Fundamenta Informaticae 130 (2014) 317–342
DOI 10.3233/FI-2014-994
IOS Press

Centrality Measures, Upper Bound, and Influence Maximization in Large Scale Directed Social Networks*

Sankar K. Pal, Suman Kundu†, C. A. Murthy
Center for Soft Computing Research
Indian Statistical Institute, Kolkata, India
sankar@isical.ac.in; suman@sumankundu.info; murthy@isical.ac.in

Abstract. The paper addresses the problem of finding top k influential nodes in large scale directed social networks. We propose two new centrality measures, Diffusion Degree for independent cascade model of information diffusion and Maximum Influence Degree. Unlike other existing centrality measures, diffusion degree considers neighbors' contributions in addition to the degree of a node. The measure also works flawlessly with non uniform propagation probability distributions. On the other hand, Maximum Influence Degree provides the maximum theoretically possible influence (Upper Bound) for a node. Extensive experiments are performed with five different real life large scale directed social networks. With independent cascade model, we perform experiments for both uniform and non uniform propagation probabilities. We use Diffusion Degree Heuristic (DiDH) and Maximum Influence Degree Heuristic (MIDH), to find the top k influential individuals. k seeds obtained through these for both the setups show superior influence compared to the seeds obtained by high degree heuristics, degree discount heuristics, different variants of set covering greedy algorithms and Prefix excluding Maximum Influence Arborescence (PMIA) algorithm. The superiority of the proposed method is also found to be statistically significant as per T-test.

Keywords: Centrality Measure, Social Network, Influence Maximization, Independent Cascade Model, Statistical Significance

*A preliminary version of a part of the investigation is published in PReMI'11, Moscow, Russia, LNCS (Springer Verlag) 6744, pp. 242-247, 2011.

†Address for correspondence: Center for Soft Computing Research, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata, India - 700108.

Received February 2013; revised April 2013.

1. Introduction

A Social Network is made up of social ties among individuals. Friends, family members, colleagues are connected to each other in the social paradigms. A new product or innovation can touch or influence thousands of people with the help of social ties. Before buying, people take advice from their friends and families. So, marketing persons always put their eyes on social happenings. After the introduction of web, people are enlarging their social boundary by electronic means. Hyperlinks and email communications show early social ties in the electronic media. Thereby, the formation of online social networks started.

In recent years, large scale online social networks have become extremely popular. Twitter, Facebook, Orkut, LinkedIn are a few examples. These social networks have millions of users. Similar to the social structure found in society, people around the globe are connected with the purpose of common interest. As a result, these applications are becoming a huge marketing platform of products and services, specially for spreading of innovations to a large number of people in a short amount of time. Marketing persons usually target few influential individuals for marketing their products. These individuals, in turn, influence their friends and families. However, the most important question arises, "How to select the influential individuals quickly?". That is, how to select the set of initial influential individuals for which the influence spread over the network is maximum. This problem is known as influence maximization problem for social networks. Besides its main application to marketing or spreading innovation, solution to this problem can also be used in other domains such as in detecting top stories in the news networks and ranking the top articles in the blog sphere.

The natural solution to the problem will be to select those persons having higher numbers of neighbors. That is, select the persons based on their centrality scores. Domingos and Richardson were the first to study this as an algorithmic problem and proposed some probabilistic methods [9, 26]. In [18] Kempe et al. formulated the problem as one of discrete optimization and showed that the problem is NP hard. They also proposed a greedy hill climbing approach, which provides $(1 - 1/e - \epsilon)$ approximation of the optimal solution. Finally, they showed through experiment that their approach provides significant improvement over those based on the classical degree and centrality based heuristic. However, for large scale graphs, the greedy approach is time consuming. It may take days to compute even on a moderate size graph of 30K nodes as reported in [6]. To overcome the drawback, several algorithms were proposed in the last few years. In [21], Leskovec et al. presented a "lazy-forward" optimization method in selecting the seed nodes and showed experimentally that this method runs 700 times faster than the greedy algorithm of Kempe et al. They called this algorithm "Cost-Effective Lazy Forward" (CELF). However, as reported in [5], this "lazy-forward" method still takes hours to generate 50 seeds. Some other approaches in this line were reported in [5, 15, 10].

In recent years, several heuristic algorithms ([6], [5], [4]) were proposed to deal with the said problem for improving the performance. These algorithms, unlike the traditional centrality based heuristics, consider the underlying principle of information diffusion process in the network. Broadly, there are two types of diffusion models available in the literature, threshold model of diffusion [16] and cascade model of diffusion [13, 14]. In [19], Kimura et al. provided a shortest path based influence cascade model and an efficient algorithm to compute the information spread under this model. In [5], the authors described their degree discount heuristic algorithm for independent cascade model. In [24], Narayanam et al. provided a game theory based approach for linear threshold diffusion model. All these algorithms are found to suffer from high execution time [6]. Recently, Chen et al. [6] described their LDAG algorithm

for linear threshold model. This model uses the local structure of the network to make the influence computation tractable and reduce the computation cost.

From the above mentioned discussion, one may note that the greedy solution to the problem provides a good estimation. However, these solutions are time consuming for large scale social networks. On the other hand, centrality based heuristic models run very fast but their solution (set of seeds) may result in less influence over the network. The reason behind it might be, the traditional centrality measures do not consider the effect of neighborhood. They also do not incorporate the principle of the information diffusion among the neighbors. Therefore, judicious integration of the concept of neighborhood and the principle of information diffusion process with the classical centrality measure seems to be appropriate for providing an efficient solution in terms of both performance and computation time.

The present paper describes such an attempt where we propose a new centrality measure, called *diffusion degree*, for Independent Cascade Model, and we use it to find the top k influential individuals in large scale directed social networks using Diffusion Degree Heuristic (DiDH). Further more, the existing centrality measures assume the propagation probability to be uniform throughout the network. That is, each node influences their neighbors with the same probability. But, in social relations the trust of each tie may not be the same. Our centrality measure takes care of this accordingly and works flawlessly for such nonuniform propagation probabilities. Besides these, we have defined mathematically the upper bound of a node's influence based on the network structure. Accordingly, a new centrality score of nodes, called *Maximum Influence Degree* (MID), is defined. Though it is computationally heavy to determine MID, yet it provides a good estimation to the upper bound of the influence over the entire network.

In our experiment we consider five different large scale social network e.g., Twitter following-follower network, Amazon co-purchasing network, Slashdot friendship network and web graph of Berkeley and Standford (Web-BerkStan) university and Notre Dame University (Web-NotreDame). First we determine the top k influential nodes of a network using the proposed DiDH. Then we estimated the information spread over it by Monte Carlo simulation and compared the results (# of nodes influenced) with other available solutions extensively. Our solution shows significant improvement over those of other methods. Additionally, our model is seen to run significantly faster compare to the greedy algorithms even on networks with millions of nodes. Furthermore, we estimated the influence through simulation for the top k nodes selected based on MID score and the results were found to corroborate to those obtained by DiDH.

Rest of the paper is organized as follows: Section 2 describes the motivation behind this research work. Section 3 briefly explains the problem. Preliminaries related to our theory are mentioned in Section 4. The proposed centrality measure *Diffusion Degree* with its characteristics is defined in Section 5, the proposed measure of upper bound of influence is illustrated in Section 6 and the proposed *Maximum Influence Degree* is reported in Section 7. Experiment and results are listed in Section 8. Finally, in Section 9 we conclude the research findings.

2. Motivation

Introducing a new product, application or innovation to the people is one of the major jobs of marketing. In case of direct marketing, the marketer takes the decision of whether or not market to a person, based on her characteristics and in case of mass marketing the marketer targets a segment of population based on their common characteristics. This decision may lead to a sub-optimal marketing decision by not

considering the effect of people on each other's buying decision. In fact, most products spread more effectively due to consumer to consumer dialogue. Word-of-Mouth has a greater effect as because this marketing channel has more trust over mass media marketing. It has been found that, innovation spreads over the social and geographical networks gradually like a domino effect. That is, in the beginning few people adopt the product, then their neighbors adopt it, then their neighbors and so on. This is because, the decision of an individual heavily depends upon their interpersonal ties.

In case of viral marketing, selecting the initial set of users by whom the influence flow will be maximum is very important. The problem is known as the influence maximization problem. That is, select the set of initial influential persons for whom the innovation spread is maximum. These motivate us to research in the topic.

3. Problem Statement

Suppose $G(V, E)$ represents a social network where V is the set of all nodes in the network and E is the set of all edges in the network. Before we introduce the problem statement, we provide the definition of the influence of an individual in a network and briefly describe the different diffusion models mostly used therein.

