Zhan *et al. J Big Data (2017) 4:16*
DOI 10.1186/s40537-017-0076-5

Journal of Big Data

RESEARCH Open Access

Identification of top-K nodes in large networks using Katz centrality

Justin Zhan*, Sweta Gurung† and Sai Phani Krishna Parsa†

*Correspondence:
justin.zhan@unlv.edu
†Sweta Gurung and Sai Phani Krishna Parsa contributed equally to this work
Department of Computer Science, College of Engineering, University of Nevada Las Vegas, Las Vegas, NV 89154, USA

Abstract

Network theory concepts form the core of algorithms that are designed to uncover valuable insights from various datasets. Especially, network centrality measures such as Eigenvector centrality, Katz centrality, PageRank centrality etc., are used in retrieving top-K viral information propagators in social networks,while web page ranking in efficient information retrieval, etc. In this paper, we propose a novel method for identifying top-K viral information propagators from a reduced search space. Our algorithm computes the Katz centrality and Local average centrality values of each node and tests the values against two threshold (constraints) values. Only those nodes, which satisfy these constraints, form the search space for top-K propagators. Our proposed algorithm is tested against four datasets and the results show that the proposed algorithm is capable of reducing the number of nodes in search space at least by 70%. We also considered the parameter (α and β) dependency of Katz centrality values in our experiments and established a relationship between the α values, number of nodes in search space and network characteristics. Later, we compare the top-K results of our approach against the top-K results of degree centrality.

Keywords: Top-K nodes, Katz centrality, Social networks

Introduction

With the outbreak of social networking platforms such as Facebook, Twitter, Instagram, etc., there has been a trend in big data community researches to engage themselves in social network studies, as these networks have proven to be the excellent sources of hidden information patterns [1–3]. Extensive studies have been conducted on identifying characteristics and heuristics of these network types, by transforming these datasets into a network graph. Undoubtedly, network theory enabled researchers to address many real time complex problems such as product recommendations in e-commerce, friendship in social networks, computing the personalized PageRank between two nodes in quick time, web page ranking for efficient information retrieval, etc. Of these, one of the major concerns for network theory researchers is to find the optimal solutions to maximize node interactions and develop algorithms with minimal time complexity. For example, instead of sending news about an offer to all the actors in the network, a set of precisely chosen actors capable of efficiently spreading the information, can be identified as information propagators to ingest the offer message into the network. For this purpose, finding popular or most influential nodes in the network has proven to be helpful [4]. Such

nodes can act as a hub or a medium to optimize internal as well as external communication. In the same way, one can get to know about the latest trending topics in the network, by identifying the actors who are capable of receiving the information from multiple actors in the network.

The trends of searching most influential nodes are not just popular in social networks but also in many other fields. For instance, in Biology, there are many biological networks, such as Protein–protein interaction networks, Cell signaling networks, Gene regulatory networks, etc. where identifying the influential nodes plays a vital role in a new discovery. Recent research on these areas [5–7] have used various forms of topological centralities such as weighted sum of loads Eigenvector centrality (WSL-EC) [5], Motif-based centrality [7], etc. to either capture the important proteins or identify important features of the genes. Using concepts from graph theory such as cliques formation, centralities, etc. along with data mining algorithms like K-means, Random Forest, Naive Bayes, etc., many scientists have been successful in identifying proteins involved in many life-threatening diseases: cancers, AIDS, and many others.

Generally in specific types of research, the researchers have a tentative idea on what they are looking for and what they want to get as the output. During these types of studies, the researchers are concerned for a specific set of nodes instead of the entire node lists. However, analyzing the entire list of nodes is time-consuming, when only particular set of nodes having particular characteristics is of interest. In general, researchers are aware of the insights that can be deduced from a dataset by considering the characteristics of the dataset. By utilizing these characteristics, one can eliminate the unwanted lists of data narrows the space on which the top-K nodes query is to be executed. There are various network reduction algorithms such as disparity filter, k-core decomposition, etc. which prune the unwanted edges on the basis of certain filter functions. Similarly, we can apply some filtering strategies on nodes so that the focus is only on the desired lists of nodes.

In this paper, we propose a novel method for identifying top-K viral information propagators from a reduced search space. Our algorithm is based on network topology and constraints. The algorithm starts by identifying the user-defined constraints, and applies those constraints on the nodes to extract only those nodes that satisfy them. Giving an option to filter out unwanted lists of data (which in our case is out-of-the-scope nodes) will make the important nodes or top-K nodes much more efficient and desirable. Katz centrality was used as a measure of topological centrality that helps to discover the relative influence of each node on the network. Given the global Katz centrality, users were required to provide the desired centrality for initial filtering of the nodes. Once the candidate nodes' list was collected, the top-K nodes were identified based on their local influence (i.e., local Katz centrality) and on a global scenario (i.e. global Katz centrality). Our proposed algorithm is tested against four datasets and the results show that the proposed algorithm is capable of reducing the number of nodes in search space at least by 70%. We also considered the parameter (α and β) dependency of Katz centrality values in our experiments and established a relationship between the α values, number of nodes in search space and network characteristics. Later, we compare the top-K results of our approach against the top-K results of degree centrality.

Related work

A significant amount of research has been conducted towards the identification of top-K influential nodes and search space reduction in a given network. Centrality theory [8–12], diffusion models [13], heat diffusion theory [14], evidence theory [15] etc., are the most frequently used techniques for obtaining top-K influential nodes in a network.

In a recent paper by Li et al. [15], a method based on evidence theory was proposed to identify influential nodes in a network of networks (NON) by dividing the complex network into sub-networks such as a series of similarity networks. For each of these individual networks, distance matrix (D) is computed which represents the similarity among the nodes. This matrix D is further used to compute similarity networks which in turn assist in finding basic probability assignment (BPA). The nodes with high similarity value in the fused similarity network are considered to be influential nodes in NON.

Kimura et al. [16] came up with a method that used the theory of bond percolation along with graph theory to extract influential nodes. The purpose of their method was to maximize the influence for information diffusion. Their method begins by finding a set of nodes A, for initial activation by using greedy hill-climbing algorithm. Using A, the initial set of nodes and deterministic diffusion model on Gr, F(A, Gr) is computed, where F(A, Gr) is the final set of active nodes.

Doo et al. [14] came up with an activity-based social influence model, using the concept of heat diffusion for measuring the influence diffusion in networks. Every interaction between any two nodes are labeled as either heat diffused $DH_i(\delta t)$ or heat received $RH_i(\delta t)$ based on the direction of an edge. Similarly, activities like comments, likes, shares, etc between any two nodes, v_i and v_j can be classified as interactive activities (IA_{ij}) and activities like status update, photo uploads, etc are classified as non-interactive activities (NIA_{ij}). Each of these activities is assigned a weight of 1 unit. The higher the number of non-interactive activities at node v_i, the higher the amount of heat collected at v_i and the slower the heat diffusion to its neighbors. Based on the $DH_i(\delta t)$ and $RH_i(\delta t)$ values, heat diffusion is calculated for each node, v_i and finally influences coverage (IC_i)—the list of nodes which are influenced by v_i are generated. Finally, the top-K nodes are selected on the basis of the $|IC_i|$.

