any two nodes is dependent on the communication level between those two nodes. To calculate the propagation probability $p_{u,v}$ between nodes u and v, the past interactions between u and v have been taken into consideration. Based on the activity performed by the node in the past, the following formula to calculate the propagation probability has been defined:

$$p_{u,v} = \frac{I(u,v)}{I_{\text{Max}}} \tag{4}$$

where

$$I(u,v) = \text{total interactions from } u \text{ } to \text{ } v$$
$$I_{\text{Max}} = \text{highest number of interactions initiated by any node in the network}$$

Consider a scenario where there are four nodes a, b, c, and d such that a initiates an interaction *two* times with b, six times with c, and never with d. It can be clearly seen that a is interacting more with c as compared to b and d. Out of total eight interactions that a has initiated only two are toward b and none are toward d. Thus, it can be said that the probability of information propagating from a to c is higher than the probability of information propagating from a to b or a to c. So, considering a constant (same valued) propagation probability on each edge does not seem realistic as it is not modeling the real scenario aptly.

In AbIC model, diffusion starts with an initial set of k active nodes. Activity-based propagation probabilities are assigned to all the edges. Each active node u gets a single chance to activate each of its inactive neighbors v with the probability $p_{u,v}$ computed as mentioned in (4). The remaining diffusion process continues as in IC model.

AbLT is a user activity-incorporated variation of the LT model. In LT model, the influence (weight) value associated with each edge (which depicts the influence being exerted over that edge) is computed based on the in-degree of the target node. The influence on an edge is equal to the reciprocal of the in-degree of the target node. In AbLT model, the edge influence has been computed based on the user activity carried over that edge. To calculate edge influence $\text{EI}(u, v)$ between nodes u and v, past interactions between u and v have been considered. The following formula has been defined to calculate edge influence:

$$\text{EI}(u,v) = \frac{I(u,v)}{\sum_{w \in N_{\text{in}}(v)} I(w,v)} \tag{5}$$

where

$$I(u,v) = \text{total interactions from } u \text{ to } v$$
$$I(w,v) = \text{total interactions from } w \text{ to } v$$
$$N_{\text{in}}(v) = \text{set of neighbours of node } v \text{ that have edges directed towards } v$$

In AbLT model, activity-based influence has then been assigned to all the edges as per (5). Diffusion starts with an initial set of k active nodes. As in LT model, each neighbor of an inactive node influences it with the influence assigned to their connecting edge. A node v gets influenced when the sum of influence exerted by its active neighbors exceeds its threshold value.

7 Experiments

To demonstrate and evaluate the performance of UAC-Rank algorithm, and AbIC and AbLT models, the performance of proposed models has been compared with different models and algorithms over four publicly available real-world network datasets. Performance of AbIC model has been compared with that of IC, performance of AbLT has been compared with that of LT, and UAC-Rank has been compared with five existing algorithms having similar goal of identifying top k initial adopters, namely—Random (Kempe et al. 2003), Degree (Kempe et al. 2003), SingleDiscount (Chen et al. 2009), DegreeDiscountIC (Chen et al. 2009), and PRDiscount (Wang et al. 2014). The following section gives a description of the datasets being used for the evaluation of the work.

All experiments have been conducted on Windows with Python. Machine settings are Inter(R) Core(TM) i5-8250U CPU @ 1.6 GHz having 8 GB RAM.

7.1 Datasets

Experiments have been performed over four real-world social network datasets. All the four datasets are communication networks, i.e., they are networks representing some sort of communication (activity) that is taking place between the users of the network. In terms of number of users and number of edges, networks with varied sizes (both small and large) have been selected. Following are the details of the four datasets used:

- *UC Irvine messages*—A directed network of messages exchanged between users of an online community of students from the University of California, Irvine. There are 1,899 nodes with 20,296 static edges and 59,835 temporal edges, and each edge represents a message being exchanged between two nodes. There are multiple edges between same pair of nodes representing that multiple messages have been exchanged between those two nodes (available at http://konect.uni-koblenz.de/).
- *Math Overflow*—A directed network of interactions on the stack exchange web site Math Overflow. There are 24,818 nodes where each node denotes a user of the web site. Each edge denotes that either a user replied to another user's question or commented on it. There are a total of 2,39,978 static edges and 5,06,550 temporal edges in the dataset which also include multiple edges between same node pair (available at http://snap.stanford.edu/data/sx-mathoverflow.html).
- *Facebook wall posts*—A directed network representing a subset of posts written by a user on another user's wall. There are 46,952 nodes with 274,086 static edges and 876,993 temporal edges wherein each edge represents a post. As a user may write multiple posts to the same user, multiple edges connecting same pair of nodes exist in the network (available at http://konect.uni-koblenz.de/).
- *Email network of a manufacturing company*—A directed internal email network being used by the employees of a mid-sized manufacturing company. There are 167 nodes (employees) with 5,784 static edges and 82,927 temporal edges in the network. Each edge represents an email communication between two employees. Since, one employee can send multiple emails to another employee, the network contains multiple edges between same pair of nodes (available at http://konect.uni-koblenz.de/).

