Information Sciences 586 (2022) 1–23

Contents lists available at ScienceDirect

Information Sciences

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

Maximizing spreading in complex networks with risk in node activation

Leyang Xue [a,b], Peng Zhang [a], An Zeng [c,*]

[a] School of Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
[b] International Academic Center of Complex Systems, Beijing Normal University, Zhuhai 519087, China
[c] School of Systems Science, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

Article history:
Received 6 August 2020
Received in revised form 16 November 2021
Accepted 18 November 2021
Available online 1 December 2021

Keywords:
Activation risk
Effective spreaders
Maximizing spreading
Influence maximization

ABSTRACT

It is widely acknowledged that the initial spreaders play an important role in the spread of information in complex networks. Thus, a variety of centrality-based methods have been proposed for identifying the most influential spreaders. However, most existing studies overlook the fact that, in real social networks, it is more costly and difficult to convince influential individuals to act as initial spreaders, resulting in a high risk to maximal spreading. In this paper, we address this problem on the basis of the assumption that the activation of large-degree nodes carries a higher risk than that of small-degree nodes. We aim to identify the initial spreaders that most effectively maximize the spreading when considering both the activation risk and the outbreak size of the initial spreaders. Analysis of random networks reveals that the degree of the optimal initial spreaders does not correspond to the largest node degree in the network, but is instead determined by the infection probability and difference in activation risk among nodes with different degrees. We propose a risk-aware metric to identify the most effective spreaders in real networks. Numerical simulations show that this risk-aware metric outperforms the existing benchmark centralities in terms of maximizing the spreading.

1. Introduction

Social networks play an important role in the spreading of information, opinions, ideas, innovations, and rumors [1–3]. In social networks, identifying influential spreaders can help to control the outbreak of epidemics [4], successfully advertise new products [5], and facilitate information dissemination [6]. A recent example concerns the super-spreader events (SSEs) associated with explosive growth in the early outbreak of COVID-19 [7]. Identifying the high-risk settings of SSEs and timely implementation of interventions will help prevent and control future infectious disease outbreaks. Hence, the problem of influence maximization has received extensive attention across multiple disciplines, such as mathematics, physics, computer science, and sociology [8–10].

To date, significant efforts have been devoted to identifying influential nodes. Originally, some well-known centrality measures were used to identify the influential nodes in complex networks, such as the degree, closeness [11], betweenness [12], eigenvector [13], Katz [14], and subgraph [15] centralities. Kitsak et al. [16] argued that the location of a node in the network is more important than its immediate neighbors in evaluating the spreading influence, and used the k-shell decom-

* Corresponding author.
E-mail address: anzeng@bnu.edu.cn (A. Zeng).

https://doi.org/10.1016/j.ins.2021.11.064
0020-0255/

position to measure the location of nodes in the network. If the node is located in the core of the network, it will have a higher influence on the dissemination of information than a node located at the periphery of the network. Although some nodes with the same "coreness" are sometimes indistinguishable under the k-shell decomposition (e.g., Barabasi–Albert networks and tree-like networks), the findings of Kitsak et al. have been widely disseminated and have drawn attention to the problem. Subsequently, many new methods have been devised to identify the influential spreaders [17]. For example, Zeng et al. presented a mixed degree decomposition (MDD) procedure to rank the spreaders by considering the residual degree and exhausted degree (in the k_s layer decomposition, nodes with a degree smaller than the k_s value of the current layer are successively pruned until no more such nodes remain; in this case, for each remaining node, the residual degree refers to the number of links to other remaining nodes, and the exhausted degree denotes the number of links connecting to removed nodes). In several applications, the MDD has been shown to outperform the k-shell decomposition [18]. Lü et al. found that the H-index provided a better quantification of a node's influence than the degree or coreness (i.e., k-shell) [19], while Zareie et al. proposed an improved cluster rank approach for identifying influential nodes by considering the common hierarchy of nodes and their neighborhood sets [20].

The main implicit assumption in most relevant studies is that the probability of influential individuals acting as initial spreaders is independent of their personal influence (i.e., status level in the social network, referred to as the rank, deference, or popularity). When dealing with real applications involving influence maximization, however, some realistic factors need to be considered (e.g., the cost and accessibility of influential individuals). In related work, researchers have proposed a variety of methods for identifying influential individuals whom marketing managers should try to seed to promote their products on online social networks. The high-influence seeding targets may produce higher returns, but are associated with the following risks. (1) There is a constraint on the budget of promotional activities. In marketing, hiring higher-influence individuals to promote products incurs higher costs than hiring average people. (2) Although marketing managers may be willing to pay this higher cost, celebrities may not wish to promote products because of duty or time constraints [21], resulting in a high risk associated with maximizing the spreading. Related empirical studies have confirmed that the probability of responding to an endorsement request is dependent on status, and declines sharply with the status difference in a user-generated network [22]. To make the problem closer to the real-world situation, we relax certain assumptions and attempt to capture the reality inherent in the activation risk of initial spreaders. Here, we assume that individuals with higher influence tend to be activated (i.e., agree to act as initial spreaders) at a relatively high risk. When considering both the outbreak size and the activation risk of the target individual, selecting initial spreaders based on existing centrality metrics may fail to maximize spreading in complex networks.

In this paper, we generalize the traditional problem of identifying influential spreaders by considering the risk of initial node activation. We assume that the probability of activating nodes to act as initial spreaders decreases with increasing node degree. For simplicity, we use the exponential decay function to approximate the relation between the activation probability and the node degree. The expected value of the outbreak size over the activated probability quantifies the effective spreading coverage of nodes (i.e., the expected value of the number of infected nodes in a spreading event initiated by a single node). Using random networks, we analyze the degree value of the optimal initial spreaders using the bond percolation model. The results suggest that the optimal seed policy (i.e., optimal initial spreaders for maximizing the effective spreading) does not focus on the largest-degree node in the network. Moreover, we verify that existing centrality metrics are correlated with the degree using a real-world network. Simply discounting for the degree in existing centralities might not be sufficient to identify the most effective spreaders. Therefore, it is necessary to devise a new method for this case. We then propose a risk-aware method that identifies the most effective spreader, further maximizing the effective spread of information. The performance of the risk-aware centrality is tested on disparate real networks using a susceptible–infected–recovered (SIR) model. Numerical simulations show that our method outperforms existing benchmark centralities (i.e., the ratio between existing centralities and degree).

2. Method

2.1. Maximizing spreading with risk assigned to node activation

We briefly describe the problem of spreading when there is a risk associated with node activation. The basic idea is that it is difficult to convince an individual with more followers in a social network (i.e., a larger-degree node) to act as the initial spreader than an individual with fewer followers (i.e., smaller-degree node). We denote this as the risk of activating the node for spreading. Although the node with a larger degree could disseminate information to a large fraction of the population, this node may refuse to initiate contagion due to the higher activation risk. A natural problem is to identify which node should be selected as the initial spreader so as to maximize the spreading under this assumed risk of node activation.