Zhang et al. in [17] and Leung et al. in [18] gave importance to user preferences for identifying the top-K nodes. Zhang et al. in [17] proposed a two-staged mining algorithm (GAUP), for finding the top-K nodes in the social networks by considering the users' preferences. The initial stage involves estimating user preferences with a set of latent items for a specific topic by adopting the Latent Semantic Indexing (LSI) method. These estimations are then used in the second stage, which is based on Extended Independent Cascade Model to maximize the influences on active nodes and then discover a selected set S of top-K nodes.

Leung et al. in [18] proposed an algorithm that reduces the search space based on the user-specified constraints and uses the MapReduce model to discover interesting patterns from uncertain data which satisfies those constraints. The algorithm first mines frequent singleton patterns followed by non-singleton patterns. Map function computes individual existential probability for each item in a transaction. Reduce function then filters the items which satisfy the user-specified constraints and then computes the expected support, *expSup* for each item and compares them with the minimum support,

minSup. Only those items whose *expSup* $\geq$ *minSup*, are selected for the singleton pattern. Using the individual expSup, non-singleton patterns are discovered.

Other works include that of He et al. [19], where top-K nodes are identified using influence maximization strategies in complex networks using community structure and in Liu et al. [20] propose a method to identify the top-K nodes based on a specific topic.

Besides the above-mentioned theories and methods, centrality has also been widely used in research related to network analysis. Centrality approaches are generally classified into two types, classical centrality measures and parameterized centrality measures. Classical centrality measures include Degree centrality, Closeness centrality and Betweenness centrality, whereas Eigenvector centrality, Katz centrality, PageRank centrality, etc are the parameterized centrality measures. Cupertino et al. came up with a network-based method that uses Katz centrality to predict the pattern class the given group of invariant transformations of the same pattern belongs to [21]. Using another measure of network centrality called Principal Component Centrality (PCC), Ilyas et al. [22] identified groups of nodes, *social hubs,* in the network which are at the center of influential neighborhoods. They then compared their results with the ones from the method using Eigenvector centrality (EVC). To further enhance the usage of α-centrality, Ghosh et al. introduced a normalized version of this centrality by generalizing a modularity maximization-based approach. Their method identified not just the local communities but also the global ones [23].

Constraint-based data mining [24] has been widely used for finding frequent items or patterns in a given pool of data [17, 18]. Various types of constraints can be used in mining, such as knowledge type constraints, data constraints, dimension constraints, interestingness constraints, rule constraints, etc. Providing a means to apply certain constraints on the data allows users to be specific in their search, so that only those datasets satisfying the constraints are looked for in the database. These types of searches improve speed and reduce unnecessary computations. At the same time, only favorable and desired outputs are received. Our proposed approach ("using Katz broadcast centrality"), falls under the category of centrality approach and constraint-based mining as the user has control over the choice of threshold used for the first level filtering.

Preliminaries

Graph definitions

Directed graph

A graph $\boldsymbol{G = (V, E)}$ where V is the set of nodes or actors (say "n") and E is the set of edges or connections (say "m") is directed, if E is a set of ordered pairs meaning that $(v_1, v_2) \neq (v_2, v_1)$, where $(v_1, v_2) \in E$ and $v_1, v_2 \in V$.

Undirected graph

A graph $\boldsymbol{G = (V, E)}$ where V is the set of vertices or nodes or points (say "n") and E is the set of edges (say "m") is undirected, if E is a set of unordered pairs meaning that $(v_1, v_2) = (v_2, v_1)$, where $(v_1, v_2) \in E$ and $v_1, v_2 \in V$.

Walk of length l

For a graph $\boldsymbol{G = (V, E)}$, a walk of length l denotes a set of nodes $\{v_1, v_2, v_3, \ldots, v_l\}$ such that there exists an edge between v_i and v_{i+1}, $\forall 1 \leq i < l$.

Centrality measures

Degree centrality

Degree centrality equals the number of ties that a vertex has with other vertices. The equation for it is as follows:

$$C_D(n_i) = d(n_i) \tag{1}$$

where $d(n_i)$ is the degree of node n_i.

In a directed network, there are two separate measures of Degree centrality, namely in-degree and out-degree. In-degree is a count of the number of ties directed to the node and out-degree is the number of ties that the node directs to others. Accordingly, the equations for them are as follows:

$$C_{in}(v_i) = d_{in}(v_i) \tag{2}$$

$$C_{out}(v_j) = d_{out}(v_j) \tag{3}$$

where $d_{in}(v_i)$ (in Eq. 2) and $d_{out}(v_j)$ (in Eq. 3) are the corresponding in-degree and out-degree centralities of nodes v_i and v_j.

Closeness centrality

Closeness centrality of a node is the average length of the shortest path between the node and all other nodes in the graph. It can be regarded as a measure of how much time it takes to spread information into the network from a given vertex. It can be used to identify nodes which are capable of quickly spreading a rumor into the network. The equation for it is as follows:

$$C_c(v_i) = \sum_{j=1}^{N} \frac{1}{d(v_i, v_j)} \tag{4}$$

where C_c is the closeness centrality of a node v_i.

Betweenness centrality

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Being between means that a vertex has the ability to control the flow of knowledge between other nodes in the network. Linton Freeman introduced this and as per his conception [25], vertices that have a high probability to occur on a randomly chosen shortest path between two randomly chosen vertices have a high betweenness. The betweenness of a node v_i is given by the formula in Eq. (5).

$$C_B(v_i) = \sum_{j=1,k\neq 1} \frac{g_{jik}}{g_{jk}} \tag{5}$$

where g_{jik} is all geodesics linking node j and node k which pass through node i; g_{jk} is the geodesic distance between the vertices of j and k.

Eigenvector centrality

Eigenvector centrality is an extension of degree centrality. In the degree centrality, all node connections are credited of equal importance. But in real life, each node may have different importance. For example, a node connected to highly important nodes itself is an important node. Thus, Eigenvector centrality provides a relative score to each node depending on the type of nodes (high-scoring and low-scoring) it is connected to. For a given graph $G = (V, E)$ containing n nodes, let A be the adjacency matrix of G and λ be the eigenvalue. Then Eigenvector centrality is given by:

$$C_e(v_i) = \frac{1}{\lambda} \sum_{j=1}^{n} d_i a_{j,i} C_e(v_j) \tag{6}$$

Katz centrality

Katz centrality measures the relative influence of each node in a given network by taking into account it's immediate neighboring nodes as well as non-immediate neighboring nodes that are connected through immediate neighboring nodes. The Katz centrality of a node v_i is computed as:

$$C_{Katz}(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} C_{Katz}(v_j) + \beta \tag{7}$$

where α is a constant called the damping factor, usually considered to be less than the largest eigenvalue, λ i.e. $\alpha < 1/\lambda$ and β is a bias constant, also called the exogenous vector, used to avoid the zero centrality values. With $\alpha \geq \lambda$, the centrality tends to diverge.