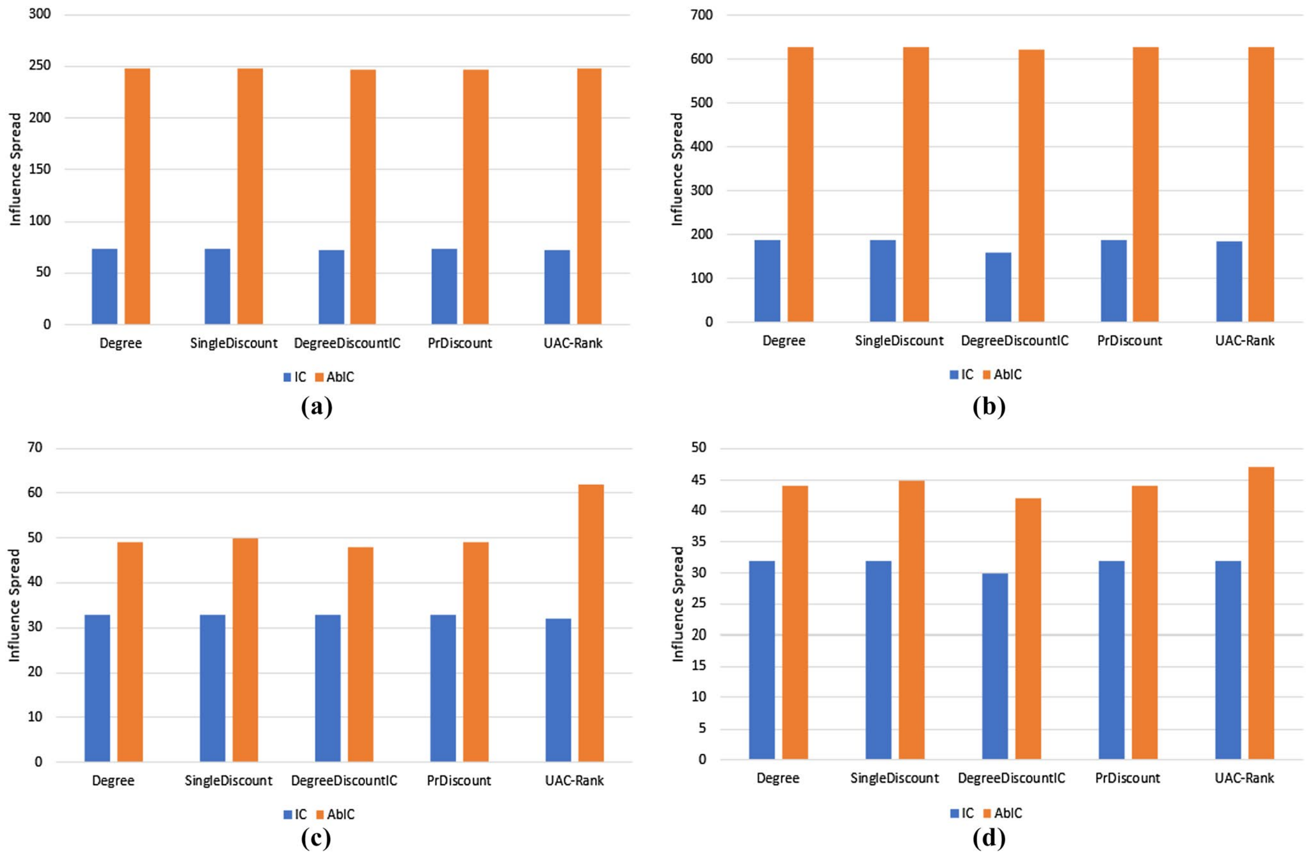
For evaluation purpose, the considered time span for *UC Irvine* dataset is 193 days (15-04-2004 to 26-10-2004), for *Math Overflow* dataset is 2350 days (29-09-2009 to 06-03-2016), for *Facebook Posts* dataset is 1591 days (14-09-2004 to 22-01-2009), and for *Email Network* dataset, it is 242 days (02-01-2010 to 01-10-2010). Different time spans have been taken for different datasets, as performance of proposed work is not being compared across datasets. However, when comparing the performance of AbIC with IC, AbLT with LT, and UAC-Rank with five existing seed identification algorithms using a particular dataset, same time spans have been considered for comparison.

7.2 Models and algorithms compared

Performance of AbIC model has been compared with that of IC model. In IC model, propagation probabilities are same across the network. For evaluation purpose, the edge probabilities in IC model has been set as $p_{u,v}$ = mode probability of EAWs (as computed under AbIC model) for all u, v. Thus, IC model has been made to work at a probability which is most frequent in AbIC. Performance of AbLT model has been compared with that of LT model.

