Physica A 436 (2015) 833–845

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Improving detection of influential nodes in complex networks

Amir Sheikhahmadi [a], Mohammad Ali Nematbakhsh [a,*], Arman Shokrollahi [b]

[a] Department of Computer Engineering, University of Isfahan, Isfahan, Iran

[b] Department of Mathematics and Computer Science, West Virginia University, Morgantown, WV 26506, USA

HIGHLIGHTS

- The goal is to identify the most influential nodes in complex networks.
- We propose DegreeDistance and improve it in two phases, FIDD and SIDD.
- We take into account distance of seeds as well as the influence score.
- We investigate the rate of unique nodes influenced by our methods.
- The SIDD outperforms other measures by choosing a more appropriate seed set.

ARTICLE INFO

Article history:
Received 17 December 2014
Received in revised form 29 March 2015
Available online 5 May 2015

Keywords:
Centrality measure
DegreeDistance centrality
IC model
Influential nodes
Complex networks

ABSTRACT

Recently an increasing amount of research is devoted to the question of how the most influential nodes (seeds) can be found effectively in a complex network. There are a number of measures proposed for this purpose, for instance, high-degree centrality measure reflects the importance of the network topology and has a reasonable runtime performance to find a set of nodes with highest degree, but they do not have a satisfactory dissemination potentiality in the network due to having many common neighbors ($CN^{(1)}$) and common neighbors of neighbors ($CN^{(2)}$). This flaw holds in other measures as well. In this paper, we compare high-degree centrality measure with other well-known measures using ten datasets in order to find a proportion for the common seeds in the seed sets obtained by them. We, thereof, propose an improved high-degree centrality measure (named *DegreeDistance*) and improve it to enhance accuracy in two phases, FIDD and SIDD, by put a threshold on the number of common neighbors of already-selected seed nodes and a non-seed node which is under investigation to be selected as a seed as well as considering the influence score of seed nodes directly or through their common neighbors over the non-seed node. To evaluate the accuracy and runtime performance of DegreeDistance, FIDD, and SIDD, they are applied to eight large-scale networks and it finally turns out that SIDD dramatically outperforms other well-known measures and evinces comparatively more accurate performance in identifying the most influential nodes.

* Corresponding author.
E-mail address: nematbakhsh@eng.ui.ac.ir (M.A. Nematbakhsh).

http://dx.doi.org/10.1016/j.physa.2015.04.035
0378-4371/

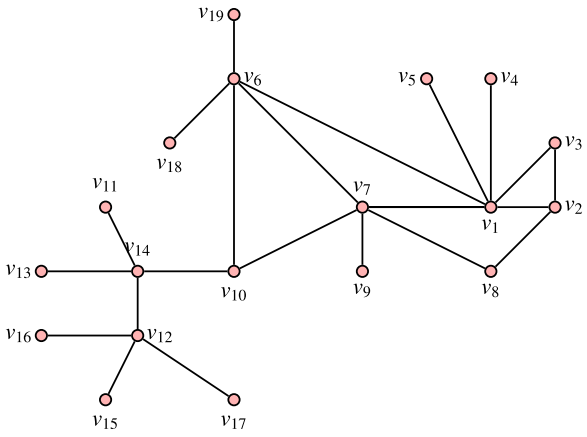


Fig. 1. A sample network which demonstrates that we get better propagation if the seed nodes (v_1 and v_{14}) are chosen in an appropriate distance of each other.

1. Introduction

Identifying the most influential nodes is a pivotal challenge and is of high importance due to its efficacious applications in complex networks, such as proliferation or ceasing the influence over social and economic networks or giving publicity to a product, organization, or venture [1–4], prevention and control of infectious diseases, understanding the function of the human brain and mental disorders [5,6], ranking web pages properly in search engines results [7,8], further analysis of the most enriched processes in biological systems and therapeutic targets [9]. Typically in social networks where the number of users is considerably increasing, one of the goals is maximizing or minimizing the spread of influence through influential nodes. The compulsive, entertaining environments of these networks and the wide diversity of services these systems provide, are making them a proper place for amusement, training, propaganda, etc. [10]. Everyday, we see a huge amount of goods and products advertisements, campaign people ads, and, etc. over these networks. Accepting an advertisement by a user and sharing it with friends and again friends with their friends actively publicize it and facilitates propagation [11–13]. It basically takes advantage of users to advertise products without too much sustained efforts rather than direct interaction which is very costly. On the other hand, the result of this process may be more efficient if friends have confidence in one another [14–18]. This interactive marketing technique is known as "viral marketing" which induces social networking services and other technologies to pass along a marketing message by finding and convincing the most influential individuals [11–17,19,20]. Shortly after, some immediate questions come up like what is the influential node? and how can they be identified? Indeed it is not practically feasible to select all these typical nodes to start propagation due to a shortage of funds and time-consuming, expensive process. Accordingly, the problem is to find an optimal subset of nodes within the network that are able to spread the influence and information as efficient and effective as possible. Previous literature address the maximization problem as "maximizing the spread of influence" [21,22].

Any complex network can be modeled as a directed or undirected network (or graph) consisting of nodes (vertices) and links (edges). Due to conspicuous lack of information about nodes in some complex networks (e.g. social networks), a fairly large amount of scientific studies have considered the structural parameters [23–26,18,27,28]. Then, nodes have been ranked based on the topology of the network and the location of each node in the network. In these approaches, nodes have been evaluated based on measures such as high-degree (or simply degree), betweenness, closeness, etc., and those with the highest/lowest measure have been taken as influential nodes (seeds) to start any desired propagation activities over the network. In this paper, we first scrutinize these measures and figure out a rate of intersection of the seed sets obtained by these measures. Another noteworthy observation is that if seeds in these seed sets are not identical, they are very close to one another so that they are either neighbors or neighbors of neighbors of each other. So, we perceive that the neighborhood overlapping of seeds of different seed sets obtained by these measures is prominent. Hence, these seed sets influence almost the same collection of nodes in the network. Fig. 1 displays a small network and, as we can see, nodes v_1, v_2, v_6, v_7 show high-degree centrality which are adjacent to each other, however by choosing v_1 and v_{14} which are in an appropriate distance of each other, we can achieve a more effective propagation.

Hereinafter, we use the following concepts and notations throughout the paper: The *distance* between two nodes v and w, denoted by $d(v, w)$, is the length of a shortest path between them. We say that a node w is an ith *neighbor* ($i \in \mathbb{Z}^+$) of nodes $v_1, v_2, \ldots, v_r$, $r \geq 1$, if $d(v_1, w) = d(v_2, w) = \cdots = d(v_r, w) = i$. Let $\mathrm{N}^{(i)}(v_1, v_2, \ldots, v_r)$ denote the family of all ith neighbors of nodes $v_1, v_2, \ldots, v_r$, and $\mathrm{N}^{(i)}$ if nodes are not specified. If $A = \{v_1, v_2, \ldots, v_r\}$, we use the short notation $\mathrm{N}^{(i)}(A)$. In some network science and graph theory texts, $\mathrm{N}^{(1)}(v)$ and $\mathrm{N}^{(2)}(v)$ are referred to as neighbors of v and neighbors of neighbors (second order contiguity) of v, respectively. A node z is said to be an ith *common neighbor* of nodes $v_1, v_2, \ldots, v_r$, $r \geq 1$, if $z \in \bigcap_{h=1}^{r} \mathrm{N}^{(i)}(v_h)$. We denote the set of all ith common neighbors of nodes $v_1, v_2, \ldots, v_r$ by $\mathrm{CN}^{(i)}(v_1, v_2, \ldots, v_r)$, and $\mathrm{CN}^{(i)}$ if v_h's ($h = 1, 2, \ldots, r$) are not specified. We define $\mathrm{CN}^{(1,2)} = \mathrm{CN}^{(1)} \cup \mathrm{CN}^{(2)}$. A node w is said to be in *distance threshold*, d_{td},

from v if $w \in \mathrm{N}^{(r)}(v)$ for some $r \geq d_{td}$. A node w is said to be *unique* in sets $X_1, X_2, \ldots, X_r,\ r \geq 1$, if there exists one and only one $h \in \{1, 2, \ldots, r\}$ such that $w \in X_h$. Lastly, let k be the seed set size.

