Sādhanā (2019)44:205
https://doi.org/10.1007/s12046-019-1189-7

Influence maximization in online social network using different centrality measures as seed node of information propagation

PARAMITA DEY, AGNEET CHATERJEE and SARBANI ROY*

Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
e-mail: dey.paramita77@gmail.com; agneet257@gmail.com; sarbani.roy@jadavpuruniversity.in

MS received 22 March 2019; revised 15 June 2019; accepted 29 July 2019

Abstract. Information propagation in the network is probabilistic in nature; simultaneously, it depends on the connecting paths of the propagation. Selection of seed nodes plays an important role in determining the levels and depth of the contagion in the network. This paper presents a comparative study when seed nodes for information propagation are selected through the properties of different centrality measures in the social network. This study captures the interaction measures of nodes in the social network, selects seed nodes based on five centrality measures, i.e. degree distribution, betweenness centrality, closeness centrality, Eigenvector and PageRank, and compares the affected nodes and levels of propagation within the network. We demonstrate the performance of the different centrality measures by processing three datasets of social network: Twitter network, Bitcoin network and author collaborative network. For the propagation of the information, we use breadth-first search (BFS) and susceptible–infectious–recovered (SIR) model and a detailed comparative study is also presented for each of the seed nodes selected using aforementioned network properties. Results show that the Eigenvector centrality and PageRank centrality measures outperform other centrality measures in all test cases in terms of propagation level and affected nodes during information propagation. Both Eigenvector and PageRank network data processing required a high computational overhead. For this reason we propose a hybrid model where using *k*-core the network is degenerated into a smaller network and centrality nodes are extracted from the smaller network. These centrality nodes, as compared to original centrality nodes, perform almost in the same manner in terms of influence maximization when *k* is chosen in a rational way.

Keywords. Online social network; influence maximization; centrality measures; eigenvector; pageRank; seed node; information propagation; degenerative network.

1. Introduction

Finding contagion in a network is an interesting problem, especially in large networks. Different nodes exhibit different roles in the network, such as being active (spreader) or neutral (ignorant) [1]. As the propagation is dependent on the topology of the network, some nodes get advantages for the impact of contagion. For example, the nodes that act as a hub in the network: it is always advantageous for that node to propagate the information [2]. Similarly, nodes with maximum connections seem to play a pivotal role, but the results may not be impressive if it is on the boundary of the network or generates a local maximum.

Viral marketing through social media has now become imperative for the increasing use of social network as the medium of communication. Competitive companies are very eager to find out the probable markets of interested target customers so that they can maximize the awareness of a new product and increase their revenue by increasing sales in the market [3]. Similarly, to influence people for some social or political interest is also an useful aspect of social media [4]. Social emotion plays an important role for influencing groups of people [5]. Study of sentiment analysis for short texts generated from online review summaries can be used for promotional purposes [6]. Similarly, significant social issues can be identified through social media [7, 8]. Now, increasing the awareness among the people or campaigning it is required to reach maximum number of users in the social network.

This paper studies propagation of information, based on the idea of selection of influential nodes as seed nodes of the propagation. Influential nodes in the network are decided using network property called centrality. We have considered five most popular measures of centrality, namely betweenness centrality, closeness centrality, degree distribution, Eigenvector and PageRank for the selection of seed nodes. Now, information is propagated from these nodes, i.e., these seed nodes are considered as target nodes for disseminating information or promoting new products

*For correspondence

Published online: 05 September 2019

in the social networks. This paper also describes how the seed nodes in all five cases can possibly affect the maximum number of nodes in the network. For propagation we use breadth-first search (BFS), which is a simplistic model, and propagate with a uniform probability to the connected neighbours. We use another model SIR (susceptible-infectious-recovered), which is probabilistic in nature and more symmetric with real life network where probability of participation is non-uniform.

This paper also highlights the importance of different modes of interactions between nodes and presents a model that captures the interaction measures of nodes in the social network and implements it on Twitter dataset. A detailed comparative study of the seed nodes has been made for Twitter, author collaborative network and Bitcoin network, to analyse the performance of information spreading with respect to these five different social network metrics. As centrality measures derivation involves large computational complexity, a degenerating model using *k*-mean is used for reducing the network and we find that working of new centrality measures is comparable to that of the original centrality measure in spreading contagion in the network. A large scale social network can be visualized and processed using modern graph analytical tools like Gephi, Pig, Hadoop, Map-Reduce, Spark, Hive, Base and Cassandra. The proposed information propagation algorithm has been implemented using the Apache Spark framework [9, 10]. It is an open source framework for processing a large scale graph, popular for its efficiency.

Rest of the paper is organized as follows. Related works are discussed in section 2. Section 3 presents the proposed approach of information propagation. Results and analysis are demonstrated in section 4. Centrality measures using *k*-core degenerating model is discussed in section 5. Finally, this paper is concluded in section 6.

2. Related work

Depending on the types of information, it can have a positive or negative impact on the society. Information about the possibility of natural disasters, terrorist attacks, disease or even natural calamity can spread easily to a large scale population very quickly through social interactions, which happen in social networks like Facebook, Twitter, etc. [11].

It is important to study not only the types of information and its transmission, but also its probable consequences on the society or the community. Community detection of targeted groups of people is also important in this context [12, 13]. As some research work suggests, some information spreading models behave like an independent cascade model. At the same time, anti-spread models that behave like a delayed start model (similar to a spread model with a time delay) or a beacon model (where some nodes start an opposite information after detecting a wrong campaign) also play an important role to resist information that is harmful in some aspect [1]. Some results show that seed nodes identified by *k*-core resist information spreading and act as a sink instead of a source [14, 15]. In [16], the authors highlight the dynamical strength of social ties in information spreading.

The concept of viral marketing [17] has become popular in this competitive era. For example, if a company has launched a new product and wants to promote this product in the market, the aim is to find out the most influential target customers from a group of people and increase awareness and sales by spreading information through the network via target customers.

On the other hand, to increase disease awareness in the society, it is very important to target those people in the network who have some social influence on the society, are proactive in nature and have considerable knowledge of the disease spread. These types of people play an important role in controlling the outbreak of the disease by increasing awareness in a large scale population. For both the cases we need to maximize the influence. The influence maximization problem is discussed in [18] to select a specific set of *N* nodes in a social network so that the total number of affected or influenced nodes can be maximized (*N*-maximization problem). The nodes having a high degree and a high clustering coefficient get preference over the low degree, low clustered nodes in the network for spreading information [19]. Moreover, for influence minimization problem the aim is to select lazy nodes in the network, which are very silent in nature and have the capability of blocking information propagation in the network [20]. However, it has not been established whether the seed nodes identified by *k*-core or node degree centrality play an important role for propagation of data through the network. Unlike the existing work, this paper is focused on the effect of centrality measures as seed node in information propagation.

In this work, we use five measures of centrality and find that Eigenvector centrality can be a good measure to identify seed nodes when the propagation of information is important. In existing approaches, interactions among the nodes are not considered. This paper presents an approach for modelling the bonding between nodes in Twitter network, which is determined in terms of interactions among them and this interaction is represented in Twitter dataset. Later this interaction measure is used to determine the weight of path for closeness and betweenness centrality measures of nodes in the network. Moreover, in the later section of the paper, we derive that the centrality measures from degenerative network also can be comparable to centrality measures acquired from the original large network, which can cost less time complexity but are capable of being important spreaders. For the degenerating network, *k*-core methodology is used in the paper.

3. Proposed approach for information propagation

3.1 *Problem statement*

For a given graph $G(V, E, w)$ where V represents set of nodes in the network, E represents set of edges in the network and w denotes the weight of the each such edge, we aim to analyse the flow of information in the network. For this purpose, we use the five most widely used centrality measures to generate root nodes for information propagation in the network and study their comparative influence both in terms of the level reached and the number of affected nodes.

