International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 14, No.1 (Aug-2023)

http://dx.doi.org/10.12785/ijcds/140144

Seed Selection Algorithm using Centrality based Entropy

Kinjal Rabadiya[1] **and Ritesh Patel**[2]

[1,2]*U & P U Patel Department of Computer Engineering, CHARUSAT, Changa, Anand-388421, Gujarat (India)*

Received 01 Jul. 2022, Revised 08 Jul. 2023, Accepted 10 Jul. 2023, Published 01 Aug. 2023

Abstract: Recently, Business expansion, marketing and advertisement is more fast and convenient process through social network analytic. In this paper, the influence maximization problem is addressed, which is the process of selecting the best suitable initial users or customers or spreaders who can use or advertise or spread the product information in such way that in their own social network maximum people can receive the information about that product. Still, the seed selection problem is NP-hard problem and to date none of the algorithm has focus on combination of various centrality of nodes that can significantly impact on seed selection process. In this paper, we propose the novel seed selection algorithm which can fill the gap and achieve the diffusion speed by combining five centrality of node. For that, We conduct simulations to evaluate the diffusion speed of our proposed algorithm and existing benchmark seed selection algorithms using real-world authors collaboration networks. Experimental results show that our proposed algorithm outperforms various existing benchmark seed selection algorithms by achieving optimal diffusion speed.
Keywords: Influence maximization, Seed selection, Social network, Diffusion, Centrality, Entropy

1. Introduction

Recently, online business expansion and advertisements are at the peak. In this era, influence maximization is a quite important phase in social network applications, such as online business campaign, product launching and so on. Influence maximization is to select group of such people from the social network who can target a large community in the network. In general, influence maximization problem is to select the initial people who can help in the maximization of the influence in their network. There are lots of seed selection related studies available for influence maximization using various information diffusion models, such as Independent Cascade (IC) model [1] and Linear Threshold (LT) model [2]. According to these two models, each social network consists of nodes having states "Active" or "Inactive". If the set of nodes have accepted the information provided by their neighbor and also actively diffusing information in their network then the set of nodes is called active nodes, otherwise, it is inactive node. Initially, Kempe et al. [3] proposed the seed selection problem which is a discrete optimization problem. There is a vast range of important applications of information diffusion in viral marketing and the problem is explored so well still many demands are not satisfied yet. According to "No Free Launch"[4] theorem, if one optimization algorithm performs well for a specific set of problems then that doesn't guarantee to solve all other optimization problems. Hence, by following the NFL theorem, researchers can propose some novel optimization algorithms to solve the problems in various fields and may exist multiple optimization algorithms. So, we take various centralities of nodes in consideration to improve the diffusion speed of information maximization. This novel algorithm is entitled, as a *Seed selection Algorithm using Centrality based Entropy*.

In this study, this algorithm is applied for selection of initial nodes which are known as seed nodes. Seed nodes help us to achieve influence maximization in network. It motivates us to propagate our study with the following objectives, as: (a) Increasing in diffusion and (b) Reduction in diffusion time. Centrality measures can be given as, Degree Centrality [5], Closeness Centrality[5], Betweenness Centrality[5], Eigenvector Centrality[5], PageRank Centrality [6]. Hence, the contributions of this article can be given, as: We have proposed a seed selection algorithm to maximize the speed of information diffusion in the network and the proposed algorithm is verified with benchmark real-world data sets as well as existing seed selection algorithms. As shown in experimental results, the proposed algorithm gives better performance over the various existing algorithms.

Analysis of large social network for information diffusion is very tedious task for data analyst. There are various social, emotional, economical factors that can maximize or minimize the influence diffusion. Thus, influence is highly depends on advertisement and profit-loss of particular product, so information maximization in social networking is

E-mail address: kinjalrabadiya20@gmail.com, riteshpatel.ce@charusat.ac.in

challenging task. As social network has a nature type dynamic, seed selection is extremely crucial task for diffusion. Besides having all these factors, an efficient seed selection algorithm is proposed based on various centrality measures, and compared with various benchmark data sets. The proposed algorithm can be applied on directed as well as undirected networks. Finally, we provide the diffusion speed analysis of benchmark algorithms v/s the proposed algorithm.

The remainder of this article is organized as follows. Section 2 includes Influence Maximization related work. Section 3 briefly discuss the preliminaries required for the study. The proposed seed selection approach is discussed in Section 4. The empirical analysis is discussed in Section 5. Finally, the conclusion is presented in Section 6.

2. Related work

The influence maximization problem is a most recent issue in the field of social network analytic. One of the important application of influence maximization is for company to promote products online using word-of-mouth effects in social networks. It should be cost-effective for company and people both. Company gives discount for connecting few more in the particular scheme. At the last potentially large cascade is generated by initial adopters of product. To generate cascade in maximum as possible way, selection of initial adopters of products is one of the crucial task [7]. Domingos and Richardson et al. [8] proposed the influence maximization as application of viral marketing. The influence maximization as an optimization problem as well as NP-hard under independent cascade model and linear threshold model was proposed by Kempe et al. [3], [9].