For example, in Fig. 1, $\mathrm{N}^{(2)}(v_7) = \{v_2, v_3, v_4, v_5, v_{14}, v_{18}, v_{19}\}$; $v_2 \in \mathrm{CN}^{(3)}(v_{10}, v_{18}, v_{19})$; v_1 is a unique node in $\mathrm{N}^{(1)}(v_6)$, $\mathrm{N}^{(2)}(v_7)$, and $\mathrm{CN}^{(3)}(v_{10}, v_{19})$ because $v_1 \in \mathrm{N}^{(1)}(v_6)$ only; if we want to take a node in distance threshold $d_{td} = 2$ from v_{13}, we can choose any node in the network but v_{14}, similarly there is no node in distance threshold $d_{td} = 4$ of v_{10}.

In this study, we first investigate structural measures including high-degree, betweenness, closeness, eigenvector, PageRank, LeaderRank, and k-shell to show that regardless of the type of the measure and performance variety, the seed sets they produce have many seeds in common. We then verify that these structural measures usually search and select the nodes in the least distance within the network. Finally, we propose a method (named *DegreeDistance*) to find the most influential nodes by reforming high-degree centrality measure. Roughly speaking, we discuss and present: (1) *DegreeDistance*: an improved high-degree centrality measure in order to select the seed set, (2) *FIDD (First Improvement of DegreeDistance)*: an improvement of DegreeDistance by analyzing the number of common neighbors of seeds up to a distance threshold $d_{td} \in \{2, 3\}$, (3) *SIDD (Second Improvement of DegreeDistance)*: an improvement of FIDD by applying the influence score of the already-selected nodes in the seed set and their neighbors over a new potential node which is under investigation to be selected as a seed.

The main advantage of our proposed methods is greater performance in maximizing influence propagation with reasonable running time.

The rest of this paper is organized as follows: Section 2 briefly overviews well-known structural measures which build the basis of our discussions. In Section 3, we present the steps of *DegreeDistance* which is similar to high-degree centrality in spirit, and its improvements, FIDD and SIDD, to effectively and efficiently select the most influential nodes. In Section 4, we compare our methods with other measures, and in the last section, we summarize the main conclusions and suggest possible future directions.

2. Structural measures

The problem of identifying the most influential nodes in order to spread information over complex networks has been already studied in Refs. [18–30]. There are well-known measures that mostly deal with the location of nodes in the network. We use some of them to show that their seed sets contain partially the same seeds, and the seeds in a seed set have a significantly large amount of $\mathrm{CN}^{(1,2)}$. We also utilize the best measures among them to test the performance of our proposed methods. In the following, we briefly sketch them.

2.1. High-degree centrality

In this method, simply the nodes with the highest degree in the network should be marked as seeds. The reason behind this strategy is that these nodes can influence more nodes effectively due to having the greatest number of neighbors [31, Ch. 3]. High-degree centrality has been considered as a measure to study complex networks and the importance of nodes in (un)weighted networks [23–25].

L. Katz [32] developed this concept and introduced *Katz centrality* to measure the degree of influence of a node which takes into account the total number of walks. Each connection with distance j will be penalized by β^j where $0 \leq \beta \leq 1$. The formula to compute this measure is as follows,

$$C_i^{\mathrm{Katz}} = e_i^T \left(\sum_{j=1}^{\infty} (\beta \mathbf{A})^j \right) \mathbf{I}, \tag{1}$$

where e_i is a column vector whose entries are all zero except the ith entry which is 1, and $\mathbf{I}$ is the identity matrix. The disadvantage of using high-degree centrality measure is that it considers a node locally, i.e. based on its location, and in graphs with multiple components, the seeds are likely to be selected only from a big component.

2.2. Closeness centrality

The *farness* of a node u is the sum of the distances of u to all other nodes, and its closeness is the reciprocal of the farness. Hence, the closeness can be interpreted as a measure indicating how long it will take to spread information from a node u to all other nodes sequentially, another words, u is taken as an influential (central) node by the closeness strategy if its total distance to all other nodes is lowest. These nodes have greater influence due to the least number of intermediaries. This centrality measure can be computed by counting the shortest paths, and the following is one of the well-known expressions that is attributed to sociologist L. Freeman [26],

$$C_i^{\mathrm{CLO}} = e_i^T \mathbf{S} \mathbf{I}, \tag{2}$$

where $\mathbf{S}$ is the matrix whose (i, j)-th entry represents the length of a shortest path from node i to node j. The closeness measure needs to travel over the whole network, and clearly it is time-consuming and inappropriate for large-scale networks.

2.3. Betweenness centrality

By this indicator, influential nodes are those that are visited by the largest number of shortest paths from all nodes to all others within the network. L. Freeman [26] has introduced the expression below to compute this centrality,

$$C_i^{\mathrm{BET}} = \sum_{j \neq r \neq i} \frac{g_{jr}(i)}{g_{jr}}, \tag{3}$$

where g_{jr} is the number of shortest paths between nodes j and r, and $g_{jr}(i)$ is the number of shortest paths between j and r passing through the node i.

The nodes with the highest betweenness are sometimes called *bottlenecks* [33], or *intermediaries* [34], or *structural holes* [27].

2.4. Eigenvector centrality

This measure is closely related to Katz centrality and was introduced first by P. Bonacich [28]. It tries to find the influence of a node by assigning a score to every node based on the adjacency of that node to high-scoring nodes.

2.5. PageRank

PageRank is an algorithm which is used in Google search engine to rank web pages [35]. A web page linking to more important web pages has higher rank. Thus, a page with fewer neighbors might have a higher PageRank than another page with more neighbors. S. Fortunato et al. [36] and J. Heidemann et al. [37] separately used this centrality measure to rank nodes in social networks.

2.6. DegreeDiscount centrality

In 2009, W. Chen et al. [22] proposed the DegreeDiscount heuristic algorithm. When a node is selected as a seed, another node with highest degree can be potentially selected as a new seed, but the edge between these two should not be counted towards its degree [38, Ch. 4]. Another words, if a node u has degree d_u, and d'_u of them are already selected as seeds, we need to discount $d(u)$ by $2d'_u + (d_u - d'_u)d'_u p$, where p is a small propagation probability. This model does not maximize the total information flow in the network.