3.2 *Determination of path vector through interaction classification*

Interactions in social networks play a critical part, enabling smooth spread of data, information and statistics. Websites such as Facebook and Twitter capture this essence by allowing people to connect in a variety of ways. To capture this bonding between nodes in the social network, we consider different kinds of interactions. Furthermore, two users almost always have different kinds and frequencies of interactions, compared with another pair of users. This essentially translates such an interaction network into a weighted directed graph, where the cost/weight of interaction plays an important role. Such a graph is directed, because interactions might not be reciprocated back between a pair of users. Take for example, the relationship between a user and his followers (in a Twitter-like network). The followers may retweet most of the user's tweets but the same cannot be guaranteed the other way round. Hence, the concept of a directed graph has been applied here keeping in mind this context.

For two nodes p and q in Twitter network, we consider the count of retweet, favourite and like, mentioned as interactions between the two users. Therefore, bonding between node p and q can be expressed as $IC_{pq} = w_{retweet}A_{pq_{retweet}} + w_{favourite}A_{pq_{favourite}} + w_{like}A_{xy_{like}}$ where $w_{retweet}$, $w_{favourite}$ and w_{like} are the respective weights allocated to each interaction. For the given network G, IC_xy is the strength of the bond in the $x \rightarrow y$ direction. The weight of $(w_{retweet} + w_{favourite} + w_{like})$ must add to 1 to ensure that all interactions are amalgamated into one entity, for each directed pair of user. The reciprocal of IC_xy determines the cost to the edge between x and y. Thus, edge with a low cost value represents a stronger bonding between the pair of nodes. In the computation of shortest paths of all pairs for closeness and betweenness centrality measures, the cost of the edges will play an important role in reflecting the interactions.

3.3 *Centrality measures*

For a given directed graph $G(V, E)$, where V represents set of nodes in the network and E represents set of edges in the network, centrality measures are used to identify the significant or active nodes in the network [21]. Centrality measures enable us to find source nodes from where we can emanate information successfully.

3.3a *Degree centrality*: Node degree [22] information deals with the number of connections among the nodes in the network. The degree centrality for a node $n \in V$ is the fraction of nodes it is connected to and is denoted by deg_n. High-degree nodes are connected with maximum number of other nodes in the social network. A node is important if it has many neighbours and can be interpreted to have a high probability of catching whatever is flowing through the network or information, in our context. Calculating degree centrality for all the nodes in a graph takes $O(V^2)$ in a dense adjacency matrix representation.

3.3b *Closeness centrality*: For a node $n \in V$, closeness centrality of a node is the reciprocal of the sum of the shortest path distances from v to all other nodes. The closeness centrality of a node [22] is defined as $Cl_v = \frac{n-1}{\sum_u d(u,v)}$, where $d(u, v)$ is the shortest distance between u and v, and n is the number of nodes in the network, which represents the normalized form of closeness centrality. As we work with weighted networks throughout the course of this work, it is noteworthy to mention that all the shortest path distances are computed using the Dijkstra's algorithm, keeping in consideration the weights associated with each edge.

3.3c *Betweenness centrality*: The betweenness centrality [22] of a node quantifies the number of times a node appears in the shortest path between two other nodes. A node that has a high betweenness centrality is capable of transferring rumours in the network, as it can control the flow of information passing amongst different nodes in the network. A vertex lying between two sets of nodes might have a less number of neighbours and/or might be far away from certain nodes in each group but has a high betweenness centrality as all paths between the two entities must pass through this node to facilitate communication. This further elucidates the importance of nodes with a high betweenness centrality. To calculate the betweenness centrality of a node, all source shortest paths are determined for every combination of nodes δ_{xy} in the network. For every ordered pair of vertices (x, y), the number of shortest paths that pass through the vertex v is denoted by δ^v_{xy}. The betweenness of a node is defined as $B_v = \sum_{x\neq y\neq v\in V}\frac{\delta^v_{xy}}{\delta_{xy}}$.

3.3d *PageRank*: PageRank [23] is a link analysis algorithm developed by Lary Page, one of the founder members of Google, and it assigns a weighted sum to each node in a graph, with the purpose of measuring its relative importance within the network. The PageRank of a node $n \in V$ is defined as $prank(n) = \frac{1-d}{|V|} + d\sum_{v:(v,n)\in E} \frac{prank(u)}{|\{v:(v,u)\in E\}|}$, where d is a dampening factor.

3.3e *Eigenvector*: A node can be more central if it is in relation with other entities that are themselves central. The centrality of some node depends not only on the number of its adjacent nodes but also on their value of centrality. Eigenvector centrality of a node is its summed connections to others weighed by its centrality. The centrality c of a node i belonging to an adjacency matrix M is given as $\lambda e_i = \sum M_{ij}e_j$, where λ is the associated Eigenvalue of the centrality. In general, the largest Eigenvalue is preferred as the associated Eigenvalue measures the accuracy with which it can reproduce M.

3.4 *Information propagation*

Information can spread very fast through social networks. In this work, to simulate the propagation of information, we use the BFS and SIR algorithms.

3.4a *BFS*: Given a graph G, and a prominent source node r, BFS explores all edges of G with a fixed probability to discover every vertex that is reachable from the source node. BFS produces a breadth-first tree with root r that contains all reachable vertices if the probability is considered as 1. The vertices in each level of the tree compose a frontier. Frontier propagation checks every neighbour of a frontier vertex to see whether it is visited already; if not, the neighbour is added into the new frontier. The time complexity of the BFS algorithm is $O(V + E)$.

After establishing the root nodes from the centrality measures discussed in section 3.3, we aim to explore the variance in propagation of information. This is established by feeding in the root nodes, from each centrality, to the BFS algorithm, and comparing them on the basis of their effect, both in terms of level and affected nodes.

3.4b *SIR model*: Three-state susceptible–infected–refractory model (SIR) [24] is a probabilistic method for information propagation. Let there be some information to be spread in the network. The three status of the node will be: the node who has knowledge about the information and wishes to spread it is in the infected status, the one who is not aware about the information is in susceptible status and the one who was aware about the information but is no longer interested in spreading it is in the refractory status. In a detailed spreading process on network, the propagation can be implemented by randomly choosing a node to hold the information at the beginning and assuming that only the neighbours of the node with the information have a probability to contact the information. As time goes on, information can infect the susceptible nodes that are connected to the node with the information and make them infected. The propagation process is over when there is no infected node in the network. Similar to BFS propagation, this algorithm takes the value of the level K and the seed node as the input and generates the list of visited nodes, i.e. nodes affected by the propagation up to level K.

3.5 *Proposed framework*

In figure 1, we show our proposed information propagation framework for the comparative study of different centrality measures as seed node. It has two main steps, which are discussed in following subsections.

3.5a *Step 1. seed node selection*: In this initial step, we find out five seed nodes based on the centrality measures, which have been discussed to be good measures for the influence maximization property of a network. We select one node with the highest value for each centrality, as the seed/root node for information propagation. Let degree distribution N_{degree}, closeness centrality N_{close}, betweenness centrality N_{bet}, Eigenvector centrality N_{ev} and PageRank centrality N_{prank} represent the node with maximum centrality. Let D_{max}, Cl_{max}, B_{max}, E_{max} and P_{max} represent the metric value of these corresponding nodes. Nodes selected as seeds have the maximum value of the said network properties as shown in figure 1. Closeness and betweenness centrality depend on all possible shortest paths of the nodes in the network. Unlike closeness and betweenness centrality, degree centrality, does not depend on the cost/weight of an edge, as it is purely a metric to measure connections. Eigenvector and PageRank are dependent on the Eigenmatrix of the nodes in the network. Table 1 lists the relationship matrix.