Besides these centrality measures, there are PageRank centrality, Sub-graph centrality, Evidential centrality and Total communicability etc. According to the concept of centrality, a node is identified as "important"?, if its centrality value is higher than that of other nodes in the network. Many papers have been written to show co-relations among various centrality measures and to date, research has been carried out to answers these questions: which centrality measure is best for obtaining the top-K influential nodes in a given network, which centrality measure is the best fit for a given type of data etc. Yan et al. [8] studied the correlation between degree centrality, closeness centrality, betweenness centrality and PageRank centrality in a co authorship network. Benzi et al. [9] studied the correlation between sub-graph centrality and total communicability.

Discussion on Katz centrality for top-K node analysis

In our proposed algorithm, we use Katz broadcast centrality and local average centrality (LAC) for obtaining top-K influential nodes in a given network. Based on Katz broadcast centrality and LAC, all the nodes are ranked in the order of importance. Katz broadcast centrality captures the behavior of spreading a rumor into the network and a high value of Katz broadcast centrality means that the given node is efficient in spreading out a rumor/marketing message into the network. The concept of using Katz centrality to rank the actors in a social graph was first proposed by Katz [26]. The very fact that a human's influence in his/her social group decreases as one moves further from his/her close connections to loosely connected distant members forms the base of Katz

centrality. Moreover it is also proven that Katz centrality is computationally efficient for filtering out the most central nodes, especially in the case of large directed networks [18]. Katz centrality of a node i counts all walks beginning at node i, instead of the usual shortest path approach, such that the longer walks are penalized through the attenuation factor α. The immediate neighbors, i.e. walk of length 1, are given the value α^1, whereas the farther neighbors, i.e. walk of length k, are assigned as α^k with the notion that k-step walk has α^k probability of being effective. Thus, farther the neighbors from the node in consideration, lesser is its influence on them.

$$(I - \alpha A)^{-1} = I + \alpha A + \alpha^2 A^2 + \cdots + \alpha^k A^k + \cdots = \sum_{k=0}^{\infty} \alpha^k A^k, \quad 0 < \alpha < 1/\lambda \tag{8}$$

Equation (7) can be generalized for the entire graph as [28]:

$$C_{Katz} = \beta(I - \alpha A^T)^{-1} \cdot \mathbf{1} \tag{9}$$

where $\mathbf{1}$ is a column vector of ones.

From the Eq. (8) it is also evident that Katz centrality is a parameter dependent index, i.e. it depends on α and β. Their values play a decisive role in getting fluctuating Katz centrality values. Benzi et al. [27] in their paper, showed that different choices of α and β lead to different centrality values. For instance, if $\alpha \to 0+$, then Katz centrality reduces to degree centrality. The degree centrality of a node i gives importance to connections that are one step away starting from i. If $\alpha \to (1/\lambda)-$, then it reduces to Eigenvector centrality, for example, if $\alpha = (1/\lambda)$ and $\beta = 0$, then Katz centrality becomes equal to Eigenvector centrality. In short, the degree centrality of node i measures the *local* influence of a node and the Eigenvector centrality measures the *global* influence of a node within the network. On the other hand, Katz centrality covers both the *local* and *global* influence of i. Hence, these parameters can be taken as a medium to tune between the rankings of nodes based on either local influence (short walks) or global influence (long walks).

In the case of a directed network graph, there are two centrality measures, which are Katz broadcast centrality and Katz receive centrality. Let $\boldsymbol{G = (V, E)}$ be a strongly connected, directed and unweighted network representing actors and their ties, represented by adjacency matrix $\boldsymbol{A}$.

Katz broadcast centrality

Katz broadcast centrality of a node i is calculated as:

$$Katz_i^b = \beta(I - \alpha A)^{-1} \cdot \mathbf{1} \tag{10}$$

Katz receive centrality

Katz receive centrality of a node i is calculated as:

$$Katz_i^r = \beta(I - \alpha A^T)^{-1} \cdot \mathbf{1} \tag{11}$$

Clearly from the Eqs. (9) and (10), it is evident that we are considering row sums to obtain the outboundness of a node and column sums to obtain the inboundness of a node. Graph theory notations and symbols are summarized in Table 1.

Table 1 Notations and symbols

Notation	Definition and description
G	Given graph
V	Set of nodes
E	Set of edges
v_1, v_2, v_i, v_j, n_i	Nodes
C_D	Degree centrality
C_{in}	In-degree centrality
C_{out}	Out-degree centrality
C_c	Closeness centrality
C_B	Betweenness centrality
g_{jik}, g_{jk}	Geodesic paths
α	Damping factor or attenuation factor
β	Exogenous vector
λ	Eigenvector
C_{Katz}	Katz centrality
A	Adjacency matrix
LAC_{Katz}	Local average centrality
GAC_{Katz}	Global average centrality
Const	User defined constant
$isim_k$	Intersection similarity

Algorithm

In this section, we will discuss our algorithm to find the top-K influential nodes. For a given network data, first the Katz broadcast centralities of all the nodes are computed. The nodes are then ranked in descending order of their centrality values. On the basis of their rankings, the least important/influential nodes from a centrality point of view are eliminated in order to reduce the size of search space, by using filtering constraints. The first filtering constraint is user defined and it can be varied as per the users' choice. This constraint tests whether the Katz centrality of a node is greater than the user defined threshold value or not. The second constraint is tested for only those nodes, which satisfy the first or user defined constraint. The second filtering constraint tests whether the average of centrality values of a node and its immediate neighbors (LAC) is greater than the average centrality of the all the nodes in the network. Nodes, which satisfy both the constraints are included into the search space. The first filtering constraint helps the users to focus only on the nodes of interest, while the second constraint further refines the set of nodes which satisfy the first constraint. Thus a finely refined set of nodes are returned to the user for executing the top-K query. The first filtering constraint denoted as *Const* keeps the users' in control on the choice of nodes they are interested in and it shows the succinctness property. While the second filtering constraint prioritizes the nodes with more number of immediate highly connected nodes.

Succint set: A set is said to be a succinct set if it is formed as result of a selection operation, $\sigma_{a\theta b}$ where a and b are attributes and θ, a binary operation [18, 24].