2.7. LeaderRank

In 2011, L. Lü et al. [39] proposed a variant of PageRank known as LeaderRank. Weighted LeaderRank is a slightly improved version of LeaderRank [40].

2.8. k-shell decomposition

M. Kitsak et al. [41] presented this measure which basically deals with the location of nodes in the network and assigns a k_S index to each node. Nodes with high index are located in the innermost network core and those with low index are at the periphery of the network.

2.9. Greedy algorithm

This algorithm introduced by D. Kempe et al. [21]. An initial seed set, S, is considered and in each step of the algorithm a single node, v, is being added to S so that $S \cup \{v\}$ maximizes the spread of influence and activates a larger number of nodes in the network. This process iteratively continues until the top k nodes are chosen, i.e. $|S| = k$.

3. Our centrality measure, *DegreeDistance*, and its improvements

In this section, we first discuss this matter that the well-known measures, mentioned in the preceding section, select almost the same seed set, and then find the rate of similarity between neighbors and neighbors of neighbors of seeds of the seed set obtained by any of the measures (i.e. $\mathrm{CN}^{(1)}$ and $\mathrm{CN}^{(2)}$ of seed nodes obtained by a particular measure). Based on this argument, we build a new seed set by exclusion of neighbors of seed nodes up to a specific distance, so seeds will be in distance threshold, d_{td}, from each other, and we propose a technique to improve identifying the most influential nodes.

Table 1
The list of the real-world datasets used in this paper. Order and size are the number of nodes and edges, resp.

Dataset	Type	Order	Size	Avg degree	Max degree
Twitter lists (TL)	Directed	23 370	33 101	2.8328	239
Facebook-NIPS (EF)	Directed	2 888	2 981	2.0644	769
Google+ (GP)	Undirected	23 628	39 242	3.3217	2 771
Facebook wall posts (Ow)	Directed	46 952	876 993	37.357	2 696
Catster (Sc)	Undirected	149 700	5 449 275	72.803	80 635
Hamsterster friendships (Shf)	Undirected	1 858	12 534	3.492	272
Wikipedia conflict (CO)	Undirected	118 100	2 917 785	49.412	136 192
Advogato (AD)	Directed	6 541	51 127	15.633	943
Brightkite (BK)	Undirected	58 228	214 078	7.353	1 134
Slashdot Zoo (SZ)	Directed	79 120	515 397	13.49 028	2 543
Epinions (ES)	Directed	75 879	508 837	13.412	3 079
Flickr (Fl)	Undirected	105 938	2 316 948	43.742	5 425
Gowalla (GW)	Undirected	196 591	950 327	9.6681	14 730
Youtube friendship (CY)	Undirected	1 134 890	2 987 624	5.2650	28 754
NetHEPT	Undirected	15 233	31 399	4.12	64

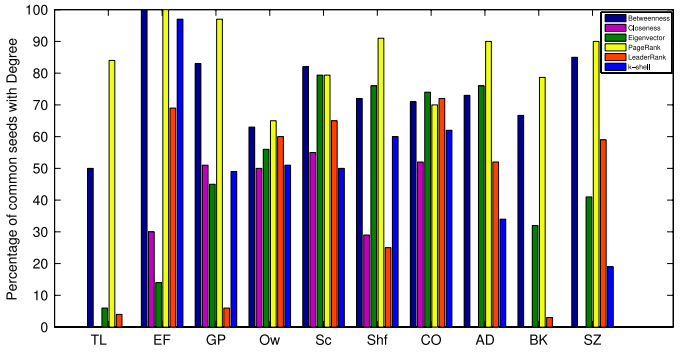


Fig. 2. The percentage of common seeds between high-degree seed set and betweenness, closeness, eigenvector, PageRank, LeaderRank, and k-shell seed sets are shown in dark blue, magenta, green, yellow, red, and light blue, respectively. Here $k = 100$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. Common seeds of different seed sets

The main question here is how many seeds do the seed sets obtained by the mentioned measures have in common? To be more clear, we can find the number of common seeds obtained by, for example, high-degree and closeness, or closeness and PageRank, etc. We also address the total cardinality of $CN^{(1,2)}$ of seeds in a seed set.

To find out the number of common seeds, we take out the first k seeds using each measure, where $k \in \{25, 50, 75, 100\}$ in our investigation, and apply the following formula,

$$COM(S_1, S_2) = \frac{|S_1 \cap S_2|}{k} \cdot 100, \tag{4}$$

where S_1 and S_2 are two seed sets obtained from two arbitrary centrality measures. To investigate this type of overlapping, we use the first ten datasets described in Table 1. All the datasets are taken from KONECT except NetHEPT which is a scientific collaboration network taken from the High Energy Physics — Theory citations from arXiv. Since we are particularly interested in high-degree centrality measure, we have examined the number of its common seeds with other measures' seed sets in Fig. 2.

3.2. $CN^{(1,2)}$ of seeds in a seed set

By computing $CN^{(1,2)}$ of seeds inside a seed set, we can easily find out how topologically close they are to each other. We want to show that the seeds selected each of the mentioned measures mostly belong to $N^{(i)}$, $i \leq 2$, of each other. This fact leads to wasting time and energy as well as ill-suited dissemination in complex networks. For instance, looking from the perspective of social networks, selecting seeds close to each other results in increasing persistence and intensity of a specific people in the network, based on the law of diminishing returns [42, Ch. 7]. Accordingly, we first find the rate of $CN^{(1)}$ and $CN^{(2)}$ (i.e. $CN^{(1,2)}$) of seeds obtained by each measure. For, we first select top k seeds ($k \in \{25, 50, 75, 100\}$) by one of the measures, and then find the $CN^{(1,2)}$ of them, put them all in F. We then compute the number of unique nodes, and find the rate by $COV = 100 - [(unique/total) \cdot 100]$. Algorithm 1 illustrates this procedure, and based on it, the rate of having $CN^{(1,2)}$ for the seeds in different seed sets is displayed in Table 2 after we introduce DegreeDistance in Algorithm

Table 2
The rate of $CN^{(1,2)}$ of seeds of different seed sets obtained by various measures on seven datasets. The last two columns belong to DegreeDistance (DD) with different distance threshold, $d_{td} = 2, 3$, between seeds.