3.5b *Step 2. information spreading*: To evaluate the efficacy of the afore-selected nodes in spreading information, BFS and susceptible–infectious–recovered algorithms are performed, considering them as the root node of the graphs. Initially it explores the neighbour of the seed and moves to the next-level neighbour. Let us consider K to be the maximum hop or depth count in the social network up to which information can be propagated. This value of K also depends on the seed node from which propagation is initiated. Let $K_{DK}, K_{CLK}, K_{BK}, K_{EK}$ and K_{PRK} denote the maximum level up to which information is propagated for different centrality measures, as shown in table 1. The corresponding number of possible affected neighbour nodes at maximum level by the seed node is denoted by C_{DK}, C_{CLK}, C_{BK}, C_{EK} and C_{PRK}, also shown in table 1.

4. Results and analysis

This section compares different sets of networks and how they behave depending on their network characterization and property.

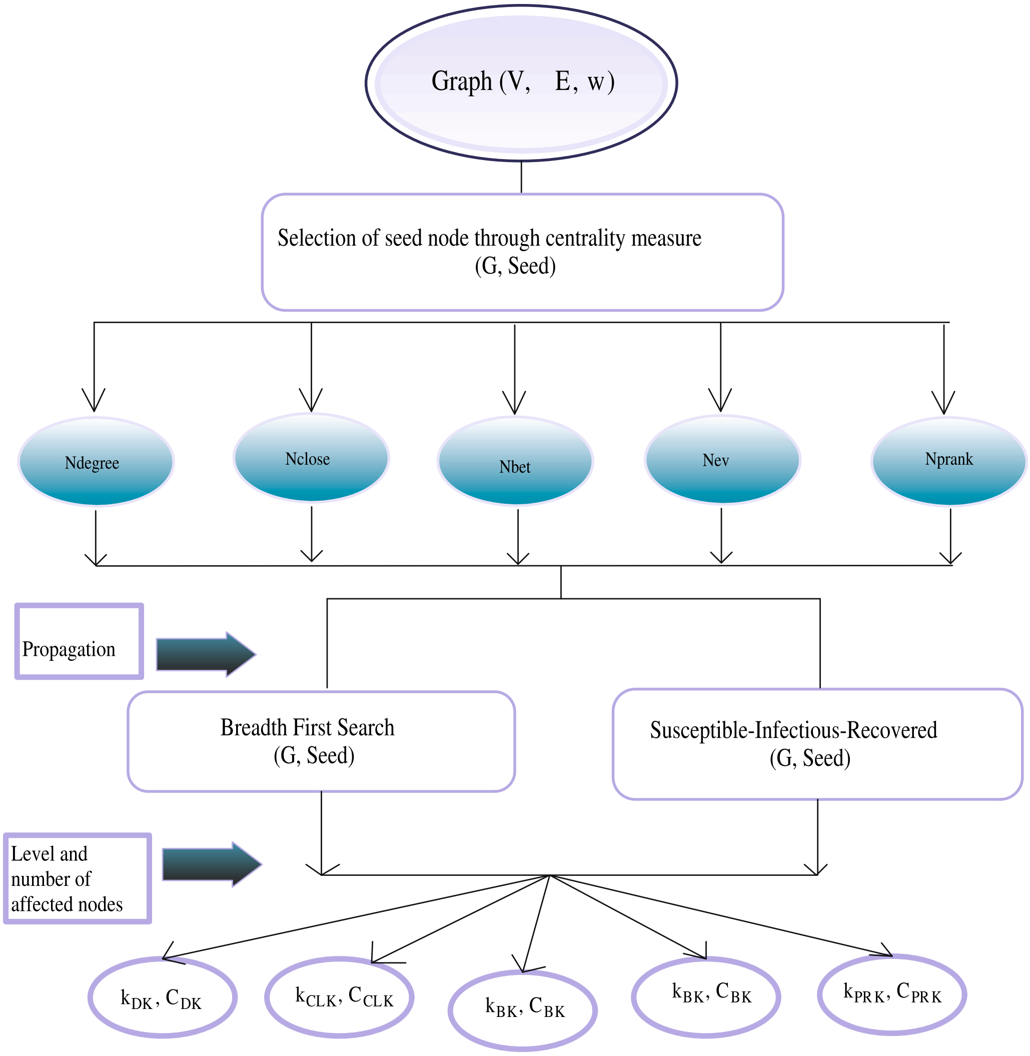
Graph (V, E, w)
Selection of seed node through centrality measure (G, Seed)
Ndegree
Nclose
Nbet
Nev
Nprank
Propagation
Breadth First Search (G, Seed)
Susceptible-Infectious-Recovered (G, Seed)
Level and number of affected nodes
k_DK, C_DK
k_CLK, C_CLK
k_BK, C_BK
k_BK, C_BK
k_PRK, C_PRK

Figure 1. Workflow of the information propagation through social network.

Table 1. Relationship matrix.

Network property	Seed node	Metric value	Maximum level and affected neighbours
Degree distribution	N_{degree}	$D_{max} = \max_{\forall v \in V}\{deg_v\}$	K_{DK}, C_{DK}
Closeness centrality	N_{close}	$Cl_{max} = \max_{v \in V}\{Cl_v\}$	K_{CLK}, C_{CLK}
Betweenness centrality	N_{bet}	$B_{max} = \max_{\forall v \in V}\{B_v\}$	K_{BK}, C_{BK}
Eigenvector	N_{ev}	$E_{max} = \max_{\forall v \in V}\{E_v\}$	K_{EK}, C_{EK}
PageRank	N_{prank}	$P_{max} = \max_{\forall v \in V}\{P_v\}$	K_{PRK}, C_{PRK}

4.1 *Twitter data*

We make use of the Higgs Twitter Dataset [25], which has been developed after monitoring the spreading process on Twitter before, during and after the announcement of the discovery of a new particle with the features of the Higgs boson. The network was sampled to 6000 nodes and 26402 edges, with an average clustering co-efficient of 0.34. It provides us with three interactions, namely retweets, favourite and like. There is a minimum of one and a maximum of three edges between two users, depending on the kind of interactions they have. To aggregate IC_{pq}

between two nodes, we set $w_{retweet}$ as 0.6, $w_{favourite} = w_{like} = 0.2$. These weights are set in accordance with the fact that retweets add more to bonding and information propagation.

4.1a *Selection of node with highest centrality in Twitter network*: Figure 2 presents the distribution function of different centrality measures. Node ID 9 has the highest degree distribution value of 0.166; hence, it is selected as N_{degree} for the twitter network. Edge costs, which are influenced by the interactions in the network, play an important part in the shortest path calculation involved in determining closeness centrality of nodes. Node ID 27 has the highest value of 0.332 and is selected as seed node N_{close} for this network. Node ID 1791 has the highest betweenness centrality value of 0.172, and is selected as N_{bet} for the next step of information propagation. Node ID 327 with Eigenvector centrality value as 0.1917 has the maximum value and emerges as N_{ev} for this network. Node ID 315 with centrality value as 0.011256 has the maximum PageRank value and emerges as N_{prank} for this network.

4.1b *Information propagation in Twitter network*: Seed nodes were selected using the five centrality measures mentioned in the previous subsections. Information propagation was then performed from each of the selected seed nodes, and a comparative study conducted for BFS and SIR propagation is shown in figures 3 and 4, respectively. At each level K for each seed node, the number of affected nodes is measured. As mentioned in earlier subsection, in our network, node 1791 is generated with the highest betweenness centrality whereas node 27 and node 9 are, respectively, the nodes with maximum closeness centrality and maximum node degree distribution. Similarly, node 327 and node 315 are, respectively, the nodes with highest Eigenvector and PageRank centrality. For comparative study, we select a random node 832. These six nodes are considered as the source node of the information propagation.