To remove unnecessary simulations, Leskovec et al. [10] proposed CELF algorithm, which uses the lazy evaluation technique based on sub-modular function and prior queue is used for implementation. To find the relevant solution of problem Chen et al. [11] discussed existing work related to influence maximization in social network. To enhances the performance of CELF, Goyal et al. [12] proposed extended version of it, which is named as CELF++. To achieve the sub-modularity of influence during the information diffusion process, Cheng et al. [13] proposed Static Greedy algorithm. With the help of Static Greedy, it is possible to achieve high accuracy, by reducing the cost of computation. Borgs et al. [14] used reverse influence sampling method and proposed novel algorithm to solve influence maximization problem, which is capable to increase the efficiency as well as it is independent from the framework of greedy algorithm. Many researchers have contributed their efforts to solve the influence maximization problem as well as extended problem of it. Initial adopters i.e., seed set demands some budget to maximize the influence in network. To reduce the allocated budget, Leskovec et al. [10] proposed budget oriented method in which selection is performed based on influence on network. As we know that social network is not static network. As it is dynamic network, structure of network can be changed at any point of time. The solution of influence maximization issues related to dynamic network was given by Zhuang et al.[15] and Chen et al.[16]. Similarly, Yang et al. [17] also proposed method named as a general coordinate descent algorithm.

Wang et al. [18] proposed a new problem called Information Coverage Maximization, in which seed selection is performed based on active nodes and informed nodes (the nodes still in inactive state). Some event must be diffuse in network as far as possible with in short period of time. Liu et al. [19] worked on same problem and proposed the time constrained influence maximization problem. Literature study shows that extensive research has been performed on the influence maximization problem [20], [21], [22], still this problem can't be answered thoroughly. In the traditional influence maximization problem, only influence extent is taken into consideration, but time required to spread the influence in network is one of the important factor [23]. Based on this study, we identified the solution for problem of the Influence Maximization problem. In proposed work, we have find out set of seed nodes which can help in maximization of the influence in social network which can be named as *Seed selection Algorithm using Centrality based Entropy.*

To check the spreading behavior in signed network Li at el. [24] proposed a simple opinion spreading model based on the susceptible-infected-recovered (SIR) epidemic model. During analysis in signed network, it is found that critical spreading rates is depends on the fraction of positive relationships. Fei at el. [25] proposed novel approach of identifying influential nodes in complex network by combining of the existing centrality measures. Experiments conducted over proposed method shows superiority of proposed work. Zhang at el. [26] also proposed seed selection algorithm which remove the edges and obtained global efficiency. Based on the global efficiency, new centrality measure can be achieved, which is more effective than the other three centrality measures. Yang at el. also [27] proposed a dynamic weighted Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) to identify the influential nodes in complex networks. It is dynamic and works based on the assignment of the appropriate weight to each attribute, based on the grey relational analysis method and the Susceptible-Infected-Recovered (SIR) model.

3. Preliminaries

In this section, we provide the some concepts related to influence maximization and seed selection, and basic definitions of terminologies used throughout the article. Initially, to analyse the various properties of social network, it is mandatory to represent the social network through one of the data structure. Here, graph data structure is most suitable, so mathematical representation of social network in terms of graph can be given as follows:

Definition 1: **(Social Network as Graph)** In a social network analysis, a graph G consists of two features as node V_i, and directed or undirected edge set E_i $(i = 1, 2, \ldots, 3)$.

TABLE I. A benchmark network: Karate data set [28].

Source → Target	Source → Target	Source → Target	Source → Target	Source → Target	Source → Target
0 → 1	4 → 6	13 → 2	19 → 0	30 → 1	32 → 23
0 → 2	5 → 6	13 → 3	19 → 1	30 → 32	32 → 29
0 → 3	8 → 0	13 → 33	19 → 33	30 → 33	32 → 33
0 → 4	8 → 2	14 → 32	21 → 0	30 → 8	32 → 8
0 → 5	9 → 2	14 → 33	21 → 1	31 → 0	33 → 20
0 → 6	10 → 0	15 → 32	23 → 25	31 → 24	33 → 22
0 → 7	10 → 4	15 → 33	23 → 27	31 → 25	33 → 23
1 → 2	10 → 5	16 → 5	25 → 24	31 → 28	33 → 26
1 → 3	11 → 0	16 → 6	27 → 2	31 → 32	33 → 27
1 → 7	12 → 0	17 → 0	27 → 24	31 → 33	33 → 28
2 → 3	12 → 3	17 → 1	28 → 2	32 → 20	33 → 29
2 → 7	13 → 0	18 → 32	29 → 23	32 → 2	33 → 8
3 → 7	13 → 1	18 → 33	29 → 26	32 → 22	33 → 9

It can be represented, as:

$$G = (V_i, E_i) \tag{1}$$

Here, V_i is the set of nodes which show person in network, and E_i is the set of edges, shows person's relationship with other people in network.

Here, graph data structure is most suitable, so mathematical representation of social network can be given in terms of graph. The edge representation of Karate dataset is as shown in Table I. By referring graph structure of social network, various analysis can be performed on it. Each analysis based on centrality measures can be defined as follows:

Definition 2: **(Degree Centrality)[5]** In social network $G = (V_i, E_i)$, the set of vertices having higher out-degree k_{out} are can be considered as more central nodes, as they are capable to generate more choices for their neighbour nodes while the set of vertices having higher in-degree k_{in} are can be considered as more prestigious nodes, as they are capable to receive more choices for their neighbour nodes.