Dataset	Top k	Degree	Betweenness	Closeness	Eigenvector	PageRank	LeaderRank	k-shell	DD, $d_{td} = 2$	DD, $d_{td} = 3$
AD	25	93.44	93.28	28.00	93.65	93.35	93.22	92.11	81.46	71.31
	50	96.52	96.42	14.00	96.59	96.48	96.17	95.74	82.24	72.55
	75	97.57	97.51	25.68	97.61	97.56	97.24	96.97	84.17	73.51
	100	98.08	98.05	44.49	98.13	98.08	97.81	97.66	85.7	74.45
Ow	25	88.04	86.39	90.53	90.35	88.04	69.6	89.16	58.05	48.33
	50	92.16	92.16	93.43	94.38	92.16	74.62	92.86	59.5	51.35
	75	93.36	93.36	94.93	96.04	93.36	82.42	94.05	61.34	53.25
	100	94.33	94.33	95.87	96.94	94.33	89.44	94.38	62.12	55.39
GP	25	82.67	79.12	74.4	86.49	78.56	59.27	87.46	52.88	41.13
	50	88.79	87.87	90.06	93.03	87.76	86.53	93.61	54.39	43.19
	75	91.02	91.6	93.22	95.36	90.95	91.09	95.53	56.98	45.25
	100	92.94	93.67	95.38	96.46	92.77	92.78	96.48	58.22	47.89
TL	25	35.58	46.92	30.51	91.13	27.23	54.45	14.78	26.78	19.62
	50	40.82	62.29	47.5	95.02	39.03	63.6	26.2	27.11	21.18
	75	53.05	65.37	52.61	96.3	44.3	66.77	41.65	29.01	22.89
	100	56.18	70.23	59.03	96.95	50.48	72.53	47.58	31.25	24.11
BK	25	88.26	88.50	0.00	89.42	87.31	50.9	64.76	58.2	50.25
	50	92.94	93.04	0.00	92.68	93.10	60.21	75.7	59.24	54.18
	75	94.71	95.00	0.00	94.32	94.97	66.07	80.41	61.33	57.31
	100	95.20	95.29	50.00	94.20	95.27	70.22	84.02	62.14	59.55
Sc	25	92.77	92.54	41.51	93.26	92.6	91.71	92.14	71.11	60.21
	50	96.078	95.81	52.7	96.3	95.93	94.29	95.68	73.25	63.42
	75	97.3	96.87	74.03	97.39	97.16	95.27	97.06	76.41	65.25
	100	97.87	97.53	93.92	97.93	97.79	96.3	97.81	78.52	68.19
SZ	25	81.55	81.07	0.00	80.99	80.95	87.59	87.33	68.45	59.94
	50	88.46	88.09	0.00	87.84	88.20	92.49	92.61	71.45	63.66
	75	91.21	90.75	0.00	90.50	90.89	94.36	94.64	73.67	65.15
	100	96.17	96.07	50.00	96.01	96.01	95.52	95.84	76.18	68.38

2. From the table, we can see that the DegreeDistance seeds with $d_{td} = 3$ have the least value of $CN^{(1,2)}$, which means the seeds are in an appropriate distance of each other, and hence, they influence a larger number of unique nodes within the network, as depicted in Fig. 5. To be more clear, seeds not too close to each other can influence other nodes in the network rather than influencing a specific set of nodes repeatedly, though in the continuation of the paper, we show that the value of *COV* for seeds is not the only factor which matters and this brings some improvements into conversation.

Algorithm 1 Computing the rate of $CN^{(1,2)}$ of k seeds

Input: S_1, k ▷ S_1 is the seed set
Output: The percentage of $CN^{(1,2)}$ of the seeds in S_1
$i \leftarrow 0$ ▷ i is the number of selected seeds
$total \leftarrow 0$
$F \leftarrow \emptyset$
while $i < k$ **do**
  $s' \leftarrow \text{Top}(S_1)$
  $S_1 \leftarrow S_1 \setminus \{s'\}$
  $F_i \leftarrow N^{(1)}(s') \cup N^{(2)}(s')$
  $total = total + |F_i|$
  $i \leftarrow i + 1$
end while
$unique \leftarrow |F|$
$COV \leftarrow 100 - [(unique/total) * 100]$

3.3. DegreeDistance: Improved high-degree centrality measure

As we discussed, one of the main issues with most of the widely-used measures such as high-degree, betweenness, closeness, eigenvector, PageRank, LeaderRank, and k-shell to select an appropriate seed set is that the seed nodes have a remarkable amount of $CN^{(1,2)}$ with one another. Therefore, due to this fact and high speed selection of seeds by high-degree centrality, the next logical step is to improve this measure in order to end up with a more effective seed set whose elements have the least value of $CN^{(1,2)}$. In our proposed method which is described in Algorithm 2, we first compute the degree of each node in the network and select a node with the highest degree and add it to a predefined selection set (*Sel*). To reduce the number of elements of $CN^{(1,2)}$ of the selected nodes in *Sel*, once we add a node to *Sel*, we take a distance threshold, d_{td}, to select the next seed, namely we remove the candidacy of the neighbors of the node in distance up to d_{td}; for instance, if a node v is already selected as a seed and $d_{td} = 3$, the nodes in $N^{(1)}(v)$ and $N^{(2)}(v)$ will not be checked for selecting more seeds. As a matter of fact, in social networks, we nominate a person for being a seed if its ith neighbors ($i = 1, 2$), who have the highest confidence in them, have the least overlapping with the ith neighbors of already-selected people.

Algorithm 2 DegreeDistance centrality measure

Input: G, k, d_{td} ▷ G is the given network, and d_{td} is the distance threshold
Output: S ▷ seed set

$S \leftarrow \emptyset$
Compute degree of all nodes in G
$L \leftarrow$ Descending list of nodes based on their degree
while $|S| < k$ **do**
  $s' \leftarrow \max(L)$
  $Sel \leftarrow$ True
  for all $v \in S$ **do**
    if $d(s', v) < d_{td}$ **then**
      $Sel \leftarrow$ False
      break
    end if
  end for
  if *Sel* **then**
    $S \leftarrow S \cup \{s'\}$
  end if
  $L \leftarrow L \setminus \{s'\}$
end while

If $d_{td} = 2$, one can replace Algorithm 2 with Algorithm 3.

Algorithm 3 DegreeDistance with threshold 2

Input: G, k
Output: S ▷ seed set

$S \leftarrow \emptyset$
Compute degree of all nodes in G
$L \leftarrow$ Descending list of nodes based on their degree
while $|S| < k$ **do**
  $s' \leftarrow \max(L)$
  $S \leftarrow S \cup \{s'\}$
  $L \leftarrow L \setminus \left\{s' \cup N^{(1)}(s')\right\}$
end while

In the last algorithm above, once we select a node, its neighbors will be removed from L, and so there exists either no path or a path of length ≥ 2 between any two seeds. Now, it is time to show the results from Section 3.2.

To clarify how to get the values of *COV* in Table 2, as a sample case, the detailed information about the cardinality of $CN^{(1,2)}$ of top k seeds of the AD dataset is displayed in Table 3. Closeness seeds apparently have the least values, this is because there are heterogeneous components in the network and the tendency of this measure to small components.

3.4. FIDD using $CN^{(1)}$

In our proposed centrality measure (i.e. DegreeDistance), if a high-degree node is selected as a seed, we then avoid selecting its neighbors up to the d_{td} which yields an increase of spreading. In this way, despite the location diversity of the

Table 3
A sample of information about the cardinality of $CN^{(1,2)}$ for different values of k on the AD dataset.