To address this problem, we first quantify the activation risk of nodes. Based on the assumption that the activation risk decreases with the node degree, we employ the exponential decay function to characterize the negative relation between degree and activation probability. This function is selected as it is tractable and agrees well with our intuition in real-world scenarios. A detailed discussion about the selection of this function form is given in [sec:Appendix F] Appendix F. The exponential decay function decreases monotonically and maps the value of the degree (i.e., the number of immediate neighbors of a node) to the range [0, 1]. The activation probability of nodes is given by

$$p_i = e^{-\frac{\lambda k_i}{\langle k \rangle}}, \tag{1}$$

where k_i, $\langle k \rangle$, and λ denote the degree of node i, the average degree of the network, and the exponential decay constant, respectively. The decay constant is called the risk parameter for determining the difference in activation probabilities among nodes with distinct degrees. When $\lambda = 0$, the problem degenerates to the original definition, because each node in the network has the same activation probability. Activation risk only emerges when $\lambda > 0$. A larger value of λ corresponds to a higher risk when trying to activate large-degree nodes as initial spreaders. p_i represents the probability that node i agrees to act as the initial spreader. The spreading coverage s_i is defined as the ratio of infected nodes to all nodes in the network when the spreading originates from node i. Taking into account the activation risk of initial spreaders, the expected value of the spreading coverage of nodes is naturally regarded as the effective spreading coverage. Thus, the effective spreading coverage of node i can be easily expressed as $\tilde{s}_i = p_i * s_i$. In this paper, we employ the effective spreading coverage as a target function to quantify the practical outbreak size of spreading initiated from a target node with a given activation risk. The problem is illustrated in Fig. 1.

2.2. SIR model

The risk of node activation is more common in real problems such as advertising. Thus, we describe the spreading process as the dissemination of information on social networks. SIR models accurately describe the information spread in social media [23,24]. Therefore, we use an SIR model to simulate this process on disparate empirical networks. In the SIR model, individuals are classed into three states: susceptible (S), infected (I), and recovered (R). S nodes do not carry the disease and can be infected. I nodes carry the disease and can infect others. R nodes either die or recover, and are immune to further infection. At the beginning of the dynamics, all nodes are in the susceptible state, except for an initial infected spreader. At each time step, nodes in the infected state infect their neighbors in the S state with probability β, then immediately transform from infected to recovered. The R nodes never change their state. The process continues until there are no infected nodes in the network. At the end of the dynamics, the total number of infected nodes is calculated by counting the number of nodes in the recovered state. The spreading coverage of nodes is calculated by the ratio of infected nodes to all nodes in the network. Due to the stochastic nature of the model, all experimental results are obtained by averaging over 1000 independent numerical simulations with the same initial conditions.

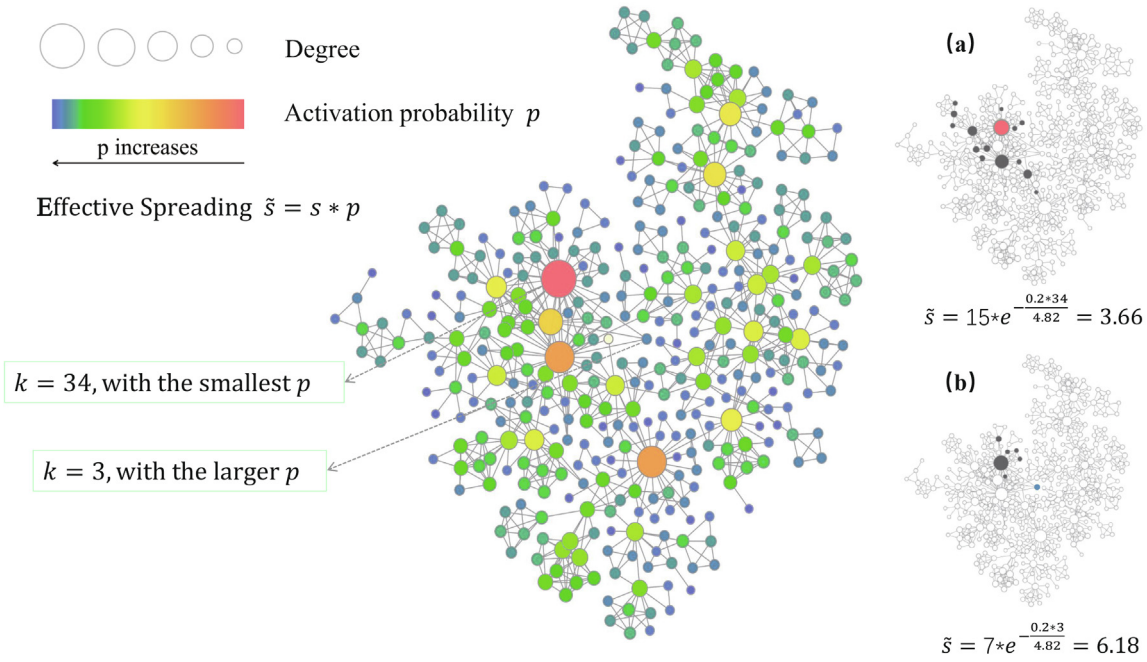

Fig. 1. Illustration of the problem for maximizing spreading with risk in node activation. We assume that the probability of initial node activation decreases with the node degree when considering realistic factors such as the cost and accessibility of a node. The effective spreading coverage is quantified by the expected value of the outbreak size over the activation probability. The degrees of two selected nodes (red and blue) in the ca-Netscience network are 34 and 3, respectively. Although using the red node as the initial spreader could infect a larger number of nodes, this node has the smallest activation probability in the network. In subplots (a) and (b), we show the contagion triggered by the two nodes under a critical infection rate using an SIR model. Obviously, the number of infected nodes is larger when originating from the red node than from the blue, but the effective spreading coverage of the red node is lower than that of the blue node ($\lambda = 0.2$).

2.3. Degree of optimal initial spreaders for maximizing the effective spreading coverage on random networks

The relation between the SIR and bond percolation models was studied by Newman [25]. The SIR model with an infection rate of β is equivalent to a bond percolation model with a bond occupation probability of T. The bond percolation model gives the exact mean size of the SIR epidemic outbreak triggered from a randomly chosen single node. Here, we investigate the maximum effective spreading coverage on a random network with a given degree distribution using the bond percolation model. We also analyze the node degree of the optimal initial spreader. To this end, we first derive the mean size of the outbreak that originates from a single node with a degree of k. The formula can be written as follows:

$$\langle s_k \rangle = 1 + \frac{k\beta}{1 - \beta G_1'(1)}, G_1'(1) = \frac{\left\langle k^2 \right\rangle - \langle k \rangle}{\langle k \rangle}, if \quad \beta < \beta_c, \tag{2}$$

where $\langle k \rangle, \left\langle k^2 \right\rangle$, and β_c are the average degree of the network, second moment of the degree, and critical infection rate, respectively, and $\beta_c = \frac{\langle k \rangle}{\left\langle k^2 \right\rangle - \langle k \rangle}$. $G_1(x)$ denotes the generation function of the degree distribution of nodes reached by following a randomly chosen edge, $G_1(x) = \sum_{k=1}^{\infty} \frac{kp_k}{\langle k \rangle} x^{k-1}$. $G_1'(1)$ is the derivative of $G_1(x)$ at $x = 1$. A detailed description of $G_1(x)$ and the derivation of Eq. (2) are given in [sec: Appendix A] Appendix A. Here, we only consider cases where the infection rate is less than the critical infection rate ($\beta < \beta_c$). In the case of larger β, the role of individual nodes is no longer important, as the final spreading coverage is independent of the location from which the infection originated. Combining Eqs. (1) and (2), we have

$$\langle \tilde{s_k} \rangle = e^{-\frac{\lambda k}{\langle k \rangle}} \left[1 + \frac{k\beta}{1 - \beta G_1'(1)} \right]. \tag{3}$$