To evaluate the performance of UAC-Rank algorithm for initial seed selection, its performance has been compared with the following five seed selection algorithms:

- *Random:* A basic scheme that picks k random nodes from the graph (Kempe et al. 2003).
- *Degree:* A simple heuristic scheme that picks k-nodes with largest degree as the initial seed set (Kempe et al. 2003).
- *SingleDiscount:* A simple degree discount heuristic scheme wherein the degree of each neighbor of a chosen node is discounted by one (Chen et al. 2009).
- *DegreeDiscountIC:* A heuristic-based scheme in which the degree of the neighbors of chosen nodes is discounted based on the degree of the node and the number of neighbors that already have been chosen into the initial seed set (Chen et al. 2009).
- *PRDiscount:* A PageRank inspired heuristic algorithm in which an influential power is assigned to each node



(a) (b) (c) (d)

Fig. 2 **a** Average influence spread achieved under IC and AbIC models for seed sets picked from UC Irvine dataset. (Average spread value has been shown based on values obtained for five seed set sizes $k = 10$, 20, 30, 40, and 50.) **b** Average influence spread achieved under IC and AbIC models for seed sets picked from Math Overflow dataset. (Average spread value has been shown based on values obtained for five seed set sizes $k = 10$, 20, 30, 40, and 50.) **c** Average influence spread achieved under IC and AbIC models for seed sets picked from Facebook Posts dataset. (Average spread value has been shown based on values obtained for five seed set sizes $k = 10$, 20, 30, 40, and 50.) **d** Average influence spread achieved under IC and AbIC models for seed sets picked from Email Network dataset. (Average spread value has been shown based on values obtained for five seed set sizes $k = 10$, 20, 30, 40, and 50)

which depends upon the influential power of its neighbors (Wang et al. 2014). When a node gets selected into the initial seed set, the degree of its neighbors is discounted accordingly.

Influence spread of seed sets picked by UAC-Rank algorithm along with those picked by using the aforementioned heuristic algorithms has been studied using the AbIC model of diffusion as described above.

7.3 Experimental results

7.3.1 Evaluation of AbIC and AbLT models

The performance of AbIC model has been compared with that of IC using the aforementioned four datasets. Initial seed sets of sizes 10, 20, 30, 40, and 50 have been generated using Degree, SingleDiscount, DegreeDiscountIC, PRDiscount, and UAC-Rank algorithms. These seed sets have then been given as input to both IC and AbIC models, and their influence spread has been compared. Influence spread achieved means how many nodes get influenced by the end of the diffusion process, when starting with k-sized initial seed set.

For each of the aforesaid algorithms, influence spread values have been computed for the chosen seed sets of varying sizes under both IC and AbIC models. Firstly, influence spread has been computed separately for seed sets of size, $k = 10$, $k = 20$, $k = 30$, $k = 40$, and $k = 50$. Thereafter, for representation purpose, average value for the spread achieved has been computed based on the values obtained for the five seed set sizes ($k = 10$, 20, 30, 40, and 50). The results pertaining to the average influence spread achieved, under both IC and AbIC models, by the five algorithms under consideration are illustrated in Fig. 2a, b, c, d.

The four figures are corresponding to the four datasets used. On the basis of the results obtained, it has been