Dataset	Top k		Degree	Betweenness	Closeness	Eigenvector	PageRank	LeaderRank	k-shell	DD, $d_{td}=2$	DD, $d_{td}=3$
AD	25	*Total*:	74 487	73 031	50	76 421	73 497	57 040	68 429	27 650	18 545
		Unique:	4 887	4 905	36	4 853	4 890	4 501	4 641	5 125	5 320
	50	*Total*:	141 671	138 247	100	143 725	140 506	109 983	124 113	29 780	20 200
		Unique:	4 933	4 944	86	4 904	4 939	4 679	4 749	5 289	5 545
	75	*Total*:	204 171	200 079	148	207 220	204 059	158 952	175 365	33 750	22 350
		Unique:	4 966	4 973	110	4 946	4 969	4 808	4 842	5 340	5 920
	100	*Total*:	260 044	256 520	236	266 417	260 010	221 245	209 872	37 890	24 321
		Unique:	4 990	5 010	131	4 982	4 992	4 845	4 911	5 421	6 213

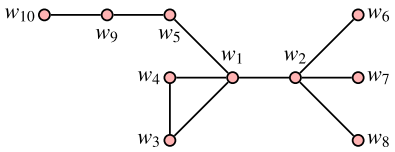


Fig. 3. DegreeDistance may remove neighbors of a seed which exert a powerful influence. By choosing w_1 and $d_{td}=2$, the node w_2 will be removed. We present FIDD to overcome this drawback.

selected nodes, we may practically remove nodes that have a highly influential neighbor in the seed set, though their connection might be weak. For example, in Fig. 3, the node w_1 with highest degree is chosen as a seed, and if the distance threshold is $d_{td}=2$, the nodes in $N^{(1)}(w_1)$ are practically put aside and the next seed will be w_9. Therefore, we see that the high degree node w_2 is removed and since there is only one path between w_1 and w_2, the subsequent nodes of w_2 will never get the chance of being influenced.

Based on the argument above, to nominate a new node (with highest degree among non-seed nodes) to be a seed, we need to evaluate $|CN^{(1)}|$ of seed nodes and the node which is in question. If it falls below a threshold θ, the node can be chosen as a seed, otherwise the influence is more likely to be easily propagated through these common neighbors, and therefore we do not select the node. This improvement is presented in Algorithm 4.

Algorithm 4 FIDD

Input: G, k, d_{td}, θ ▷ $d_{td} \in \{2, 3\}$, and θ is the threshold for $|CN^{(1)}|$
Output: S

```
S ← ∅
L ← Descending list of nodes based on their degree
while |S| < k do
    s' ← max(L)
    L ← L \ {s'}
    Sel ← True
    for all v ∈ S do
        if d(s', v) < d_td then
            No ← |CN^(1)(s', v)|
            if No ≥ θ then
                Sel ← False
                break
            end if
        end if
    end for
    if Sel then
        S ← S ∪ {s'}
    end if
end while
```

3.5. SIDD using $CN^{(1)}$ and the influence of seeds and their neighbors

The point missing in the last algorithm above is that how much may a non-seed node be influenced by seed nodes and their neighbors? In this regard, we present Algorithm 5.

Algorithm 5 SIDD

Input: D, k, d_{td}, θ ▷ $d_{td} \in \{2, 3\}$, and θ is the threshold for $|\mathrm{CN}^{(1)}|$
Output: S

$S \leftarrow \emptyset$
$L \leftarrow$ Descending list of nodes based on their degree
while $|S| < k$ **do**
  $s' \leftarrow \max(L)$
  $L \leftarrow L \setminus \{s'\}$
  $Sel \leftarrow$ True
  inf $\leftarrow 0$
  for all $v \in S$ **do**
    if $d(s', v) < d_{td}$ **then**
      $No \leftarrow |\mathrm{CN}^{(1)}(s', v)|$
      inf $\leftarrow \mathbb{P}(v, s') + \sum_{w \in \mathrm{CN}^{(1)}(s', v)} \left(\mathbb{P}(v, w) * \mathbb{P}(w, s')\right)$
      if $No \geq \theta$&inf $\geq \beta$ **then**
        $Sel \leftarrow$ False
        break
      end if
    end if
  end for
  if Sel **then**
    $S \leftarrow S \cup \{s'\}$
  end if
end while

Table 4
The Pearson correlation between SIDD and other measures on three datasets.

Dataset	Degree	Closeness	PageRank	DegreeDiscount	LeaderRank	k-shell
BK	0.431	−0.343	0.35	0.53	−0.009	0.27
ES	0.39	−0.341	0.18	0.22	−0.0421	0.31
SZ	0.53	0.28	0.35	0.66	0.19	0.39

In SIDD measure, to determine whether is not a new node, s', with highest degree should be selected as a seed, we add one more condition to FIDD which is the influence score and can be computed via the following expression,

$$\mathrm{inf} = \mathbb{P}(v, s') + \sum_{w \in \mathrm{CN}^{(1)}(s', v)} \Big(\mathbb{P}(v, w) \cdot \mathbb{P}(w, s')\Big). \tag{5}$$

Applying this expression, the activation probability of the in-question node, s', by a seed node v such that $d(s', v) < d_{td}$ through nodes $w \in \mathrm{CN}^{(1)}(s', v)$, can be determined. If this score is large enough, we can remove s' and give the chance of being a seed to another node which has little possibility to be influenced by seed nodes directly or through their neighbors.

4. Evaluation and experimental results

In this section, we assess the rate of having $\mathrm{CN}^{(1,2)}$ of DegreeDistance seeds and the rate of the number of seeds that DegreeDistance seed set and other measures' have in common. We also assess the runtime performance and spread ability of influence by DegreeDistance, FIDD, and SIDD seeds, then compare them with some other well-known measures. The proposed measures in this paper can be applied to any complex networks, albeit here we have mostly conducted the experiments on social networks and networks of this sort.

In Fig. 4, we have compared the number of common seeds between DegreeDistance, FIDD, SIDD seed sets and other measures' for $k = 100$. By looking back at Fig. 2, we can see that the rate of having common seeds between our measures and other measures is looked up, and our methods choose almost different seeds.

The relationship between SIDD and some other measures is evaluated using Pearson's correlation on three real-world datasets presented in Table 4.

4.1. Unique nodes influenced by DegreeDistance, FIDD, SIDD, and high-degree seeds

To evaluate DegreeDistance seeds in distance threshold $d_{td} = \{2, 3\}$ from each other, FIDD and SIDD seeds, we check the percentage of the unique nodes in the network that are influenced via them. From Fig. 5, it is clear that in large-scale networks, DegreeDistance seeds with $d_{td} = 3$ and SIDD cover significantly more unique nodes in comparison with high-degree.

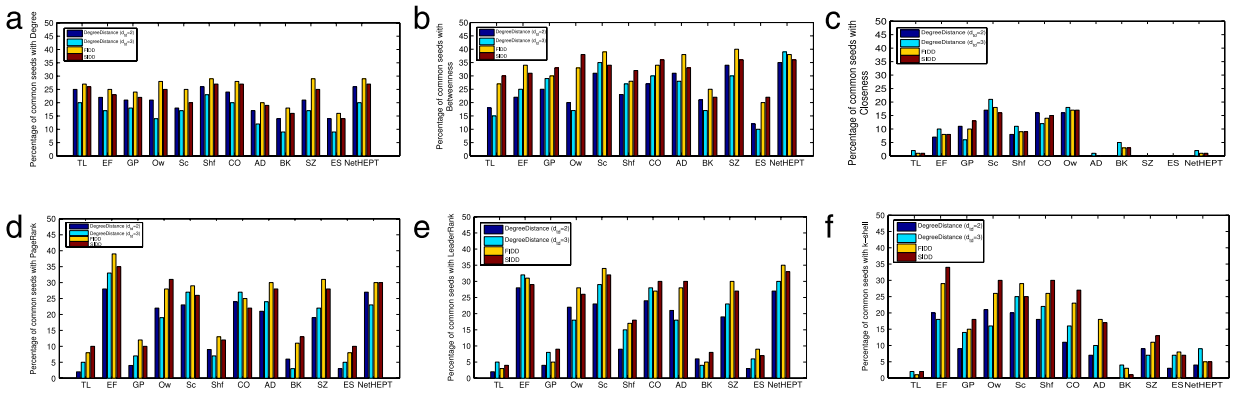


Fig. 4. The number of common seeds of the seed sets of DegreeDistance, FIDD, SIDD, and the seed sets of (a) high-degree, (b) betweenness, (c) closeness, (d) PageRank, (e) LeaderRank, and (f) k-shell. Here the seed set size $k = 100$.

