Identifying and quantifying potential super-spreaders in social networks

Dayong Zhang1, Yang Wang2 & Zhaoxin Zhang2

Quantifying the nodal spreading abilities and identifying the potential influential spreaders has been one of the most engaging topics recently, which is essential and beneficial to facilitate information flow and ensure the stabilization operations of social networks. However, most of the existing algorithms just consider a fundamental quantification through combining a certain attribute of the nodes to measure the nodes’ importance. Moreover, reaching a balance between the accuracy and the simplicity of these algorithms is difficult. In order to accurately identify the potential super-spreaders, the CumulativeRank algorithm is proposed in the present study. This algorithm combines the local and global performances of nodes for measuring the nodal spreading abilities. In local performances, the proposed algorithm considers both the direct influence from the node’s neighbourhoods and the indirect influence from the nearest and the next nearest neighbours. On the other hand, in the global performances, the concept of the tenacity is introduced to assess the node’s prominent position in maintaining the network connectivity. Extensive experiments carried out with the Susceptible-Infected-Recovered (SIR) model on real-world social networks demonstrate the accuracy and stability of the proposed algorithm. Furthermore, the comparison of the proposed algorithm with the existing well-known algorithms shows that the proposed algorithm has lower time complexity and can be applicable to large-scale networks.

Spreading process is a ubiquitous phenomenon in the nature1,2. In most of the real-world networks, the spreading process is the results of the interactions between infected individuals and uninfected individuals. Some spreading processes, including virus propagation and rumor spreading have profound negative economical and social impacts3,4. In order to prevent potential disruptions and ensure social stability, external intervention is widely essential as the most effective way to control epidemic transmission or rumor spreading5,6. Specifically, influential users, which are source spreaders, can obtain higher information diffusion levels on the network than non-influential users7. Thus, the optimized solution for the influence maximization problem is the most common and effective intervention method. However, quantifying the individual spreading abilities in complex networks, especially identifying the potential super-spreaders in a large-scale social network, is still a big challenge nowadays8–11.

The term “super-spreaders” refers to those who are particularly effective in transmitting infectious diseases or spreading information. In epidemiology, a super-spreader is an infected organism that infects disproportionally more secondary contacts than others who are also infected with the same disease12. Similarly, in information diffusion, super-spreaders play a more important role than normal individuals in promoting the information diffusion or determining the emergence of hot topics, including opinion leaders who have the ability to influence others to share and retweet their messages13, information brokers who connect different groups of users or have strong ties with the influential followers14. In fact, identifying individuals with the ability to be super-spreaders during virus propagation and rumor spreading can help us to better prevent the epidemic or public events bursting15,16.

The ranking algorithms were proposed initially to investigate the influence or prestige of individuals in social networks by several social scientists17,18. Nowadays, the ranking algorithms are introduced to study real-world issues. Moreover, they are utilized for novel applications, including optimizing communication networks19, finding social leaders20, and assessing network vulnerability21. However, these conventional centrality indices just
consider a fundamental quantification through combing geodesics between individuals and they are not suitable to describe the pivotal positions of individuals from multiple angles. Thus, some researchers have attempted to redefine the concept of the individual influence. Stephenson and Zelen introduced the concept of the information centrality to capture the information contained in all possible paths of a connected network. Kitsak et al. proposed the k-shell decomposition algorithm to evaluate the nodal spreading capabilities. Wang and Zhao suggested a multi-attribute integrated index based on the degree centrality, the closeness centrality, the clustering coefficient and the topology potential. Sheikhahmadi and Nematabakhsh introduced a hybrid algorithm called the MCDE algorithm. Ahajjam and Badir provided a new centrality for the identification of influential nodes in networks based on the improved coreness centrality and the eigenvector centrality.

Furthermore, with the explosive data growth, designing efficient and effective ranking algorithms on large-scale networks has attracted considerable attentions. Some diffusion algorithms based on random-walk were proposed, including the well-known PageRank, HITS scores, LeaderRank, ClusterRank and other improved PageRank algorithms. These representative algorithms have a common assumption that a node is expected to be of high influence if it points to many highly influential neighbours. So, they work well for directed networks, however they have a poor performance for undirected networks.

In fact, many researches regarding artificial network dataset and real social networks have demonstrated that when investigating the effectiveness of a ranking algorithm, it must be combined with the structural properties of networks and a certain functional goal. For example, identifying influential nodes according to their roles in maintaining the network connectivity or facilitating information flow. In addition, the existing ranking algorithms are difficult to reach a balance between accuracy and simplicity. In other words, some measures perform very simple but limit the accuracy, such as the local centrality indices, whereas others with a high computational complexity perform accurately but are incapable to be applied in large-scale networks, such as the global centrality indices. Thus, in the present study it is intended to fill this gap by exploring a new algorithm to quantify the nodal spreading abilities and identify potential super-spreaders in real-world social networks.

In the present study, a new ranking algorithm named CumulativeRank is proposed to quantify the nodal spreading abilities, which combines the local and global performances of each node in a given social network. Finally, the SIR model is applied to simulate the spreading processes in real social networks. The experimental results show that the proposed algorithm outperforms other well-known ranking algorithms in terms of accuracy and simplicity. The contributions of this study can be summarized as the following:

- The CumulativeRank algorithm is a three-step implementation algorithm that combined the local and global performances of each node.
- The improved network constraint coefficient (INCC) is proposed to assess the local performances of each node.
- The concept of the tenacity is introduced to measure the node’s prominent position in maintaining the network’s connectivity.
- The experimental results verify the outperformance of the proposed algorithm in terms of accuracy and simplicity.