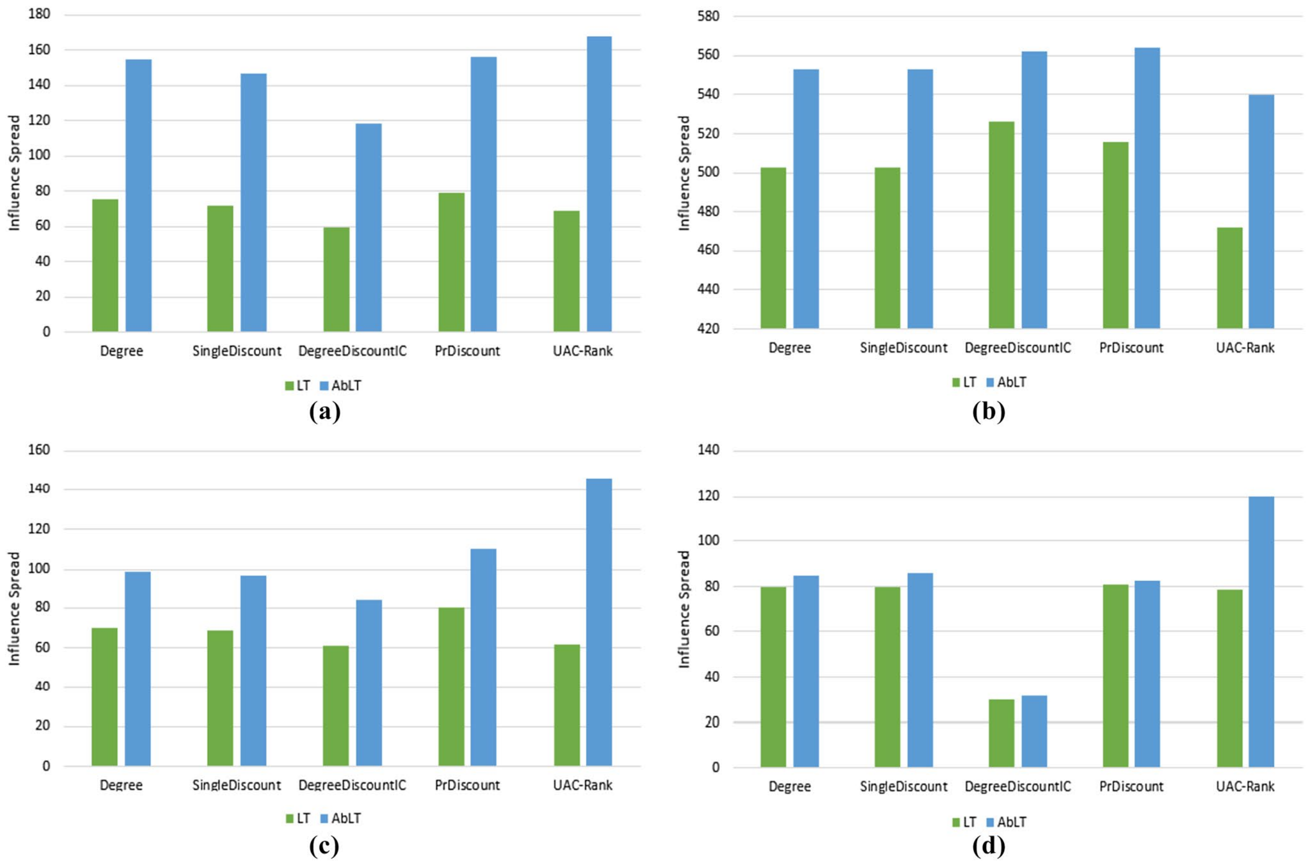


Fig. 3 **a** Average influence spread achieved under LT and AbLT models for seed sets picked from UC Irvine dataset. (Average spread value has been shown based on values obtained for five seed set sizes $k=10$, 20, 30, 40, and 50.) **b** Average influence spread achieved under LT and AbLT models for seed sets picked from Math Overflow dataset. (Average spread value has been shown based on values obtained for five seed set sizes $k=10$, 20, 30, 40, and 50.) **c** Average influence spread achieved under LT and AbLT models for seed sets picked from Facebook Posts dataset. (Average spread value has been shown based on values obtained for five seed set sizes $k=10$, 20, 30, 40, and 50.) **d** Average influence spread achieved under LT and AbLT models for seed sets picked from Email dataset. (Average spread value has been shown based on values obtained for five seed set sizes $k=10$, 20, 30, 40, and 50)