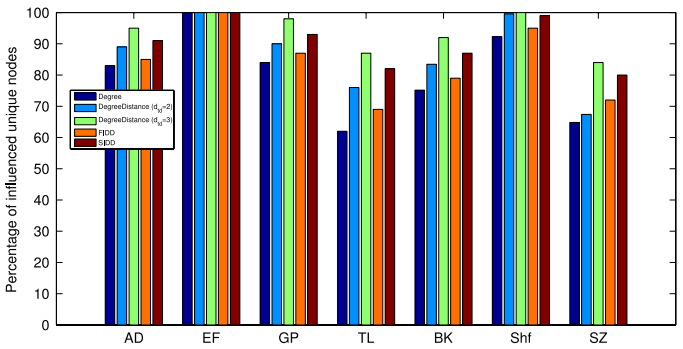


Fig. 5. The percentage of unique nodes influenced by DegreeDistance seeds with $d_{td} \in \{2, 3\}$, FIDD, SIDD, and high-degree seeds.

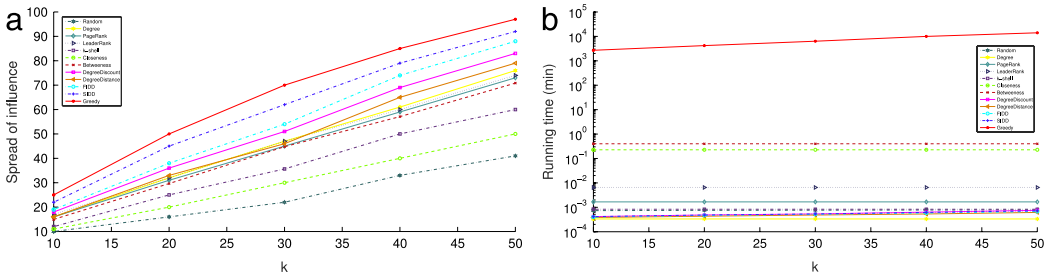


Fig. 6. (a) Comparison of spread of influence by seeds obtained from DegreeDistance, FIDD, and SIDD with other measures on NetHEPT dataset. (b) Comparison of runtime performance in order to identify seeds using DegreeDistance, FIDD, and SIDD with other centrality measures on the same dataset.

4.2. Runtime performance and spread of influence by DegreeDistance, FIDD, and SIDD

To evaluate the spread ability of DegreeDistance, FIDD, and SIDD, we compare them not only with other well-known measures, but with random method (k random nodes form the seed set) under the independent cascade (IC) model [21] to simulate the influence propagation with a 10,000-iteration process for each seed set and take the average of all the influence spreads. To analyze the spread efficiency of the mentioned methods, which are depicted in Figs. 6, 7, and Table 5, we apply them to some large-scale datasets from Table 1. The value of θ is assumed to be equal to the average degree of the network. Figs. 6 and 7 show the spread effectiveness and runtime efficiency on NetHEPT and BK datasets, respectively.

In our experiments, the influence score of a seed, v, on each $w \in \mathrm{N}^{(1)}(v)$ is set to be the fixed value 0.01, that is Eq. (5) becomes

$$\mathrm{inf} = \begin{cases} 0.01 + \left((0.01)^2 \cdot |\mathrm{CN}^{(1)}(s', v)|\right), & \text{if } v \text{ and } s' \text{ are adjacent,} \\ (0.01)^2 \cdot |\mathrm{CN}^{(1)}(s', v)|, & \text{otherwise.} \end{cases}$$

From this figure, one can find out that in spite of random model which has the lowest spread ability of influence, greedy method has the highest propensity. Clearly, greedy method is exceedingly time-consuming and is not an appropriate

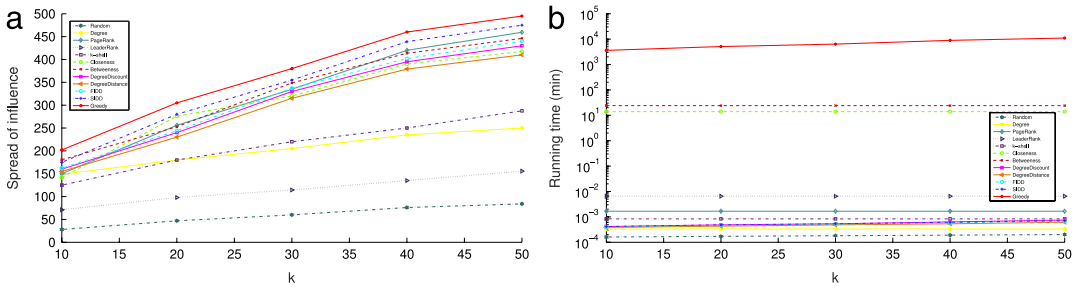


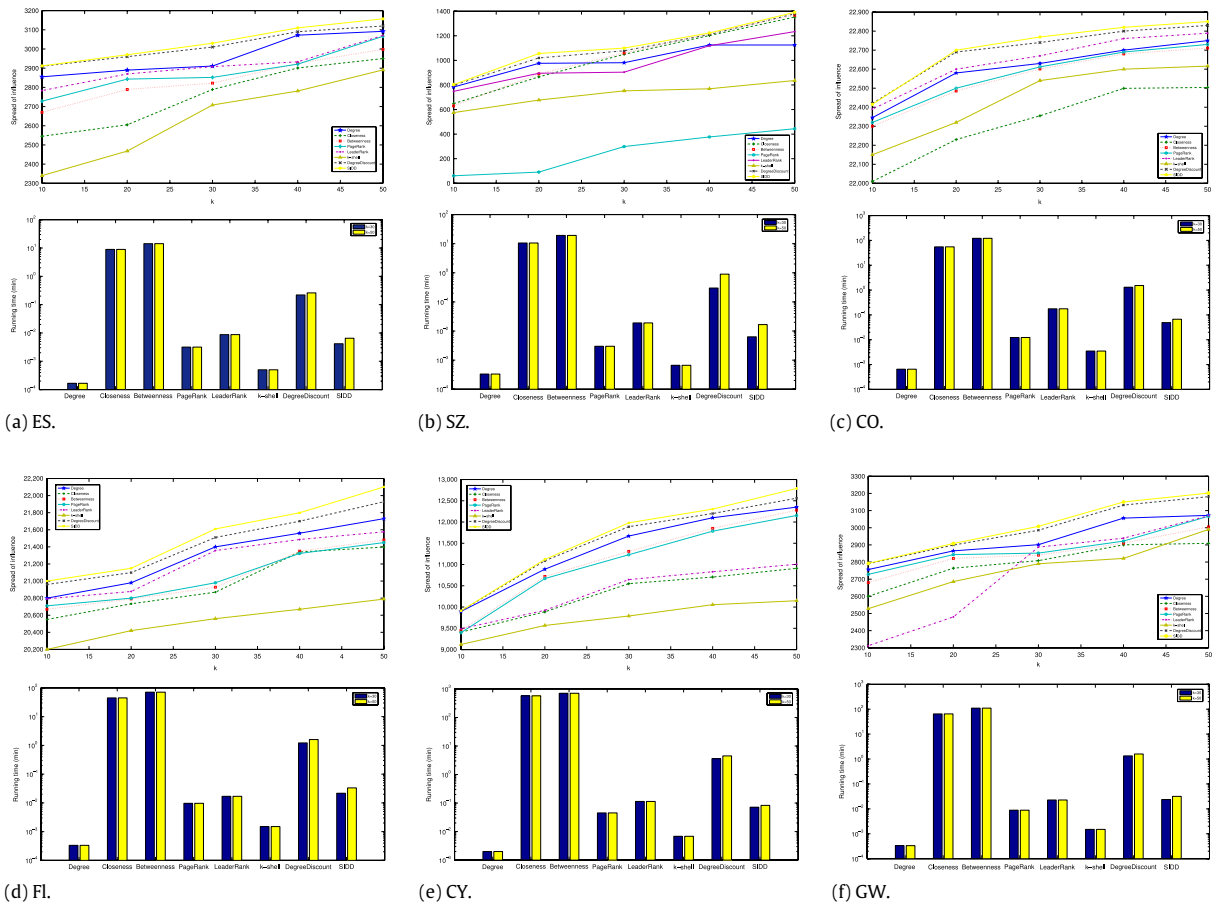
Fig. 7. (a) Comparison of spread of influence by seeds obtained from DegreeDistance, FIDD, and SIDD with other measures on BK dataset. (b) Comparison of runtime performance in order to identify seeds using DegreeDistance and its improvements with other measures on the same dataset.