Definition 3: **(Closeness Centrality)[5]** In social network $G = (V_i, E_i)$, the farness of node v_i can be given by sum of the distance of node v_i to all other nodes in V_i. The closeness of node v_i is the inverse of the farness. Mathematically, closeness of node v_i can be given as:

$$Closeness(v_i) = \frac{1}{\sum_{v_i \neq u_i} d_{u_i}} \tag{2}$$

Closeness of node shows how much time will it take to spread the information in network from v_i to all other nodes u_i $(i = 1, 2, \ldots, n)$, in network. More closeness of nodes shows less time will it take to diffuse the information.

Definition 4: **(Betweenness Centrality)[5]** In social network $G = (V_i, E_i)$, the betweenness of node v_i can be given by sum of fraction of all pairs of shortest paths from u to w that pass through node v_i to all pairs of shortest paths from u to w. Mathematically, betweenness of node v_i can be given as:

$$Betweenness(v_i) = \sum_{\forall u,w \in V_i} \frac{P(u,w)|v_i}{P(u,w)} \tag{3}$$

Definition 5: **(Eigenvector Centrality)[5]** In social network $G = (V_i, E_i)$, the eigenvector centrality of node v_i is the centrality for a node based on the centrality of its neighbors. Nodes become important if its neighbors have strong connections.

$$Eigenvector(v_i) = \lambda \sum_{\forall v_j} (W * X(v_j)) \tag{4}$$

where W is weight of each neighbours of v_i and v_j are the set of nodes which are neighbours of v_i.

Definition 6: **(PageRank Centrality) [6]** In social network $G = (V_i, E_i)$, the PageRank centrality of node v_i ranks the importance of nodes in a graph based on how likely they are to be reached when traversing a graph.

Recently, most of the research trends are towards the solution of NP-hard problems. Influence maximization is one of the NP-hard problem. The detailed description of influence maximization and seed selection problem are as given below.

A. *Influence Maximization*

When information is diffused in network to maximize the spread, basically there are four models are used for diffusion, as: (1) Linear Threshold model (LTM), (2) Independent Cascade model (ICM), (3) Heat diffusion model (HDM), (4) Epidemic model. Detailed description of linear threshold model and independent cascade model are as given below:

1) **Linear threshold model:** Initially, Granovetter and Schelling [2] proposed the linear threshold model. In this model, each individual has a its own threshold to adopt the behavior of group from where they belongs. The threshold of individual nodes v_i can be indicated using $\theta_i \in [0, 1]$. If more number of nodes join to the same behavior then automatically social pressure increases on the nodes who are still agreed to the same behaviour. In this model, once the node v_i get activated, it will remain activated throughout the diffusion [20]. At each stage of diffusion, all inactive nodes v_j compare their own threshold θ_j with other active nodes and if they found that their own threshold θ_j is less than other activated nodes then inactive nodes change their status to activated. As well as newly activated nodes remain activated and also tries to activate all other inactive nodes. Also they can tries more than one attempt to activate other inactive neighbour nodes.
2) **Independent cascade model:** Initially, the independent cascade model was proposed by Goldenberg [1]. This model works based on activation probability of other neighbor nodes and a node has single chance to get activated. For example, in network node v_i is active node and it tries to get activate neighbor node v_j. Node v_i can attempt to activate node v_j only once. Activation of node v_j may results in fail or successful activation through v_i. The edge between node v_i and v_j has probability p so node v_i has single chance to

activate node v_j with probability value p. If edges have some weight assigned then weighted cascade model is taken in to account.

B. Seed Selection

Domingos and Richardson proposed the influence maximization problem for probabilistic methods. Kempe et al. proposed the model for the discrete optimization problem[29], [30], [31]. Greedy algorithm is computationally inefficient for large networks[32]. Thus, two overcome various limitations two major types of algorithms have been proposed as solutions:(i) Heuristic algorithm, (ii) Greedy algorithms. To improve the efficiency of seed selection many heuristic algorithms have been proposed i.e. Degree Discount, SIMPATH, ShortestPath etc., whereas optimization algorithms have been proposed to improve the running time i.e. CELF, CELF++, NewGreedy and MixedGreedy [33]. Some other seed selection methods are as explained below:

1) Random: In this method, it assigns seeds uniformly at random.
2) Greedy: It is the result when algorithm picks seeds independently from each others [34], [35].
3) Degree: It is a heuristic algorithm based on degree centrality in which high degree nodes selected as influential ones. The seeds are the nodes with the k highest out-degrees [36], [37].
4) Degree Centrality: In this method the highest degree nodes are used as seeds. The higher degree nodes take less time to reach more nodes. [38].
5) Degree Discount [39] : A degree discount heuristic algorithm which provides much better results than the classical degree and centrality based heuristic algorithms..
6) CELF [10] : Cost Effective Lazy Forward (CELF) is focus on contaminant detection for water distribution network, finding important stories in a blog network and 700 times faster than the greedy algorithm.
7) CELF++ [12] : It is extended version of CELF which is 35% to 55% faster than CELF.
8) NewGreedy [39] : This algorithm is specially made for independent cascade model with 20000 simulation rounds.
9) MixedGreedy [39] : The MixedGreedy algorithm is made specially for independent cascade model. In this algorithm first round uses NewGreedy and then it uses CELFGreedy algorithm.
10) SIMPATH [40] : This algorithm is based on vertex cover of nodes. SIMPATH works on the CELF optimization that iteratively selects seeds in a lazy forward manner.
11) ShortestPath [41] : This model is based on shortest path. The node having shortest path between other node, it can be selected as seed node to influence other nodes.