The average centrality values of a node and its neighbors is denoted as LAC_{Katz} (local average centrality) and the average centrality value for the entire network is denoted as GAC_{Katz} (global average centrality).

$$LAC_{Katz}(v_i) = \frac{C_{Katz}(v_i) + \sum_{j=1}^{n_i} C_{Katz}(v_{ij})}{n_i + 1} = \frac{LSC_{Katz}(v_i)}{n_i + 1} \tag{12}$$

$$GAC_{Katz} = \frac{\sum_{i=1}^{n} C_{Katz}(v_i)}{n} \tag{13}$$

where n_i is the number of neighbors of a node, v_i and n, the total number of nodes in the network.

We present our approach in thre different sections. In Algorithm 1, Katz broadcast centrality values are calculated. Algorithm 2 shows the identification of neighboring nodes for each node. In Algorithm 3, it returns the reduced search space for identifying the top-K nodes.

Algorithm 1 Compute Individual Katz Centralities

```
Read the network input file
Form Adjacency matrix, A of the network
User desired values for α and β
for each node v_i ∈ V do
    Calculate Katz Centrality, C_Katz(v_i)
    Return each node, v_i and its respective C_Katz(v_i)
end for
```

Algorithm 2 List of Neighbors' Centralities

```
for each node v_i ∈ V do
    Find a list of its neighbors v_ij and their C_Katz(v_ij)
    Return v_i and list of its neighbors' C_Katz(v_ij)
end for
```

Algorithm 3 Extract Top-K Nodes

```
Const ← User Desired Katz Centrality for filtering purpose
GAC_Katz ← Σ_{j=1}^{n} C_Katz(v_j)/n
for each (v_i, listof(v_ij, C_Katz(v_ij))) do
    if v_i satisfies Const then
        set LAC_Katz(v_i) ← Katz(v_i)
        set LSC_Katz(v_i) ← 0
        set ngbrCount ← 0
        for each v_ij ∈ listof(v_ij, C_Katz(v_ij)) do
            LSC_Katz(v_i) ← LSC_Katz(v_i) + C_Katz(v_ij)
            ngbrCount ← ngbrCount + 1
        end for
        LAC_Katz(v_i) ← LSC_Katz(v_i)/(ngbrCount + 1)
        if LAC_Katz(v_i) ≥ GAC_Katz then
            Return v_i and its C_Katz(v_i)
        end if
    end if
end for
```

Working of the algorithm on karate club dataset

The karate club dataset shows data related to friendships among members of a karate club at a US university in the 1970s. This undirected network dataset consists of 34 nodes and 78 edges. This dataset is obtained from Mark Newman network datasets repository [29].

The visualization in Fig. 1 was created using D3.js (Data Driven Documents) [30].

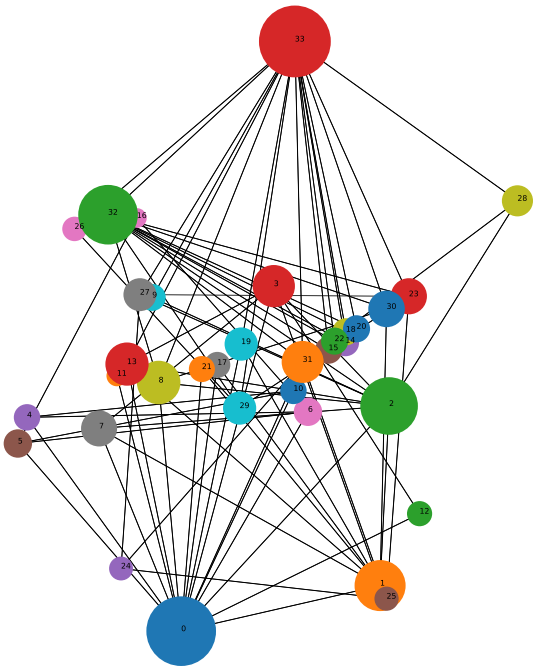


Fig. 1 Karate club membership network

Let **A** represent the adjacency matrix of this network. The largest eigenvalue λ obtained for A is 6.725 and to satisfy the constraint that $\alpha < \frac{1}{\lambda}$, α values chosen are less than 0.148. We also tested the algorithm with various values for α, while keeping β value constant as 1. Constraint *Const* denotes the user's interest in finding only those nodes whose Katz centrality (C_{Katz}) is greater than or equal to *Const*. Here, *Const* shows the succinctness property and its value which is used for filtering out the nodes/actors in the network, is chosen as the sum of standard deviation and average of Katz centralities of all the nodes in the network. As mentioned earlier, the second filtering constraint is the average centralities of all the nodes in the network. The user can run the top-K node query from the set of nodes obtained after filtering. For all the datasets, we used the same formula as the first filtering constraint, kept β value as 1 and varied the α values and studied the effect of varying α values coupled with two sets of filtering contraints on the number of nodes in search space and ordering of the nodes obtained upon querying for top-K nodes. Below table captures the number of nodes obtained by varying α values.

Figure 2, clearly depicts that the filtering condition is effective in filtering the nodes, thereby the user has more control over the search space from which the top-K nodes are retrieved. The top-5 nodes obtained using the proposed algorithm are 33, 0, 32, 2, and 1. The top node is node 33, which corresponds to the president of the karate club, and the second is node 0, which corresponds to the instructor. These were the two most influential members of the club and their fight with each other eventually led the club to split into two factions aligned around each of them [31].

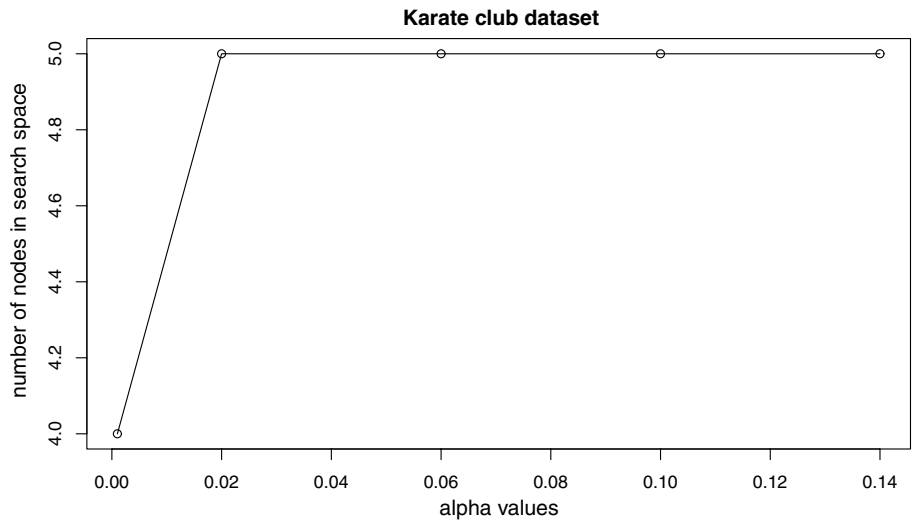


Fig. 2 Figure showing the relationship between α values and the number of nodes in search space

Experimentation

For the implementation of our approach, the code was written in Java using Jblas [32] and Graph-stream [33] packages. The algorithms were written using the java data structures like Lists and Hashmaps.