Table 5
Spreading effectiveness of seeds by different seed sets.

Dataset	Top k	Degree	Closeness	Betweenness	DegreeDiscount	PageRank	LeaderRank	k-shell	SIDD
ES	$k=10$	2855	2545	2670	2910	2728	2783	2340	2912
	$k=20$	2890	2605	2789	2960	2843	2870	2468	2970
	$k=30$	2911	2789	2822	3010	2852	2908	2708	3030
	$k=40$	3072	2901	2925	3090	2921	2933	2781	3110
	$k=50$	3092	2950	2998	3120	3066	3071	2891	3157
SZ	$k=10$	783	648	630	798	61	748	575	803
	$k=20$	976	866	890	1020	91	895	677	1056
	$k=30$	981	1049	1060	1078	299	904	752	1100
	$k=40$	1125	1202	1220	1210	377	1119	769	1223
	$k=50$	1125	1352	1367	1379	444	1234	836	1391
CO	$k=10$	23865	23521	23814	24031	23836	23920	23661	24091
	$k=20$	24100	23741	23999	24301	24016	24130	23831	24371
	$k=30$	24150	23866	24115	24351	24128	24200	24031	24440
	$k=40$	24220	24010	24194	24411	24206	24291	24111	24491
	$k=50$	24270	24015	24225	24441	24236	24319	24127	24541
FI	$k=10$	20800	20550	20670	20960	20710	20792	20200	21000
	$k=20$	20980	20734	20791	21099	20800	20878	20420	21149
	$k=30$	21400	20870	20928	21510	20980	21357	20560	21610
	$k=40$	21560	21340	21350	21700	21323	21486	20670	21799
	$k=50$	21730	21398	21481	21928	21450	21576	20789	22101
CY	$k=10$	9897	9410	9450	9910	9398	9489	9120	9912
	$k=20$	10890	9887	10720	11089	10670	9923	9567	11120
	$k=30$	11670	10550	11310	11890	11230	10645	9789	11980
	$k=40$	12100	10705	11850	12200	11785	10830	10056	12304
	$k=50$	12350	10910	12279	12560	12154	11007	10148	12789
GW	$k=10$	2756	2600	2680	2790	2728	2312	2528	2792
	$k=20$	2865	2764	2821	2899	2843	2480	2686	2908
	$k=30$	2911	2808	2840	2985	2852	2887	2790	3008
	$k=40$	3072	2900	2910	3132	2922	2939	2821	3150
	$k=50$	3092	2909	3005	3181	3067	3070	2990	3202

measure for large-scale networks. Therefore, we have not taken these two measures any farther. In addition to the high speed performance of DegreeDistance (especially SIDD), it has a satisfactorily close spread ability of influence compared to greedy method. The running time of each algorithm is illustrated in Fig. 6. The experiments are carried out on a state-of-the-art desktop machine with Intel Core i7 3.4 GHz CPU and 4 GB RAM.

From the above arguments, we conclude that our proposed centrality measure and its improvements have a satisfying, acceptable performance in comparison to other methods.



(a) ES. (b) SZ. (c) CO.

(d) Fl. (e) CY. (f) GW.

Fig. 8. SIDD outperforms other measures with respect to maximizing spread of influence, which demonstrates a more precise selection of seeds. Its running time is quite legitimate for $k = 30, 50$.

As we mentioned, the influence score is set as 0.01. By increasing this value, we can avoid selecting seed nodes that have a high influence on one another and consequently we observe more difference between FIDD and SIDD.

Based on the above discussion and evaluation of our method on the two datasets, NetHEPT and BK, SIDD outperforms the previous two versions. Thereof, for further evaluation, we have considered SIDD only and carried out the same experiments as the previous two ones on more datasets (see Fig. 8. Here the running time is computed for $k = 30, 50$) as well as the numerical comparison of the spread effectiveness of SIDD with other methods on the same datasets (Table 5).

The results assert the superiority of SIDD over the other methods. Due to the close distance of nodes in the seed sets obtained from other measures, by increasing the size of the dataset, we do not see much spreading progress; for example, by applying the closeness measure on Gowalla, when we change $k = 40$ to $k = 50$, only nine more nodes got activated, or similarly k-shell decomposition does not show satisfactory promotion, the reason is that it gives the key users topologically in the inner-core of the network. Although the seed nodes (with high k-shell index) have high spread ability individually, we observe that these nodes are mostly in close neighborhood of one another, and they hence all together (top k) do not display a good spreading effectiveness compared to other commonly used measures of influence.

5. Conclusions and future directions

In this paper, we presented and overviewed some well-known measures such as high-degree, betweenness, closeness, eigenvector, PageRank, DegreeDiscount, LeaderRank, and k-shell. Using ten datasets, we verified that the seed sets obtained by these measures have many seeds in common. We also showed that in the seed sets, the cardinalities $|CN^{(1)}|$ and $|CN^{(2)}|$ are significantly large, another words, some nodes within the network can however be influenced by more than one seed. According to this fact and the similarity of seed sets obtained by high-degree and other measures, we proposed a new centrality measure, DegreeDistance, which would choose high-degree seeds in an appropriate distance of each other. We then improved this measure by inspecting the distance of the non-seed node of highest degree and seed nodes, and if the distance fell below the distance threshold, which was set as 2 and 3, the number of common neighbors (if applicable) of the node and a single seed in each step would determine whether the node could be a seed or not; we put a threshold θ for this

value which was taken the same as the average degree of each dataset in our experiments. On the other hand, since each node has influence over its neighbors, we considered the influence of its neighbors as a factor to keep or remove the in-question node. The experiments showed that the proposed measures are promising as they outperformed other measures on large-scale networks in terms of maximizing the spread of influence with acceptable running time.