For BFS propagation, for node degree the maximum level of propagation is 14 and affected nodes are 1418, whereas for other centralities it reaches the maximum of 13th level of propagation. Closeness and betweenness centrality propagation reaches up to 1529 and 1607 nodes,

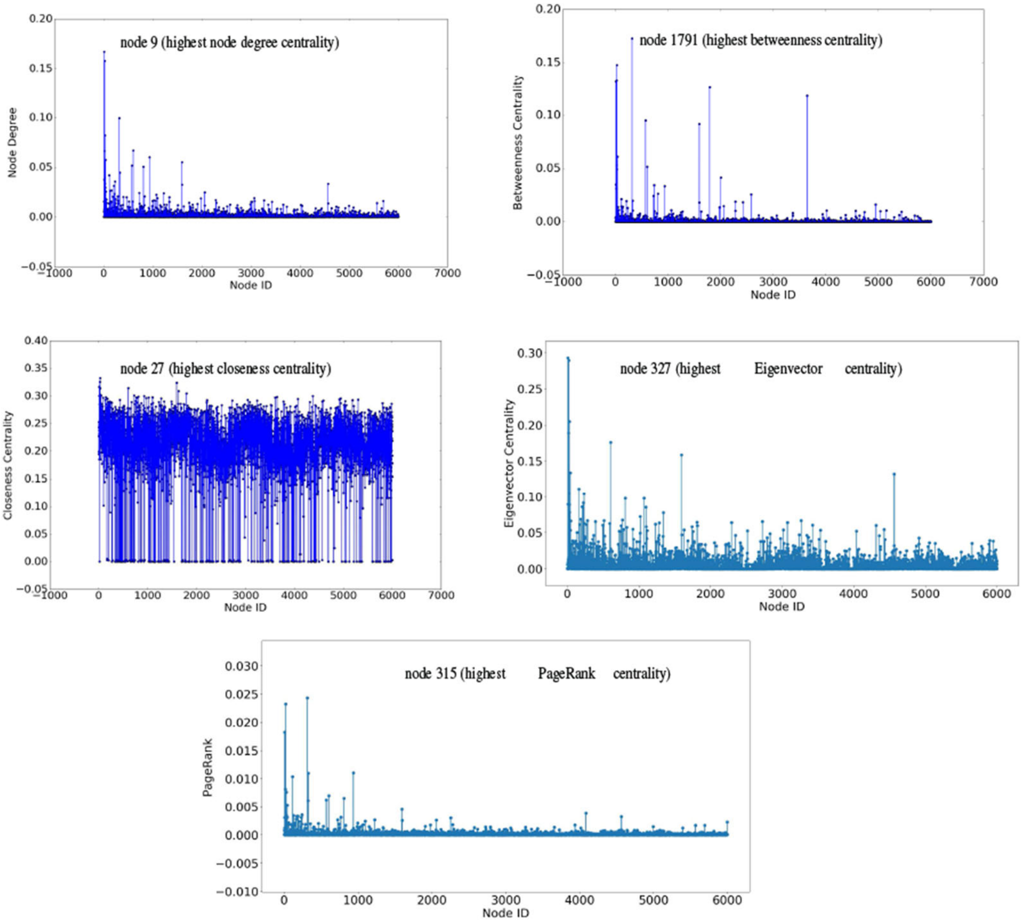

Figure 2. Distribution of different centrality measures in Twitter network.

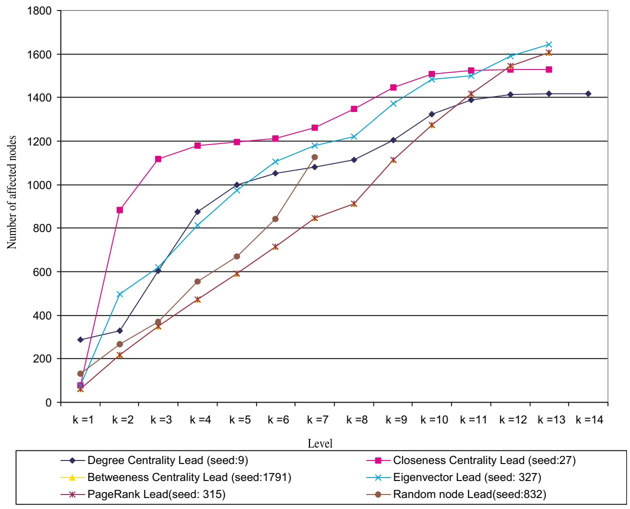

Figure 3. Propagation in Twitter network using BFS algorithm.

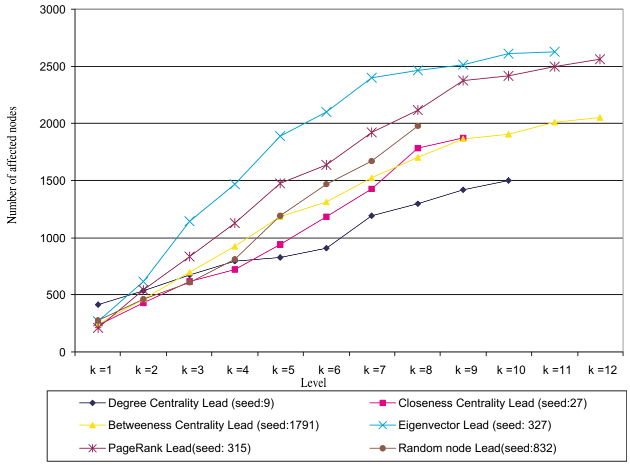

Figure 4. Propagation in Twitter network using SIR algorithm.

respectively. For Eigenvector and PageRank centrality, the number of affected nodes is 1645 and 1607, respectively. For the randomly selected node, the propagation reaches the maximum up to 7th level and affected nodes are 1124, which is very less compared with other centrality nodes. For SIR propagation, for node degree the maximum level of propagation is 10 and affected nodes are 1501. For closeness centrality and betweenness centrality, propagation reaches up to 1874 and 2054 nodes and 9th and 12th level, respectively. For Eigenvector and PageRank centrality, the number of affected nodes is 2623 in 11th level and 2564 in 12th level, respectively. For the randomly selected node, the propagation reaches the maximum up to 8th level and affected nodes are 1978. It can be noted that in SIR model, which is a probabilistic model, affected nodes are much less for random node as compared with other centrality nodes. The corresponding propagation measures are shown in figures 3 and 4.

4.2 *Coauthor network*

We have collected data of coauthor network [26], where total number of nodes in the input graph is 6234 and total number of edges is 63310, with an average clustering coefficient of 0.257. As discussed in the previous section, five nodes with highest centrality measures and one random node are selected as seed nodes and propagation through these selected nodes are considered for comparative study.

4.2a *Selection of node with highest centrality in coauthor network*: From the normalized degree distribution graph, node ID 4163 is selected as N_{degree}, which has the highest degree and will be used as the seed node, in the next step. Node ID 772 has the highest closeness centrality value, leading to its being selected as N_{close}. The shortest path calculations for betweenness centrality are made similar to the calculations made for closeness centrality as in earlier subsection. From normalized betweenness centrality distribution of the network, node ID 929 with the highest betweenness centrality value is selected as N_{bet}. Node ID 3914 with Eigenvector centrality value as 0.1917 has the maximum value and emerges as N_{ev} for this network. Node ID 219 with PageRank centrality value as 0.011256 has the maximum value and emerges as N_{prank} for this network.

4.2b *Propagation of information in coauthor network*: To simulate the information propagation in the network, we use the BFS method, considering seed node as the root. The comparison of these cases in terms of reachability and affected nodes has been done. In BFS, when node degree centrality is selected as seed node, the maximum level of propagation is 11 and affected nodes are 3410, whereas for closeness it reaches the maximum of 9th level with 3841 affected nodes. For betweenness centrality, propagation reaches up to 9th level and affected 3918 nodes. For Eigenvector, maximum affected nodes are 4157 at 10th level. For PageRank centrality, 4085 nodes are affected at maximum at 8th level. For a randomly selected node with node ID 832, maximum affected nodes are 1347 at 7th level. A comparative chart for all propagations through BFS is shown in figure 5.