The rest of the paper is organized as the following: In section 2, several widely-used ranking algorithms are reviewed briefly and the SIR model is introduced as the evaluation metric. In Section 3, the proposed algorithm is described in detail. In Section 4, the SIR model is applied to compare the performances of the proposed algorithm and six well-known ranking algorithms in real-world datasets. Finally, conclusions are given in Section 5.

**Background**

**Ranking algorithms.** A social network is a social structure made up of social members and their relationships. From the view of graph theory, most of social networks can be written as a graph \( G = (V, E) \), where \( V = \{v_1, v_2, \ldots, v_n\} \) represents a node set and \( E = \{e_1, e_2, \ldots, e_m\} \) represents an edge set. A node \( v_i \in V \) denotes an individual or organizations in the social network. Moreover, an edge \( e_i \in E \) denotes a possible social interaction between individuals or organizations, including communication or collaboration between members of a social group. The number of elements in \( V \) and \( E \) are presented by \( n \) and \( m \), respectively. Nowadays, with the emergence of web2.0, people can share their opinions, communicate and relate to one another anytime and anywhere. In fact, some individuals play an important role in facilitating information flow and ensuring the stabilization operations of the whole network. In this section, five widely-used ranking algorithms, including degree centrality, betweenness centrality, eigenvector centrality, LocalRank, PageRank and a hybrid algorithm defined by Fu et al. are introduced as benchmark algorithms.

**Degree centrality (DC).** The DC is obtained by calculating the ratio of the number of edges of node \( i \) to the maximum possible number of edges, which reflects the ability of a node to connect directly with other nodes.

**Betweenness centrality (BC).** The BC measures a node’s influence through the ratio of the shortest path over the nodes to the number of all paths. The BC considers the global structure information of a given graph. The higher the BC value of a node, the stronger its controlling or spreading abilities.

**Eigenvector centrality (EC).** Another important index in the category of global centrality indices is the EC, which considers that the centrality of a node depends not only on the number of its neighbours, but also on the centrality of its neighbours.
LocalRank (LR). Most of the global centrality indices have rather high computational complexity, which restricts their applications to large-scale networks. To overcome this problem, Chen et al. [33] proposed a semi-local centrality, named the LocalRank, as a tradeoff between the local centrality indices and the global centrality indices. The LR value of node $i$ is written as the following:

$$LR_i = \sum_{j \in \Gamma(i)} Q_j$$  \hspace{1cm} (1)$$

$$Q_j = \sum_{w \in \Gamma(j)} N_w$$  \hspace{1cm} (2)$$

where $\Gamma(i)$ and $N_w$ are the total number of first-degree neighbours of the node $i$ and the total number of first- and second-degree neighbours of node $w$, respectively.

PageRank (PR). The PR is an application of the random walker model on a Markov chain [24], where nodes are the web pages and edges represent the links from one page to another. The PR value of a page $i$ at $t$ step is described as the following:

$$PR_i^t = a \sum_{j \in \Gamma_i} PR_j^{t-1} \frac{k_{jout}}{k_j} + \frac{1-a}{n}$$  \hspace{1cm} (3)$$

where the damping factor $a \in [0, 1]$ is usually set to be around 0.85. Moreover, $\Gamma_i^{in}$, $k_{jout}$ and $n$ are the sets of pages that link to page $i$, the out-degree of page $j$ and is the total number of pages, respectively.

Fu et al.’s algorithm (FA). Fu et al. [32] proposed a hybrid algorithm that combines global diversity and local features to identify the most influential network nodes. They used the k-shell entropy to measure the global connecting capability of each node and considered the local degree values of its neighbours. The FA value of node $i$ is defined as the following:

$$FA_i = E_i \cdot L_i$$  \hspace{1cm} (4)$$

where $E$ and $L$ are the k-shell entropy value of node $i$ and the sum of all neighbours’ degree centrality values within the two-step distance of node $i$, respectively.

SIR model. Centrality measures provide a method to quantify the nodal influence values, however the numeric values may not be directly interpretable [33]. As the information diffusion is similar to the epidemic spreading, a series of epidemic models are proposed to track the information spreading process and identify the influential spreaders. In the present study, the standard SIR model is utilized to estimate the spreading abilities of selected individuals and illustrate competitive advantages of the proposed algorithm over the conventional well-known algorithms. It should be noted that the standard SIR model assumes that nodes in a network can be in one of three possible states, including susceptible (denoted by S), infected (denoted by I) and recovered (denoted by R). Only a few individuals are set to be infective initially, while other individuals are in the susceptible state. The initial infected individuals are the originators of diseases, which can be obtained by various ranking indices. Once susceptible nodes get in contact with one or more infected neighbours, they become infected with the infection probability of $\beta$. Meanwhile, the infected individuals can be cured with the recovery probability of $\gamma$. The epidemic spreading is repeated until there are no infected individuals in the network and the network reaches a stable state. Since the large infection probability $\beta$ makes the spreading cover almost all of the network, where the role of the individual is no longer important, the $\beta$ value is set to be slightly larger than the epidemic threshold $\beta^* \approx \langle k \rangle / \langle k^2 \rangle$, where $\langle k \rangle$ and $\langle k^2 \rangle$ represent the average degree and the second order average degree [35], respectively. Moreover, the recovery probability is $\gamma = 0.3$.

In order to characterize the spreading abilities of individuals, the spreading scope $F(t)$ at time $t$ is presented as the following:

$$F(t) = \frac{n_{I(t)} + n_{R(t)}}{n}$$  \hspace{1cm} (5)$$

Where $n$ is the number of individuals in a given network. Moreover, $n_{I(t)}$ and $n_{R(t)}$ represent the number of infected and recovered nodes at time $t$, respectively. It should be noted that all of the simulations are carried out on the real social networks. When all infected individuals are converted to the recovered state, the spreading process ends and the final spreading scope $F(t)$ is equal to the maximum values of the recovered individuals. Generally, within the same network, the larger the final spreading scope $F(t)$ triggered by initial spreaders, the stronger their spreading abilities.