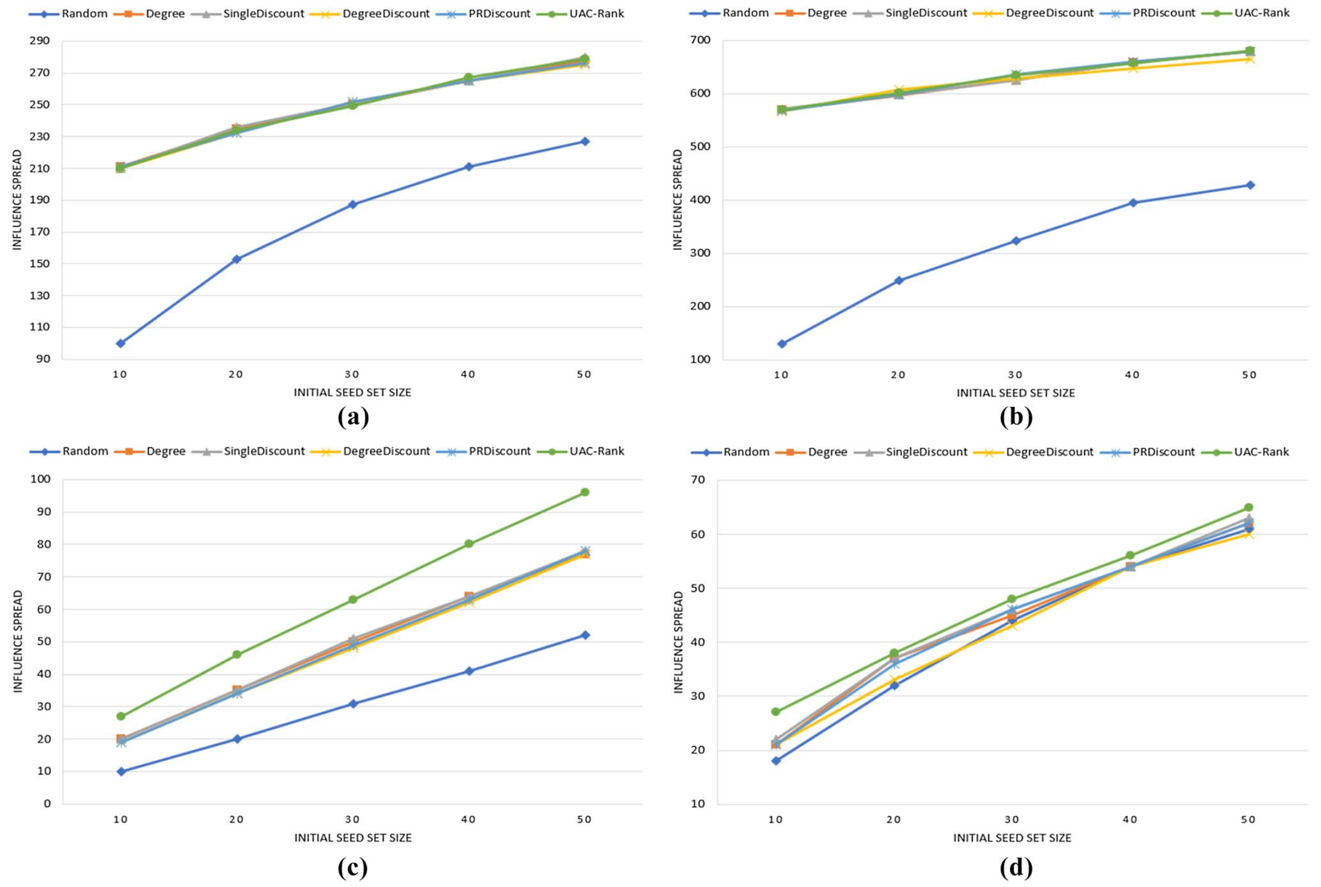
observed that influence spread reached by the various seed sets is much higher under AbIC model of diffusion as compared to IC model. It has been found that compared to IC model, a higher number of nodes get influenced under AbIC when starting from the seed sets generated using the five algorithms.

From Fig. 2a, it can be observed that, spread achieved by seed set generated using Degree algorithm under AbIC is 240% more than what has been achieved under IC, for SingleDiscount seed set, spread is 240% more under AbIC, for DegreeDiscountIC seed set, spread is 243% more, for PRDiscount seed set, spread is 238% more, and for UAC-Rank seed set, spread is 244% more under AbIC. Figure 2b, c, d also affirms the attainment of higher influence spread under AbIC when compared to IC model.

Furthermore, the seed sets generated using the five algorithms have been given as input to both LT and AbLT models of diffusion. Subsequently, for each algorithm, an average value for the spread achieved has been computed based on the spread values obtained for five seed set sizes ($k=10$, 20, 30, 40, and 50). Figure 3a, b, c, d illustrates the average influence spread achieved by the five algorithms under LT and AbLT models.

It has been observed that the influence spread achieved by the same seed set is larger under AbLT compared to the spread achieved under LT. Figure 3a illustrates the total influence spread achieved by the seed sets selected from the UC Irvine dataset, under LT and AbLT models. It has been found that spread for Degree algorithm seed set is 107% more under AbLT, for SingleDiscount seed set, it is 104% more, for DegreeDiscountIC seed set, it is 100% more, for PRDiscount seed set, it is 97% more, and for UAC-Rank seed set, spread is 143% more under AbLT. Similarly, Fig. 3b, c, d illustrates the average influence spread achieved under LT and AbLT models, by seed sets from Math Overflow, Facebook Wall Posts, and Email Network datasets, respectively.

Based on the results obtained, it has been observed that performance of activity-based models, AbIC and AbLT, is



(a)

(b)

(c)

(d)

Fig. 4 **a** Influence spreads of different algorithms for UC Irvine dataset under AbIC model. **b** Influence spreads of different algorithms for Math Overflow dataset under AbIC model. **c** Influence spreads of different algorithms for Facebook Posts dataset under AbIC model. **d** Influence spreads of different algorithms for Email Network dataset under AbIC model

better (comparable in some cases) than the performance of IC and LT models, respectively, for all four datasets and for all seed sets picked by the five algorithms. In other words, the performance of both base models improves when activity is incorporated into the considerations.