For SIR propagation, for node degree the maximum level of propagation is 11 and affected nodes are 3891. For closeness centrality and betweenness centrality, propagation reaches up to 4032 and 4080 nodes in 11th level, respectively. For Eigenvector and PageRank centrality, the number of affected nodes is 3017 in 12th level and 2514 in 12th level, respectively. For the randomly selected node, the propagation reaches a maximum up to 6th level and affected nodes are 1070. This comparative study is shown in figure 6.

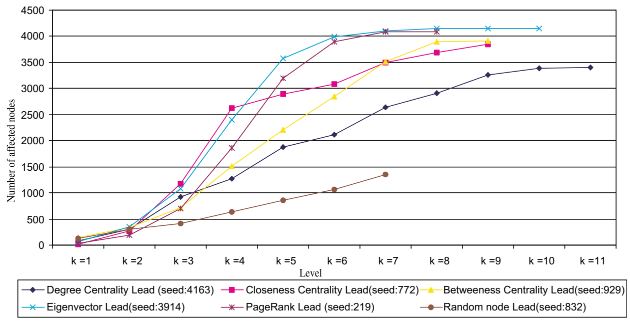

Figure 5. Propagation in author collaborative network using BFS propagation.

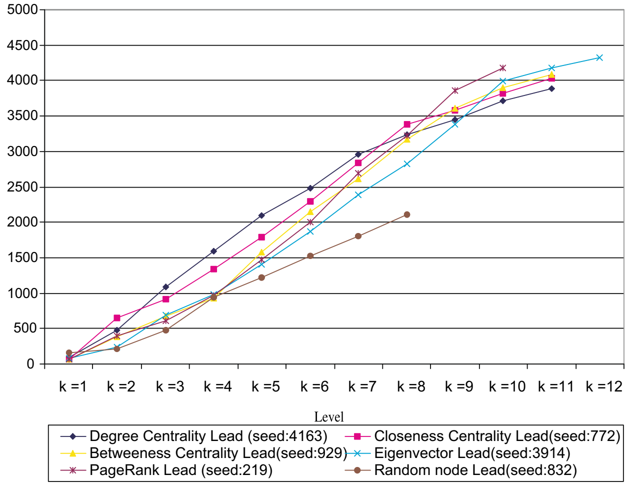

Figure 6. Propagation in author collaborative network using SIR propagation.

4.3 *Bitcoin network*

We also use a Bitcoin Trust Network [27], where each interaction is modelled as a factor of trust. It is a weighted, directed network, where each weight defines how much a user, trusts another user in the network, in a range of [–10,10]. We consider only positive rating to ensure snap. Information propagation plays a crucial part in such a network, as the extent of true information is imperative when it comes to investing in a cryptocurrency such as Bitcoin. This network has 5581 nodes and 34,294 edges, with an average clustering coefficient of 0.1785. Here, IC_{pq} signifies the bond between two nodes, indicating how much p trusts q. Next, we move on to validate our approach on this network.

4.3a *Selection of node with highest centrality in Bitcoin network*: Node ID 1810 has the maximum node degree of 0.0746, and results in being selected as N_{degree} for the next step of information propagation. N_{close} for this network is the node with ID 2388 having a centrality value of 0.2611. Node ID 3129 with centrality value as 0.0914 has the maximum value and emerges as N_{bet} for Bitcoin network. Node ID 905 with centrality value as 0.1917 has the maximum Eigenvector value and emerges as N_{ev} for this network whereas node ID 2125 with centrality value as 0.011256 has the maximum PageRank value and emerges as N_{prank} for the network.

4.3b *Information propagation in Bitcoin network*: Information propagations from three centrality nodes are performed from each of the selected seed nodes and a comparative study is conducted. At each level K for each seed node, the number of affected nodes is measured. Five seed nodes 1810, 2388, 3129, 905 and 2125 are considered as the source node of the information propagation. Moreover one random node is selected as 225. In BFS propagation for node degree the maximum level of propagation is 11 and affected nodes are 975, whereas for closeness it reaches the maximum of 17th level with 1150 affected nodes. For betweenness centrality, propagation reaches up to 26th level and affects 1315 nodes. For Eigenvector centrality in 20th level, 1487 nodes are affected at the maximum. For PageRank centrality, the maximum level of propagation is 16 with 1326 nodes affected as shown in figure 7. It can be noted that for randomly selected node, the number of affected nodes is 602 at the maximum of 7th level.

In SIR propagation for node degree the maximum level of propagation is 10 and affected nodes are 1724, whereas for closeness it reaches maximum of 13th level with 2109 affected nodes. For betweenness centrality, propagation reaches up to 15th level and affects 2462 nodes. For both Eigenvector and PageRank, maximum level is 12 and number of affected nodes is, respectively, 3017 and 2514 as shown in figure 8. For randomly selected node at the maximum, affected nodes are 1070 at 6th level.

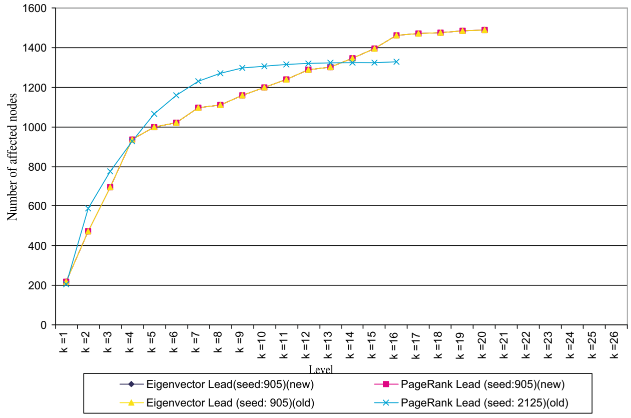

Figure 7. Propagation in Bitcoin network using BFS propagation.

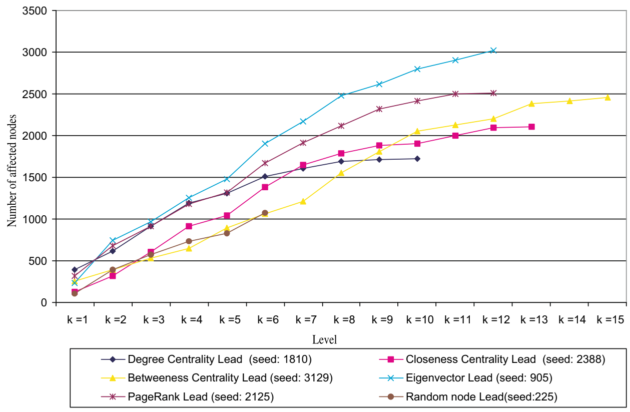

Figure 8. Propagation in Bitcoin network using SIR propagation.

5. Degenerative network

Degenerative network is a representative network of original network that reduces the sparseness of the network. It can be mentioned as representative as graph properties like centrality or clustering coefficient remain unchanged or change minimally after the decomposition. We consider k-core for generating a degenerative network. Figure 9 presents the information propagation when seed node is selected through degenerative network using k-core.

The k-core remains after the removal of vertices with degree less than k and edges incident to them recursively from G until no vertex has degree less than k. The $(k + 1)$-core can be computed iteratively from the k-core as the $(k + 1)$-core is a subgraph of the k-core. Likewise, by computing k-core sequentially from $k = 1$ to $k = k_{max}$, we divide all vertices according to their k-core value. This process, called core decomposition, has time complexity of O(V+E), where V and E represent the vertices and edges, respectively.