The Proposed Algorithm

From the previous section, it is resulted that a reasonable algorithm for the identification of the key individuals should focus on two aspects, including accuracy and simplicity. The accuracy is mainly reflected in the computational accuracy and stability, while simplicity is mainly correlated to the computational runtime. Especially for a social network, which usually contains millions of nodes, the key issue is how to reduce the complexity and improve the computational efficiency. In order to fill this gap, a novel algorithm is required to effectively reach a balance between the accuracy and the simplicity. According to Burt’s structural holes theory [36], the individuals’
structural position in the social network is more important than the corresponding external relationship strength. Certain positional advantages indicate that the individuals occupying these positions have more information, resources and power than others. The positional advantages in the social network include local and global advantages as the following: The former advantages can be quantified by the local structural information9,37, while the latter advantages should consider global topological connections8,38. In this regard, it is intended to propose a three-step algorithm, named the CumulativeRank, in the present study. It is expected that the proposed algorithm can sufficiently combine node’s local and global performances. The details of the proposed algorithm are described as the following:

**Step 1: Quantifying the local advantage of each node.** The structural hole theory provides a novel perspective for understanding the local performance of individuals. In fact, a structural hole is a gap between two unconnected nodes. When these two unconnected nodes are connected by a third node, the bridging node usually has more information advantages and control advantages since it acts as a mediator between different nodes. In order to quantify the control advantage of bridging nodes, Burt introduced the network constraint coefficient (NCC). The NCC for ith node is described as the following:

\[ \text{NCC}_i = \sum_{j \in \Gamma(i)} \left( P_{ij} + \sum_{k=1, k \neq i}^n P_{ik}P_{kj} \right)^2 \]  

(6)

Where \( p_{ij} \) is the proportion of a given node i’s energy invested directly related to node j, which is written as the following:

\[ p_{ij} = \frac{z_{ij}}{\sum_{j \in \Gamma(i)} z_{ij}} \]  

(7)

where \( z_{ij} \) is equal to 1 when a path from i to j exists, and it is zero otherwise. \( \Gamma(i) \) is the set of the nearest neighbours i. \( \sum_{k=1, k \neq i}^n P_{ik}P_{kj} \) measures the strength of the indirect connection from i to j. The NCC value of a node generally has a negative association with its influence in a given network. Therefore, it is found that as the NCC values reduce, the formation of structural holes is enhanced and subsequently the influence of nodes increases.

Equation (6) indicates that the NCC value of a node is calculated based on its neighbourhood topology, including the number of neighbours and the the the corresponding closeness between them. However, the NCC has similar disadvantages to the DC. It only collects information from the nearest neighbours, while the structural information from farther neighbours is ignored. In fact, the NCC is ineffective when it faces the nodes bridging the same number of non-redundant contacts. For example, Fig. 1 illustrates that nodes C and F act as bridges between nodes G and H, and between nodes L and G, respectively. Based on Eq. (6), nodes C and F have the same value (i.e. NCCC = NCCF = 0.46). In other words, the two nodes have the same local influence. However, Fig. 1 shows that in addition to the common neighbour node G and its neighbours, node C has a high-order neighbour like node H, while node F only has one-order neighbour, which is entitled by node L. In fact, it is found that the spreading ability of each node highly depends on its neighbours. For example, although nodes C and F have the same NCC values, node C has stronger spreading ability, which originates from wide range of contacts of high-order neighbours. Therefore, it is concluded that the NCC cannot accurately quantify the difference between nodes C and F in the abovementioned sample network.

This analysis shows that the NCC only collects information from the nearest neighbours, which leads to low resolution. In order to increase the accuracy of the method, more local structural information should be considered. Therefore, an improved network constraint coefficient (INCC) is proposed in the present study. It should be indicated that the INCC scheme is inspired by Chen’s semi-local centrality index33. In the proposed method, the local influence of a node combines the direct and indirect influences on its nearest and next nearest neighbours. Compared with the Burt’s NCC, INCC can provide richer connection information that individual has established, which gives a full understanding of the node influence in facilitating the information flow. The INCC value of node i is defined as the following:

\[ \text{INCC}_i = \sum_{j \in \Gamma(i)} \left( P'_{ij} + \sum_{k=1, k \neq i}^n P'_{ik}P'_{kj} \right)^2 \]  

(8)

Figure 1. An example network consisted of 15 nodes and 19 edges.
and denote the giant component size after destruction and the number of components of the remaining

\[ w_{Gi} \]

is the cost of node \( i \). The basic values of the example network are presented in Table 1. According to Eqs (1) and (2), we can calculate \( Q_A \) and \( Q_B \), respectively. Take node A as an example,

\[
INCC_A = \frac{Q_A}{LR_A} + \frac{Q_B}{LR_B} \times \frac{Q_A}{LR_A}^2 + \left( \frac{Q_A}{LR_B} + \sum_{i \in \{E,F\}} \frac{Q_i}{LR_i} \right)^2 + \frac{Q_E}{LR_E} + \frac{Q_B}{LR_B} \times \frac{Q_A}{LR_A}^2.
\]

As shown in Fig. 1, node C has three nearest neighbours, including nodes B, G and H and seven next nearest neighbours, including nodes A, D, E, F, I, J, and K, thus \( \text{NCC}_A = 10 \). The N values of the other nodes are presented in the third column of Table 1. According to Eqs (1) and (2), we can calculate \( \text{NCC}_C = 24, \text{LR}_C = 24, \text{NCC}_F = 82, \text{LR}_F = 82 \). Similarly, \( \text{NCC}_D = 92, \text{LR}_D = 92, \text{NCC}_H = 71, \text{LR}_H = 71 \). Thus, it is easy to know the difference between node C and F by the corresponding INCC values, where INCC_C = 0.566 and INCC_F = 0.771. INCC_C > INCC_F, that indicates node C has more spreading influence than node F. Table 1 shows that, node H has the lowest INCC value, which indicates that it has the largest local influence in the example network, while nodes L, M, N and O have the largest INCC values, indicating that they have the lowest influence. In a descending order of the nodal influence, the ranking result is H, G, K, C, F, B, A, D, E, I = J, L = M = N = O. It is found that the INCC can more accurately quantify the differences between node C and F, node A and B, node D and E, in comparison to the performance of the NCC method.