Experimental setup

The code was run on 64-bit Windows 7 Enterprise (Server Pack 1) with Intel(R) Xenon(R) CPU E5-1607 0 with 3.00GHz and 16GB RAM. Eclipse IDE with Java 1.7 was used for programming.

Datasets

The networks' data, as shown in Table 2, were collected from Mark Newmann datasets [29], SNAP Stanford Large Network Database Collection [34] and ILAB-Data Centre [35].

Facebook dataset

Facebook is an online social networking platform [36], where nodes represent the users and edges represent the relationship among the users. Our Facebook dataset consists of 1034 nodes and 53,498 edges. The largest eigenvalue of the network is ≈123.215. Keeping the fact that the value of α should be less than $\frac{1}{\lambda}$ (0.008 in this case) in mind, the

Table 2 Network dataset summary

Dataset	Type	Number of nodes	Connectivity
Facebook-I	Undirected	1034	53,498
CA-GrQc	Undirected	5242	14,496
Epinions-I	Directed	1247	51,558
Epinions-II	Directed	1799	61,037

values for the parameter α values are varied as 0.0005, 0.001, 0.0015, 0.002, 0.0025, 0.003, $0.0035 \cdots 0.008$.

Collaboration network dataset

CA-GrQc dataset covers the scientific collaboration between authors' papers submitted to General Relativity and Quantum Cosmology category from January 1993 to April 2003 (124 months). If an author **i** co-authored a paper with author **j**, the graph contains an undirected edge from i to j. This dataset consists of 5242 nodes, 14,496 edges and the largest eigenvalue is $\approx$45.616. As, the value of α should be less than 0.021, α values are varied as 0.005, 0.01, 0.015 and 0.02.

Epinions network datasets

Epinions.com [37] is a who-trusts-whom social network of general consumer review site. Members of the site can decide whether to "trust" each other or not. A Web of Trust is formed basing up all the trust relationships interactions and then combined with review ratings to determine which reviews are shown to the user. Epinions-I and Epinions-II are directed network datasets. Epinions-I dataset consists of 1247 nodes and 51,558 edges. Epinions-II dataset consists of 1799 nodes and 61,037 edges. As, the largest eigenvalues of the two networks are $\approx$83.751, α values should be less than 0.011. For both the datasets, α values are varied as 0.001, 0.004, 0.007, and 0.011 and results are analyzed.

Results and discussion

For all the datasets, we analyzed the results from two different perspectives. Firstly, we analyzed the relationship between α values, number of nodes obtained for each α value and dataset characteristics. Secondly, we compared the top-K results of our algorithm with the top-K results of degree centrality algorithm using intersection similarity as a measure and analyzed the significance of our algorithm. Intersection similarity (Intersection distance) captures the notion of union minus the intersection. Previously, Benzi et al. used intersection similarity measure in their research in [9].

Let x_k and y_k be the top-K ranked items in two ranked lists x and y respectively. Then, the top-K intersection similarity can be computed as:

$$isim_k(x, y) := \frac{1}{k} \sum_{i=1}^{k} \frac{| x_i \Delta y_i |}{2i} \tag{14}$$

where Δ is the symmetric difference operator between the two sets. If the lists are identical, then $isim_k$(x, y) = 0 for all k. If the two sequences are disjoint, then $isim_k$= 1 [9].

Figure 3a, shows the relationship between α values and the number of nodes in search space for Facebook dataset. The number of nodes in search space followed an increase–decrease pattern for α values between 0.0005 and 0.004. For α values between 0.004 and 0.008, the number of nodes in search space increased with an increase in α values. On the whole, there has been an increase in the number of nodes, with an increase in α value. Figure 3c, shows the relationship between α values and the number of nodes in search space for CA-GrQc dataset and unlike Facebook dataset, the number of nodes in search space decreased with an increase in *alpha* values (on the whole). Figure 3b and d

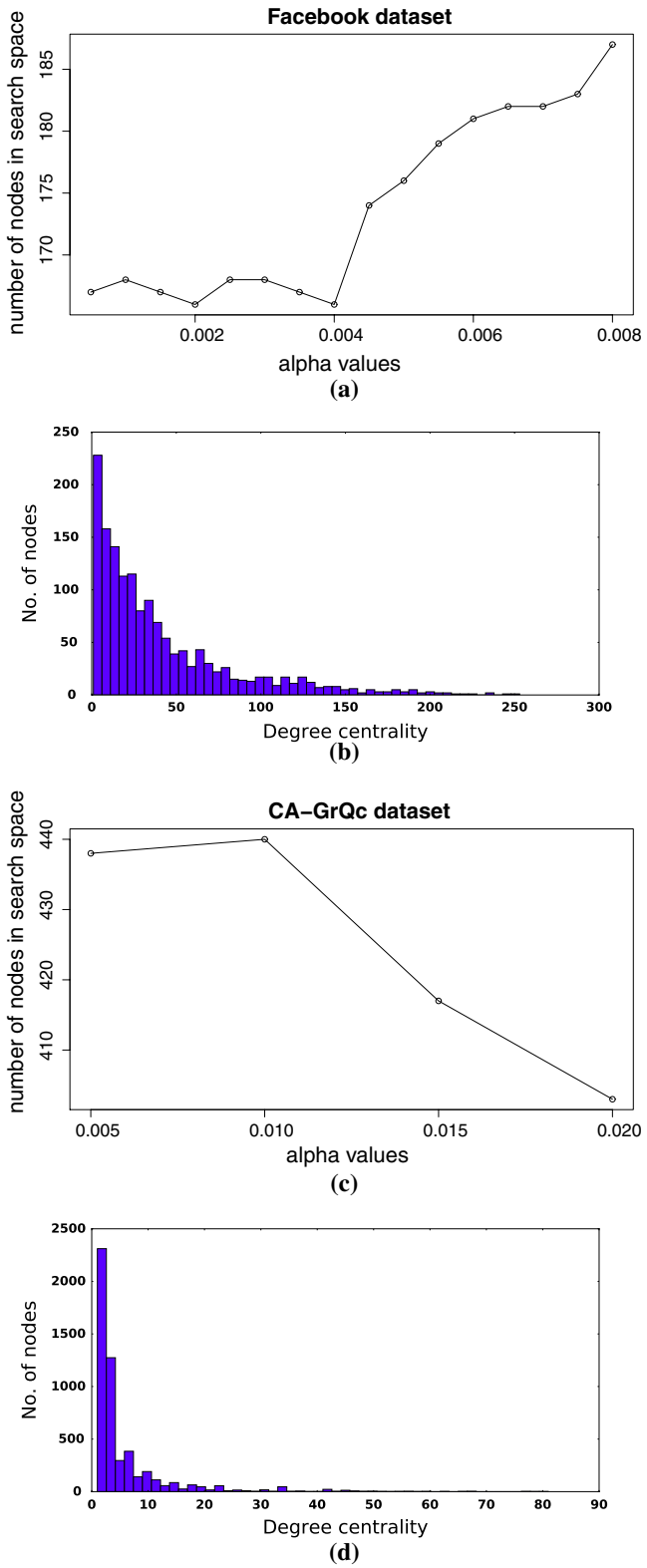
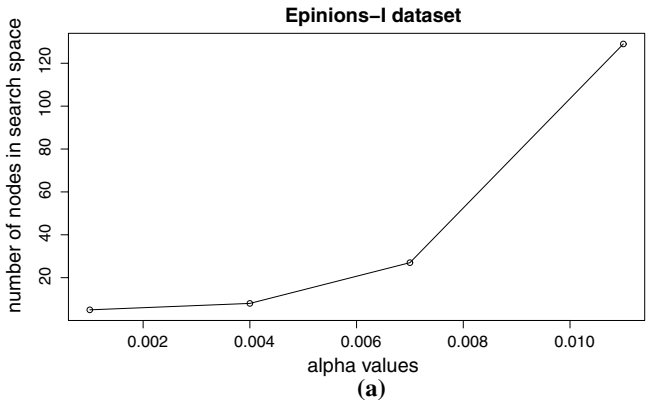