Eq. (3) represents the mean size of the effective spreading coverage initiated from a single node with degree k on a random network. By differentiating with respect to k, we obtain

$$\frac{\partial \langle \tilde{s_k} \rangle}{\partial k} = e^{\frac{\lambda k}{\langle k \rangle}} \left[-\frac{\lambda}{\langle k \rangle} \left(1 + \frac{k\beta}{1 - \beta G_1'(1)} \right) + \frac{\beta}{1 - \beta G_1'(1)} \right]. \tag{4}$$

The degree of the optimal node (k^*) corresponding to the maximum effective spreading coverage $\langle \tilde{s}^* \rangle$ can be obtained by setting $\frac{\partial \langle \tilde{s_k} \rangle}{\partial x} = 0$:

$$k^* = \frac{\langle k \rangle}{\lambda} - \frac{1 - \beta G_1'(1)}{\beta}. \tag{5}$$

Substituting $G_1'(1) = \frac{1}{\beta_c}$ into Eq. (5), we further simplify this equation as

$$k^* = \frac{\langle k \rangle}{\lambda} - \frac{1 - \frac{\beta}{\beta_c}}{\beta}. \tag{6}$$

Thus, the degree of the optimal node depends on the infection rate β and the risk parameter λ, because $\langle k \rangle$ and β_c are constant for a given degree distribution. This result suggests that the optimal initial spreaders for maximizing the effective spreading coverage are different from those in the traditional problem (i.e., influence maximization problem on a random network). Additionally, substituting Eq. (6) into Eq. (3), we find the mean size of the maximum effective spreading coverage $\tilde{s}^*$ triggered from nodes with k^* on a random network with a given degree distribution to be

$$\langle \tilde{s}^* \rangle = e^{-\left[1 - \frac{\lambda}{\beta \langle k \rangle} + \frac{\lambda}{\beta_c \langle k \rangle}\right]} \frac{\beta_c \beta \langle k \rangle}{\lambda(\beta_c - \beta)}. \tag{7}$$

We further analyze k^*. In Eq. (6), k^* is codetermined by λ and β. Specifically, we wish to analyze the effect of these variables on the degree of the optimal spreader by fixing a parameter. (1) If we take the limit of β as $\lim_{\beta \to \beta_c} \frac{\langle k \rangle}{\lambda} - \frac{1 - \frac{\beta}{\beta_c}}{\beta}$, the optimal degree $k^* = \frac{\langle k \rangle}{\lambda}$. As $\lambda \to 0, k^* \to \infty$; as $\lambda \to \infty, k^* \to 0$. The optimal degree of the nodes decreases with λ for a fixed β. This tells us that, in a higher-risk condition, the optimal spreaders shift to lower-degree nodes. (2) By setting $\lambda = \langle k \rangle$, the optimal degree $k^* = \frac{1}{\beta_c} - \frac{1}{\beta}$. As k^* increases monotonically with β, larger-degree nodes have a comparative advantage in terms of maximizing the effective spreading coverage as the infection rate increases.

2.4. Risk-aware metrics

In related studies, various centrality metrics have been proposed to identify influential spreaders or important nodes [16,26,27]. When risk is incorporated into node activation, analysis of the optimal initial spreaders within random networks reveals that the optimal seed policy is obviously distinct from the largest-degree node selected in the original problem. This suggests that it might be not efficient to use existing metrics to identify the initial spreaders in a real network. Moreover,

correlation analysis between the degree and other metrics has verified that existing centrality metrics do not perform sufficiently well in the identification of effective spreaders because they are correlated with the degree (for this correlation analysis, see [sec: Appendix C] Appendix C). Therefore, to maximize the effective spreading, it is necessary to design a new measure for real networks.

The analytic results show that the degree value of the optimal spreaders is inversely proportional to the risk parameter λ and has a positive relation with the infection rate β, suggesting that small-degree nodes linked to many hubs are more likely to maximize the effective spreading. Inspired by this analytic result, we consider two factors in designing the new metric, namely the outbreak size and activation risk. First, we expect the metric to select nodes linked to many hubs, because the spreading initiated from such nodes will cover more nodes with the help of the neighboring hub under a larger infection rate β, although this requires that the node degree itself is not too small (this corresponds to the analytical solution where the optimal degree value has a positive relation with the infection rate β). Second, the initial spreaders should be small-degree nodes when the activation risk λ is relatively high (this corresponds to the analytical solution in which the optimal degree value is inversely proportional to the risk parameter λ). Therefore, the effective spreaders could be better characterized by the following two aspects: (1) spreaders with a strong spreading ability (e.g., possessing many high-influence neighbors) and (2) spreaders associated with lower activation risk (e.g., smaller-degree nodes).

In this paper, we propose the risk-aware (*RA*) metric to identify the most efficient spreaders by rewarding nodes with higher-degree neighbors and penalizing higher-degree nodes. The risk-aware metric of node i is defined as follows:

$$RA_i = \sum_{j\in\tau(i)} \left(\frac{k_j}{k_i + k_j}\right)^{\theta}, \tag{8}$$

where $k_i, \tau(i), k_j$ are the degree of node i, the neighbor set of node i, and the degree of node j, respectively. $\frac{k_j}{k_i+k_j}$ denotes the potential influence of node i obtained from neighbor j. The potential influence refers to the node's neighbors having sufficient influence to initiate spreading, although the node itself has lower influence. If $k_i = k_j, \frac{k_j}{k_i+k_j} = 1/2$; if $k_i \ll k_j, \frac{k_j}{k_i+k_j} \approx 1$; and if $k_i \gg k_j$, then $\frac{k_j}{k_i+k_j} \approx 0$. As a consequence, $\frac{k_j}{k_i+k_j} \in (0, 1)$. In this paper, the risk parameter λ determining the risk difference among nodes with distinct degrees is incorporated into the defined problem. To identify the most effective spreader under different conditions (i.e., various λ), the parameter θ is introduced to adjust the potential influence obtained from different neighbors. When $\theta = 0$, the potential influence obtained from different neighbors is equal, and the risk-aware metric degenerates to the degree. When $\theta > 0$, the lower potential influence obtained from neighbors will be largely weakened by an increase in θ. In other words, small-degree nodes are likely to have larger potential influence, because the term $\left(\frac{k_j}{k_i+k_j}\right)^{\theta}$ plays a more important role in the contribution to *RA* than the number of neighbors. Intuitively, a node has a large *RA* value if it is connected to many other nodes that have higher degrees. In fact, it is hard to estimate which nodes have large *RA* values. For instance, when the degree of node i is rather small, the potential influence obtained from each neighbor is relatively large. However, the overall sum of the potential influence from neighbors is small as the node has few neighbors. To further understand the *RA* metric, we show the node ranking for different values of θ in Fig. 2. Clearly, the most highly ranked node identified by *RA* varies with θ. As θ increases, the top ranked nodes are likely to be small-degree nodes with higher-degree neighbors. Actually, θ determines the ability to identify the potential influence of nodes. Larger values of θ imply that the identified node has greater potential influence. The *RA* value is significantly different from traditional centrality metrics, which measure the importance of nodes.