In our analysis, we find that apart from node degree centrality, all centrality measures are involved with high computational complexity. It can be noted that for computation of shortest paths of all pairs, time complexity for a graph with n number of nodes is on the order of O(n^2). Thus, for betweenness and closeness centrality, which involve computation of all possible shortest paths, time complexity becomes high for large scale network. Similarly, the Eigenvalue computation of each node involves time complexity of O(n^2). Table 2 presents all centrality measures along with the time complexity.

For reducing the computational time involved to find out the influential spreader, we consider a degenerative network G' after k-core decomposition where k = average node degree of graph + standard deviation of node degree. For the selection of k, it is to be considered that the maximum number of nodes remains in the range of average node degree of graph ± standard deviation of node degree. Now, to reduce the graph, the rule should be follow that without affecting average node degree centrality and clustering coefficient, we have to reduce significant number of nodes in the network and thus the value of k is determined empirically in the experiment.

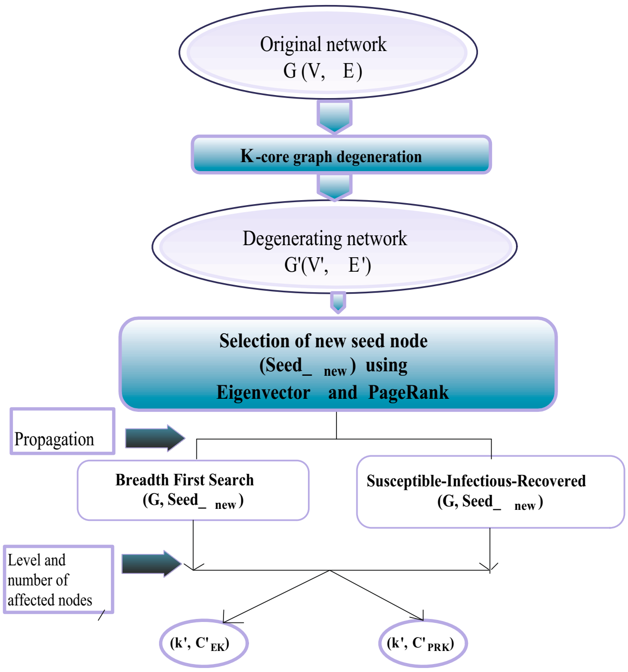

Figure 9. Workflow of the information propagation with seed node selected through degenerative network.

Table 2. Time complexity.

Network property	Time complexity
Degree distribution	O(n)
Closeness centrality	O(n^2)
Betweenness centrality	O(n^3)
Eigenvector	O(n^2)
PageRank	O(n^2)

Hypothesis: Let $G'(V', E')$ be the degeneracy-core of graph G. Then, for each vertex v in V', centrality in G is defined as v 's centrality in G'.

After k-core degeneration, number of nodes is 1115 and number of edges is 12357 for Twitter network (table 3). For collaborative author network, number of nodes is 1273 and number of edges is 13817. After k-core degeneration, number of nodes is 919 and number of edges is 10577 for Bitcoin network. Among these centrality measures, Eigenvector centrality and PageRank centrality are considered since these two centrality measures give better results for seed node in our previous subsection. Distribution functions for both the centrality measures are shown in figure 10. Figures 11–16 present comparative study of

Table 3. Number of nodes in original and degenerative network.

	Original n/w	Degenerative n/w
Twitter network	6000	1115
Coauthor network	6234	1273
Bitcoin network	5581	919

propagation in terms of affected nodes in three networks for BFS and SIR propagation.

Figures 11 and 12 present the propagation of information in the Twitter network for Eigenvector and PageRank centrality when these nodes are simulated through BFS and SIR nodes, respectively. Here the horizontal axis represents level of propagation from seed node and vertical axis represents the number of affected nodes up to that level. Eigenvector Lead (old) and Eigenvector Lead (new) represent the nodes with highest Eigenvector centrality in original and degenerated graph, respectively. Similarly,

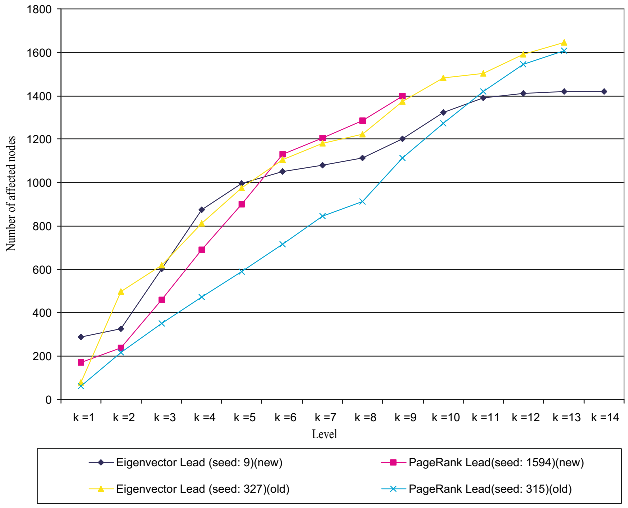

Figure 11. Propagation of information in Twitter network using BFS propagation.

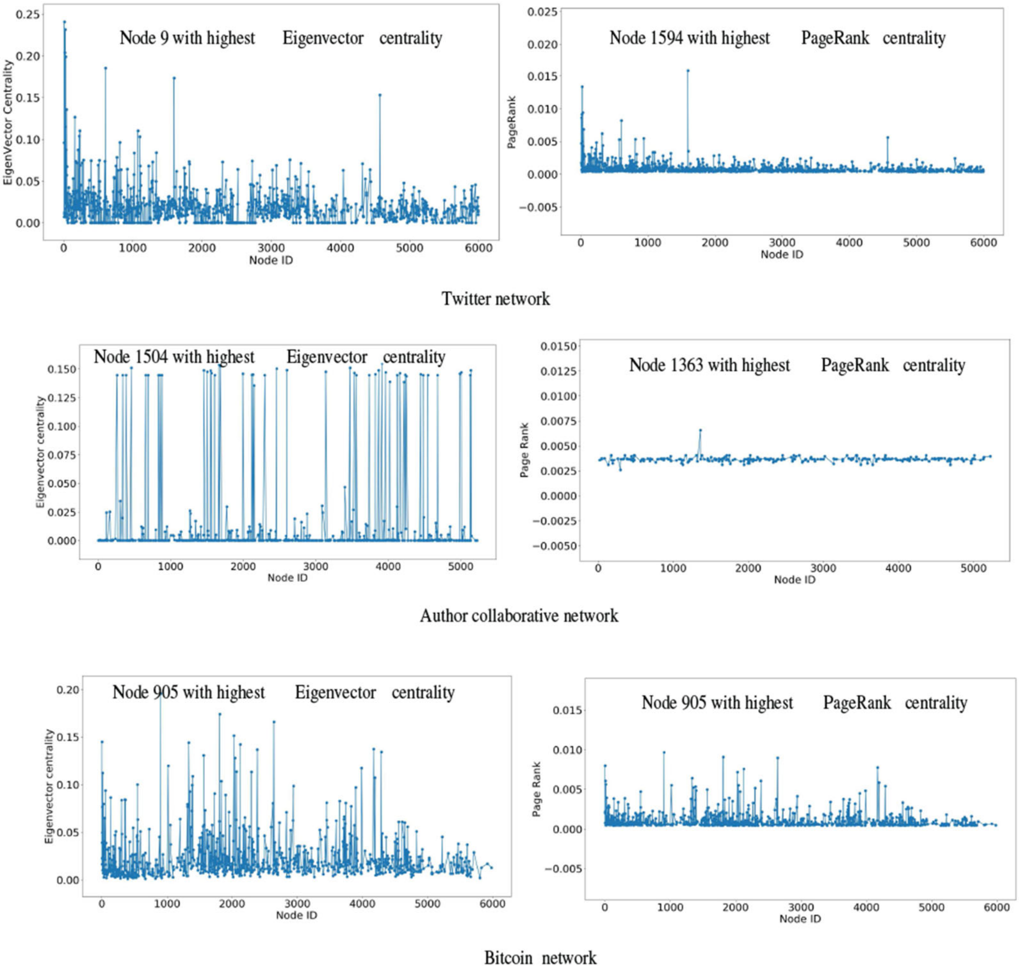

Figure 10. Eigenvector and PageRank distribution after *k*-core decomposition.