Step 2: Quantifying the global advantages of each node. Generally, the influential nodes also play a crucial role in maintaining the network connectivity. If these top influential nodes are removed or not involved in the spreading process, the final spreading scope and the spreading efficiency are reduced. Consequently, the global performances of nodes are considered on maintaining the network connectivity and facilitating the information flow. In general, if the removal of a node leads to the network remaining more components and smaller connected components, the removed node is important in maintaining the network connectivity. The inherent attachment mechanism of social networks often leads to their excessive sensitivity to the removal of key nodes. In order to measure the vulnerability of a given network, Cozzens et al. proposed the concept of the tenacity. As a vulnerability parameter of graph, the tenacity integrates three criteria, including the cost of the network breakage, the number of components and the size of largest connected component.

In this paper, the tenacity are introduced to assess the individual prominent position in maintaining the network connectivity. Due to the inhomogeneity of general networks, most nodes don't belong to the cut-set of a given network, so random removal of these nodes cannot change the balance of the network structure or directly trigger the collapse of the network. Thus, we redefine the tenacity of each node, which is denoted as the following:

\[
R_j = \min \left\{ \frac{|I| + m(G - i)}{w(G - i)} : i \in V, w(G - i) \geq 2 \right\}
\]

(9)

Where \( |I| \) is the cost of node \( i \) removed, when the targeted removal obeys one by one model, \( |I| = 1, m(G - i) \) and \( w(G - i) \) denote the giant component size after destruction and the number of components of the remaining network \( G - i \). In Supplementary Note, an iterative algorithm is proposed for evaluating the nodes’ global performances. If node \( i \) plays a key role in maintaining the network connectivity, it either is the root with at least two descendants, or has a descendant \( u \) whose lowest depth-first number (lfn) is not less than its depth-first number (dfn). According to the above rule, we can identify five key nodes in the example network, which are H, G, K, C and F, respectively. The identification process is shown in Fig. 2.

Step 3: Identifying the influential spreaders in the network. Based on the basic definitions mentioned above, a new ranking algorithm named the CumulativeRank algorithm is presented in this study, which

| j | DCi | Ni | Qi | LRi | NCCi | INCCi | Ri | CRi |
|---|---|---|---|---|---|---|---|---|
| A | 3 | 7 | 21 | 76 | 0.676 | 0.886 | 15 | 0.168 |
| B | 3 | 7 | 24 | 82 | 0.676 | 0.829 | 15 | 0.164 |
| C | 3 | 10 | 24 | 92 | 0.460 | 0.566 | 4 | 0.055 |
| D | 2 | 6 | 15 | 67 | 0.785 | 0.928 | 15 | 0.172 |
| E | 2 | 6 | 15 | 64 | 0.785 | 0.937 | 15 | 0.173 |
| F | 3 | 7 | 18 | 71 | 0.460 | 0.771 | 7 | 0.096 |
| G | 6 | 8 | 43 | 117 | 0.384 | 0.328 | 5.5 | 0.047 |
| H | 4 | 9 | 25 | 71 | 0.406 | 0.323 | 3 | 0.027 |
| I | 2 | 4 | 13 | 38 | 0.953 | 0.994 | 15 | 0.177 |
| J | 2 | 4 | 13 | 38 | 0.953 | 0.994 | 15 | 0.177 |
| K | 4 | 7 | 21 | 46 | 0.25 | 0.365 | 3 | 0.031 |
| L | 1 | 3 | 7 | 18 | 1 | 1 | 15 | 0.178 |
| M | 1 | 4 | 7 | 21 | 1 | 1 | 15 | 0.178 |
| N | 1 | 4 | 7 | 21 | 1 | 1 | 15 | 0.178 |
| O | 1 | 4 | 7 | 21 | 1 | 1 | 15 | 0.178 |
measures the node spreading ability from two layers, including locally, the proposed algorithm considers both
the local influence of nodes on their neighbours and globally, the algorithm considers the nodes’ prominent
position in maintaining the network connectivity. The precise definition of the CumulativeRank is defined as the
following:

\[
CR(i) = \frac{\text{INCC}_i}{\sqrt{\sum_{j=1}^{N} \text{INCC}_j}} + \frac{\text{TC}_i}{\sqrt{\sum_{j=1}^{N} \text{TC}_j}}
\]

(10)

Where INCC\(_i\) is calculated by the Eq. (8). Moreover, TC\(_i\) is the normalized tenacity value of a given node i,
0 ≤ TC\(_i\) ≤ 1 and it is defined as the following:

\[
TC_i = \frac{R_i - \min_{j=1}^{N} \{R_j\}}{\max_{j=1}^{N} \{R_j\} - \min_{j=1}^{N} \{R_j\}}
\]

(11)

According to the above algorithm, nodes with the lowest CR values have the largest influence in facilitating
information flow and maintaining the network connectivity. In a descending order of the nodal spreading ability,
the ranking result of the example network is H, K, G, C, F, B, A, D, E, I, J, L, M, N, O.