3.1. Influence of an Individual

In reality, everyone has his/her own opinion about things near to them. However, when it comes about an unknown, people usually tend to rely on others' opinions. Ideas, innovations or information do not always spread at once, but it gradually spreads over social networks. In a social economic structure, few persons are relatively more influencing compared to others. In their paper [8], Dolecek et. al. describes influence of an individual under two different information dynamics, namely, short term influence or first impression and long term influence or equilibrium. In case of short term influence, a node v is considered to be influential when node $u \in V \setminus (\Gamma(v) \cup v)$ shares the same opinion that originally was held by v. Here $\Gamma(v)$ denotes the set of neighbors of v. On the other hand, in case of long term influence, an individual v is said to be influential if after a long time period other agents in the network remain attentive to the opinion of v.

3.2. Information Diffusion

Information diffusion in the social network is well studied in sociology. There have been extensive experiments to understand the effect of "word-of-mouth" in spreading innovations. During diffusion in a social network, at any given time there exist two sets of nodes. Members of one set have already adapted the innovation (i.e., active nodes), whereas the members of the remaining set have not (i.e., inactive nodes). In literature, there are two fundamental processes by which nodes adapt the innovations. These processes are as follows:

3.2.1. Threshold Model

Threshold model of diffusion was first proposed by Granovetter [16]. According to this model, one inactive node becomes active based on the proportion of neighbors already activated. Typically, node

$v \in V$ chooses a threshold value $\theta_v \in [0, 1]$ selected randomly from a probability distribution. Each edge of v is assigned with a non negative edge weight $\omega_{v,u}$ where $\sum_{u \in \Gamma(v)} \omega_{v,u} \leq 1$. A node v is activated or influenced if and only if $\theta_v \leq \sum_{active\ u \in \Gamma(v)} \omega_{v,u}$.

3.2.2. Cascade Model

Goldenberg et al. in [13, 14] inspected Cascade Model in marketing perspective. In this model a node u is influenced by its neighbor v with a probability $\lambda_{u,v}$. This probability is called propagation probability or diffusion probability.

The simplest and popular form of the cascade model is *Independent Cascade (IC) Model* of [13]. The IC model runs in discrete time. In the process, there are two sets of nodes, the active nodes which have already adapted the behavior and inactive nodes which are prospected to adapt the behavior in future. Initially, a few nodes are activated. At each successive steps, active nodes will try to activate one of its inactive neighbors. However, the node will get only one chance to activate it and if it fails there will be no further chance to activate the same node again. The process terminates when no further activation is possible. Edge $e(u, v) \in E$ is assigned with a non negative probability $\lambda_{u,v}$. This probability indicates the probability at which node u is activated by v.

3.3. Problem Statement

For a given social network $G(V, E)$, we are interested to mine a set of top-k influential individuals S such that $\alpha = |\cup_{v \in S} I_v|$ is maximized under Independent Cascade Model of diffusion. Here, I_v denotes the set of nodes influenced by v.

4. Preliminaries

Before we describe the new centrality measure in Section 5, we provide here a few preliminary definitions.

4.1. Centrality

In a social network, centrality is considered to be a measure of relative importance of the nodes in the network. From the years of research on centrality, it is clear that it is an important structural attribute [12]. However, there is no defined agreement about what it is. Scientists provided different measures for finding the most central nodes in a network. Two such measures are as follows:

4.1.1. Degree Centrality

One of the classic measures of centrality is degree centrality. Nieminen [25] provided a simple and natural measure for centrality based on the count of degree or number of links. The degree of a node v is mathematically represented as

$$C_D(v) = \sum_{i=1}^{n} \sigma(u_i, v) \tag{1}$$

where the function $\sigma(u_i, v)$ is defined as

$$\begin{aligned} \sigma(u_i, v) &= 1 \textit{ if and only if } u_i \textit{ and } v \textit{ are connected} \\ &= 0 \textit{ otherwise.} \end{aligned}$$

4.1.2. Betweenness Centrality

Another classic centrality measure is betweenness. This measure is based on the frequency at which a node falls between other nodes [12]. Mathematically, the betweenness score of a node v is

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2}$$

where σ_{st} is the number of shortest paths between s and t, and $\sigma_{st}(v)$ is the number of shortest paths from s to t passing through v.

4.2. Directed Social Network vs Undirected Social Network

Based on the type of ties, we can broadly classify social networks into two categories, namely, directed social network and undirected social network. In case of an undirected network, a tie defines both way communications. On the other hand, for directed social networks, each tie defines one way communication. For example, in case of a blog network supose a person A follows a blog B. Here, the author of the blog B may not follow back the blog of A. In this type of network a both way communication is represented by two different edges in the network.

For directed social networks in-degree defines the number of followers and the out-degree means the number of nodes it follows. Here a node (or a person) gets influenced only by the nodes (or persons) it is following not by its followers. That means, it is only the in-degree, not the out-degree, that should be considered as an index for quantifying the significance of an individual, and for characterizing its influencing channel.

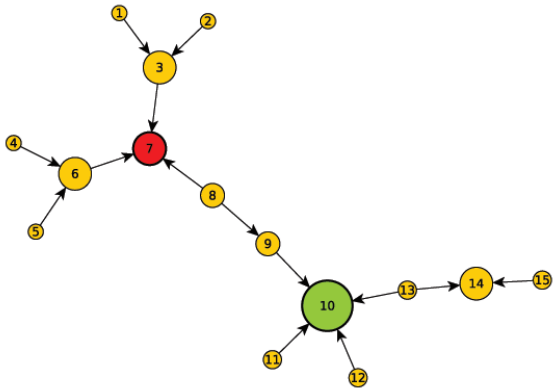
5. A new centrality measure: Diffusion Degree

As we described in Section 3, information in a social network flows through its structural dynamics and one of the well known models for information diffusion is Independent Cascade (IC) model of diffusion. In traditional approach, solutions to the problem of influence maximization focused mainly on finding the most central nodes, i.e., how close they are to the center of action. However, in case of real life social networks, influential capability of one person is boosted by its neighbors' contributions. In this section, we will define a new centrality measure named "Diffusion Degree" for directed social networks with IC model as the underlying diffusion model.

5.1. Philosophy and Criteria

In a social network, it is found that higher degree nodes (i.e., elite users) are connected with lower degree nodes, specially in directed networks. On the other hand, nodes with comparatively lower degree than

elite users might be connected with similar nodes having higher neighborhood size. In degree centrality approach, those nodes in the latter situation get lower centrality value compared to the former. When the information diffusion starts propagating in the network, it begins from the seeding nodes and tries to influence its neighbors. Influenced neighbors, will then try to influence further their neighbors, and so on. Thus, the effect of a node on the network depends not only on its own degree but also on the neighbors' degree.



Nodes	In-Degree	Out-Degree
1, 2, 4, 5, 11, 12, 15	0	1
8, 13	0	2
3, 6	2	1
9	1	1
14	2	0
7	3	0
10	4	0

Figure 1. A sample network. Nodes are sized as per their in-degree scores. The highest degree node is 10 with 4 in-degree.

Consider the following-follower social network shown in Figure 1. A link directed from a node u (say, 11) to a node v (say, 10) means node u is following node v. Here node 10 has highest in-degree. For the time being, consider that one can influence all of its followers. If we select node 10 as the seed node based on its degree, it can directly influence 4 nodes (nodes 9, 11, 12 and 13) in the network. Then in the next step node 9 will influence node 8 and no further activation is possible as nodes 8 and 13 do not have any follower. That is, in total, 5 nodes will be activated by selecting node 10 as the seed. However, if we consider the node 7 as seed, it can influence 7 nodes (3, 6 and 8 directly and 1, 2, 4 and 5 indirectly) in total. So, in the above example, a relatively lower degree node can activate more nodes in the network. The reason behind such a behavior is because node 7 is connected with relatively higher degree nodes like nodes 3 and 6 as compared to those of the node 10.

Now, in the same example, consider a different situation with IC model of diffusion. Suppose, the node 6 can influence both of its neighbors with the propagation probability 1, while the other nodes in the network influence their each neighbor with a probability 0.25. In this scenario selecting node 10 as the seed, might activate all the four or even none of its followers. However node 6 as a seed will definitely influence its two follower nodes. That is, accurately determining whether a node would be influenced or not relies on the diffusion probability in case of IC diffusion model. Accordingly, in the given example, the highest degree node 10 is not the best as a seed. The above points have been addressed while defining the centrality measure for influence maximization problem. In the following section, we will define the proposed centrality measure mathematically.

5.2. Definitions

Postulate 5.1. Influence of a node gradually decreases with the increased distance and the influence is maximum when the distance is one, that is, the influenced node is one of its neighbors.

In our social relations, we take advice from our friends and families and it is observed that, usually we form our opinion on a problem/topic either of our own or we refer to some known persons, mostly friends and family members. Sometimes, we trust a person if (s)he is a friend of our friends, even though (s)he is not a direct friend of us. However, as this relation-distance increases, e.g., friend of a friend of friends, the trust usually decreases, and so the influence. Considering this scenario of social relations we believe, the Postulate 5.1 is true.