The computational complexity of the *RA* metric is now analyzed. The procedure RiskAwareMetric(G, θ), described in Algorithm 1, returns the *RA* value of each node in the network G. The computational complexity of traversing the neighbors of a node is simply the average degree of the network $\langle k\rangle$. If one estimates the potential influence of each node in a network using the *RA* metric, the computational complexity is $O(M + N\langle k\rangle)$, where M, N are the number of edges and nodes in the network, respectively.

Based on the idea of rewarding nodes connected to higher-degree nodes and penalizing high-degree nodes, we further consider other metrics to identify the most effective spreaders in different functional forms. The first potential influence metric (*PI_1*) and second potential influence metric (*PI_2*) are defined as

$$PI\_1 = \sum_{j\in\tau(i)} (k_j - k_i), \tag{9}$$

$$PI\_2 = \sum_{j\in\tau(i)} e^{k_j - k_i}, \tag{10}$$

where $k_i, \tau(i), k_j$ are the degree of node i, the neighbor set of node i, and the degree of node j, respectively. These two potential influence metrics are parameter-free. $PI\_1 \in (-\infty, +\infty), PI\_2 \in (0, +\infty)$. Although *PI_1* might be negative, we can still rank the nodes.

Algorithm 1: RiskAwareMetric(G, θ)

$degree \leftarrow 0, RA \leftarrow 0$
for *each* $e \in G.edges()$ **do**
  $degree[e.vertex] \leftarrow degree[e.vertex] + 1$
  $degree[e.anothervertex] \leftarrow degree[e.anothervertex] + 1$
end
for *each* $v \in G.nodes()$ **do**
  for *each* $neigh \in G.neighbors(v)$ **do**
    $RA[v] \leftarrow RA[v] + (\frac{degree[neigh]}{degree[v]+degree[neigh]})^{\theta}$
  end
end
return RA

2.5. Centrality metrics

To confirm the effectiveness of the new metrics, we need to compare them with some baseline methods. As there is a higher correlation between existing centrality metrics and the node degree (for this correlation analysis, see [sec: Appendix C] Appendix C), it is not fair to use existing metrics directly as baseline methods for comparison with the new metrics. Here, we calculate the ratio between existing centrality metrics and the node degree as baseline methods (e.g., Katz score divided by degree). In this way, it is reasonable to compare the new metrics with the baseline methods, as proved in the correlation analysis (see Fig. 6 in [sec: Appendix C] Appendix C). In this paper, we employ some representative centrality metrics to obtain the baseline method. We briefly introduce these metrics.

(1) Degree. The degree of node i is defined as the number of immediate neighbors. $k(i) = \sum_j^N a_{ij}$, where a_{ij} is a component of the network's adjacency matrix and N is the number of nodes in the network. The degree reflects the direct influence of this node. The computational complexity of the degree is $O(M)$, where M is the number of edges.

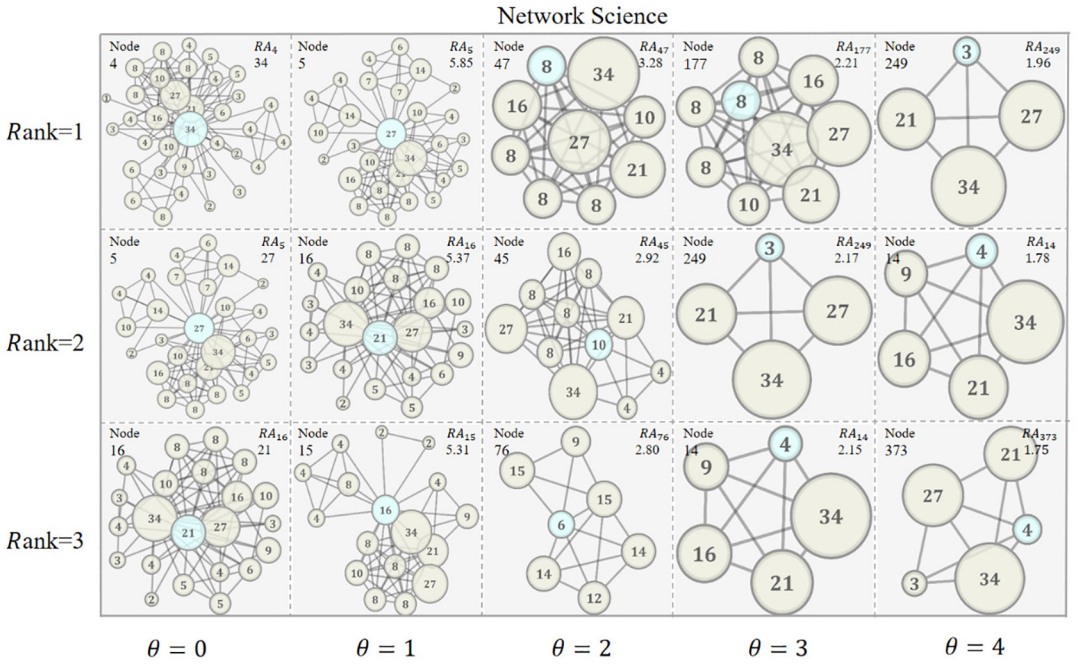

Fig. 2. Illustration of top-3 nodes ranked by the *RA* metric for different values of θ in the ca-Netscience network. In each inset, we display the identified target node (blue node) and its neighbors (yellow nodes) extracted from the original network, as well as the links between them. The numbers in each node and its size denote the degree in the original network (note that not all links to neighbor nodes are shown in each inset). The numbers in the upper-left and upper-right corners specify the node label and *RA* score, respectively. By observing the local structure of each identified node, it is easier to understand the concept of the *RA* metric. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(2) Coreness (also called k-shell, *KS*) [16]. The k-shell reflects the location of a node, which is regarded as being more important than the degree in evaluating its spreading influence. If a node is located in the core part of the network, the influence of that node will be higher than that of a node that is located at the periphery. Nodes are assigned to the k-shell by the following steps: (1) In the first step, all nodes with degree $k = 1$ are removed, which reduces the degree of the remaining nodes. This process continues until no more nodes with degree $k = 1$ remain. All removed nodes are then assigned to the 1-shell. In this process, for each remaining node, the residual degree refers to the number of links connecting a node to the other remaining nodes, and the exhausted degree denotes the number of links connecting a node to removed nodes. (2) In the second step, all remaining nodes with degree $k = 2$ are removed. The process is iteratively updated until the residual degree of all remaining nodes is greater than 2, and the 2-shell is formed from the nodes removed in the second step. (3) The process of decomposition continues until all nodes in the network have been removed. The k-shell of each node corresponds to its shell layer. The computational complexity of k-shell decomposition is $O(M)$ [28].

(3) Closeness [11]. The closeness centrality (CC) is defined as the inverse of the mean geodesic distance from a node to all other nodes. $CC_i = \frac{1}{N-1}\sum_{j,j\neq i}^{N}\frac{1}{d_{ij}}$, where N is the number of nodes in the network and d_{ij} denotes the length of the shortest path between node i and node j. Obviously, a larger closeness value corresponds to a more central node. The computational complexity of this measure is $O(MN)$ [29].