**Experimental Evaluation**

**Dataset.** To validate the effectiveness of the proposed algorithm, the algorithm is evaluated in nine real social
networks. All of the data except for OClinks can be downloaded from the Stanford network dataset. Table 2
indicates that the six real social network include: (1) OClinks, which is a representative online community
network, where users are from the University of California, Irvine; (2) Ego-Facebook, which is an ego network
consisting of friends lists from Facebook; (3) Soc-Epinions, which is a who-trust-whom online social network of
a general consumer review site Epinions.com; (4) Wiki-Vote, which is a network containing all of the Wikipedia
voting data from the inception of Wikipedia till January 2008; (5) Ca-HepPh, which is a collaboration network
covering scientific collaborations between authors papers submitted to High Energy Physics from January 1993
to April 2003; (6) Email-Enron, which is an email communication network from Enron posted to the web by
the Federal Energy Regulatory Commission during its investigation; (7) Ca-CondMat, which is a collaboration
network covering scientific collaborations between authors papers submitted to Condense Matter category; (8)
Ca-GrQc, which is a collaboration network covering scientific collaborations between authors papers submitted

| Network   | n     | m     | ⟨k⟩   | L    | C    | H    | βth |
|-----------|-------|-------|-------|------|------|------|-----|
| OClinks   | 1899  | 20296 | 21.375| 3.197| 0.085| 3.773| 0.012|
| Ego-Facebook | 4039 | 88234 | 43.691| 3.693| 0.606| 2.439| 0.009|
| Soc-Epinions | 75879| 508837| 6.706 | 11.549| 0.138| 16.569| 0.009|
| Wiki-Vote | 7115  | 103689| 14.573| 3.341| 0.141| 9.803| 0.007|
| Ca-HepPh  | 12008 | 118521| 19.74 | 4.673| 0.612| 28.21| 0.002|
| Email-Enron | 36692| 183831| 10.732| 4.025| 0.497| 13.265| 0.007|
| Ca-CondMat | 23133| 93497 | 8.0835| 5.352| 0.706| 2.734| 0.045|
| Ca-GrQc   | 5242  | 14496 | 5.531 | 6.049| 0.687| 3.049| 0.059|
| Email-Eu-core | 1005| 25571 | 33.246| 2.653| 0.372| 5.614| 0.005|

Table 2. The basic topological properties of the six social networks, including number of nodes n and edges m
within the networks, average degree ⟨k⟩, characteristic path length L, clustering coefficient C, degree
heterogeneity H = ⟨k²⟩/⟨k⟩², epidemic threshold βth ≈ ⟨k⟩/(⟨k²⟩).
The top-10 ranked nodes by the proposed method and their corresponding ranks by degree centrality (DC), betweenness centrality (BC), eigenvector centrality (EC), LocalRank (LR), PageRank (PR), Fu et al.'s algorithm (FA) and CumulativeRank(CR).

| OLinks | Rank | DC | BC | EC | LR | PR | FA | CR |
|--------|------|----|----|----|----|----|----|----|
| OClinks | 1 | 444 | 55 | 444 | 49 | 35 | 443 | 718 |
| Ego-Facebook | 1 | 4 | 107 | 1912 | 107 | 3457 | 106 | 5437 |
| Soc-Epinions | 1 | 363 | 14 | 14 | 14 | 14 | 15 | 1677 |
| Wiki-Vote | 1 | 363 | 14 | 14 | 14 | 14 | 14 | 1677 |
| Ca-CondMat, as none of the 10 key nodes identified by CR appear in the top 10 nodes for DC, BC, EC, PR and CR.

Experimental results. In general, the super-spreader in a social network is regarded as an individual that has greater spreading abilities. In other words, it is capable of widely spreading the message to other recipient individuals infinitely. In this section, the standard SIR model is utilized to simulate the spreading processes in real social networks.

The ranking order of each node is calculated in the above mentioned six networks initially according to DC, BC, EC, LR, FA and CR, respectively. In a sample network, if nodes have the same calculated scores according to the same ranking algorithm, they will have the same rank. In the present study, we are more interested in the super-spreaders instead of all nodes in the network. Thus, only the top nodes of the ranking list are considered. For this purpose, the top-10 nodes of each ranking algorithm are selected as the initial spreaders and all of the other nodes in the network are marked as susceptible nodes.

Table 3 shows that the ranking results of the proposed algorithm are quite different from those of other ranking algorithms. In Wiki-Vote, it is observed that the proposed algorithm suggests node 326 as the most influential node, followed by the nodes 656 and 488. The same data is not found for the other ranking algorithms. Indeed, DC, BC and LR suggest node 699 as the most influential node, followed by nodes 286 and 2374 (third for DC and LR) or 409 (third for BC). It is indicated that the CR gives the ranking results with the most significant difference from other five well-known ranking algorithms in Ca-HepPh and Ca-CondMat, as none of the 10 key nodes identified by CR appear in the top 10 nodes for DC, BC, EC, PR and CR, respectively.
In addition, it is observed that both the proposed algorithm and the FA algorithm combine the global and local attributes of the network, however there are great differences between the two algorithms in terms of ranking results.