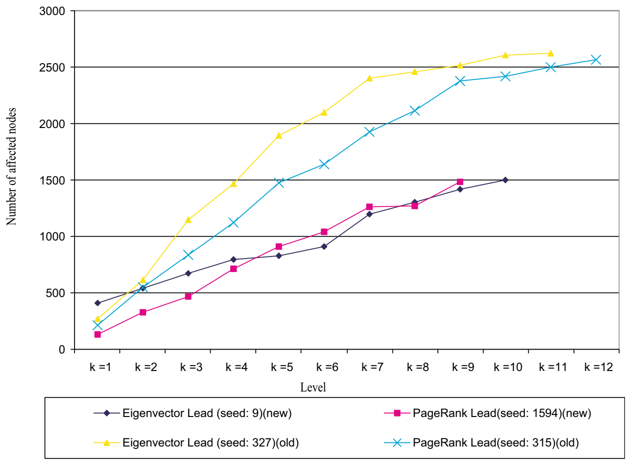

Figure 12. Propagation of information in Twitter network using SIR propagation.

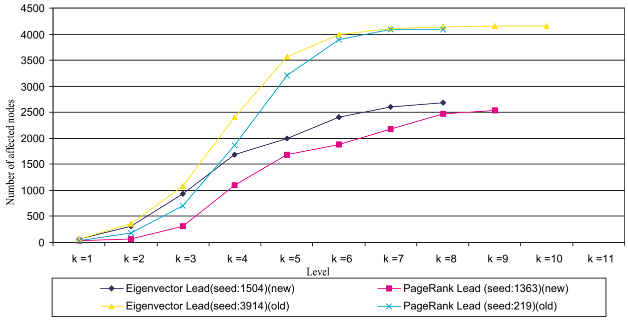

Figure 13. Propagation of information in author cooperative network using BFS propagation.

PageRank Lead (old) and PageRank Lead (new) represent the nodes with highest PageRank centrality in original and degenerated graph, respectively.

Node ID 327 and 9 are calculated as the nodes with highest Eigenvector centrality from the original and degenerated network, respectively. Now, for BFS propagation, 1420 nodes are affected through the propagation as compared with 1600 affected nodes when seed nodes are node ID 327 and 9, respectively. Similarly for PageRank, node ID 1504 (PageRank new) affected 1400 nodes as compared with seed node 315 (PageRank old), which affected 1600 nodes. From the graph, we find that though at the initial level, it performs better than PageRank old, level of propagation is restricted to 9 whereas in the other case it propagates up to 13th level.

For SIR propagation in Twitter network, affected nodes are 1500 in both PageRank new and Eigenvector new as compared with 2500 nodes affected through PageRank old and Eigenvector old seed node.

Figures 13 and 14 present the propagation of information for Eigenvector and PageRank centrality when these nodes are simulated through BFS and SIR nodes, respectively, in the author cooperative network for original and degenerated network. In coauthor network, propagation due to seed node derived from degenerative network both for Eigenvector and PageRank is approximately 2700 affected nodes, as compared with 4000 affected nodes when seed

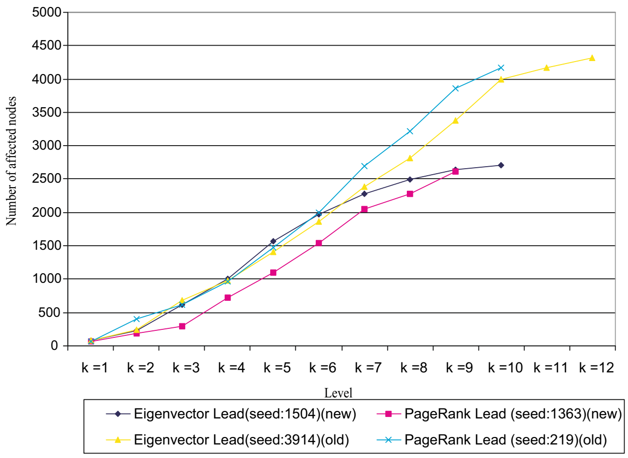

Figure 14. Propagation of information in author cooperative network using SIR propagation.

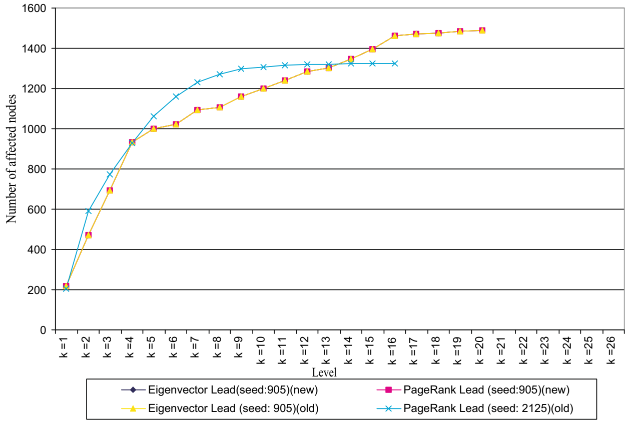

Figure 15. Propagation in Bitcoin network using BFS propagation.

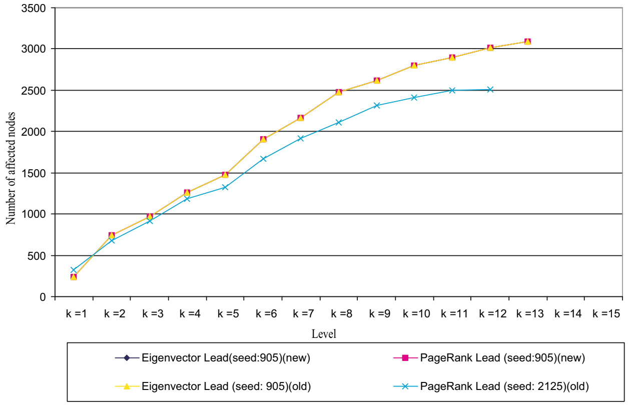

Figure 16. Propagation in Bitcoin network using SIR propagation.

Table 4. Computation time (in seconds) in original and degenerated network.

Centrality									
Network name	Eigenvector		PageRank		Betweenness		Closeness		*k*-core computation time
	Original	*k*-core	Original	*k*-core	Original	*k*-core	Original	*k*-core	
Bitcoin	5.64	2.38	5.91	2.2	18.523	5.26	12.64	5.43	0.593
Coauthor	9.08	3.59	5.16	2.24	12.225	4.84	10.876	4.8	0.385
Twitter	10.551	3.343	9.681	3.201	10.845	5.04	8.912	4.78	0.789

nodes are computed from the original network. Here node ID for Eigenvector (old), Eigenvector (new), PageRank (old) and PageRank (new) is 3914, 1504, 219 and 1363, respectively.

Figures 15 and 16 present the propagation of information for Eigenvector and PageRank centrality when these nodes are simulated through BFS and SIR nodes, respectively, in the Bitcoin network for original and degenerated graph. Here, node ID for Eigenvector (old) and PageRank (old) is 905 and 2125, respectively. Eventually, Eigenvector (new) and PageRank (new) derive the same node ID 905, which is the same as that for Eigenvector (old). As with all results, it is evident that Eigenvector performs the best as centrality measure among all five; PageRank (new) performance is better than that of PageRank (old).