Besides the above solutions of seed selection algorithms, we proposed the seed selection from the view of social community, which is based on centrality measures for mining top-k influential nodes.

4. The Proposed Seed Selection Algorithm

In social network, each node has some properties as they have the connection to other people. The set of best influential nodes can be selected using their edge properties. As per the Definitions (2) - (6), various conclusion can be derived. Higher degree centrality shows higher connections. So the nodes having higher degree are suitable for seed selection. Higher value of closeness centrality shows more closeness of nodes with each other. So the nodes having more closeness are suitable for seed selection. Higher betweenness centrality shows higher connections as intermediate node for shortest path. So the nodes having higher betweenness are suitable for seed selection. Higher value of eigenvector centrality shows more connections of its neighbour nodes. So the nodes having more eigonvector are suitable for seed selection. Information can be spread by selecting the nodes having neighbour with strong connections.

Here, problem can be described as, based on above measures of each nodes, it can be derived that the nodes having higher values of degree centrality may have not the higher reach to network in directed network. Because higher degree of nodes may consists of higher in-degree and out-degree may be less. So, here node doesn't give higher betweenness centrality. Both degree centrality and betweenness centrality may be high for same node is not possible and it can happened with all centrality measures. So, selection of nodes having all optimal centrality is solution for influence maximization. Separately, these all centralities have different importance in network. So, the proposed method uses equation which can be used for selecting optimal seed selection using all centrality, as:

$$Entropy = \exp BC + \exp CC + \exp DC + \exp EC + \exp PR \quad (5)$$

here, BC= Betweenness Centrality, CC= Closeness Centrality, DC= Degree Centrality, EC= Eigonvector Centrality and PR=PageRank Centrality.

This Equation makes all five centrality values bigger by exponential function. So we can select optimal nodes from the network. It is looks like if we want see difference between two small dots then we have to see them from microscopic view. By making dots larger from microscope, it can be easily sort out by size. This same fundamental works for seed selection. So, we have proposed novel algorithm based on Eq. 5, as:

Algorithm 1 shows the complete entropy measures based seed selection algorithm. Line 1 shows Input as graph G, G consists of V and E where V = set of vertex, E = set of edges. Line 2 shows set of seed S where each seed is belongs to vertex. Line 3 shows that Line 4 to 7 perform for each vertex v_i in set V. Line 4 shows degree centrality, Closeness centrality, Betweenness centrality, Eigenvector centrality, PageRank based on Definitions 2- 6. Calculate

Algorithm 1 Algorithm for centrality measures based seed selection algorithm.

1: **Input:** Graph $G = (V, E)$, where V = set of vertex, E = set of edges.
2: **Output:** Seed set S, where $\forall s_i \in V$.
3: **for** $\forall v_i \in V$ **do**
4: Obtain the degree centrality, Closeness centrality, Betweenness centrality, Eigenvector centrality and PageRank for node v_i (based on Definitions (2)-(6)).
5: Calculate entropy measure based on Eq. 5.
6: Sort all the vertices in decreasing order.
7: Prepare seed set S, where $\forall s_i \in S$, having maximum energy value.
8: **end for**
9: **return** Seed set S, where $\forall s_i \in V$.

TABLE II. Various centrality measure and entropy for each node of Karate data set.

Node	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigen Vector Centrality	PageRank Centrality	Entropy
0	0.9696	0.4376	0.5689	0.3554	0.0970	8.4813
1	0.5454	0.0539	0.4852	0.2659	0.0528	6.7644
2	0.6060	0.1436	0.5593	0.3171	0.0570	7.1691
3	0.3636	0.0119	0.4647	0.2111	0.0358	6.3138
4	0.1818	0.0006	0.3793	0.0759	0.0219	5.7624
5	0.2424	0.0299	0.3837	0.0794	0.0291	5.8847
6	0.2424	0.0299	0.3837	0.0794	0.0291	5.8847
7	0.2424	0	0.44	0.1709	0.0244	6.0382
8	0.3030	0.0559	0.5156	0.2274	0.0297	6.3717
9	0.1212	0.0008	0.4342	0.1026	0.0143	5.7959
10	0.1818	0.0006	0.3793	0.0759	0.0219	5.7624
11	0.0606	0	0.3666	0.0528	0.0095	5.5692
12	0.1212	0	0.3707	0.0842	0.0146	5.6803
13	0.3030	0.0458	0.5156	0.2264	0.0295	6.3597
14	0.1212	0	0.3707	0.1014	0.0145	5.6991
15	0.1212	0	0.3707	0.1014	0.0145	5.6991
16	0.1212	0	0.2844	0.0236	0.0167	5.4987
17	0.1212	0	0.375	0.0923	0.0145	5.6953
18	0.1212	0	0.3707	0.1014	0.0145	5.6991
19	0.1818	0.0324	0.5	0.1479	0.0196	6.0603
20	0.1212	0	0.3707	0.1014	0.0145	5.6991
21	0.1212	0	0.375	0.0923	0.0145	5.6953
22	0.1212	0	0.3707	0.1014	0.0145	5.6991
23	0.3030	0.0176	0.3928	0.1501	0.0315	6.0469
24	0.1818	0.0022	0.375	0.0570	0.0210	5.7366
25	0.1818	0.0038	0.375	0.0592	0.0210	5.7404
26	0.1212	0	0.3626	0.0755	0.0150	5.6596
27	0.2424	0.0223	0.4583	0.1334	0.0256	6.0471
28	0.1818	0.0017	0.4520	0.1310	0.0195	5.9325
29	0.2424	0.0029	0.3837	0.1349	0.0262	5.9161
30	0.2424	0.0144	0.4583	0.1747	0.0245	6.0861
31	0.3636	0.1382	0.5409	0.1910	0.0371	6.5528
32	0.7272	0.1452	0.5156	0.3086	0.0716	7.3363
33	1.0303	0.3040	0.55	0.3733	0.1009	8.4493

TABLE III. A benchmark network data set [42], [28].