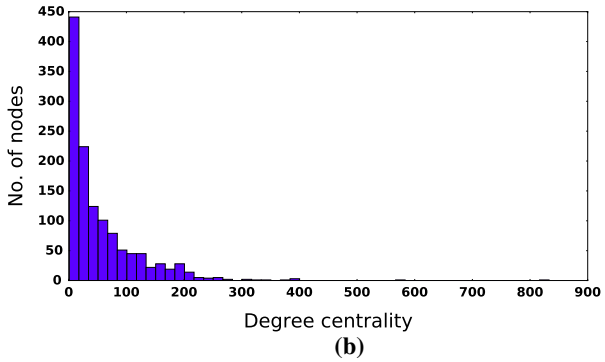
Fig. 3 Experimental results of undirected datasets. **a** α vs number of nodes in search space for Facebook dataset. **b** Degree centrality distribution of Facebook dataset. **c** α vs number of nodes in search space for CA-GrQc dataset. **d** Degree centrality distribution of CA-GrQc dataset

shows the degree distribution frequencies of Facebook and CA-GrQc datasets respectively. For Facebook dataset, a large number of nodes have small degree values or no degree values and yet there are a considerable number of nodes with high degree values. But, in case of CA-GrQc dataset a very large number of nodes have smaller degree values and the number of nodes with high degree values are negligible when compared to this. As mentioned before, Katz centrality is a measure which captures both the local and global influences of a node. If the value of $\alpha \rightarrow 0^{+}$, then Katz centrality is approximately equal to that of degree centrality. And as α values start moving from 0^{+} to $\frac{1}{\lambda}^{-}$ Katz centrality values starts capturing the global influences of the nodes as well. As there are a very less number of nodes with high degree values, compared to the number of nodes with smaller degree values in case of CA-GrQc, the number of nodes that can exhibit local and global influence are very less than the number of nodes which can exhibit local influence (as $\alpha \rightarrow 0^{+}$). Hence, there is a decrease in the number of nodes in search space with an increase in α value. The converse of this can be observed in case of Facebook dataset, where there are a considerable number of nodes with higher degree values in comparison to those with a smaller degree values. Figure 4a and c show the relationship between α values and the number of nodes in search space for Epinions dataset. It can be seen that there is an overall increase in the number of nodes in search space with an increase in α value. The degree distribution frequencies of Epinions dataset, in Fig. 4b and d, are similar to that of Facebook dataset in Fig. 3b. Hence, the relationship between α values and the number of nodes in search space is similar to that of in Facebook dataset.

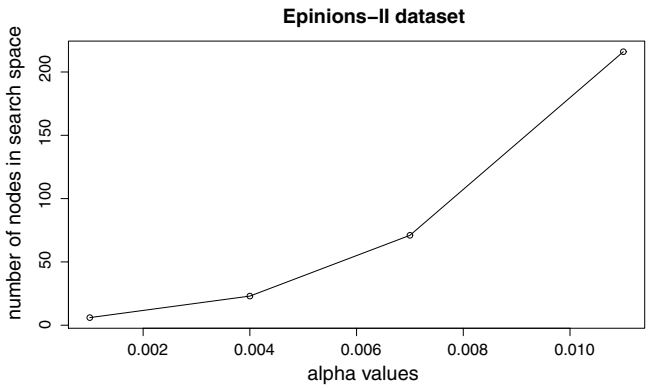
Figure 5a–d show the intersection similarity values for top-K nodes between degree centrality and our algorithm. For Facebook dataset, intersection similarity values are computed for top 20, 50, 80, 100, 120, 150 and 166 nodes, with α value as 0.004. It can be observed from Fig. 5a, that in all of the cases, intersection similarity values are around 0.2. For CA-GrQc dataset, intersection similarity values are computed for top 40, 140, 240, 340 nodes, with α value as 0.015. It can be observed from Fig. 5b, that in all the cases, intersection similarities are around 0.4. For Epinions-I dataset, intersection similarities are computed for top 5, 10, 15, 24, 25 and 27 nodes, with α value as 0.007. Intersection similarity values are increased with an increase in k value, with a maximum values around 0.6. For Epinions-II dataset, intersection similarities are computed for top 20, 40, 60 and 71 nodes, with an α value as 0.007. On the whole, the intersection similarity values are around 0.8. Except Facebook dataset, experiments performed on the other datasets show that, intersection similarity values are more than 0.4. This highlights the fact that there is a significant difference in the rankings produced by degree centrality measure and our algorithm. Moreover, the top-3 nodes obtained in each case are same, but there is a considerable difference in rankings of the remaining nodes as our concept of giving importance to a node, based on LAC proved to give importance to nodes with high local and global influence rather than nodes with high degree values. This confirms that the results obtained by using degree centrality and our algorithm are different and both the approaches capture different perspectives in giving importance to nodes. Also, our approach gives more power to the user in choosing the parameters and narrowing the search space for running the top-K query. These results support our algorithm as a new method for ranking nodes in a given network.