Definition 5.2. (*Diffusion Degree of Node*)
In IC model, let the propagation probability of a link $e(u, v)$ from node u to v be denoted by $\lambda_{u,v}$, that is, node u who follows node v will get activated by v with probability $\lambda_{u,v}$. Suppose, a node v has m neighbors denoted by the set $\Gamma(v) = \{u_1, u_2, ..., u_m\}$ which are connected with the links $e_v = \{e_1(u_1, v), e_2(u_2, v), ..., e_m(u_m, v)\}$. Let us also consider that the corresponding propagation probabilities of these links are denoted by the set $\Lambda_v = \{\lambda_{u_1,v}, \lambda_{u_2,v}, ..., \lambda_{u_m,v}\}$. In the diffusion process, the expected number of nodes activated or influenced by v can then be defined as,

$$Exp(v) = \sum_{u \in \Gamma(v)} \lambda_{u,v}. \tag{3}$$

When the diffusion propagates further, active neighbors of v will activate their inactive neighbors. The expected number of nodes activated in distance two, i.e., number of nodes activated by active neighbors of v is,

$$Exp(v^{(2)}) = \sum_{u \in \Gamma(v)} (\lambda_{u,v} \times \sum_{i \in \Gamma(u)} \lambda_{i,u}). \tag{4}$$

The diffusion degree of a node is defined as the cumulative contribution of the node itself and contributions due to its neighbors. Considering only the effect of its immediate followers (Postulate 5.1) we can define the diffusion degree C_{DD} of node v as,

$$\begin{aligned} C_{DD}(v) &= Exp(v) + Exp(v^{(2)}) && (5)\\ &= \sum_{u \in \Gamma(v)} (\lambda_{u,v} + \lambda_{u,v} \times \sum_{i \in \Gamma(u)} \lambda_{i,u}) && (6)\\ &= \sum_{u \in \Gamma(v)} \lambda_{u,v} \times (1 + \sum_{i \in \Gamma(u)} \lambda_{i,u}) && (7) \end{aligned}$$

5.3. Algorithm

The pseudo-code for calculating the diffusion degree is shown in Algorithm 1.

Algorithm 1 Diffusion Degree Calculation

```
1: function DIFFUSIONDEGREE(v)
2:     sum ← 0
3:     for all u ∈ neighbor(v) do
4:         Exp_u ← 0
5:         for all i ∈ neighbor(u) do
6:             Exp_u ← Exp_u + λ_{i,u}
7:         end for
8:         sum ← sum + λ_{u,v} * (1 + Exp_u)
9:     end for
10:    return sum
11: end function
```

5.4. Notes

5.4.1. On Complexity

The diffusion degree measure (Equation 7) of a node depends upon its in-degree, in-degree of its followers and propagation probabilities of the links between the node itself and its followers. Unlike others, the computation does not depend on the in-degree of a node already selected before as seeds. Thus the diffusion degree for every node in a network could be determined in $O(E + E)$ time where E is the number of edges in the network.

5.4.2. On Overlapping Neighborhood

The IC model is highly stochastic process and in the model, an active node gets chance to activate each inactive neighbor only once. But the reverse statement is not true, i.e., an inactive node will get at most one activation tries from all of its neighbors not hold for information diffusion in IC model. So, if one of the inactive nodes (say u) gets the information from one of its active neighbor (say v) and fail to activate at that time step. It is possible that it gets activated in the further time steps by any active neighbors other than v. This property of the model also supported by the sociological findings reported in [27]. In [27], the author describes that social influence is a complex process which completed over time in phases and there are agents who accept the changes only when several others adopt the same. To consider the aforesaid facts of the model and the social system, the diffusion degree measure (Equation 7) did not discount the overlapping neighbors among a node and its neighbors.

6. Upper Bound of Influence

In this section we formulate a measure to determine theoretically the maximum possible influence by a node at a distance n. Let us consider the social network is represented by adjacency matrix A and the elements of the matrix are defined as,

$$a_{u,v} = 1 \textit{ if there is a link from node u to node v} \quad (8)$$

$$= 0 \textit{ otherwise.} \quad (9)$$

The total number of edges of the network $|E| = \sum_{u \in V} \sum_{v \in V} a_{u,v}$, and in-degree of a node $v = \sum_{u \in V} a_{u,v}$.

To get the upper bound of influence for a node v at distance n we compute A^n. A positive value at position (u, v) in Matrix A^n means, there exists at least a path connecting the node u to node v having length n and the value itself is the number of such n-length paths in the network. It is obvious that v can influence u only if there exists a path from u to v. Thus the highest possible influence for a node at distance n is the number of nodes connected to it with a path having length n. So, the upper bound of influence at distance n is the number of elements having positive value in the columns of matrix A^n, and the nodes influenced are the set of nodes identified by the row indices.

Mathematically, the influence of a node v at distance n is any subset of,

$$\xi_v^{(n)} = \{u \in V | a_{u,v}^{(n)} > 0, A^n = ((a_{u,v}^{(n)}))\}. \tag{10}$$

So, the Upper Bound of Influence for node v at distance n is,

$$\alpha^{(n)}(v) = |\xi_v^{(n)}|. \tag{11}$$

Now, if a node v influences all possible nodes then, the changes of influence at distance n will be any subset of,

$$\Delta\xi_v^{(n)} = \xi_v^{(n)} \setminus \cup_{i=1}^{n-1} \xi_v^{(i)} \tag{12}$$

and

$$\Delta\alpha^{(n)}(v) = |\Delta\xi_v^{(n)}|. \tag{13}$$

7. A new centrality measure: Maximum Influence Degree

Based on the upper bound of influence discussed in Section 6, we now propose a new centrality measure Maximum Influence Degree of a node. Intuitively, it means the maximum possible influence of a node in the network.

Definition 7.1. (*Maximum Influence Degree of Node*)
Maximum Influence Degree (MID) of a node is the sum of the upper bound of influence in each distance. For a node v it may be theoretically defined as

$$C_{MID}(v) = \sum_{n=1}^{\infty} \alpha^{(n)}(v). \tag{14}$$

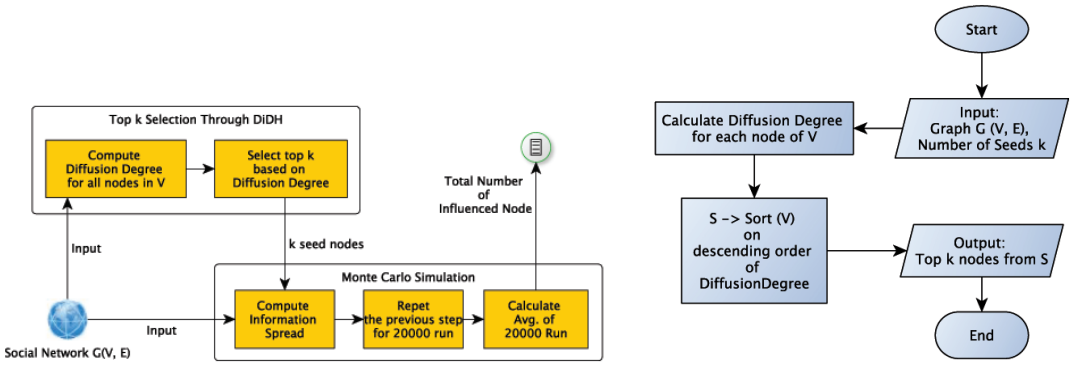
Here, $\alpha^{(n)}(v)$ is calculated as described in Equation 11.

One may note here that computing the MID up to $n = \infty$ may not be necessary and practical as the required information on node coverage in a network is almost contained with $n = D$, the network diameter. Accordingly, Equation 14 reduces to

$$C_{MID}(v) = \sum_{n=1}^{D} \alpha^{(n)}(v). \tag{15}$$

8. Experiment and Results

We conducted experiments with five different types of live social networks. A block diagram of the experiment is given in Figure 2(a). We extracted the top k influential individuals using Diffusion Degree Heuristics (DiDH). A flow chart corresponding to DiDH algorithm is shown in Figure 2(b). We computed the information spread (i.e., the number of nodes directly or indirectly influenced by those k seeding nodes) over the network using Monte Carlo simulation. Similarly, we extracted the top k nodes using other available algorithms and computed the number of influenced nodes. We then compared their results with the results obtained by the proposed DiDH. We also verified the significance of the difference in results statistically using T-Test. Following subsections describe the data sets, experimental consideration, comparative results and the test of significance.



(a) Block diagram of the experiment (b) Flow chart of DiDH

Figure 2. Block diagram & flow chart

8.1. Description of Data Sets

We have collected different types of directed social network data sets for our experiment. These are friendship networks of Twitter [7] and Slashdot [22], Amazon co-purchasing network [20], web graph of Berkeley and Standford (Web-BerkStan) university collected in 2002 [22] and web graph of Notre Dame University (Web-NotreDame) [17]. Properties of these data sets are listed in Table 1.