(4) Betweenness [12]. The betweenness measures the number of times that a node acts as a bridge along the shortest path between two other nodes. It is defined as $BC_i = \sum_{j,j\neq i}^{N}\sum_{k,k\neq i,k\neq j}^{N}\frac{\sigma_{jk|i}}{\sigma_{jk}}$, where σ_{jk} denotes the number of shortest paths between node j and node k and $\sigma_{jk|i}$ denotes the number of path passing through node i among all shortest paths between node j and node k. N is the number of nodes in the network. This is a global centrality measure and its computational complexity is $O(MN)$ [30].

(5) Eigenvector [13]. The eigenvector centrality considers not only the number of immediate neighbors, but also the influence of each neighbor. The eigenvector centrality of node i is denoted by $x_i = \frac{1}{\lambda}\sum_{j=1}^{N}a_{ij}x_j$, where λ and a_{ij} are the largest eigenvalue and a component of the adjacency matrix A, respectively, and x_j denotes the eigenvector centrality of node j, which can be described in matrix form as $\overrightarrow{x} = \frac{1}{\lambda}A\overrightarrow{x}$. The eigenvector centrality is widely used to measure the influence of nodes, and its computational complexity is $O\left(N^2\right)$ [29].

(6) Subgraph [15]. The subgraph centrality is defined as a weighted sum of the number of all closed walks starting and ending at node i. The subgraph centrality of node i is defined as $SC_i = \sum_{k}^{\infty}\frac{\left(A^k\right)_{ii}}{k!}$, where k is the length of a closed walk. $\left(A^k\right)_{ii}$ represents the number of closed walks of length k starting and ending at node i, which corresponds to i-th diagonal entry of the matrix A^k. Closed walks with shorter lengths have more influence on the centrality than longer closed walks in the subgraph centrality. The computational complexity is the same as that of the eigenvector centrality, $O\left(N^2\right)$ [29].

(7) Katz [14]. The Katz centrality of nodes is defined by considering all walks of different lengths, although shorter walks are assumed to be more important than long walks. The specific formula is $KC_i = \sum_{j}^{N}\sum_{k}^{\infty}s^k\left(A^k\right)_{ij}$, where $s \in (0,1)$ is a tunable parameter and s^k denotes the weight of a walk of length k. For the power method to converge, the value of the attenuation factor s has to be set such that it is smaller than the reciprocal of the absolute value of the largest eigenvalue of A. According to the original definition, the computational complexity of the exact Katz centrality is $O(n^3)$ [31]. An efficient approach for computing the Katz centrality is the Cholesky decomposition, which has a computational complexity of $O\left(N^2\right)$ [32].

(8) Collective influence [8]. The collective influence (CI) attempts to maximize the overall influence of multiple spreaders. This may output the minimal set of spreaders as the objective. The CI algorithm removes nodes progressively according to the current CI value, defined as $CI_\ell(i) = (k_i - 1)\sum_{j\in\partial B(i,\ell)}(k_j - 1)$, where k_i is the degree of node i, $B(i,\ell)$ is the ball of radius ℓ centered on node i, and $\partial B(i,\ell)$ is the frontier of the ball, that is, the set of nodes with the shortest path length ℓ from node i. Here, ℓ is a nonnegative integer that cannot exceed the network diameter. The computational complexity is $O(NlogN)$ [33].

(9) Nonbacktracking [34,35]. The nonbacktracking (NB) centrality mainly considers the local effect of the eigenvector centrality, because a hub node with a higher eigenvector centrality distributes the centrality to its neighbors, which then distribute it back again to inflate the hub's centrality. The NB centrality prevents this reflection and excludes self-effects in summing over neighbors. It can be calculated from the nonbacktracking matrix B (the nonbacktracking matrix can be computed from the adjacency matrix A, see Refs. [36,35]). The nonbacktracking centrality x_j of node j is defined to be the sum of centralities over the neighbors of j, $x_j = \sum_i A_{ij}v_{i\to j}$, where $v_{i\to j}$ is an element of the leading eigenvector of the nonbacktracking matrix B and gives the centrality of node i ignoring any contribution from j. The computational complexity of NB is $O\left(N^2\right)$ [36].

In the simulation experiments, the degree, KS, closeness, betweenness, eigenvector, subgraph, and Katz centralities are implemented by the *networkx* package in Python. We implement the CI and NB centralities based on the original definition of these metrics. In the CI centrality, we set the radius $\ell = 3$, because ℓ must be less than the diameter of the network (the smallest diameter among disparate networks in this paper is 5, see Table 2 in [sec: Network] Appendix B).

Table 1
Improvement rate of risk-aware metric compared with other benchmark centralities in terms of the overall performance of the average Kendall rank correlation $\langle\tau\rangle$ and top-1, −10, and −20 nodes' effective spreading coverage $\langle\bar{s_n}\rangle$. The smallest improvement among the different baseline centralities is displayed in bold. In most cases, the smallest improvement rates are still positive, indicating the advantage of the risk-aware metric.

	$NS_{\langle\tau\rangle}$		$NS_{\langle\bar{s_1}\rangle}$		$NS_{\langle\bar{s_{10}}\rangle}$		$NS_{\langle\bar{s_{20}}\rangle}$	
Benchmark centralities	$RA(*)$	$RA(2.5)$	$RA(*)$	$RA(2.5)$	$RA(*)$	$RA(2.5)$	$RA(*)$	$RA(2.5)$
betweenness/k	186.16	164.66	216.03	212.30	184.77	184.01	147.09	152.07
closeness/k	259.06	232.09	181.49	178.16	207.72	206.90	214.23	220.56
eigenvector/k	12.51	4.06	82.21	80.06	33.92	33.56	20.46	22.88
NB/k	17.93	9.07	76.80	74.71	39.71	39.34	26.57	29.12
Katz/k	250.59	224.26	153.81	150.81	187.46	186.70	192.58	198.48
subgraph/k	**8.24**	**0.11**	**21.74**	**20.30**	**6.88**	**6.59**	**−1.39**	**0.60**
KS/k	205.06	182.15	293.63	288.97	269.67	268.69	298.11	306.14
CI(l = 3)/k	169.17	148.95	416.20	410.10	320.75	319.63	249.00	256.03
degree	108.41	92.75	83.22	81.05	84.44	83.95	69.80	73.22

Table 2
Structural properities on real networks. Structural properities include the number of nodes in the network (N), average degree ($\langle k\rangle$), second order moment of degree($\langle k^2\rangle$), maximum degree (k_{max}), network diameter (d), assortativity coefficient (r) and critical infection rate (β_c).