Dataset	Nodes	Edges	Maximum Degree	Avg. Clustering	Density	Avg. Shortest Path Length
Karate	34	78	17	0.57063	0.1390	2.3374
GrQc	5242	14496	81	0.5296	0.00105	6.0466
CondMat	23133	93497	281	0.6334	0.00034	5.3518
HepPh	9877	25998	504	0.4714	0.00053	5.9444
AstroPh	18772	198110	504	0.6305	0.00112	4.1937
Facebook	4039	88234	1045	0.6055	0.0108	3.6915

entropy measure of each node based on Eq. 5 as shown in Line 5. Sort all vertices in decreasing order and choose first k - nodes from them as shown in Line 6 and 7 of Algorithm 1. Line 9 gives seed set S as a result of Algorithm 1.

TABLE IV. A Diffusion Speed for all Dataset using different size of Seed nodes for Linear Threshold Model.

	No. of Seed Nodes	Random	Degree Centrality Centrality	Betweenness Centrality	Closeness Centrality	Eigen Vector Centrality	PageRank Centrality	The Proposed Algorithm
Karate Dataset	5	45.351	1036.269	348.028	1294.879	1085.142	1244.019	**1859.799**
	7	1000.667	1187.215	2170.284	2085.071	2168.474	2365.787	**2898.551**
	10	787.176	2337.229	2309.469	619.490	788.309	1222.707	**2367.942**
	15	1334.159	430.652	2430.307	1867.787	2418.682	1890.254	**2638.763**
GrQc	15	40.793	96.808	66.038	47.526	34.027	49.484	**118.679**
	20	62.944	131.137	102.770	151.147	48.599	117.928	**201.628**
	25	74.723	41.841	76.673	118.309	74.722	129.273	**132.960**
	30	82.133	211.306	174.466	107.175	51.114	130.180	**215.874**
	50	142.153	252.817	230.191	151.796	108.748	245.966	**523.752**
	100	268.707	419.067	311.402	330.274	168.804	596.825	**648.686**
CondMat	15	5.440	19.334	13.283	11.854	23.297	34.871	**35.191**
	20	7.998	24.547	33.279	9.701	31.074	34.488	**37.268**
	25	11.036	26.814	21.718	11.757	21.627	30.809	**31.405**
	30	13.436	36.551	54.808	20.780	46.989	48.748	**59.597**
	50	19.535	63.142	66.821	18.067	89.747	78.133	**116.010**
	100	42.371	127.372	80.803	35.627	152.238	141.083	**154.832**
HepPh	15	19.290	58.193	16.641	16.603	20.763	**113.393**	58.299
	20	31.270	73.029	22.359	30.066	40.688	**111.981**	104.951
	25	38.291	88.475	34.786	21.880	34.188	**133.367**	97.653
	30	49.323	110.059	30.973	30.957	43.992	**186.986**	130.468
	50	83.554	220.834	58.920	68.817	74.912	304.556	**311.060**
	100	137.402	350.588	111.867	147.337	163.402	563.200	**654.297**
AstroPh	15	2.555	8.397	8.439	4.630	5.553	7.286	**11.266**
	20	3.921	10.465	12.117	6.063	5.628	7.801	**21.615**
	25	4.331	12.180	14.695	6.111	5.702	8.180	**24.913**
	30	5.294	13.511	16.451	8.046	7.523	12.066	**25.037**
	50	9.489	20.472	38.217	19.679	10.874	17.041	**45.157**
	100	18.054	36.841	49.846	19.081	27.808	39.622	**63.812**
Facebook	15	5.737	75.818	96.044	35.899	16.048	113.002	**217.740**
	20	8.615	77.952	110.988	35.985	17.182	103.790	**229.870**
	25	11.793	81.137	168.439	36.550	17.621	101.179	**218.969**
	30	12.471	80.278	141.819	56.640	21.290	108.135	**251.572**
	50	26.677	81.567	122.507	55.088	27.648	134.711	**251.268**
	100	41.691	80.389	148.490	64.902	80.784	161.691	**222.334**

Table II shows all Centrality for all nodes in Karate dataset. Entropy is calculated based on all Centrality of node which shows each centrality measures are independent from each other and each have different importance in entropy function shown in Eq. 5.

5. Empirical Analysis

A. Data sets description

In order to evaluate the proposed seed selection algorithm, we conducted experiments on 6 real networks of various sizes from the Stanford Large Network Data set Collection (SNAP) [42], and Newman's Network data [28]. The main characteristics of the studied networks are shown in Table III.