8.2. Consideration

In social networks, it is observed that the propagation probability of each tie is different. Each individual node gets influenced by its neighbors with different probabilities. Unlike other previous investigations, where the propagation probability was considered to be the same for all the ties, we additionally experimented on networks considering that the propagation probabilities are different for different ties. Moreover, only with the knowledge of the network structure it is not possible to provide a solution to the problem of predicting the propagation probabilities. The reason behind this is that the propagation probability of a tie also depends on the communication dynamics of the nodes. In the absence of such

Table 1. Features of Data Sets

Property	Twitter	Amazon	Slashdot	Web-BerkStan	Web-NotreDame
Nodes	455818	400727	82168	685230	325729
Edges	822487	3200440	948464	7600595	1497134
Nodes in Largest WCC[1]	455818	400727	82168	654782	325729
Edges in Largest WCC	822487	3200440	948464	7499425	1497134
Nodes in Largest SCC [2]	2208	380167	71307	334857	53968
Edges in Largest SCC	10401	3069889	912381	4523232	304685
Avg. Clustering Coefficient	0.0175	0.4113	0.0617	0.6149	0.454
Number of Triangles	57769	3686467	602592	64690980	8910005
Fraction of Closed Triangles	0.0002781	0.1605	0.02411	0.08769	0.08767
Diameter	7	18	12	669	46
90-Percentile Effective Diameter	4	7.7	4.7	10	9.3

[1] Weakly Connected Component

[2] Strongly Connected Component

knowledge, we have assumed different distributions for the propagation probabilities of the network in our experiment. Two different setups used are as follows

8.2.1. Propagation Probability Setups

Uniform Propagation Probability (UPP) We assign uniform propagation probabilities for all edges in the network i.e., $\lambda_{u,v}$ is considered to be the same $\forall u, v$. The assigned values $\forall \lambda_{u,v}$ are reported with corresponding results in Section 8.5. As mentioned before, this setup is very unlikely to be true for a real life social network, yet we consider it in our experiment to make an unbiased comparison of the results.

Non-Uniform Propagation Probability (NUPP) In this setup we assign non-uniform propagation probabilities for all edges of the network, i.e., $\lambda_{u,v}$ values are different for different ties. We use three different methods to generate the non-uniform values of propagation probabilities. These are,

i) *Random with Uniform Distribution:* We assign the propagation probabilities for all edges of the network randomly, generated from an uniform distribution. The range of the distribution is varied to generate more than one such propagation probability. Corresponding ranges have been reported in Section 8.5 along with the distribution graph.

ii) *Random with Normal Distribution:* In this method the propagation probabilities are generated randomly from a Normal or Gaussian distribution. The mean of the distribution is varied to generate more than one such setting for the same network. Graphs mentioning the distribution along with the parameters of the distribution are shown in Section 8.5.

iii) *Random with Power Law Distribution:* The propagation probabilities in this method have been generated randomly from a Power Law distribution. The parameters of the distribution are var-

ied to generate more than one such propagation probability. The parameters of the Power Law distribution are reported along with the graphs in Section 8.5.

8.3. Baseline Algorithms for Comparison

In the experiment, we used Monte-Carlo simulation technique for estimating the number of influenced nodes in a information diffusion. We simulated for 20000 runs and took the average to get the accurate estimation. Number of seeds, i.e., k is varied from 0 to 100 depending upon the data set.

We have compared results of the proposed DiDH and MIDH with those of the following two heuristic methods and two greedy algorithms.

- *High Degree Heuristics (HDH):* Nodes are ranked according to their degrees. We ranked the influential nodes based on their in-degree and computed the influential capabilities of the seeds. This method is a special case of DiDH where the neighborhood and the diffusion model are not considered in the centrality measure.
- *Degree Discount Heuristics (DDH):* We implemented the degree discount algorithm as described in [5]. The influential nodes are ranked according to the degree discount score.
- *Prefix excluding Maximum Influence Arborescence Model (PMIA):* We implemented PMIA algorithm as described in [4]. Best possible value of the threshold (θ) had been experimentally decided as noted in the original paper.
- *Set Covering Greedy (SGA):* We considered SGA to compare with the proposed algorithm as it uses the neighborhood concept with greedy approach. We implemented the basic set covering greedy algorithm proposed in [11]. We used three different values for the neighborhood size m ($m \in \{1, 2, 3\}$) as mentioned in [11].

8.4. Baseline Parameters for Comparison

We compared the algorithms in terms of the following three parameters:

- *No. of Active Nodes (α):* With the help of Monte Carlo simulation we estimated the total number of nodes (α) influenced by k seeds. In influence maximization, for a given k, higher value of α signifies that the algorithm provides better seeds as compared to those having lower α-values.
- *Execution Time:* Execution time is one of the vital parameters to check when working with the large scale social networks. We measured the execution time for all the algorithms and made a comparative analysis.
- *No. of Active Nodes with Distance:* We estimated the total number of influenced nodes by an individual with increased distance. This enables to make a comparison in terms of information spreading.

8.5. Results

The experiment has been conducted intensively on five data sets (Table 1). The salient features of the observations with UPP and NUPP setups are described here:

8.5.1. Results for Twitter Data

In Twitter following-follower network, each node (u) represents a Twitter user. If a user v is followed by a user u there is a connected link from node u to v in the graph. Here, α denotes how many people of the network have the knowledge or the idea originally started propagating through the seed nodes. That is, if an information is given to k persons in the network, at the end, how many persons got to know the same information. So, in this type of network, higher value of α defines the higher quality of selected seeds.

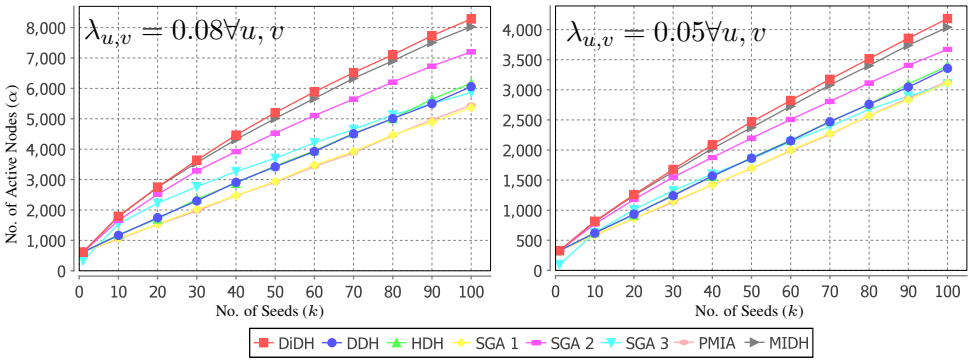


Figure 3. Plot of α with k for Twitter network on UPP setup

UPP Setup In UPP setup different values of $\lambda_{u,v}$ are taken from the interval [0.01, 0.10] with 0.01 step. Figure 3 shows two such results graphically. The graph shows the variation of the number of nodes influenced (α) with different values of k using the proposed methods (DiDH and MIDH) and two other heuristics algorithms (High Degree Heuristics (HDH) and Degree Discount Heuristics (DDH)). The graph also shows the comparative results with different variants (viz., SGA1, SGA2 and SGA3) of the Set Covering Greedy algorithm (SGA) and Prefix excluding Maximum Influence Arborescence Model (PMIA). Each entry shown here corresponds to an average value of α computed over 20000 runs. As expected, α increases consistently with k for all methods and for different values of $\lambda_{u,v}$. Interestingly, the difference in α values between the proposed methods and other algorithms is seen to increase with k. HDH, being a special case of the proposed DiDH, improvement in α of the latter can be viewed as the effect of including the neighborhood nodes in the centrality measure. It is also visible from the graphs that with the higher value of $\lambda_{u,v}$, α increases. Performance-wise the proposed methods are significantly higher compared to other methods in all the cases, and the ordering may be made as follows:

$$\begin{aligned} \alpha_{SGA1} \approxeq \alpha_{PMIA} &< \alpha_{SGA3} \approxeq \alpha_{HDH} \approxeq \alpha_{DDH} \\ &< \alpha_{SGA2} \\ &< \alpha_{DiDH} \approxeq \alpha_{MIDH}. \end{aligned}$$

NUPP Setup: In this setup we used non uniform values of propagation probabilities. As discussed in Section 8.2.1, we used three different methods (distributions) to generate it. In the experiment with Twitter network, 10 different sets of values for $\lambda_{u,v} \forall u, v$ have been generated for each of the three