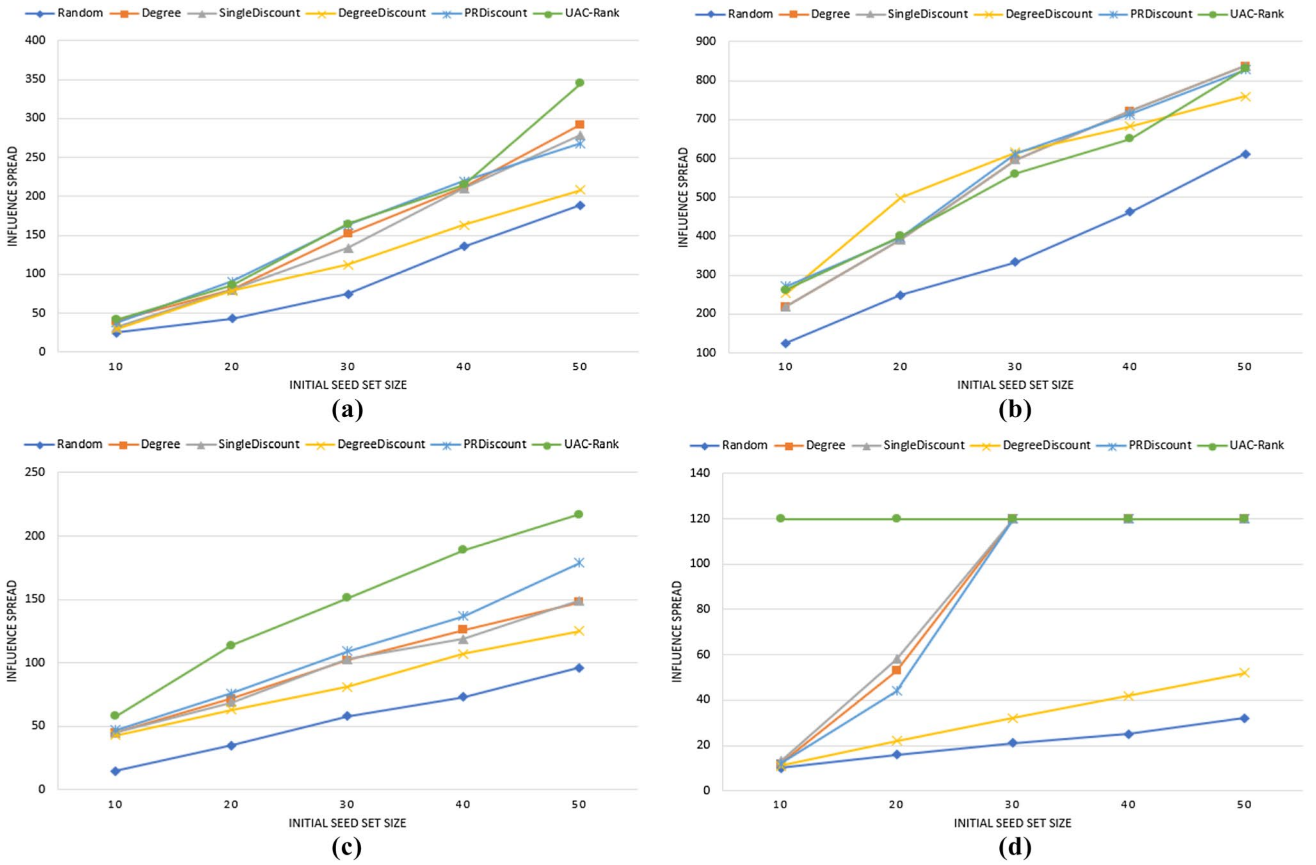
7.3.2 Evaluation of UAC-Rank algorithm

Performance of UAC-Rank algorithm has been compared with five other seed set generating algorithms, namely—Random, Degree, SingleDiscount, DegreeDiscountIC, and PRDiscount. Influence spread has been compared for initial seed sets with different sizes ranging from $k=10$ to $k=50$.

Figure 4a, b, c, d illustrate the performance of Random, Degree, SingleDiscount, DegreeDiscountIC, PRDiscount, and UAC-Rank algorithms when run under AbIC model of diffusion on UC Irvine, Math Overflow, Facebook Wall Posts, and Email Network datasets, respectively.

Figure 4a illustrates that the performance of UAC-Rank algorithm is much better than Random but close to that of the other algorithms under consideration, for UC Irvine dataset, when run under AbIC model. Figure 4b shows the influence spread achieved for Math Overflow dataset. It can be seen that for seed set size of 30 and beyond, influence spread achieved by UAC-Rank is slightly better than the spread achieved by DegreeDiscountIC. From Fig. 4c, it can be seen that performance of UAC-Rank is better than all the other algorithms under consideration, for Facebook Wall Posts dataset. Additionally, as the seed set size increases, improvement in performance of UAC-Rank increases. Figure 4d is for Email Network dataset. Here again, UAC-Rank algorithm has performed better than all the other algorithms being considered.