In table 4, computation times for each algorithm for original and degenerated network are given. Unit of time is seconds. For the derivation of *k*-core, degenerating network is also presented in the table. It is evident from table 4 that computation time reduces to almost 40% of its original value of computation time for all centrality measures. For example, in Twitter network, calculation of Eigenvector centrality in original graph takes 10.55 s. However, for degenerating network, the time reduced to (3.343 + 0.789) s = 4.132 s.

5.1 *Analysis*

Social networks thrive on propagation of information. The extent to which information flows is an important parameter in understanding the property of a network. Certain information may be needed to be sent to a wide majority of people or to only a few people up to a certain strata. We model both of these cases. At each level of propagation from each seed node, the number of affected nodes is measured. If propagation emanating from a node can reach the maximum number of levels and can affect maximum number of nodes, then that node might be ideally selected as the root node for propagation. In some cases, a trade-off might be needed between the number of levels and the number of nodes affected, which is explained here.

Degree distribution of a node is a very important property for the analysis of social networks, but it has a few major drawbacks that arise by not considering interaction measures among the nodes in the network. A node in the network can have many neighbours. However, this node may be very silent in nature and it may rarely interact with the other nodes in the network. Hence, this node cannot be chosen as an efficient seed node in spreading information in the network. Thus, as a next step, strength of interaction measure among the nodes is considered, which is represented as the cost of the edge. Depending on the interaction measures, closeness centrality and betweenness centrality of a node are calculated. However, in some cases it may happen that seed nodes based on closeness centrality are not capable of spreading awareness efficiently in the network. The reason behind such a behaviour is that the node in the network can have a large number of interactions, but it has made a major part of that interaction with a specific node in the network; hence, in this case, though the node has made a sizeable number of interactions, it is capable of influencing only very few nodes in the network. In large networks, it can be seen that degree centrality outperforms the other measures in the initial stages when it comes to affecting nodes. This is fairly obvious because of the innate definition of node degree centrality. If a certain information is to be spread in only a few initial hops, then degree centrality can be preferred. For example, in the Twitter network, degree centrality leads the other measures till level 6, after which the influences of the more comprehensive measures take over. As seen, the results go to show that nodes chosen on the basis of Eigenvector centrality seem to show the best results for all the networks. This definitively goes to show that a node that is more related to other vertices should be the most ideal node as seed node to propagate information.

Moreover, degenerative network is used for finding centrality nodes of the network, which reduces computational complexity and it is used for contagion spreading.

6. Conclusion

The main focus of this paper was to investigate information propagation in real-life social networks using different centrality measures as seed node. We demonstrated the simulation of the proposed algorithm on three social networks using five centrality measures. The study shows that nodes selected on the basis of Eigenvector centrality perform better as seed nodes compared with other centrality measures, in information propagation. In *k*-core degenerative network, we find out the centrality measures; we find that propagation is comparable to the centrality measures derived from the original network. However, it reduces computational complexity. Therefore measuring centrality after degenerating the original network can be effectively used for derivation of seed node of propagation for influence maximization.

References

[1] Tripathy R M, Bagchi A and Mehta S 2010 A study of rumor control strategies on social networks. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, ACM, pp. 1817–1820

[2] Grando F, Noble D and Lamb L C 2016 An analysis of centrality measures for complex and social networks. In: *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, pp. 1–6

[3] Leskovec J, Adamic L A and Huberman B A 2006 The dynamics of viral marketing. In: *Proceedings of the ACM Conference on Electronic Commerce*, pp. 228–237

[4] Vicario M D, Zollo F, Caldarelli G and Scala A and Quattrociocchi W 2017 Mapping social dynamics on Facebook: the Brexit debate. *Social Networks* 50: 6–16

[5] Li X, Rao Y, Xie H, Liu X, Wong T and Wang F L 2017 Social emotion classification based on noise-aware training. *Data & Knowledge Engineering*. https://doi.org/10.1016/j.datak.2017.07.008

[6] Amplayo R K and Song M 2017 Adaptable fine-grained sentiment analysis for summarization of multiple short online reviews. *Data & Knowledge Engineering* 110: 54–67

[7] Vavliakis K N, Symeonidis A L and Mitkas P A 2013 Event identification in web social media through named entity recognition and topic modeling. *Data & Knowledge Engineering* 88: 1–24

[8] Reyes A, Rosso P and Buscaldi D 2012 From humor recognition to irony detection: the figurative language of social media. *Data & Knowledge Engineering* 74: 1–12

[9] Zaharia M, Chowdhury M, Franklin M J, Shenker S and Stoica I 2010. In: *Proceedings of HotCloud*

[10] Apache Spark. http://spark.apache.org/

[11] Mendoza M, Poblete B and Castillo C 2010 Twitter under crisis: can we trust what we get? In: *Proceedings of the First Workshop on Social Media Analytics*, pp. 71–79

[12] Yang B, Di J, Liu J and Liu D 2013 Hierarchical community detection with applications to real-world network analysis. *Data & Knowledge Engineering* 83: 20–38

[13] Shi C, Cai Y, Fu D, Dong Y and Wu B 2013 A link clustering based overlapping community detection algorithm. *Data & Knowledge Engineering* 87: 394–404

[14] Holthoefer J B and Moreno Y 2012 Absence of influential spreaders in rumor dynamics. *Physical Review E* 85: 026116

[15] Arruda G F, Barbieri A L, Rodriguez P M, Rodrigues F A, Moreno Y and Costa L F 2014 The role of centrality for the identification of influential spreaders in complex networks. *Physical Review E* 90(3): 032812–032829

[16] Miritello G, Moro E and Lara R 2010 The dynamical strength of social ties in information spreading. *Physics Review E* 83: 045102

[17] Wenjing Y, Brenner L and Giua A 2019 Influence maximization in independent cascade networks based on activation probability computation. *IEEE Access* PP(99): 1. https://doi.org/10.1109/ACCESS.2019.2894073

[18] Kempe D, Kleinberg J and Tardos E 2003 Maximizing the spread of influence in a social network. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 137–146

[19] Sun J and Tang J 2011 A survey of models and algorithms for social influence analysis. *Social network data analytics*. Springer, Boston, MA, pp. 177–214

[20] Yao Q, Shi R, Zhou C, Wang P and Guo L 2015 Topic-aware social influence minimization. In: *Proceedings of the 24th International Conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, pp. 139–140

[21] Dey P and Roy S 2016 Social network analysis. In: *Advanced Methods for Complex Network Analysis*. IGI, Hershey, Pennsylvania, USA, pp. 237–265

[22] Guzman J D, Deckro R F, Robbins M J, Morris J F and Ballester N A 2014 An analytical comparison of social network measures. *IEEE Transactions on Computational Social Systems* 1(1): 35–45

[23] Page L, Brin S, Motwani R and Winograd T 1999 *The PageRank citation ranking: bringing order to the web*. Tech. Report, Stanford InfoLab

[24] Jiang J, Wen S, Liu B, Yu S, Xiang Y and Zhou W 2019 Identifying propagation source in time-varying networks. *Malicious attack propagation and source identification*. Cham: Springer, pp. 117–137

[25] Domenico M D, Lima A, Mougel P and Musolesi M 2013 The anatomy of a scientific rumor. *(Nature Open Access) Scientific Reports* 3: 2980

[26] Newman M E J 2004 Coauthorship networks and patterns of scientific collaboration. In: *Proceedings of the National Academy of Science*, pp. 5200–5205

[27] Kumar S, Spezzano F, Subrahmanian VS and Faloutsos C 2016 Edge weight prediction in weighted signed networks. In: *Proceedings of the IEEE International Conference on Data Mining, ICDM*