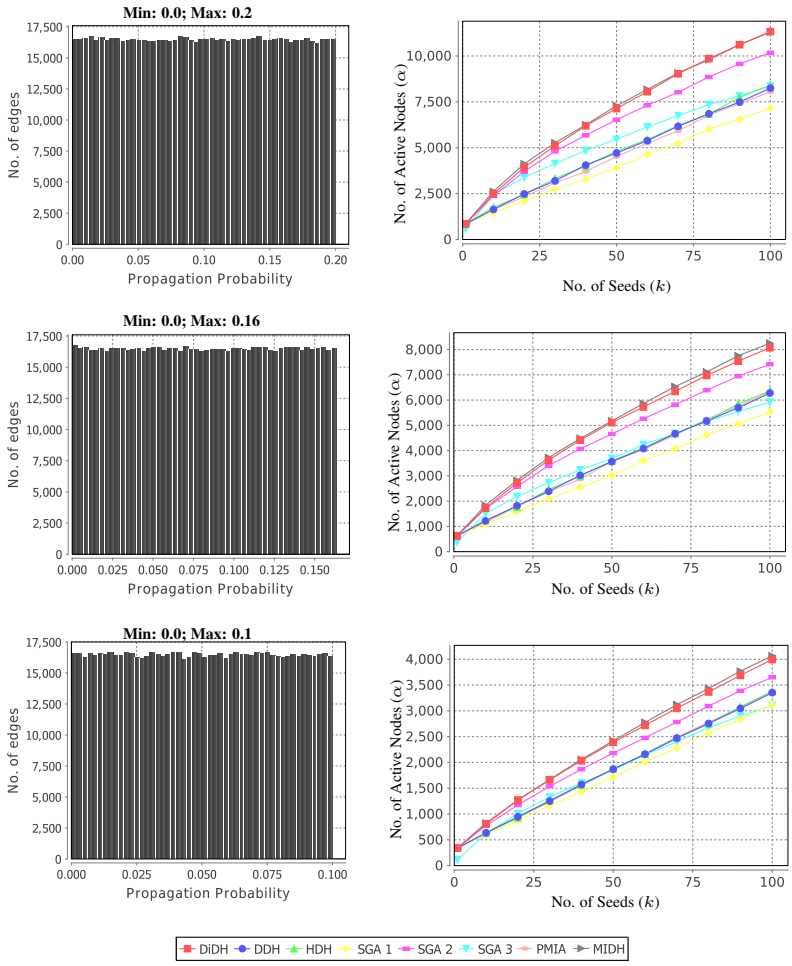


Figure 4. Assigned Propagation Probability Distribution (PPD) and corresponding results for Twitter

different methods. We observed similar outcome for all the simulations. In this section we have reported the outcome of three such simulations for each of the three different methods.

Figures 4, 5 and 6 show the histograms of assigned propagation probabilities for NUPP setup along with corresponding comparative results. The assigned propagation probabilities are generated randomly from uniform distribution, normal distribution and power law distribution respectively. Parameter values of the distributions are labeled in the diagram.

The plots of Figures 4, 5 and 6 show the variation of the number of nodes influenced (α) with different values of k for DiDH, HDH, DDH, SGA1, SGA2, SGA3 and PMIA. As expected, the value of α increases with k. As in the UPP setup, for each setup of NUPP we found that the proposed method significantly outperforms other comparing methods and the difference increases as k increase. In contrast to the UPP, the improvement with DiDH here is not only due to the consideration of the neighborhood but also for inclusion of the propagation probabilities of the links. We explained this phenomenon earlier with an example in Section 5.1. For each case discussed above, we also included the results found when

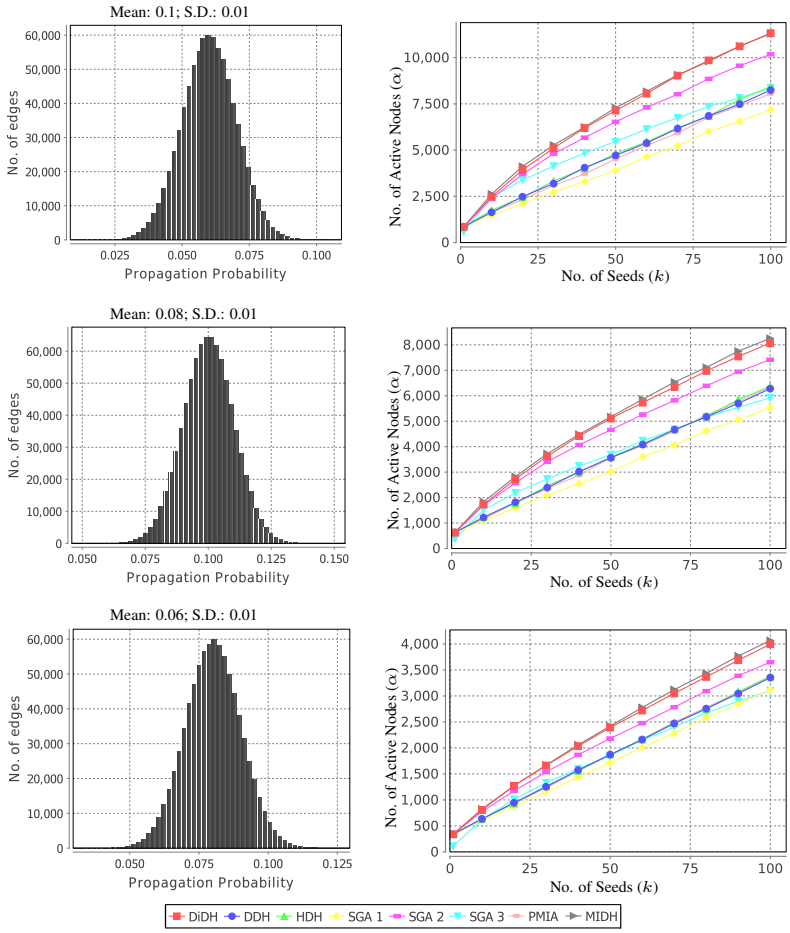


Figure 5. Assigned PPD and corresponding results for Twitter

using the top k nodes selected using MIDH. Interestingly, similar to the UPP setup, the results of DiDH are equivalent to the outcome of MIDH.

For a typical NUPP setup, the variation of the number of nodes (α) influenced by an individual seed with increasing distance for different heuristic algorithms is shown in Figure 7. We have shown some selected seed nodes, as example, from the 50 seeds to generate the graphs. It is clear from the graphs that the influence of an individual seed decreases with the distance. Interestingly, we can see an increased influence at distances 2 and 3 for the proposed DiDH algorithm compare to the HDH algorithm where the influence is either flat or it decreases after distance 1. On the other hand, DDH had a mixed behavior i.e., some of the seeds show increased influence at distances 2 and 3, while the others show the reverse. We found similar behavior for most of the top 100 seeds in our experiment. The said increased influence at distances 2 and 3 for the proposed method is due to the effect of the neighbors' contribution in the centrality measure. The results of Figure 7 also support the proposed Postulate 5.1, that is, "influence of an individual gradually decreases as distance increases".