From the proposed measures, one may improve other centrality measures in a similar way as well as the semi-local centrality measure [23,43]. Another interesting direction is finding a way to pick one seed from a set of nodes all of equal degree. We investigate DegreeDistance for the distance threshold $d_{td} \in \{2, 3\}$, it might be interesting to study the case of $d_{td} \geq 4$ and come up with the best distance threshold possible, though it highly depends on the type of networks.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. A. Shokrollahi thanks Aaron Clauset for his useful discussion on Pearson's correlation.

References

[1] P.A. Estevez, P. Vera, K. Saito, Selecting the most influential nodes in social networks, in: International Joint Conference on Neural Networks, 2007, IJCNN 2007, IEEE, 2007, pp. 2397–2402.
[2] M. Kimura, K. Saito, R. Nakano, Extracting influential nodes for information diffusion on a social network, in: AAAI, Vol. 7, 2007, pp. 1371–1376.
[3] M. Kimura, K. Saito, R. Nakano, H. Motoda, Extracting influential nodes on a social network for information diffusion, Data Min. Knowl. Discov. 20 (1) (2010) 70–97.
[4] Y. Zhang, Z. Wang, C. Xia, Identifying key users for targeted marketing by mining online social network, in: 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA), IEEE, 2010, pp. 644–649.
[5] M.N. Moussa, C.D. Vechlekar, J.H. Burdette, M.R. Steen, C.E. Hugenschmidt, P.J. Laurienti, Changes in cognitive state alter human functional brain networks, Front. Hum. Neurosci. 5.
[6] O. Sporns, Structure and function of complex brain networks, Dialogues Clin. Neurosci. 15 (3) (2013) 247.
[7] A.N. Langville, C.D. Meyer, Google's PageRank and Beyond: The Science of Search Engine Rankings, Princeton University Press, 2011.
[8] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the web.
[9] T. Arodz, D. Bonchev, Identifying influential nodes in a wound healing-related network of biological processes using mean first-passage time, New J. Phys. 17 (2) (2015) 025002.
[10] J. Heidemann, M. Klier, F. Probst, Online social networks: A survey of a global phenomenon, Comput. Netw. 56 (18) (2012) 3866–3878.
[11] T.N. Dinh, D.T. Nguyen, M.T. Thai, Cheap, easy, and massively effective viral marketing in social networks: truth or fiction? in: Proceedings of the 23rd ACM Conference on Hypertext and Social Media, ACM, 2012, pp. 165–174.
[12] O. Hinz, C. Schulze, C. Takac, New product adoption in social networks: Why direction matters, J. Bus. Res. 67 (1) (2014) 2836–2844.
[13] R. Iyengar, C. Van den Bulte, T.W. Valente, Opinion leadership and social contagion in new product diffusion, Mark. Sci. 30 (2) (2011) 195–212.
[14] M.Y. Cheung, C. Luo, C.L. Sia, H. Chen, Credibility of electronic word-of-mouth: informational and normative determinants of on-line consumer recommendations, Int. J. Electron. Commer. 13 (4) (2009) 9–38.
[15] J.A. Chevalier, D. Mayzlin, The effect of word of mouth on sales: Online book reviews, J. Mark. Res. 43 (3) (2006) 345–354.
[16] N.S. Koh, N. Hu, E.K. Clemons, Do online reviews reflect a product's true perceived quality? an investigation of online movie reviews across cultures, Electron. Commer. Res. Appl. 9 (5) (2010) 374–385.
[17] D.-H. Park, J. Lee, I. Han, The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement, Int. J. Electron. Commer. 11 (4) (2007) 125–148.
[18] J. Yang, C. Yao, W. Ma, G. Chen, A study of the spreading scheme for viral marketing based on a complex network model, Physica A 389 (4) (2010) 859–870.
[19] M. Richardson, P. Domingos, Mining knowledge-sharing sites for viral marketing, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2002, pp. 61–70.
[20] F. Probst, D.-K.L. Grosswiele, D.-K.R. Pfleger, Who will lead and who will follow: Identifying influential users in online social networks, Bus. Inf. Syst. Eng. 5 (3) (2013) 179–193.
[21] D. Kempe, J. Kleinberg, 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, ACM, 2003, pp. 137–146.
[22] W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 199–208.
[23] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, T. Zhou, Identifying influential nodes in complex networks, Physica A 391 (4) (2012) 1777–1787.
[24] C. Gao, D. Wei, Y. Hu, S. Mahadevan, Y. Deng, A modified evidential methodology of identifying influential nodes in weighted networks, Physica A 392 (21) (2013) 5490–5500.
[25] D. Wei, X. Deng, X. Zhang, Y. Deng, S. Mahadevan, Identifying influential nodes in weighted networks based on evidence theory, Physica A 392 (10) (2013) 2564–2575.
[26] L.C. Freeman, Centrality in social networks conceptual clarification, Soc. Networks 1 (3) (1979) 215–239.
[27] R.S. Burt, Structural Holes: The social Structure of Competition, Harvard university press, 2009.
[28] P. Bonacich, P. Lloyd, Eigenvector-like measures of centrality for asymmetric relations, Social Networks 23 (3) (2001) 191–201.
[29] D. Katsaros, P. Basaras, Detecting influential nodes in complex networks with range probabilistic control centrality, in: Coordination Control of Distributed Systems, Springer, 2015, pp. 265–272.
[30] J. Liu, Y. Li, Z. Ruan, G. Fu, X. Chen, R. Sadiq, Y. Deng, A new method to construct co-author networks, Physica A 419 (2015) 29–39.
[31] J. Golbeck, Analyzing the Social Web, Newnes, 2013.
[32] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.
[33] H. Yu, P.M. Kim, E. Sprecher, V. Trifonov, M. Gerstein, The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics, PLoS Comput. Biol. 3 (4) (2007) e59.
[34] J. Scott, P.J. Carrington, The SAGE Handbook of Social Network Analysis, SAGE Publications, 2011.
[35] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1) (1998) 107–117.
[36] S. Fortunato, M. Boguna, A. Flammini, F. Menczer, How to make the top ten: Approximating pagerank from in-degree. ArXiv Preprint cs/0511016.
[37] J. Heidemann, M. Klier, F. Probst, Identifying key users in online social networks: a PageRank based approach.
[38] C.C. Aggarwal, An Introduction to Social Network Data Analytics, Springer, 2011.
[39] L. Lü, Y.-C. Zhang, C.H. Yeung, T. Zhou, Leaders in social networks, the delicious case, PLoS One 6 (6) (2011) e21202.
[40] Q. Li, T. Zhou, L. Lü, D. Chen, Identifying influential spreaders by weighted leaderrank, Physica A 404 (2014) 47–55.
[41] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks, Nat. Phys. 6 (11) (2010) 888–893.
[42] M. Hirschey, Managerial Economics, Cengage Learning, 2008.
[43] S. Gao, J. Ma, Z. Chen, G. Wang, C. Xing, Ranking the spreading ability of nodes in complex networks based on local structure, Physica A 403 (2014) 130–147.