- Zachary's Karate Club network: It is about that due to the conflicts between the instructor Mr. Hi and the administrator Mr. John A, the karate club was divided into two parts. In includes 34 nodes and 78 edges between nodes, which was studied by Wayne W.
- Facebook Network: The circles are collected by surveying friend list from Facebook app. By this survey, the node profile, friends list as circles and ego network information are collected.
- Collaboration networks: AstroPh network, CondMat network, GrQc network, HepPh network are collaboration network of Astro Physics, Condense Matter Physics, General Relativity and Quantum Cosmology, and High Energy Physics - Phenomenology, respectively. The data includes papers of the duration from January 1993 to April 2003.

TABLE V. A Diffusion Speed for all Dataset using different size of Seed nodes for Independent Cascade Model.

	No. of Seed Nodes	Random	Degree Centrality Centrality	Betweenness Centrality	Closeness Centrality	Eigen Vector Centrality	PageRank Centrality	The Proposed Algorithm
Karate Dataset	5	469.799	239.976	604.026	270.270	765.156	667.408	**3768.844**
	7	428.980	1000	750.469	800.533	800.453	387.346	**1828.335**
	10	766.058	385.604	850.851	583.819	1215.153	800.533	**2857.142**
	15	833.809	1111.729	873.362	836.820	1053.740	1125.703	**1320.607**
GrQc	15	46.835	122.304	109.947	60.094	78.872	100.203	**473.721**
	20	47.741	105.112	113.797	91.675	136.888	133.202	**384.102**
	25	77.450	146.546	134.826	127.111	98.685	161.984	**608.155**
	30	85.235	232.668	96.438	156.593	108.790	223.049	**253.014**
	50	128.645	362.196	158.701	367.209	125.583	312.487	**474.793**
	100	284.745	414.175	335.959	486.902	230.287	439.029	**938.436**
CondMat	15	8.833	104.110	85.623	60.124	45.937	60.626	**219.683**
	20	11.060	142.263	176.469	29.929	44.031	82.430	**184.198**
	25	11.801	155.393	88.828	44.154	65.680	99.662	**188.727**
	30	20.560	155.643	146.599	56.312	72.528	97.875	**205.557**
	50	31.068	230.277	288.663	93.278	90.840	133.243	**449.649**
	100	50.852	368.852	305.647	251.703	151.549	287.199	**448.375**
HepPh	15	32.787	**130.909**	108.728	32.031	28.081	94.097	70.091
	20	33.879	**156.295**	135.115	46.575	29.840	82.430	133.812
	25	31.300	220.763	**340.980**	81.056	37.367	133.622	76.595
	30	60.168	**204.082**	232.887	141.083	33.010	169.165	115.740
	50	81.725	335.888	**357.598**	141.434	61.595	243.797	277.777
	100	153.379	567.757	417.447	264.028	128.305	373.984	**1206.584**
AstroPh	15	6.737	88.519	45.482	37.500	43.631	88.112	**106.201**
	20	7.556	145.170	44.010	88.925	65.732	85.783	**243.394**
	25	15.469	**120.960**	93.033	40.076	73.381	78.782	89.907
	30	12.381	130.206	95.976	75.657	99.327	76.811	**235.536**
	50	21.079	203.811	149.045	98.188	110.951	129.723	**439.432**
	100	47.609	264.177	121.088	177.782	171.407	196.837	**535.156**
Facebook	15	22.875	241.388	199.439	118.369	72.811	204.737	**278.743**
	20	37.817	234.585	87.146	47.391	100.0958	204.300	**418.284**
	25	72.656	303.142	193.888	77.883	109.717	212.973	**399.869**
	30	71.469	287.847	127.688	80.690	116.436	241.212	**413.150**
	50	79.684	507.147	267.717	152.395	109.483	268.826	**779.422**
	100	204.824	521.310	262.996	230.854	115.095	420.097	**1136.233**

B. Evaluation criteria

In this paper, *Diffusion Speed* is taken into consideration to evaluate efficiency of the proposed seed selection algorithm. Diffusion Speed can be defined as the ratio of total influenced nodes to total time taken for diffusion. The proposed a novel solution called centrality based seed selection, is used to spread influence in network as far as possible within minimum time period. We have used various benchmark seed selection algorithms such as Random and all Centrality. Finally, we conducted a series of experiments on both linear threshold model and independent cascade model, to verify the proposed seed selection algorithm. The experimental results show that proposed seed selection algorithm works far better than various existing benchmark methods. The performance of proposed seed selection algorithms is also demonstrated in the experiments.

Table IV shows various experiments performed for different seed set and seed selection algorithms. For all other datatset the proposed entropy based algorithm gives better performance for Linear Threshold Model. Reason behind this performance is that entropy is representative of the strength node and it is made up of all Centrality. Entropy gives equal importance to all centrality measures, so each centrality measures get combined. More entropy represents more connectivity and more information diffusion occurs. So the proposed entropy based seed selection outperforms with Linear threshold model and Karate, GrQc, CondMat, AstroPh, HepPh and Facebook datasets. From the Table VI, it can be say that in Linear Threshold Model, the minimum average performance gain is with PageRank Centrality and HepPh dataset i.e. -4.18% and the maximum average performance gain is with Random and Facebook Dataset i.e. 92.31%

Table V shows various experiments performed for different seed set and seed selection algorithms. For all other datatset the proposed entropy based algorithm gives better performance for Independent Cascade Model. The proposed seed selection algorithm selects the nodes which have all optimal Centrality so that maximum diffusion can be achieved. More entropy represents more connectivity so it is more suitable for seed node which gives more information diffusion occurs. So, the proposed entropy based seed selection outperforms with Independent Cascade Model and Karate, GrQc, HepPh, CondMat, AstroPh, and Facebook datasets. From the Table VII, it can be say that in Independent Cascade Model, the minimum average performance gain is with Degree Centrality and HepPh dataset i.e. 14.09% and the maximum average performance gain is with Random and AstroPh Dataset i.e. 93.28%

TABLE VI. Average Performance Gain (in %) of the proposed algorithm with respect to the existing benchmark algorithms for various network data set using Linear Threshold Model.