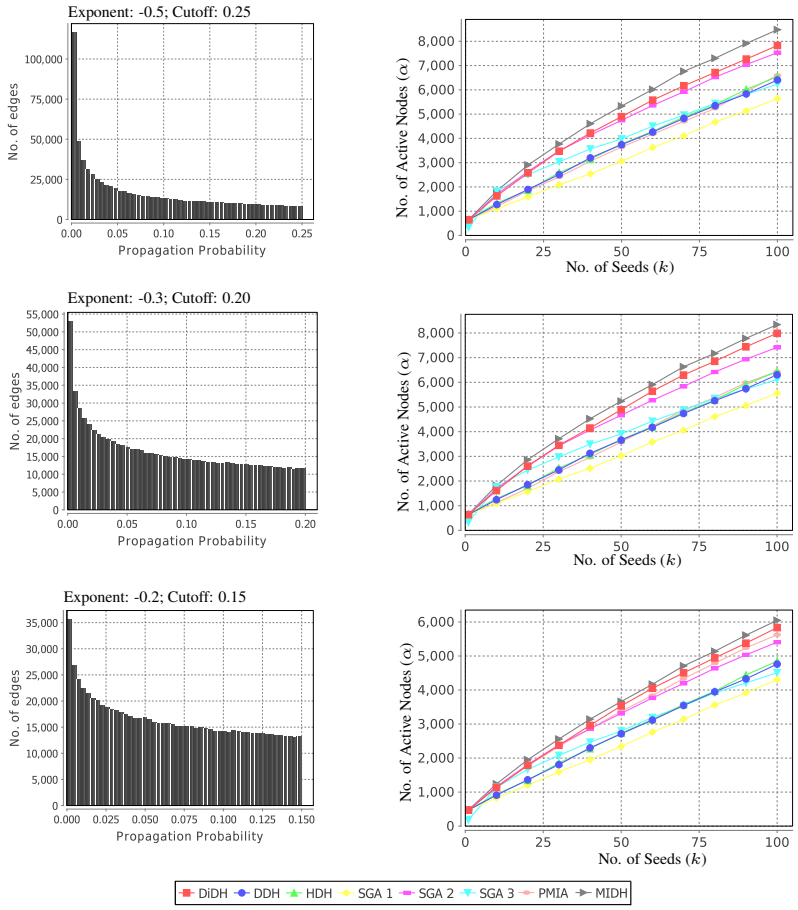


Figure 6. Assigned PPD and corresponding results for Twitter

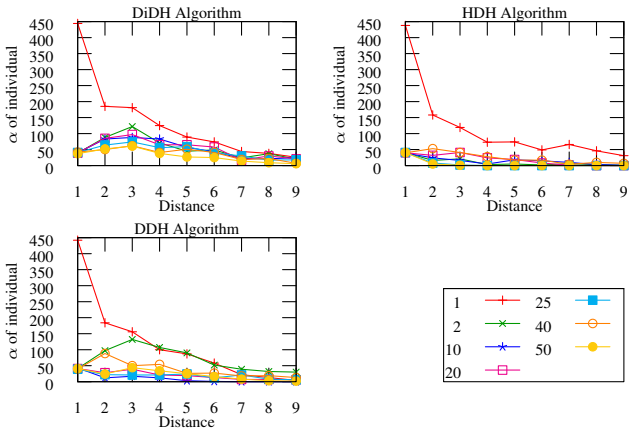


Figure 7. Plot of α for seven nodes with distance on Twitter data for different heuristic algorithms

Table 2. Maximum possible influence by selected seeds of different algorithm for Twitter

Algorithm	$\sum_{v\in S} I_v^1$	$\sum_{v\in S} I_v^2$	$\sum_{v\in S} \Delta I_v^2$	$\sum_{v\in S} I_v^3$	$\sum_{v\in S} \Delta I_v^3$	$\sum_{v\in S} I_v^4$	$\sum_{v\in S} \Delta I_v^4$	Total (% of Total Network)
HDH	27355	83550	70581	200737	124853	337713	129158	351947 (77.16%)
DDH	27407	80501	69942	194024	122716	327274	125445	345510 (75.75%)
DiDH	*23241*	*149602*	*129317*	***292760***	*155583*	*398177*	*99746*	*407887 (89.42%)*
SGA 1	**29470**	52623	45938	143527	96959	278322	**132387**	304754 (66.81%)
SGA 2	25370	**167004**	**152959**	287835	156120	**414010**	95716	**430165 (94.31%)**
SGA 3	19089	100775	94414	257362	**176696**	379550	127505	417704 (91.58%)

We computed the maximum possible influence to compare the quality of the seeds selected by different algorithms. We computed it upto the degree 4 because 90 percentile effective diameter of the network is 4 i.e., effectively to cover 90% of the network, distance 4 is sufficient. Table 2 shows the observed results. From the results it is clear that the theoretically possible maximum influence resulting by the seeds obtained from the proposed DiDH is highest among those of the other heuristic methods. However it is marginally lower than that of the SGA 2 and SGA 3.

Table 3. Centrality scores and 2^{nd} level neighborhood size for each node, and α by the top k nodes for Twitter network

k	HDH[a]				DiDH[b]				
	Node Id	In-Degree	# 2^{nd} Level Neighbors	# Influenced Nodes Upto k	Node Id	Diffusion Degree	In-Degree	# 2^{nd} Level Neighbors	# Influenced Nodes Upto k
1	0	5345	22570	1115	0	2422	5345	22570	1097
5	324342	500	5730	2026	228787	1251	496	13526	2933
10	424125	499	7216	2988	15298	1082	479	12383	4919
15	65180	499	5345	3736	399698	1012	462	11971	6566
20	297131	498	6499	4332	175776	990	433	12329	7837
25	119285	498	6421	5202	140201	963	480	10812	9021
28	106958	498	6278	5608	277261	959	322	10929	9562
30	102641	498	7867	6051	382965	949	488	10406	10036
35	2069	498	6793	6590	46373	938	496	10348	11116
37	391906	497	7596	6743	124854	937	382	10583	11516
40	286074	497	8994	7277	281580	931	488	9515	12087
45	142473	497	7661	8051	420556	922	493	10954	13065
50	55849	497	5345	8593	216476	917	478	11192	13908

[a] HDH Stat: Avg. In-Degree = 595, Avg. Neighborhood size in 2^{nd} level = 7337

[b] DiDH Stat: Avg. Diffusion Degree = 1036, Avg. In-Degree = 566, Avg. Neighborhood size in 2^{nd} level = 11625

In order to determine the effect of the neighborhood in the centrality measure, let us consider the results comparing DiDH and its spacial case HDH as shown in Table 3. Here we list the centrality scores and 2^{nd} level neighborhood size for each node, and the number of influence nodes by the top k nodes for Twitter data. For HDH, the top k nodes are selected based on their in-degree score and in DiDH, top k nodes are selected based on their diffusion degree score. Number of influenced nodes is estimated using Monte Carlo simulation. The number of nodes influenced by k nodes clearly shows that the proposed model outperforms the HDH model. The list also shows that the top k nodes selected by the proposed algorithm have higher 2^{nd} level neighbors. Even though the seed nodes in HDH have higher in-degree, the DiDH outperforms it due to the higher number of followers in 2^{nd} level. That is, the proposed measure is able to identify correctly those nodes having more influence (direct plus indirect).

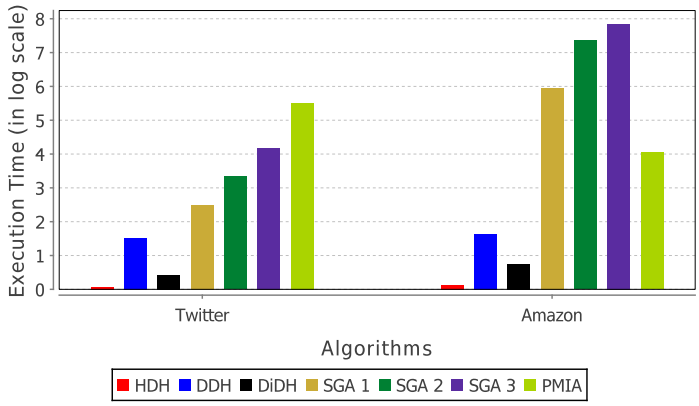


Figure 8. Comparing execution time of different algorithms for different networks

Execution Time Figure 8 shows the comparative plot of execution time for all the algorithms in sec (log scale). Being a special case of DiDH, HDH takes less time than DiDH and lowest time compare to all other methods. The DiDH takes the second lowest time. The PMIA, takes the longest time when we use an optimal threshold as per the authors guideline. Though DDH method is a heuristic model, it takes comparably longer time to execute with respect to the proposed algorithm. The reason behind it may be that the discounted degree need to be recalculated for every node after selection of each seed node i.e., selection of a node as seed is dependent on nodes already selected before. Ordering of the algorithms in terms of execution time is

$$\begin{aligned} t_{PMIA} &> t_{SGA3} \\ &> t_{SGA2} \\ &> t_{SGA1} \\ &> t_{DDH} \\ &> t_{DiDH} \\ &> t_{HDH}. \end{aligned}$$

Test of Significance We verified the significance of the difference in number of influenced nodes for each value of k using T-test. Here we perform the T-test only for DiDH vs HDH and DiDH vs DDH as these are seen to perform very close in Figures 4, 5 and 6. It is found that for $k \leq 10$ the differences among three heuristic algorithms are not significant. That is, the resulting influence by three different heuristics is very close. However, as k increases, the differences are seen to be statistically significant. The p-value and t-score for different values of k in a particular experiment, as an example, are listed in Table 4. We used Apache Commons Math library to obtain the results listed in Table 4. Each p-value returned by the procedure is the smallest significance level at which one can reject the null hypothesis that the two means are different [1] for a two-tailed T-test. That is, $(1 - p)$ signifies the probability at which one can accept that the means are different. The test does not assume that the underlying population variances are equal, and it uses Welch-Satterthwaite approximation for degrees of freedom.

Corresponding t-scores were calculated as

$$t = \frac{(m_1 - m_2)}{\sqrt{(\frac{var_1}{n_1} + \frac{var_2}{n_2})}}$$

where n_1, n_2 are the sizes, m_1, m_2 are the means and var_1, var_2 are the variances of the samples.

The minimum observed significance level (p-value) of the proposed DiDH compared to HDH is found to be 0.022 for $k = 10$. For higher value of k, this increases further. Similar is the case for DiDH compared to DDH. Thus the results of the proposed method are statistically found to be significantly higher than those of the other two.