In order to evaluate the performances of the proposed algorithm, the all-contact SIR model is applied to simulate the information spreading process. In the all-contact SIR model each infected node can contact all of its susceptible neighbours at per time step. Figure 3 shows the simulation result of the spreading scopes, $F(t)$, as a function of time for nine networks. For each initial infective node set, the SIR process is repeated 100 times to ensure the stability of the results (see the Supplementary Note). By comparing the changes of $F(t)$ under different spreading sources, it is observed that in almost all of the networks, the initial spreaders obtained by the CR algorithm spread the information faster and the final spreading scopes $F(t)$ always reach the highest value (see Table 4). In the cases of OClinks, Ego-Facebook, Ca-HepPh, Ca-GrQc and Email-Eu-core, it is observed that the performance of the CR is much better than those of other algorithms. Even though in Soc-Epinions, the initial infection sources determined by different ranking algorithms are very similar and the changes of $F(t)$ under different spreading sources perform nearly the same, but the proposed algorithm is slightly better and the final spreading scope still reaches the highest value. Relatively, other six classical ranking algorithms do not show significant consistency in the change tendency of $F(t)$. Although the performance of the EC can outperform the other five algorithms except the CR in Wiki-Vote and Email-Enron, DC and LR perform better than EC in OClinks, Ego-Facebook, Ca-HepPh and Email-Eu-core. Experimental results show that the ranking of the proposed algorithm is more accurate and stable.

![Figure 3](https://example.com/figure3.png)
| Datasets       | Algorithms | Final spreading scopes | Time steps | n  | β  | P     |
|---------------|------------|------------------------|------------|----|----|-------|
| OClinks       | DC         | 0.157                  | 39         | 1899 | 0.02 | 0.0053 |
|               | BC         | 0.14                   | 35         |      |    |       |
|               | EC         | 0.146                  | 35         |      |    |       |
|               | LR         | 0.157                  | 31         |      |    |       |
|               | FR         | 0.138                  | 42         |      |    |       |
|               | FA         | 0.163                  | 40         |      |    |       |
|               | CR         | **0.388**              | **36**     |      |    |       |
| Ego-Facebook  | DC         | 0.383                  | 54         | 4039 | 0.01 | 0.0025 |
|               | BC         | 0.381                  | 52         |      |    |       |
|               | EC         | 0.298                  | 58         |      |    |       |
|               | LR         | 0.302                  | 51         |      |    |       |
|               | FR         | 0.384                  | 40         |      |    |       |
|               | FA         | 0.374                  | 56         |      |    |       |
|               | CR         | **0.396**              | **33**     |      |    |       |
| Soc-Epinions  | DC         | 0.107                  | 41         | 75879 | 0.01 | 0.0001 |
|               | BC         | 0.105                  | 46         |      |    |       |
|               | EC         | 0.106                  | 42         |      |    |       |
|               | LR         | 0.105                  | 43         |      |    |       |
|               | PR         | 0.108                  | 41         |      |    |       |
|               | FA         | 0.105                  | 43         |      |    |       |
|               | CR         | **0.110**              | **40**     |      |    |       |
| Wiki-Vote     | DC         | 0.288                  | 35         | 7115 | 0.01 | 0.0014 |
|               | BC         | 0.292                  | 45         |      |    |       |
|               | EC         | 0.301                  | 42         |      |    |       |
|               | LR         | 0.296                  | 42         |      |    |       |
|               | PR         | 0.292                  | 41         |      |    |       |
|               | FA         | 0.296                  | 36         |      |    |       |
|               | CR         | **0.305**              | **34**     |      |    |       |
| Ca-HepPh      | DC         | 0.124                  | 41         | 12008 | 0.01 | 0.0008 |
|               | BC         | 0.123                  | 61         |      |    |       |
|               | EC         | 0.115                  | 53         |      |    |       |
|               | LR         | 0.119                  | 44         |      |    |       |
|               | PR         | 0.119                  | 55         |      |    |       |
|               | FA         | 0.120                  | 40         |      |    |       |
|               | CR         | **0.129**              | **39**     |      |    |       |
| Email-Enron   | DC         | 0.095                  | 45         | 36692 | 0.01 | 0.0003 |
|               | BC         | 0.102                  | 43         |      |    |       |
|               | EC         | 0.105                  | 40         |      |    |       |
|               | LR         | 0.102                  | 44         |      |    |       |
|               | PR         | 0.101                  | 46         |      |    |       |
|               | FA         | 0.099                  | 44         |      |    |       |
|               | CR         | **0.106**              | **41**     |      |    |       |
| Ca-CondMat    | DC         | 0.355                  | 66         | 23133 | 0.05 | 0.0004 |
|               | BC         | 0.047                  | 56         |      |    |       |
|               | EC         | 0.355                  | 48         |      |    |       |
|               | LR         | 0.349                  | 66         |      |    |       |
|               | PR         | 0.350                  | 63         |      |    |       |
|               | FA         | 0.355                  | 55         |      |    |       |
|               | CR         | **0.365**              | **48**     |      |    |       |
| Ca-GrQc       | DC         | 0.141                  | 49         | 5242  | 0.06 | 0.0019 |
|               | BC         | 0.146                  | 48         |      |    |       |
|               | EC         | 0.167                  | 48         |      |    |       |
|               | LR         | 0.172                  | 51         |      |    |       |
|               | PR         | 0.153                  | 45         |      |    |       |
|               | FA         | 0.168                  | 51         |      |    |       |
|               | CR         | 0.182                  | 41         |      |    |       |

Continued
The SIR model is based on discrete step iterations to demonstrate the spreading process of information. If information cannot continue to flow between nodes according to the model rules, iterations end and the network reaches its stable state. The final time required for the end of the spreading process is obtained from the number of iterations. In general, the higher the number of iterations, the longer information propagates within a given network. Table 4 shows that in the same network, the difference in the spreading sources can increase the number of iterations and prolong the time of information propagation in reaching a steady state. For example, in Ego-Facebook, when $\beta = 0.01$, the spreading sources selected by the CR have the least number of iterations before convergence and the time steps are 33. The PR performs next above the CR, and then LR, BC, DC and EC show ascending order of the number of iterations. In contrast, the CR has the average lesser number of iterations before convergence, however it achieves the largest final spreading scope in all sample networks, which means that the spreading sources selected by the CR effectively accelerate the spreading process of information.