Figure 5a, b, c, d illustrates the performance of Random, Degree, SingleDiscount, DegreeDiscountIC, PRDiscount, and UAC-Rank algorithms when run under AbLT model for UC Irvine, Math Overflow, Facebook Wall Posts, and Email Network datasets, respectively. Figure 5a illustrates that the performance of UAC-Rank is better than that of Random for all seed sizes and is comparable with the performance of the remaining four algorithms for smaller



(a) (b) (c) (d)

Fig. 5 **a** Influence spreads of different algorithms for UC Irvine dataset under AbLT model. **b** Influence spreads of different algorithms for Math Overflow dataset under AbLT model. **c** Influence spreads of different algorithms for Facebook Posts dataset under AbLT model. **d** Influence spreads of different algorithms for Email Network dataset under AbLT model

Table 2 Running time (in seconds) of the six seed selection algorithms under consideration when run under AbIC model

	UC Irvine	Math Overflow	Facebook Posts	Email Network
Random	0.14	2.88	0.00	0.02
Degree	0.23	9.93	0.03	0.02
Single discount	0.24	9.96	0.03	0.02
DegreeDiscountIC	0.24	10.34	0.03	0.02
PRDiscount	0.23	9.79	0.03	0.02
UAC-Rank	0.21	10.03	0.03	0.02

Table 3 Running time (in seconds) of the six seed selection algorithms under consideration when run under AbLT model

	UC Irvine	Math Overflow	Facebook Posts	Email Network
Random	0.01	0.01	0.00	0.08
Degree	1.21	4.97	0.22	0.18
Single discount	0.95	4.95	0.21	0.19
DegreeDiscountIC	0.81	3.67	0.14	0.04
PRDiscount	1.04	4.99	0.19	0.14
UAC-Rank	2.18	5.34	0.21	0.16

seed set sizes, for UC Irvine dataset. Moreover, as the seed set size increases, performance of UAC-Rank improves in comparison with the four algorithms.

Figure 5b depicts that performance of UAC-Rank is much better than Random and comparable with that of remaining four algorithms under consideration for Math Overflow dataset. Figure 5c illustrates that UAC-Rank performs better than all the other algorithms under consideration for all seed set sizes. Figure 5d depicts the performance for Email Network dataset.

Here again, it can be observed that UAC-Rank performs better than Random and DegreeDiscountIC, and its performance is comparable with that of the other algorithms under consideration.

To conclude, it has been found that performance of UAC-Rank algorithm, in terms of influence spread, is either comparable or better than the Random, Degree, SingleDiscount, DegreeDiscountIC, and PRDiscount algorithms when considering user activity-based diffusion process. We again attribute this improvement to the consideration of activity in seed selection.

7.3.2.1 Comparison of running time The running times of all six seed selection algorithms have been computed under both AbIC and AbLT models for all four aforesaid datasets. Results obtained are presented in Tables 2 and 3 for AbIC and AbLT models, respectively. It can be observed from Tables 2 and 3 that the efficiency of UAC-Rank algorithm in terms of running time is comparable to Degree, SingleDiscount, DegreeDiscountIC, and PRDiscount algorithms under both AbIC and AbLT models. Although Random algorithm takes least time to execute, it performs poorly when compared to the other five algorithms as the influence spread achieved by seeds generated using Random algorithm is quite less.

7.3.2.2 Comparing with other state-of-art algorithms In such kind of work, researcher generally does not present a comparison with other schemes as the results vary due to many parameters and are thus not directly comparable. Traces of datasets used differ in terms of the data stored and are taken at different time intervals. However, for comparison purpose, UAC-Rank's performance, in terms of influence spread percentage achieved, has been compared with four state-of-art algorithms like *Greedy Algorithm with Node Features* (Deng et al. 2015), *LPIMA* (Sheng and Zhang 2018), *Genetic Algorithm with Dynamic probabilities* (Agarwal and Mehta 2018), and *A-Greedy* (Tong et al. 2017). The average influence spread percentage has been compared when starting the influence maximization process with seed set of size 50.

In their work, Deng et al. (2015) have presented the *Greedy Algorithm with Node Features* for which, average influence spread percentage achieved is 0.3%. Sheng and Zhang (2018) have proposed *LPIMA* algorithm, for which average influence spread percentage achieved is 2.7%; Agarwal and Mehta (2018) have done work on *Genetic Algorithm with Dynamic probabilities* which has achieved 11.33% average influence spread percentage, and Tong et al. (2017) have presented *A-Greedy* algorithm, for which average influence spread percentage is 1.6%, whereas, for UAC-Rank algorithm proposed in this work, average influence spread percentage achieved is 18.8%.

8 Conclusion and future scope

In this work, the process of influence maximization in social networks based on user's connectivity and his/her past activity has been studied. The work presented is inspired by the fact that influence maximization is not just about spreading the influence to maximum number of users, but also about spreading the influence faster. Activity-based diffusion models, AbIC and AbLT, have been proposed, wherein real user behavior, i.e., the actual activity performed by a user in the past, has been used to model influence diffusion. More the activity level of a user, more is the probability of propagation. This work supports the belief that having a large number of connections should not be the only parameter under consideration, when computing the influential potential of a user. In scenarios, where influence maximization is to be done within a given time span, the frequency with which a user communicates is equally important. A user who has fewer connections but is highly active in terms of interacting with its neighbors might turn out to be more influential than a user who has larger number of connections but is dormant. Thus, when computing the influence of a node, its connectivity as well as the frequency at which it is interacting with its neighbors should be considered.