Table 4. T-test Results for Twitter Data Set

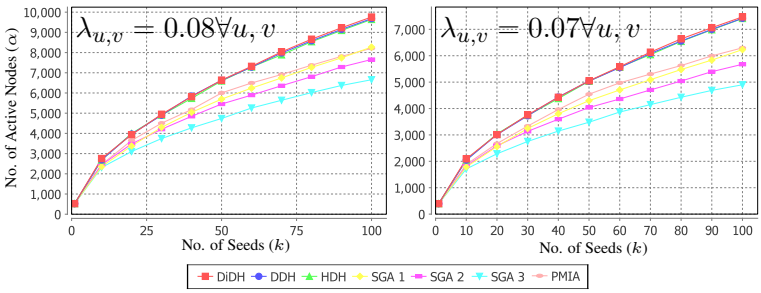
	DiDH vs HDH		DiDH vs DDH	
k (Seed#)	**p-Value**	**t-Score**	**p-Value**	**t-Score**
1	8.31×10^{-03}	-2.64	7.90×10^{-02}	-1.76
5	6.03×10^{-02}	1.88	1.60×10^{-03}	3.16
10	2.23×10^{-02}	2.29	1.55×10^{-02}	2.42
20	6.76×10^{-08}	5.42	8.10×10^{-09}	5.79
30	1.38×10^{-20}	9.41	3.52×10^{-19}	9.05
40	1.26×10^{-28}	11.30	1.73×10^{-29}	11.50
50	2.86×10^{-58}	16.70	4.89×10^{-48}	15.00

8.5.2. Results for Amazon Data

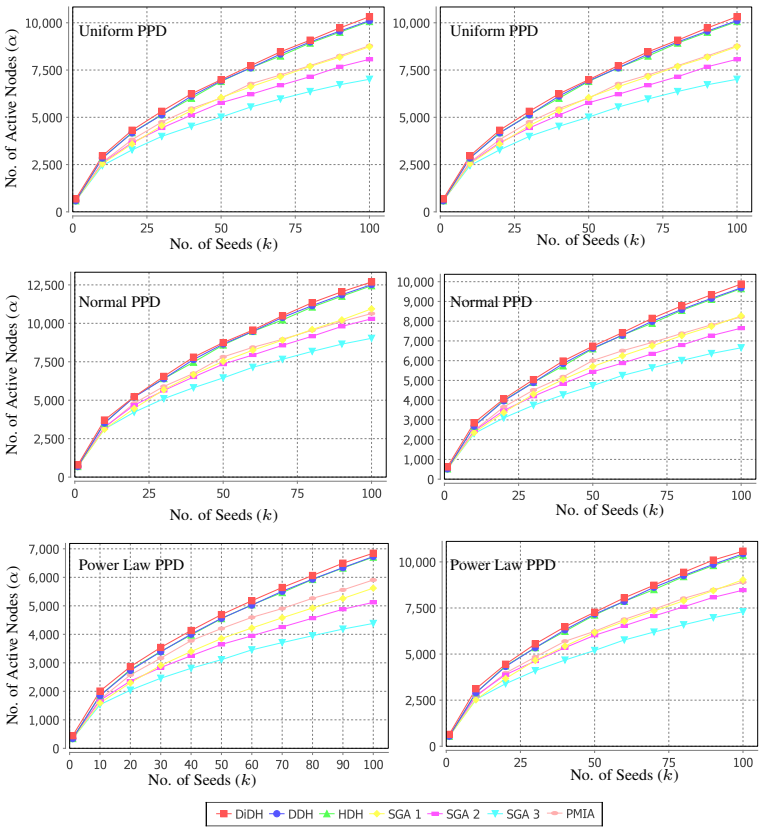
The Amazon product co-purchasing network is based on *Who Bought This Item Also Bought* feature of Amazon web site. If a product u is frequently purchased with product v the graph contains a directed link from node u to node v. Here α denotes the number of products sold through Amazon website. So, higher values of α for given k seeds mean the number of additional products sold is higher for those k products, and the performance of the corresponding algorithm is better in terms of the number of products sold.

UPP Setup Figure 9(a) shows the variation of the number of influenced nodes (α) with k for Amazon co-purchasing network. The graphs show the comparative results among DiDH, HDH, DDH, different variants of SGA and PMIA for different values of $\lambda_{u,v}$. Similar to Twitter, α values for DiDH are found to be higher than the other comparing methods, and the difference is more apparent for propagation probabilities $\lambda_{u,v} > 0.05\ \forall u, v$ as well as for higher k-value.

NUPP Setup We assigned propagation probabilities using the aforesaid distributions in the network and conducted experiments. Here we have reported few of the results as examples. Comparative plots of α with k are given in Figures 9(b). Results shows that the seeds selected by the proposed DiDH has the highest possible influence in the network. Unlike Twitter, here the other heuristic methods HDH



(a) UPP Setup



(b) NUPP Setup

Figure 9. Plot of α with k for Amazon network

and DDH provides comparatively closer results, however, the greedy algorithms SGA and PMIA are significantly lower.

Figure 10 shows the variation of α for different seed nodes with distance for an experiment with NUPP setup. Here we considered seven seed nodes from the top 50 seeds to generate the graph. Variation in α is seen to be similar for the three heuristics. Decreasing nature of α with distance supports the Postulate 5.1.

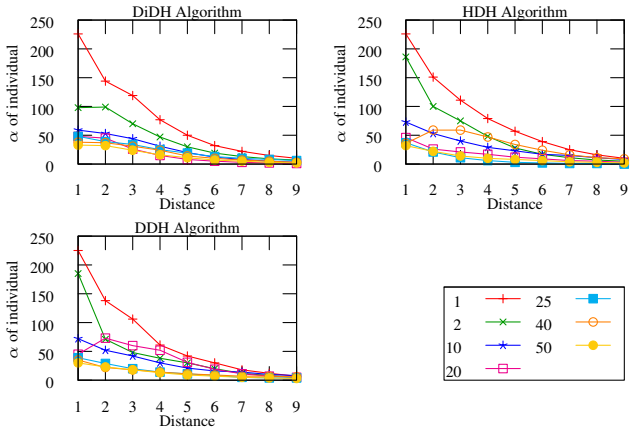


Figure 10. Plot for α for seven nodes with distance on Amazon data for different heuristic algorithms

Similar to Twitter, Table 5 lists the centrality scores and 2^{nd} level neighborhood size for each node, and the number of influence nodes by the top k nodes for Amazon network resulting from an experiment with the NUPP setup for DiDH and its spacial case HDH. Here also, we found that the selected nodes by the proposed DiDH method have higher average 2^{nd} level neighborhood. It is also notable that among the top 50 nodes, few low in-degree nodes are identified by the proposed DiDH as influencing, due to their relatively higher 2^{nd} level neighbors. Those nodes are ignored in the special case HDH. As a result the proposed algorithm is seen to have more influence than HDH.

Table 5. Centrality scores and 2^{nd} level neighborhood size for each node, and α by the top k nodes for Amazon network

k	HDH[a]				DiDH[b]				
	Node Id	In-Degree	# 2^{nd} Level Neighbors	# Influenced Nodes Upto k	Node Id	Diffusion Degree	In-Degree	# 2^{nd} Level Neighbors	# Influenced Nodes Upto k
1	32	2747	19703	627	32	1847	2747	19703	621
5	12588	1282	10663	2139	335	1068	2247	10795	2178
10	21020	875	6328	3433	2886	754	773	9104	3503
15	19878	575	5710	4409	21020	584	875	6328	4507
20	6847	530	6654	5180	11	519	413	5892	5304
25	30921	474	4582	5809	12030	499	583	5608	6225
30	159	450	4823	6351	19878	470	575	5710	6947
35	10903	426	4763	7120	1573	456	531	5377	7481
40	95401	412	3402	7765	7790	430	475	4901	8339
45	4537	388	5979	8326	7890	420	349	4886	9172
50	46024	372	3125	9184	2689	392	423	4452	9824

[a] HDH Stat: Avg. In-Degree = 673.08, Avg. Neighborhood size in 2^{nd} level = 6390.22

[b] DiDH Stat: Avg. Diffusion Degree = 620.68, Avg. In-Degree = 643.48, Avg. Neighborhood size in 2^{nd} level = 7052.10

Execution Time Figure 8 shows the chart of execution time in sec (log scale). Similar to Twitter data, we found DiDH taking comparatively less time than DDH and SGA algorithms. Being a special case, HDH takes less time than DiDH. Ordering of the algorithms in terms of execution time is similar to that found in the Twitter (Eq. 8.5.1) except PMIA taking comparatively less amount of time.

8.5.3. Other Data

We have run the same experiments over the remaining three networks (Table 1). Superiority of the proposed DiDH compared to HDH, DDH, PMIA and SGA is also seen to be valid for Web-NotreDame and Web-BerkStan web networks. In the Shashdot friendship network, the performance of DiDH is higher than DDH, SGA and PIMA, but it is comparable to the HDH. Figure 11 shows graphs comparing the performance for those networks.