Networks	N	$\langle$k$\rangle$	$\langle k^2\rangle$	k_{max}	d	r	β_c
ani-Mammalia	1430	5.45	48.88	34	18	0.012	0.126
ani-Aves-Songbird	108	19	519.80	56	6	−0.005	0.038
ani-Reptilia	496	3.97	25.38	17	21	0.345	0.185
ani-Dolphins	62	5.13	34.90	12	8	−0.044	0.172
bio-Celegans	453	8.94	358.49	237	7	−0.226	0.026
bio-Yeast	1458	2.67	19.05	56	19	−0.210	0.163
bio-Grid-Plant	1271	4.29	52.20	71	26	0.001	0.090
bio-Grid-Worm	3342	3.85	196.54	523	13	−0.169	0.020
bn-Mouse-Kasthuri	986	3.12	50.87	123	12	−0.242	0.065
ca-CSphd	1025	2.04	12.17	46	28	−0.253	0.201
ca-Erdos992	4991	2.98	48.83	61	14	−0.453	0.065
ca-GrQc	4158	6.46	116.09	81	17	0.639	0.059
ca-Netscience	379	4.82	38.69	34	17	−0.082	0.142
econ-Poli	2343	2.28	22.04	63	27	−0.335	0.115
econ-Mahindas	1258	12.03	456.55	206	8	−0.060	0.027
econ-Wm1	258	19.78	945.28	108	11	−0.037	0.021
email-Dnc	1833	4.79	354.22	404	8	−0.305	0.014
email-Corecipient	849	24.46	2276.19	368	8	−0.133	0.011
email-Enron-Only	143	8.71	112.59	42	8	−0.020	0.084
email-Univ	1133	9.62	179.82	71	8	0.078	0.057
hs-Arenas-Jazz	198	27.70	1070.24	100	6	0.020	0.027
hs-Physical	117	7.95	79.16	26	5	−0.084	0.112
hs-Zachary	34	4.59	35.65	17	5	−0.476	0.148
ia-Crime-Moreno	829	3.55	21.69	25	10	−0.165	0.196
ia-Fb-Messages	1266	10.19	279.09	112	9	−0.084	0.038
ia-Infect-Dublin	410	13.49	252.43	50	9	0.226	0.056
inf-Openflights	2905	10.77	601.45	242	14	0.049	0.018
inf-Euroroad	1039	2.51	7.75	10	62	0.090	0.479
inf-Power	4941	2.67	10.33	19	46	0.003	0.348
inf-Usair97	332	12.81	568.16	139	6	−0.208	0.023
rt-Retweet	96	2.44	12.52	17	10	−0.179	0.241
rt-Twitter-Copen	761	2.70	22.22	37	14	−0.099	0.139
socfb-Caltech36	762	43.70	3275.75	248	6	−0.066	0.014
socfb-Haverford76	1446	82.42	10480.61	375	6	0.068	0.008
socfb-Reed98	962	39.11	2784.14	313	6	0.023	0.014
socfb-Simmons81	1510	43.69	3197.12	300	7	−0.062	0.014
soc-Karate	34	4.59	35.65	17	5	−0.476	0.148
Metabolic	1038	9.13	947.24	637	6	−0.250	0.010
Protein	1646	3.06	35.71	89	14	−0.106	0.094
web-EPA	4253	4.18	118.45	175	10	−0.304	0.037

2.6. Evaluation metrics

To examine the performance of the various methods, we employ the Kendall rank correlation coefficient to estimate the ability of centrality metrics to identify the effective spreaders. In addition, we use the average of the effective spreading coverage to test the performance of each method in identifying the top-n effective spreaders. To obtain a comprehensive under-

standing of each method's performance, we define a normalized score to summarize the performance of each method in all networks considered in this paper.

(1) **Kendall rank correlation coefficient (τ)** [37]. Also known as Kendall's τ coefficient, this is used to assess statistical associations based on the ranks of the data. The τ test is a nonparametric hypothesis test for statistical dependence based on the τ coefficient. Any pair of observations (x_i, y_i) and (x_j, y_j) is said to be concordant if $(x_j - x_i)(y_j - y_i) > 0 \forall j > i, i, j = 1, 2, \ldots, n$ and discordant if $(x_j - x_i)(y_j - y_i) < 0 \forall j > i, i, j = 1, 2, \ldots, n$. The Kendall's τ coefficient is defined as $\tau = \frac{N_c - N_d}{\frac{n}{2}} = \frac{2(N_c - N_d)}{n(n-1)}$, where N_c is the number of concordant pairs and N_d is the number of discordant pairs. Here, we use the Kendall rank correlation to investigate how the ranking based on centralities is correlated to the ranking generated by the effective spreading coverage. According to the definition of the Kendall rank correlation, $-1 \leqslant \tau \leqslant 1$. Kendall's τ coefficient will be 1 if the agreement between centralities and effective spreading is perfect, indicating that the employed centrality metric accurately identifies the effective spreader.

(2) **Average effective spreading coverage ($\langle \tilde{s_n} \rangle$).** To systematically estimate the performance of each method to identify the effective spreaders under various risk values, we average the effective spreading coverage $\langle \tilde{s} \rangle$ over various λ ($\lambda \in [0, 0.9]$ with a step size of 0.1). When dealing with actual problems, we may only need to identify high-ranking effective spreaders, rather than all spreaders. To estimate the performance of each method in identifying high-ranking effective spreaders, we average the effective spreading coverage over the top-n effective spreaders. The formula is

$$\langle \tilde{s}_n \rangle = \frac{1}{|\sigma(top_n)|} \sum_{i \in \sigma(top_n)} \sum_{\lambda=0}^{\lambda=0.9} \frac{\tilde{s}_i(\lambda)}{10}, \tag{11}$$

where $\sigma(top_n)$ denotes the set of nodes whose centrality scores are ranked in the top n (e.g., $n = 1, 10, 20$) and $\tilde{s}_i(\lambda)$ is the effective spreading coverage with the risk parameter set to λ.

(3) **Normalized score (*NS*).** This is designed to summarize the performance of the metrics over all networks, providing a comprehensive understanding of the overall performance. We normalize the performance of the metrics in each network such that performance in all metrics is in the range $[0, 1]$, and then average the normalized performance across networks. The normalized score is defined as

$$NS_m(e) = \frac{\sum_{i \in \gamma(n)} \frac{c^i_m(e) - c^i_{min}(e)}{c^i_{max}(e) - c^i_{min}(e)}}{|\gamma(n)|}, m \in \gamma(c), \tag{12}$$

where e represents the evaluation metric (e.g., Kendall's τ, $\langle \tilde{s_n} \rangle$), $\gamma(n)$ and $\gamma(c)$ are the set of networks and centrality metrics, respectively, $c^i_{min}(e)$ denotes the value of the worst-performing metric among all centralities in network i according to the e evaluation metrics, $c^i_{max}(e)$ denotes the value of the best-performing metric among all centralities in network i according to the e evaluation metrics, and $c^i_m(e)$ is the value of metric m in network i according to the e evaluation metrics. $|\gamma(n)|$ is the number of networks in the dataset. Based on $NS_m(e)$, we present a normalized score for each different method according to a given estimation metric, which quantifies the overall performance in all networks. $NS_m(e) \in [0, 1]$. For Kendall's τ and $\langle \tilde{s_n} \rangle$, larger values of the normalized score indicate that the centrality metrics perform better in all networks.