The proposed work has been evaluated using four real-world online social network datasets. From the results obtained, it has been observed that influence spread achieved by activity-based models is better than the other two models under consideration. As an example, spread achieved by UAC-Rank algorithm, for UC Irvine dataset, run under AbIC model is 244% more than what it has achieved under IC model. Similarly, for Math Overflow dataset, it is 242%, for Facebook Wall Posts, it is 94%, and for Email Network dataset, the improvement is 47%. On comparing the spread achieved by UAC-Rank under LT and AbLT models, the spread achieved is 143% higher under AbLT for UC Irvine, 14% for Math Overflow, 135% for Facebook Wall Posts, and 52% for Email Network, as compared to base model LT. Hence, the user activity level when taken into consideration improves the influence spread significantly in most cases.

Looking at the performance of UAC-Rank as a seed set identification algorithm, it is found that compared to five existing algorithms, it produces better or similar seed sets from the point of view of influence spreading potential as observed under AbIC model of diffusion. This further confirms the faith in the ability of user activity consideration.

There are certain aspects that can be further explored regarding the proposed approach. First, the other models for studying influence maximization can be extended to incorporate user activity levels in their processes. Second, the AbLT model can be further extended so as to compute a node's threshold value based its activity level. Third, the proposed approach of combining user connectivity and activity can be further extended by considering other social network features like topic-level influence, homophily, etc.

References

Agarwal S, Mehta S (2018) Social influence maximization using genetic algorithm with dynamic probabilities. In: 2018 eleventh international conference on contemporary computing (IC3). https://doi.org/10.1109/ic3.2018.8530626

Alshahrani M, Fuxi Z, Sameh A, Mekouar S, Huang (2018) Top-K influential users selection based on combined Katz centrality and propagation probability. In: IEEE 3rd international conference on cloud computing and big data analysis (ICCCBDA), pp 52–56

Chen W, Wang Y, Yang S (2009) Efficient influence maximization in social networks. In: 15th ACM SIGKDD international conference on knowledge discovery and data mining—KDD'09, pp 199–208

Deng X, Pan Y, Wu Y, Gui J (2015) Credit distribution and influence maximization in online social networks using node features. In: 12th international conference on fuzzy systems and knowledge discovery (FSKD), pp 2093–2100

Domingos P, Richardson M (2001) Mining the network value of customers. In: Seventh ACM SIGKDD international conference on knowledge discovery and data mining—KDD'01, pp 57–66

Goyal A, Bonchi F, Lakshmanan LVS (2011) A data-based approach to social influence maximization. Proc VLDB Endow 5(1):73–84

Heidemann J, Klier M, Probst F (2010) Identifying key users in online social networks: a PageRank based approach. In: 31st international conference on information systems (ICIS), pp 1–22

Jianqiang Z, Xiaolin G, Feng T (2017) A new method of identifying influential users in the micro-blog networks. IEEE Access 5:3008–3015

Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: 9th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'03, pp 137–146

Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632

Kurka DB, Godoy A, Von Zuben FJ (2016) Online social network analysis: a survey of research applications in computer science. arXiv:1504.05655v2

Liu D, Jing Y, Zhao J, Wang W, Song G (2017) A fast and efficient algorithm for mining top-k nodes in complex networks. Sci Rep 7:43330

Morone F, Makse HA (2015) Influence maximization in complex networks through optimal percolation. Nature 524(7563):65–68

Page L, Brin S, Motwani R, Winograd T (1998) The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab

Peng S, Wang G, Xie D (2017) Social influence analysis in social networking big data: opportunities and challenges. IEEE Netw 31(1):11–17

Peng S, Zhou Y, Cao L, Yu S, Niu J, Jia W (2018) Influence analysis in social networks: a survey. J Netw Comput Appl 106:17–32

Riquelme F, Gonzalez-Cantergiani P, Molinero X, Serna M (2018) Centrality measure in social networks based on linear threshold model. Knowl Based Syst 140:92–102

Sheng K, Zhang Z (2018) Research on the influence maximization based on community detection. In: 13th IEEE conference on industrial electronics and applications (ICIEA). https://doi.org/10.1109/iciea.2018.8398185

Tong G, Wu W, Tang S, Du DZ (2017) Adaptive influence maximization in dynamic social networks. IEEE/ACM Trans Netw 25(1):112–125

Wang Y, Zhang B, Vasilakos AV, Ma J (2014) PRDiscount: a heuristic scheme of initial seeds selection for diffusion maximization in social networks. Springer, Cham, pp 149–161

Zhu J, Liu Y, Yin X (2017) A new structure-hole-based algorithm for influence maximization in large online social networks. IEEE Access 5:23405–23412