The total runtime of the CR consists of two parts, including the time of computing INCC values for all nodes and the time of quantifying the tenacity of each node. For the former, as $N_v$ requires traversing node v's neighbourhood within two steps and takes $O(k^2)$ time on average, the time complexity is $O(n(k^2))$, where $n$ and $k$ are the total number of nodes and the average degree in a given network, respectively. After each iteration, the number of components $w(G - i)$ and the size of the largest connected component $m(G - i)$ are re-calculated, which takes the complexity $O(n^2)$. Totally, the whole time complexity of the proposed algorithm is $O(n(k^2) + n^2)$. In contrast, the time complexity of the CR is much lower than that of the BC and the EC, which have the complexity of $O(n^2 \log n + nm)$ and $O(n^2)$ respectively. These are close to the PR with the complexity of $O(n^2 + n)$. Analysis of time complexity demonstrates that the proposed algorithm has relatively less computational burden in identifying potential super-spreaders and can be applicable to large-scale networks.

### Conclusion

In the present study, a three-step ranking algorithm, named the CumulativeRank, is proposed in order to identify and quantify potential super-spreaders in a social network. Previous studies have shown that the nodal spreading ability originate from the local prominent position and the total connectivity strength that the node obtains. The proposed algorithm sufficiently combines the node's local and global performances. Locally, inspired by Burt's structural holes theory, the improved network constraint coefficient is proposed based on the semi-local centrality index. Compared with the conventional network constraint coefficient, the improved network constraint coefficient provides richer connection information for evaluating the local performance of each node. Globally, the concept of the tenacity are introduced to evaluate the nodes’ global connectivity strengths. Furthermore, extensive experiments on real-world social networks show explicitly that the proposed algorithm outperforms the existing well-known ranking algorithms and can be applicable to large-scale networks.

### References

1. Goh, K. I. et al. The human disease network. *PNAS* **104**, 8685–8690 (2007).
2. Wang, P., Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A. L. Understanding the spreading patterns of mobile phone viruses. *Science* **324**, 1071–1076 (2009).
3. Funk, S., Gilad, E., Watkins, C. & Jansen, V. A. The spread of awareness and its impact on epidemic outbreaks. *PNAS* **106**, 6872–6877 (2009).
4. Wang, Q. S., Yang, X. & Xi, W. Y. Effects of group arguments on rumor belief and transmission in online communities: An information cascade and group polarization perspective. *Information and Management* **55**, 441–449 (2018).
5. Tian, R., Zhang, X. & Liu, Y. SSI: model for intervention of online rumors spreading. *Physica A* **427**, 181–191 (2015).
6. Jiang, J. & Zhou, T. S. Resource control of epidemic spreading through a multilayer network. *Scientific Reports* **8**, 1629 (2018).
7. Ahijam, S. & Badir, H. Identification of influential spreaders in complex networks using HybridRank algorithm. *Scientific Reports* **8**, 11932 (2018).
8. Kitsak, M., Gallos, L. K., Havlin, S. & Makse, H. A. Identification of Influential Spreaders in Complex Networks. *Nature Physics* **6**, 888–893 (2010).
9. Hu, Y. et al. Local structure can identify and quantify influential global spreaders in large scale social networks. *PNAS* **115**, 7468–7472 (2018).
10. Liu, J., Ren, Z. & Guo, Q. Ranking the spreading influence in complex networks. *Phys. A* **392**, 4154–4159 (2014).
11. Xu, S. & Wang, P. Identifying important nodes by adaptive LeaderRank. *Phys. A* **469**, 654–664 (2017).
12. Liu, Y. et al. Characterizing super-spreading in microblog: An epidemic-based information propagation model. *Phys. A* **463**, 202–218 (2016).
13. Aleahmad, A., Karisani, P., Rahgozar, M. & Oroumchian, F. OLFinder: Finding opinion leaders in online social networks. *Journal of Information Science* **42**, 659–674 (2016).

| Datasets       | Algorithms | Final spreading scopes | Time steps | $n$   | $\beta$ | $P$ |
|----------------|------------|------------------------|------------|-------|--------|-----|
| Email-Eu-core  | DC         | 0.398                  | 34         |       |        |     |
|                | BC         | 0.383                  | 35         |       |        |     |
|                | EC         | 0.357                  | 41         |       |        |     |
|                | LR         | 0.406                  | 37         | 1005  | 0.01   | 0.01|
|                | FR         | 0.362                  | 40         |       |        |     |
|                | FA         | 0.407                  | 37         |       |        |     |
|                | CR         | 0.430                  | 33         |       |        |     |