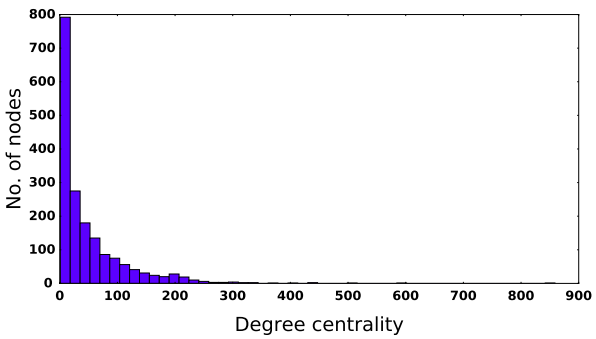
(a)



(b)



(c)



(d)

Fig. 4 Experimental results of directed datasets. **a** α vs number of nodes in search space for Epinions-I dataset. **b** Degree centrality distribution of Epinions-I dataset. **c** α vs number of nodes in search space for Epinions-II dataset. **d** Degree centrality distribution of Epinions-II dataset

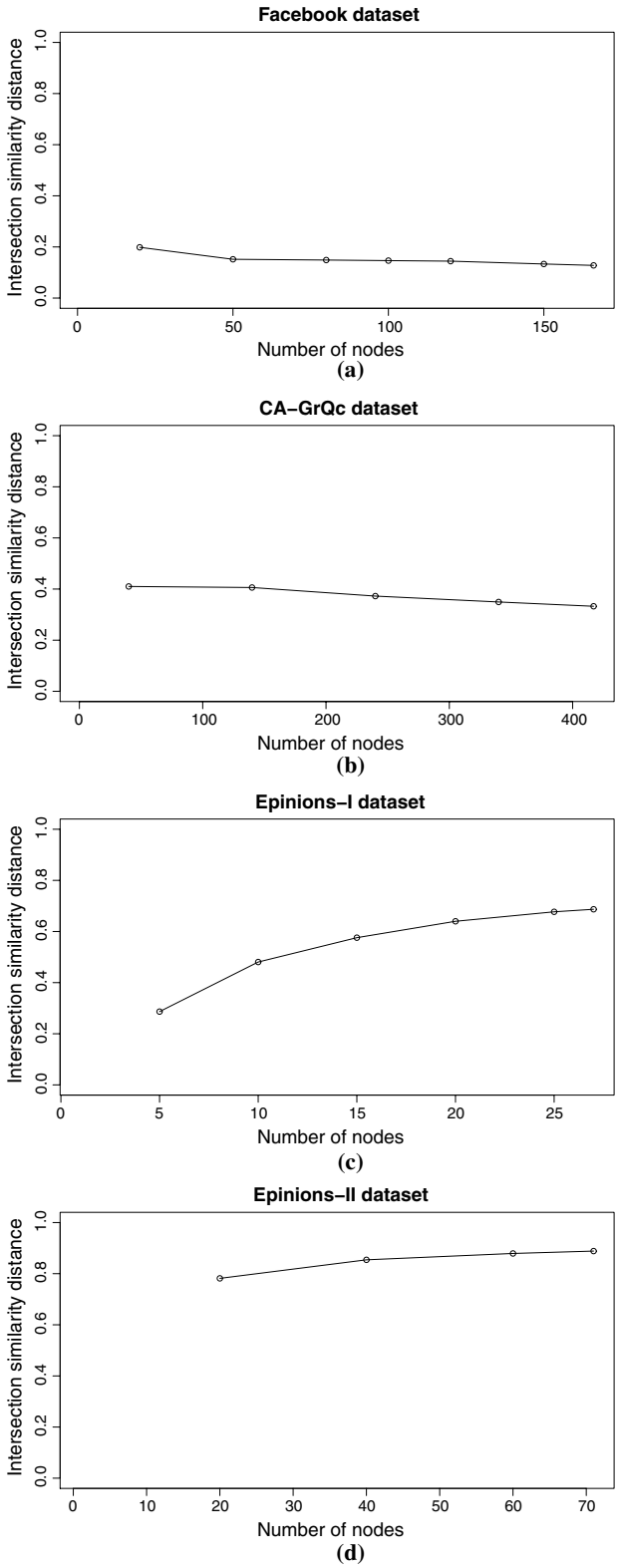


Fig. 5 Intersection similarity distances between Katz centrality and degree centrality. **a** Intersection similarity distances for Facebook dataset. **b** Intersection similarity distances for CA-GrQc dataset. **c** Intersection similarity distances for Epinions-I dataset. **d** Intersection similarity distances for Epinions-II dataset

Conclusion and future work

With growing interest in finding the most important nodes, centrality measures have been one of the sought after methods for this purpose. Our algorithm uses Katz centrality measure to discover the top-K nodes in the network. The centrality is computed using the user-preferred values for the first filtering constraint and additionally the user can choose the values of α and β. The first level of filtering constraints is controlled by the user, denoted by *Const*. The second level of filtering is done by filtering out those nodes which have higher degrees and a mixture of neighbors with much higher and much lower centralities. While the first filtering constraint can be varied by the user, the second filtering constraint is constant for a given network and varies from network to network. This is done by using the formula, LAC_{Katz} greater or equal to GAC_{Katz}. Our experimental results show that the number of nodes obtained in search space are decreased by a factor of more than 75%. This shows the effectiveness of filtering constraints in our approach. We performed experiments with various α values and analyzed the relationship between the number of nodes in search space, α values and network characteristics such as degree distribution. These experiments aid the user in choosing α value for running the algorithm. Also, we showed that there is a significant difference in rankings produced in both the approaches and this enables the users to capture two different perspectives while studying the properties of top-K nodes. Using the centrality measures gives only the topologically important nodes. But there are various other factors that affect the nodes' importance, such as how active they are, the different types of activities they perform, their overall performances, etc. Our future work will be related to incorporating activity analysis along with centrality measures in finding the top-K nodes.

Authors' contributions

SG, as the first author, performed the primary literature review, data collection and experiments, and also drafted the manuscript. JZ and SPKP worked with SG to develop the algorithm, the paper, and the framework. All authors read and approved the final manuscript.

Authors' information

Dr. Justin Zhan is the director of ILAB. He is a faculty member in the Department of Computer Science, College of Engineering, University of Nevada, Las Vegas (UNLV). He was previously a faculty member at North Carolina A&T State University, Carnegie Mellon University and National Center for the Protection of Financial Infrastructure in South Dakota State. His research interests include Big Data, Information Assurance, Social Computing, and Health Science. He is a steering chair of ASE/IEEE International Conference on Social Computing (SocialCom), ASE/IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT), and ASE/IEEE International Conference on BioMedical Computing (BioMedCom). He is currently an editor-in-chief of International Journal of Privacy, Security and Integrity, International Journal of Social Computing and Cyber-Physical Systems, and managing editor of SCIENCE Journal and HUMAN Journal. He has published 180 articles in peer-reviewed journals and conferences and delivered more than 30 keynote speeches and invited talks. He has been involved in a number of projects as a Principal Investigator (PI) or a Co-PI, which were funded by the National Science Foundation, Department of Defense, National Institute of Health, etc.