3. Dataset

To confirm the accuracy of the analytic results on random networks and validate the risk-aware method, we conduct experiments on the Erdos–Renyi network (ER) and 40 real networks. The size of these networks ranges from 34 to 4991 nodes, and their average degree varies from 2.04 to 82.42. We consider 40 small- and medium-sized networks from different systems. Specifically, these networks include six biological networks (bio-Yeast [38], bio-Celegans, bio-Grid-Plant, bio-Grid-Worm, Protein [39], Metabolic [40]), four collaboration networks (ca-Erdos992, ca-GrQc, ca-Netscience [41], ca-CSphd), four animal networks (ani-Dolphins [42], ani-Mammalia, ani-Aves-Songbird, ani-Reptilia), four email networks (email-Dnc, email-Corecipient, email-Enron-Only, email-Univ), four infrastructure networks (inf-Power [43], inf-Euroroad [44], inf-Usair97, inf-Openflights), four Facebook networks (socfb-Calrech36, socfb-Haverford76, socfb-Reed98, socfb-Simmons81), three ecology networks (econ-Mahindas, econ-Wml, econ-Poli), three human social networks (hs-Arenas-Jazz [45], hs-Physical [46], hs-Zachary [47]), three interaction networks (ia-Crime-Moreno, ia-Fb-Messages, ia-Infect-Dublin), two retweet networks (rt-Twitter-Copen, rt-Retweet), one brain network (bn-Mouse-Kasthuri), one web network (web-EPA), and one social network (soc-Karate). The networks with unlabeled references were downloaded from the network repository [48]. Details of the analyzed networks can be found in [sec: Network] Appendix B.

4. Results

4.1. Degree of the optimal initial spreaders on ER network

Taking the ER network as an example, we verify the degree of the optimal spreaders for the maximum effective spreading coverage. The degree of the ER network has a Poisson distribution for larger N, where N is the total number of nodes in the network. The mean degree of the ER network is $\langle k\rangle = Np_c$, where p_c refers to the probability of edge creation. The critical infection rate is $\beta_c = \frac{\langle k\rangle}{\langle k^2\rangle - \langle k\rangle} = \frac{1}{Np_c}$ because $\langle k^2\rangle = \langle k\rangle + \langle k\rangle^2$ on the ER network. Substituting β_c and $\langle k\rangle$ into Eq. (6), we obtain

$$k^* = Np_c + \frac{Np_c}{\lambda} - \frac{1}{\beta}. \tag{13}$$

For an ER network with given values of N and p_c, we find that k^* is determined by λ and β. In Fig. 3(a), we show the optimal degree plotted as a function of λ and β. It is clear that k^* is inversely proportional to λ and has a positive relation with β. This reveals that there is an optimal initial spreader for maximizing the effective spreading for any pair of parameters λ and β. Once risk is considered in the problem, the most effective spreader does not correspond to the largest-degree node in network, but is determined by λ and β. Moreover, the parameter region where $k^* > 0$ is indicated in Fig. 3(b), as the degree value of nodes in the connected network is greater than 0. When we take the pair of parameters located in the region where $k^* \leqslant 0$, the optimal degree value of nodes will be 1.

In Fig. 3(c) and (d), we show the effective spreading coverage $\tilde{s}$ for several pairs of parameters, as obtained from exact numerical simulations of the SIR model on ER networks (see the green and yellow stripes in Fig. 3(b)). For these simulations, N = 1000 and $p_c = 0.01$. For each pair of λ and β, we simulate the spreading triggered from a single node in the network and conduct 1000 experiments for each node. We then calculate the effective spreading coverage $\tilde{s}$ by averaging over nodes with

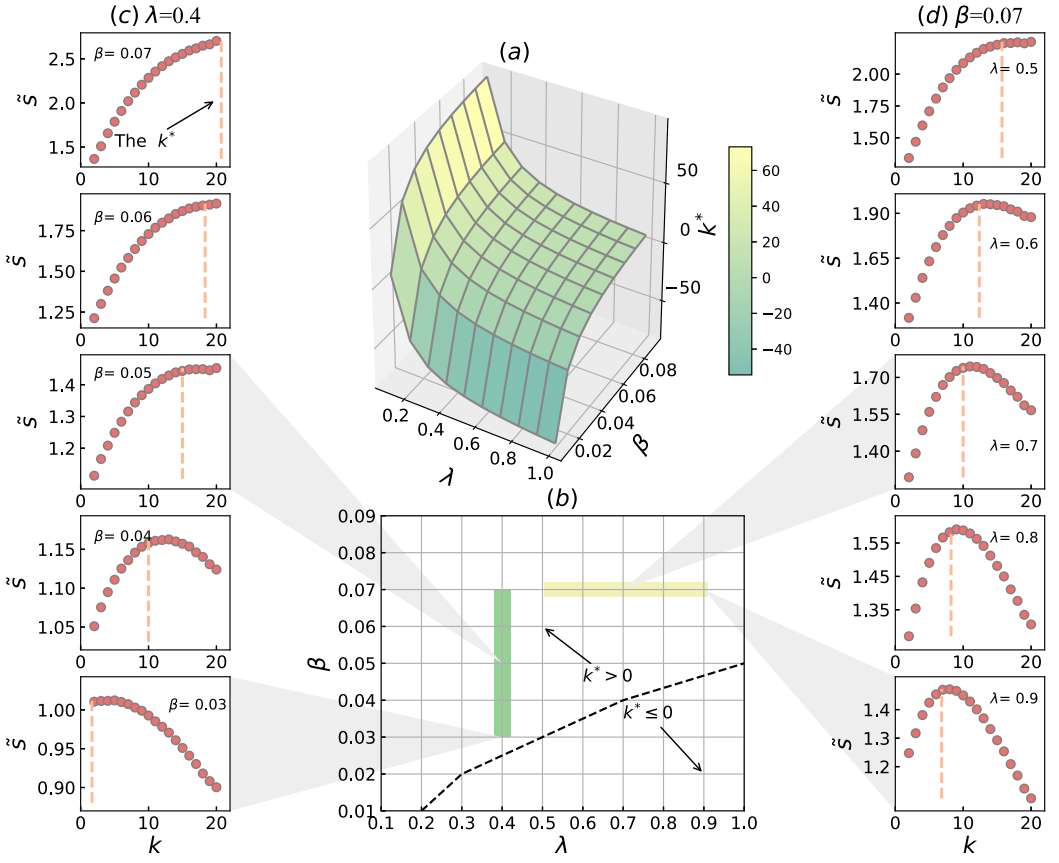

Fig. 3. Degree of optimal initial spreader for maximizing the effective spreading coverage on ER networks. We set the parameters of the ER network to $N = 1000$ and $p_c = 0.01$. The mean degree of the ER network is $\langle k\rangle = Np_c = 10$. (**a**) Optimal degree k^* plotted as a function of risk parameter λ and infection rate β. (**b**) Parameter space of λ and β; the dashed line marks the region where $k^* > 0$. (**c**), (**d**) Effective spreading obtained from numerical simulations of SIR model as a function of degree k on ER networks. The points in each subfigure are the average of 500 ER networks. The orange dashed line denotes the exact k^* calculated by Eq. (13).

the same degree. The simulation results show two important findings. First, k^* corresponding to the maximum effective spreading agrees well with our analytic results (orange dashed line). Additionally, the trend in k^* for different pairs of parameters is consistent with the analytic result, confirming the correctness of the degree of the optimal initial spreaders. The small disagreement between the simulations and the analytic results for k^* appears to be a finite size effect due to the relatively small network size in the simulations. Second, k^* increases with β when λ is fixed, suggesting that it is better to target