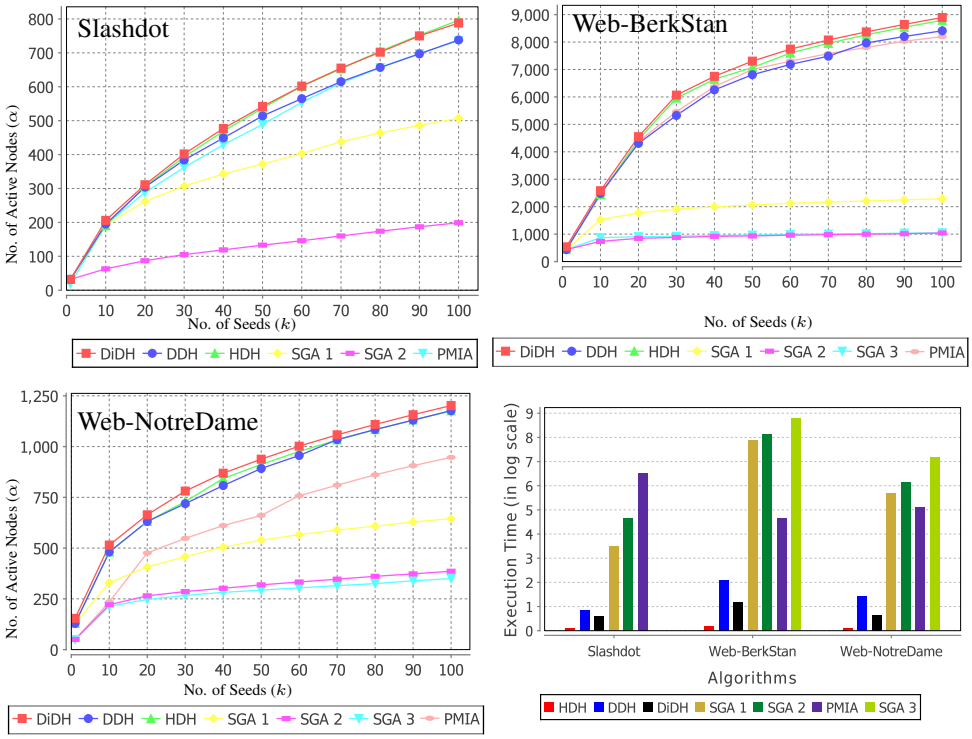


Figure 11. Plot of α with k for different networks

9. Conclusion & Discussions

In this paper, we proposed two centrality measures namely *Diffusion Degree* and *Maximum Influence Degree (MID)* which are then used in centrality based heuristic models DiDH and MIDH respectively for influence maximization in social networks. The measure *diffusion degree*, which takes in account the neighbors' contribution, is designed to work with nonuniform values of propagation probabilities. On the other hand, *maximum influence degree* provides a theoretical upper bound of influence by integrating all possible contributions from both direct and indirect neighbors. We showed through extensive experiments and statistical tests that using these measures in heuristics algorithms provides a significant improvement over the existing centrality based heuristics and different greedy algorithms like PMIA and SGAs for large scale directed social networks.

The distribution of the propagation probability is not well studied in the literature and is also dependent upon the network, relations and context. In our experiment we tried to find out the performance of our algorithm with few well known distributions characterizing non-uniform propagation probability. Improvement in performance with them demonstrates the significance of incorporating neighbors' contribution in the proposed centrality measures. One may also interested in context based analysis to acquire propagation probabilities of social networks. Readers may be refer [2, 3, 23] for more on context applied in AI, Machine Learning, Natural Language Processing and other domains.

Although our methodology has been formulated keeping the problem of influence maximization for viral marketing (i.e., applicable to friendship networks like Twitter and Slashdot) in mind, we have additionally done the experiment with other types of networks (like web graph of Web-BerkStan and Web-NotreDame, and co-purchasing network of Amazon) where we think the proposed investigation has significance too. For example, in case of web graphs, it can provide an answer to the questions like "Which web pages should we chose to advertise so that the impression maximizes?" or "Which web pages can be used for spreading social awareness?". For product co-purchasing network, the methodology can be fitted to check "Which product needs to be promoted for an increase in sell?" or it can be easily modified to figure out "Which product provides the maximum percent of revenue for an e-commerce company?".

The computation of the upper bound is time consuming, it is seen experimentally that the similar influence can be achieved by using Diffusion Degree in very less time. The running time of our DiDH algorithm matches that of the classical degree centrality based heuristic algorithm, and it is significantly faster than the PMIA and SGAs.

Acknowledgment

The authors acknowledge the Department of Science and Technology, Govt. of India for funding the Center for Soft Computing Research at Indian Statistical Institute. S. K. Pal acknowledges the J. C. Bose National Fellowship.

References

[1] Apache Commons Math (TTest) Documentation.

[2] Bettini, C., Brdiczka, O., Henricksen, K., Indulska, J., Nicklas, D., Ranganathan, A., Riboni, D.: A survey of context modelling and reasoning techniques, *Pervasive and Mobile Computing*, **6**(2), 2010, 161–180.

[3] Brézillon, P.: Context in problem solving: a survey, *The Knowledge Engineering Review*, **14**(01), 1999, 47–80.

[4] Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks, *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2010.

[5] Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2009.

[6] Chen, W., Yuan, Y., Zhang, L.: Scalable influence maximization in social networks under the linear threshold model, *2010 IEEE International Conference on Data Mining*, IEEE, 2010, ISSN 1550-4786.

[7] Choudhury, M. D., Sundaram, H., John, A., Seligmann, D. D., Kelliher, A.: "Birds of a Feather": Does User Homophily Impact Information Diffusion in Social Media?, *CoRR*, **abs/1006.1702**, 2010.

[8] Dolecek, L., Shah, D.: Influence in a large society: Interplay between information dynamics and network structure, *2009 IEEE International Symposium on Information Theory*, IEEE, June 2009, ISBN 978-1-4244-4312-3.

[9] Domingos, P., Richardson, M.: Mining the network value of customers, *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2001.

[10] Estevez, P. a., Vera, P., Saito, K.: Selecting the Most Influential Nodes in Social Networks, *2007 International Joint Conference on Neural Networks*, August 2007, 2397–2402, ISSN 1098-7576.

[11] Estevez, P. a., Vera, P., Saito, K.: Selecting the Most Influential Nodes in Social Networks, *2007 International Joint Conference on Neural Networks*, IEEE, August 2007, ISBN 978-1-4244-1379-9, ISSN 1098-7576.

[12] Freeman, L.: Centrality in social networks conceptual clarification, *Social networks*, **1**(3), 1979, 215–239, ISSN 0378-8733.

[13] Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth, *Marketing Letters*, **12**(3), 2001, 211–223, ISSN 0923-0645.

[14] Goldenberg, J., Libai, B., Muller, E.: Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata, *Academy of Marketing Science Review*, **9**(3), 2001, 1–18.

[15] Goyal, A., Lu, W., Lakshmanan, L.: CELF++: optimizing the greedy algorithm for influence maximization in social networks, *Proceedings of the 20th international conference companion on World wide web*, ACM, 2011.

[16] Granovetter, M.: Threshold models of collective behavior, *The American Journal of Sociology*, **83**(6), 1978, 1420–1443, ISSN 0002-9602.

[17] Jeong, H., Albert, R.: Diameter of the world-wide web, *Nature*, **401**(September), 1999, 398–399.

[18] Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*, ACM Press, New York, New York, USA, 2003, ISBN 1581137370.

[19] Kimura, M., Saito, K.: Tractable Models for Information Diffusion in Social Networks, *Principles of Data Mining and Knowledge Discovery*, 2006.

[20] Leskovec, J., Adamic, L. a., Huberman, B. a.: The dynamics of viral marketing, *ACM Transactions on the Web*, **1**(1), May 2007, 5–es, ISSN 15591131.

[21] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks, *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2007.

[22] Leskovec, J., Lang, K., Dasgupta, A., Mahoney, M.: Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters, *Internet Mathematics*, **6**(1), 2009, 29–123.

[23] McCarthy, J.: Notes on formalizing context, *Proceedings of the 13th international joint conference on Artifical intelligence - Volume 1*, Morgan Kaufmann Publishers Inc., Chambery, France, 1993.

[24] Narayanam, R., Narahari, Y.: A Shapley Value-Based Approach to Discover Influential Nodes in Social Networks, *IEEE Transactions on Automation Science and Engineering*, **8**(1), 2011, 130–147.

[25] Nieminen, J.: On the Centrality in a Graph, *Scandinavian Journal of Psychology*, **15**, September 1974, 332–336.

[26] Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, 2002, 61.

[27] Rogers, E. M.: *Diffusion of Innovations*, The Free Press of Glencoe, New York, 1962.