Data Set	Random	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigen Vector Centrality	PageRank Centrality
Karate	67.56%	48.89%	25.67%	39.92%	33.84%	31.15%
GrQc	63.54%	37.39%	47.79%	50.79%	73.61%	31.06%
CondMat	77.02%	31.44%	37.67%	75.18%	15.96%	15.24%
HepPh	73.53%	33.58%	79.69%	76.73%	72.14%	-4.18%
AstroPh	77.25%	46.89%	27.13%	66.84%	67.11%	52.04%
Facebook	92.31%	65.72%	43.36%	79.52%	87.03%	48.09%

TABLE VII. Average Performance Gain (in %) of the proposed algorithm with respect to the existing benchmark algorithms for various network data set using Independent Cascade Model.

Data Set	Random	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigen Vector Centrality	PageRank Centrality
Karate	74.44%	71.99%	68.50%	74.51%	60.77%	69.50%
GrQc	78.59%	55.85%	69.68%	58.83%	75.13%	56.26%
CondMat	92.09%	31.82%	35.63%	68.43%	72.26%	55.13%
HepPh	79.09%	14.09%	15.31%	62.45%	83.08%	41.66%
AstroPh	93.28%	42.24%	66.74%	68.59%	65.78%	60.23%
Facebook	85.72%	38.83%	66.75%	79.34%	81.80%	54.69%

6. Conclusion

The proposed novel seed selection algorithm entitled as Seed Selection Algorithm using Centrality based Entropy. The idea of entropy function is microscopic effect of various centrality measures on nodes. So, it gives seed set which have optimal information diffusion speed. Higher the diffusion speed shows more faster the algorithm. We conduct a series of experiments on both linear threshold model and independent cascade model and can be concluded that for both, Linear Threshold Model and Independent Cascade Model the proposed algorithm is optimal faster than existing centrality based seed selection algorithms with Karate, GrQc, HepPh, CondMat, AstroPh, and Facebook Datasets. The performance gain (in %) shows how faster the proposed

seed selection algorithm performs. From the performance gain, it can be concluded that the proposed seed selection algorithm is faster than existing centrality based seed selection algorithms with Karate, GrQc, HepPh, CondMat, AstroPh, and Facebook Datasets and linear threshold model as well as independent cascade model.

References

[1] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing letters*, vol. 12, no. 3, pp. 211–223, 2001.

[2] M. Granovetter, "Threshold models of collective behavior," *American journal of sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

[3] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 137–146.

[4] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[5] S. Segarra and A. Ribeiro, "Stability and continuity of centrality measures in weighted graphs," *IEEE Transactions on Signal Processing*, vol. 64, no. 3, pp. 543–555, 2015.

[6] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[7] S. Agarwal and S. Mehta, "Effective influence estimation in twitter using temporal, profile, structural and interaction characteristics," *Information Processing & Management*, vol. 57, no. 6, pp. 102–321, 2020.

[8] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 57–66.

[9] D. Kempe, J. Kleinberg, and É. Tardos, "Influential nodes in a diffusion model for social networks," in *International Colloquium on Automata, Languages, and Programming*, 2005, pp. 1127–1138.

[10] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 420–429.

[11] W. Chen, L. V. Lakshmanan, and C. Castillo, "Information and influence propagation in social networks," *Synthesis Lectures on Data Management*, vol. 5, no. 4, pp. 1–177, 2013.

[12] A. Goyal, W. Lu, and L. V. Lakshmanan, "Celf++: optimizing the greedy algorithm for influence maximization in social networks," in *Proceedings of the 20th international conference companion on World wide web*, 2011, pp. 47–48.

[13] S. Cheng, H. Shen, J. Huang, G. Zhang, and X. Cheng, "Staticgreedy: solving the scalability-accuracy dilemma in influence maximization," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 509–518.

[14] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, 2014, pp. 946–957.

[15] H. Zhuang, Y. Sun, J. Tang, J. Zhang, and X. Sun, "Influence maximization in dynamic social networks," in *IEEE 13th International Conference on Data Mining (ICDM), 2013*, 2013, pp. 1313–1318.

[16] X. Chen, G. Song, X. He, and K. Xie, "On influential nodes tracking in dynamic social networks," in *Proceedings of the 2015 SIAM International Conference on Data Mining*, 2015, pp. 613–621.

[17] Y. Yang, X. Mao, J. Pei, and X. He, "Continuous influence maximization: What discounts should we offer to social network users?" in *Proceedings of the 2016 International Conference on Management of Data*, 2016, pp. 727–741.