Table 4. The final spreading scopes $P(t_c)$ for the nine real networks under different ranking algorithms. $n$ is the total number of nodes, $P$ is the ratio of the number of source spreaders and $\beta$ is the infection rate.
36. Burt, R. S.
37. Gao, S.
34. Ibnoulouafi, A. & El Haziti, M. Density centrality: identifying influential nodes based on area density formula.
31. Moreno, Y., Nekovee, M. & Vespignani, A. Efficiency and reliability of epidemic data dissemination in complex networks.
32. Fu, Y. H., Huang, C.Y. & Sun, C.T. Identifying super-spreader nodes in complex networks.
40. Saito, K., Kimura, M., Ohara, K. & Motoda, H. Super mediator-A new centrality measure of node importance for information diffusion over social network.
41. Cozzens, M. B., Moazzami, D. & Stueckle, S. Tenacity of Harary Graphs.
39. Xu, S., Wang, P. & Lü, J. Iterative neighbour-information gathering for ranking nodes in complex networks.
38. Alp, Z. Z. & Ogducu, S. G. Identifying topical influencers on twitter based on user behavior and network topology. Knowledge-based Systems 141, 211–211 (2018).
37. Gao, S.
36. Burt, R. S.
35. Castellano, C. & Pastor-Satorras, R. Thresholds for epidemic spreading in networks. Phys. Rev. Lett. 105, 218701 (2010).
34. Ibnoulouafi, A. & El Haziti, M. Density centrality: identifying influential nodes based on area density formula. Chaos, Solitons and Fractals 114, 69–80 (2018).
33. Chen, D. B. et al. Identifying influential nodes in complex networks. Phys.A 391, 1777–1787 (2012).
32. Fu, Y. H., Huang, C.Y. & Sun, C.T. Identifying super-spreader nodes in complex networks. Math. Probl. Eng., 675713 (2015).
31. Moreno, Y., Nekovee, M. & Vespignani, A. Efficiency and reliability of epidemic data dissemination in complex networks. Phys. Rev. E 69, 055101 (2004).
30. Yu, E.-Y., Chen, D.-B. & Zhao, J.-Y. Identifying critical edges in complex networks. Scientific Reports 8, 14469 (2018).
29. Wang, X., Zhang, X., Yi, D. & Zhao, C. Identifying influential spreaders in complex networks through local effective spreading paths. J.Stat. Mech-Theory & E. 5, 053402 (2017).
28. Rizos, G., Papadopoulos, S. & Kompatsiaris, Y. Multilabel user classification using the community structure of online networks. Plos One 12, e0173347 (2017).
27. Rizos, G., Papadopoulos, S. & Kompatsiaris, Y. Multilabel user classification using the community structure of online networks. Plos One 12, e0173347 (2017).
26. Chen, D. B.
25. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. J. ACM 43, 329–374 (1996).
24. Panzarasa, P., Opsahl, T. & Carley, K. M. Patterns and Dynamics of users’ behavior and interaction: network analysis of an online community. J. Am. Soc. Inf. Sci.Tec. 60, 911–932 (2009).
23. Wang, S. & Zhao, J. Multi-attribute integrated measurement of node importance in complex networks. JCMCC 12, 236–252 (2018).
22. Stephenson, K. & Zelen, M. Rethinking centrality: methods and applications. Social Networks 11, 1–37 (1989).
21. Crucitti, P., Latora, V., Marchiori, M. & Rapisarda, A. Error and attack tolerance of complex networks. Phys. A 340, 388–394 (2004).
20. Lü, L. Y.
19. Tizghadam, A. & Leon-Garcia, A. Autonomic Traffic Engineering for Network Robustness. IEEE J. Sel. Area. Comm. 28, 39–50 (2010).
18. Bonacich, P. Power and centrality: a family of measures. Am. J. Sociol. 92, 1170–1182 (1987).
17. Freeman, L. C. Centrality in social networks conceptual clarification. Social Networks 1, 215–239 (1979).
16. Sheikhahmadi, A. & Nematbakhsh, M. A. Identification of multi-spreader users in social networks for viral marketing. Journal of Information Science 43, 412–423 (2017).
15. Liu, Y., Tang, M., Zhou, T. & Do, Y. Core-like groups result in invalidation of identifying super- spreader by k-shell decomposition. Scientific Reports 5, 9602 (2015).
14. Araujo, T., Neijens, P. & Vliegenthart, R. Getting the word out on Twitter: the role of influencers, information brokers and strong ties in building word-of-mouth for brands. International Journal of Advertising 36, 496–513 (2016).
13. Wang, X., Zhang, X., Yi, D. & Zhao, C. Identifying influential spreaders in complex networks through local effective spreading paths. J.Stat. Mech-Theory & E. 5, 053402 (2017).
12. Fu, Y. H., Huang, C.Y. & Sun, C.T. Identifying super-spreader nodes in complex networks. Math. Probl. Eng., 675713 (2015).
11. Moreno, Y., Nekovee, M. & Vespignani, A. Efficiency and reliability of epidemic data dissemination in complex networks. Phys. Rev. E 69, 055101 (2004).
10. Dí, L. Y. et al. Leaders in social networks, the delicious case. PLoS One 6, e21202 (2011).
9. Wang, X., Zhang, X., Yi, D. & Zhao, C. Identifying influential spreaders in complex networks through local effective spreading paths. J. Stat. Mech. Theory & Eng. 2015, 053402 (2017).
8. Yu, E.-Y., Chen, D.-B. & Zhao, J.-Y. Identifying critical edges in complex networks. Scientific Reports 8, 14469 (2018).
7. Panzarasa, P., Opsahl, T. & Carley, K. M. Patterns and Dynamics of users’ behavior and interaction: network analysis of an online community. J. Am. Soc. Inf. Sci.Tec. 60, 911–932 (2009).
6. Saito, K., Kimura, M., Ohara, K. & Motoda, H. Super mediator: A new centrality measure of node importance for information diffusion over social network. Information Sciences 329, 985–1000 (2016).
5. Castellano, C. & Pastor-Satorras, R. Thresholds for epidemic spreading in networks. Phys. Rev. Lett. 105, 218701 (2010).
4. Cozzens, M. B., Moazzami, D. & Stueckle, S. Tenacity of Harary Graphs. JCMCC 16, 33–56 (1994).
3. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. Complex networks: structure and dynamics. Physics Reports 424, 175–308 (2006).
2.1. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. J. ACM 43, 329–374 (1996).
2. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. Complex networks: structure and dynamics. Physics Reports 424, 175–308 (2006).
1. Panzarasa, P., Opsahl, T. & Carley, K. M. Patterns and Dynamics of users’ behavior and interaction: network analysis of an online community. J. Am. Soc. Inf. Sci.Tec. 60, 911–932 (2009).