Sweta Gurung is a recent Computer Science graduate of University of Nevada Las Vegas (UNLV). Her research interests include Graph and Network Theory, Data Mining and Data Analysis. She received her Bachelor of Engineering in Computer Engineering from Kathmandu University, Nepal. Prior to joining the graduate program in UNLV, she worked as Software Developer and Data Engineer in Nepal.

Sai Phani Krishna Parsa is a Master's candidate in the Department of Computer Science at the College of Engineering, University of Nevada, Las Vegas (UNLV). He has received the Bachelor of Technology in Computer Science from Sree Nidhi Institute of Science and Technology, India. He is currently working under the guidance of Dr. Zhan in Big Data Hub-ILAB at UNLV. Previously, he was Systems Engineer at Infosys Limited, the IT bellwether of India.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials
The data used in the experimentation can be found in Ilab [35].

Funding
The research has been funded by United States Department of Defense (DoD Grants #W911 NF-13-1- 0130), National Science Foundation (NSF Grant #1560625), and Oak Ridge National Laboratory (ORNL Contract #4000144962).

Publisher's Note

Received: 27 December 2016 Accepted: 20 May 2017
Published online: 30 May 2017

References

1. Hanneman RA, Riddle M. Introduction to social network methods. Riverside: Department of Sociology, University of California; 2015.
2. Newman MEJ. Networks: an introduction. Oxford: OUP Oxford; 2010.
3. Zafarani R, Abbasi MA, Liu H. Social media mining: an introduction. Cambridge: Cambridge University Press; 2014.
4. Grindrod P. Mathematical underpinnings of analytics: theory and applications. Oxford: OUP Oxford; 2015.
5. Karabekmez ME, Kirdar B. A novel topological centrality measure capturing biologically important proteins. Mol Biosyst R Soc Chem. 2015. doi:10.1039/C5MB00732A.
6. Wang P, Lu J, Yu X. Identification of important nodes in directed biological networks: a network motif approach. PLos ONE. 2014;9(8):e106132. doi:10.1371/journal.pone.0106132.
7. Koschützki D, Schreiber F. Centrality analysis methods for biological networks and their application to gene regulatory networks. Gene Regul Syst Biol. 2008;2:193–201.
8. Yan E, Ding Y. Applying centrality measures to impact analysis: a coauthorship network analysis. J Am Soc Inf Sci Technol. 2010;60(10):2107–18. doi:10.1002/asi.v60:10.
9. Benzi M, Klymko C. Total communicability as a centrality measure. J Complex Netw. 2013. doi:10.1093/comnet/cnt007.
10. Borassi M, Crescenzi P, Marino A. Fast and simple computation of top-k closeness centralities. CoRR, abs/1507.01490. 2015.
11. Bergamini E, Borassi M, Crescenzi P, Marino A, Meyerhenke H. Extracting top-K closeness centrality faster in unweighted graphs. In: Proceedings of the eighteenth workshop on algorithm engineering and experiments (ALENEX); 2016.
12. Aprahamian M, Higham DJ, Higham NJ. Matching exponential-based and resolvent-based centrality measures. J Complex Netw. 2016;4(2):157–76. doi:10.1093/comnet/cnv016.
13. Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network. Theory Comput. 2015;11(4):105–47. doi:10.4086/toc.2015.v011a004.
14. Doo M, Liu L. Extracting top-K most influential nodes by activity analysis. In: IEEE 15th international conference on information reuse and integration (IRI); 2014.
15. Li M, Zhang Q, Liu Q, Deng Y. Identification of influential nodes in network of networks. arXiv preprint http://arxiv.org/abs/1501.05714.
16. Kimura M, Saito K, Nakano R. Extracting influential nodes for information diffusion on a social network. AAAI. 2007;7:1371–6.
17. Zhang Y, Zhou J, Cheng J. Preference-based top-K influential nodes mining in social networks. In: IEEE 10th international conference on trust, security and privacy in computing and communications (TrustCom); 2011.
18. Leung CK-S, MacKinnon RK, Jiang F. Reducing the search space for big data mining for interesting patterns from uncertain data. In: IEEE international congress on big data (BigData Congress); 2014.
19. He J, Fu Y, Chen D. A novel top-k strategy for influence maximization in complex networks with community structure. PLoS ONE. 2015;10(12):e0145283.
20. Liu W, Deng Z, Cao L, Xu X, Liu H, Gong X. Mining top-K spread sources for a specific topic and a given node. IEEE Trans Cybern. 2015;45(11):2015.
21. Cupertino TH, Zhao L. Using katz centrality to classify multiple pattern transformations. In: Brazilian Symposium on Neural Networks (SBRN); 2012.
22. Ilyas MU, Radha H. Identifying influential nodes in online social networks using principal component centrality. In: IEEE international conference on communications (ICC); 2011.
23. Ghosh R, Lerman K. Parameterized centrality metric for network analysis. Phys Rev E. 2011;83(6):066–118.
24. Maimon O, Rokach L. The data mining and knowledge discovery handbook. Berlin: Springer; 2005.
25. Freeman L. A set of measures of centrality based upon betweenness. Sociometry. 1977;40:35–41. doi:10.2307/3033543.
26. Katz L. A new status index derived from sociometric analysis. Psychometrika. 1953;18(1):39–43.
27. Benzi M, Klymko C. On the limiting behavior of parameter-dependent network centrality measures. arXiv preprint, arXiv: 1312.6722; 2015.
28. Benzi M, Klymko C. A matrix analysis of different centrality measures. arXiv:1312.6722v3; 2014.
29. Newman MEJ. Network data. http://www-personal.umich.edu/~mejn/netdata/.
30. Bostock M, Ogievetsky V, Heer J. D3: Data-Driven Documents. IEEE Trans Visual Comp Graphics (Proc. InfoVis); 2011. http://vis.stanford.edu/papers/d3.

31. Zachary WW. An information flow model for conflict and fission in small groups. J Anthropol Res. 1977;33(4):452–73.
32. Braun ML. jblas: Fast Linear Algebra for JAVA. Berlin: TU Berlin; 2010. http://jblas.org. Accessed 12 Sept 2016.
33. Graph-Stream. http://graphstream-project.org/. Accessed 10 Sept 2016.
34. Leskovec J, Krevl A. SNAP datasets: stanford large network dataset collection. 2014. http://snap.stanford.edu/data. Accessed 18 Sept 2016.
35. ILAB: Interdisciplinary Research Institute. http://www.ilabsite.org/?page_id=1088. Accessed 21 Sept 2016.
36. Facebook: online social networking service; 2004. https://www.facebook.com/
37. Epinions. General consumer review site; 1999. http://www.epinions.com/. Accessed 16 Sept 2016.