[18] Z. Wang, E. Chen, Q. Liu, Y. Yang, Y. Ge, and B. Chang, "Maximizing the coverage of information propagation in social networks."

[19] B. Liu, G. Cong, D. Xu, and Y. Zeng, "Time constrained influence maximization in social networks," in *IEEE 12th International Conference on Data Mining (ICDM) 2012*, 2012, pp. 439–448.

[20] S. Wang and X. Tan, "Solving the robust influence maximization problem on multi-layer networks via a memetic algorithm," *Applied Soft Computing*, vol. 121, p. 108750, 2022.

[21] F. Kazemzadeh, A. A. Safaei, and M. Mirzarezaee, "Influence maximization in social networks using effective community detection," *Physica A: Statistical Mechanics and its Applications*, vol. 598, p. 127314, 2022.

[22] Q. He, L. Sun, X. Wang, Z. Wang, M. Huang, B. Yi, Y. Wang, and L. Ma, "Positive opinion maximization in signed social networks," *Information Sciences*, vol. 558, pp. 34–49, 2021.

[23] X. Zhang, S. Liu, and Y. Gong, "A new strategy in boosting information spread," *Entropy*, vol. 24, no. 4, p. 502, 2022.

[24] W. Li, P. Fan, P. Li, H. Wang, and Y. Pan, "An opinion spreading model in signed networks," *Modern Physics Letters B*, vol. 27, no. 12, p. 1350084, 2013.

[25] L. Fei, H. Mo, and Y. Deng, "A new method to identify influential nodes based on combining of existing centrality measures," *Modern Physics Letters B*, vol. 31, no. 26, p. 1750243, 2017.

[26] T. Zhang, B. Fang, and X. Liang, "A novel measure to identify influential nodes in complex networks based on network global efficiency," *Modern Physics Letters B*, vol. 29, no. 28, p. 1550168, 2015.

[27] P. Yang, X. Liu, and G. Xu, "A dynamic weighted topsis method for identifying influential nodes in complex networks," *Modern Physics Letters B*, p. 1850216, 2018.

[28] M. Newman, "Zachary's karate club Datasets," http://www-personal.umich.edu/mejn/netdata/, 2015.

[29] W. Hong, C. Qian, and K. Tang, "Efficient minimum cost seed selection with theoretical guarantees for competitive influence maximization," *IEEE Transactions on Cybernetics*, 2020.

[30] D. Mohapatra, A. Panda, D. Gouda, and S. S. Sahu, "A combined

approach for k-seed selection using modified independent cascade model," in *Computational Intelligence in Pattern Recognition*. Springer, 2020, pp. 775–782.

[31] A. Topîrceanu and M. Udrescu, "Fast colonization algorithm for seed selection in complex networks based on community detection," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2021, pp. 214–218.

[32] B. Nöldeke, E. Winter, and U. Grote, "Seed selection strategies for information diffusion in social networks: An agent-based model applied to rural zambia," *Journal of Artificial Societies and Social Simulation*, vol. 23, no. 4, 2020.

[33] G. Nie and M. Tang, "A multi-seed nodes selection strategy for influence maximization based on reinforcement learning algorithms," in *Journal of Physics: Conference Series*, vol. 1746, no. 1. IOP Publishing, 2021, p. 012045.

[34] B. Nettasinghe, V. Krishnamurthy, and K. Lerman, "Diffusion in social networks: Effects of monophilic contagion, friendship paradox, and reactive networks," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1121–1132, 2019.

[35] Z. Chen, "An agent-based model for information diffusion over online social networks," *Papers in Applied Geography*, vol. 5, no. 1-2, pp. 77–97, 2019.

[36] F. Wang, W. Jiang, G. Wang, and S. Guo, "Influence maximization by leveraging the crowdsensing data in information diffusion network," *Journal of Network and Computer Applications*, vol. 136, pp. 11–21, 2019.

[37] S. S. Singh, A. Kumar, K. Singh, and B. Biswas, "C2im: Community based context-aware influence maximization in social networks," *Physica a: Statistical mechanics and its applications*, vol. 514, pp. 796–818, 2019.

[38] K. Rabadiya and R. Patel, "Empirical analysis of various seed selection methods," in *International Conference on Information and Communication Technology for Intelligent Systems*. Springer, 2020, pp. 399–407.

[39] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceeding of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, 2009, pp. 199–208.

[40] A. Goyal, W. Lu, and L. V. Lakshmanan, "Simpath: An efficient algorithm for influence maximization under the linear threshold model," in *IEEE 11th International conference on data mining 2011*, 2011, pp. 211–220.

[41] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," in *European conference on principles of data mining and knowledge discovery*, 2006, pp. 259–271.

[42] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," http://snap.stanford.edu/data, Jun. 2014.

Kinjal Rabadiya She has received B. Tech. and M. Tech. degrees in 2015 from Uka Tarsadia University and 2017 from CHARUSAT university respectively. Her areas of interest include Social Network Analysis, Time series Analysis, fuzzy set, soft computing.

Dr.Ritesh Patel He has received his Doctorate degree in 2017 from CHARUSAT University. He has received B. Tech. and M. Tech. degrees in 2002 from North Gujarat University and 2004 from DDU respectively. His areas of interest include Cloud Computing, Internet of Things, Communication and Networking, Computer Architecture, Software Engineering and Cluster Computing. He is a member of Professional Societies CSI, IETF.