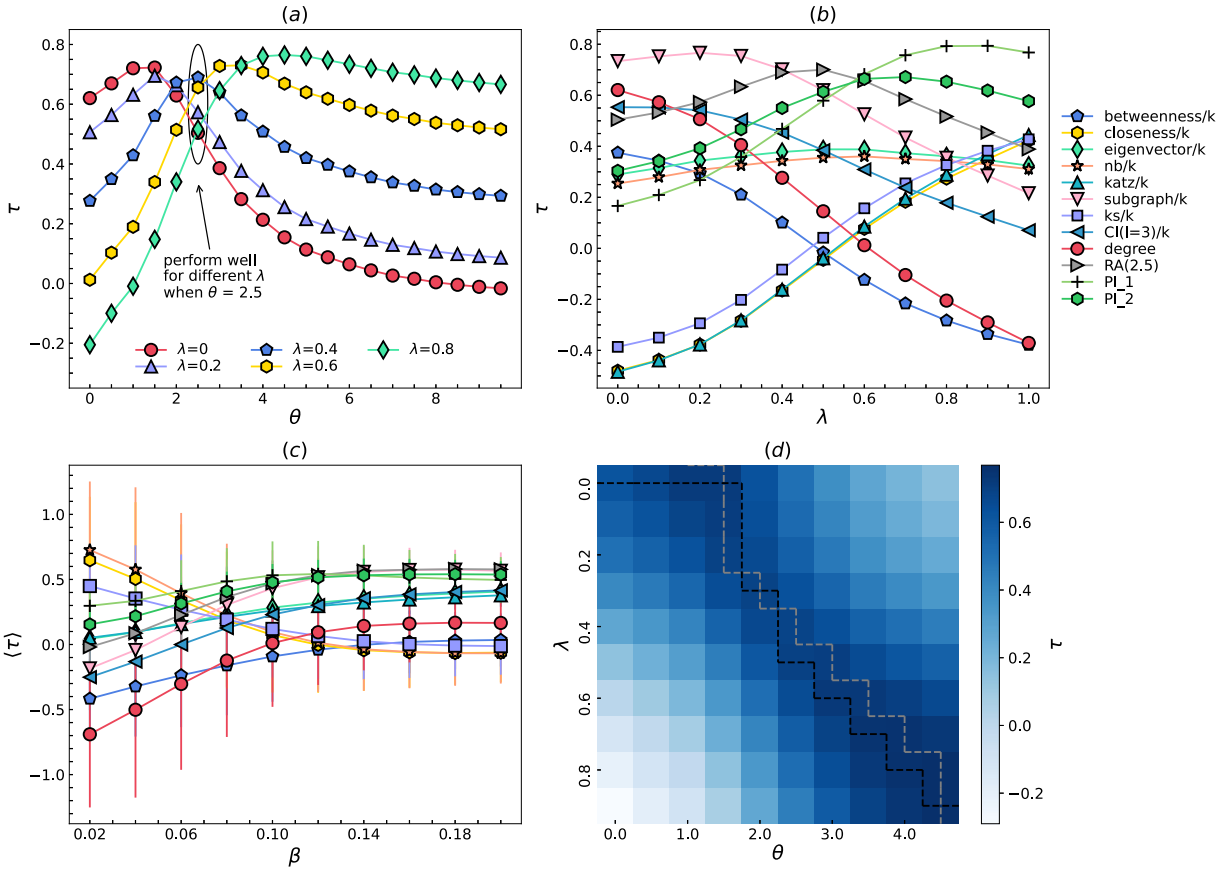

Fig. 4. Kendall rank correlation (τ) between different methods and the effective spreading coverage $\bar{s}$ on the ca-Netscience network. (**a**) τ value of risk-aware metric for various λ under distinct θ. (**b**) τ value of different methods plotted as a function of λ. (**c**) $\langle\tau\rangle$ of different methods plotted as a function of β. The error bar shows the standard deviation. (**d**) τ value plotted as a function of λ and θ. In (**a**), (**b**), and (**d**), the results are obtained with a critical infection rate β_c.

Table 3
The p-value for Kendall rank correlation between different methods and effective spreading $\bar{s}$. The number in "bold" denotes $p > 0.05$, suggesting that the correlation is not significant. The full name of methods in the table is respectively betweenness (*BTN*), closeness (*CLO*), eigenvector (*EIG*), non-backtracking (*NB*), subgraph (*SG*), k-shell (*KS*), collective influence (*CI*), risk-aware metric ($RA(\theta = 2.5)$), and potential influence (*PI*_1 and *PI*_2). The methods/k denotes the ratio between methods and degree.

λ	BTN/k	CLO/k	EIG/k	NB/k	Katz/k	SG/k	KS/k	CI(3)/k	Degree	RA(2.5)	PI_1	PI_2
0	0	0	0	0	0	0	0	0	0	0	0	0
0.1	0	0	0	0	0	0	0	0	0	0	0	0
0.2	0	0	0	0	0	0	0	0	0	0	0	0
0.3	0	0	0	0	0	0	0	0	0	0	0	0
0.4	0.0087	0	0	0	0	0	0.0322	0	0	0	0	0
0.5	**0.6717**	**0.1909**	0	0	**0.2584**	0	**0.2862**	0	0.0001	0	0	0
0.6	0.0013	0.0266	0	0	0.0131	0	0.0001	0	**0.7332**	0	0	0
0.7	0	0	0	0	0	0	0	0	0.0041	0	0	0
0.8	0	0	0	0	0	0	0	0	0	0	0	0
0.9	0	0	0	0	0	0	0	0.0003	0	0	0	0
1	0	0	0	0	0	0	0	0.0404	0	0	0	0

a high-degree node as the initial spreader if the infection rate is larger, even when there is a risk associated with node activation. On the contrary, for a given β, k^* decreases with λ, indicating that a conservative seed policy should be employed when the risk difference for nodes with various degrees is relatively large, e.g., selecting a small-degree node as the initial spreader.

4.2. Performance of the risk-aware metric on real networks

We now investigate the final ranking of nodes with the different methods. In principle, the ranking generated by a benchmark centrality metric with good performance should be as consistent as possible with the ranking based on the node's effective spreading coverage $\tilde{s}$. Here, we use the Kendall rank correlation (τ) to measure the performance of the different methods. As the parameter θ is incorporated into the risk-aware metric, we first study how θ affects the performance of *RA* under different values of λ. In Fig. 4(a), we show the τ value of $RA(\theta)$ for various λ on the ca-Netscience network. For any λ, there is an optimal parameter θ^* that maximizes τ under the current resolution of θ (step size of 0.5). We find that θ^* increases with λ, suggesting that *RA* performs better with larger θ in cases where the activation risk of nodes with a large degree is far higher than that of nodes with a small degree. This is because potentially influential nodes connected to a few high-degree nodes are assigned a higher centrality value by *RA* in the case of larger θ, producing a higher level of consistency with the effective spreading coverage of nodes when λ is larger. Although θ^* varies with λ within the range [0, 0.9], we could

Table 4
The p-value of Kendall rank correlation between *RA* and effective spreading $\tilde{s}$ under various λ and θ. The number in "bold" denotes $p > 0.05$, suggesting that the correlation is not significant.

λ	θ=0	θ=0.5	θ=1	θ=1.5	θ=2	θ=2.5	θ=3	θ=3.5	θ=4	θ=4.5
0	0	0	0	0	0	0	0	0	0	0
0.1	0	0	0	0	0	0	0	0	0	0
0.2	0	0	0	0	0	0	0	0	0	0
0.3	0	0	0	0	0	0	0	0	0	0
0.4	0	0	0	0	0	0	0	0	0	0
0.5	0.0001	0	0	0	0	0	0	0	0	0
0.6	**0.7332**	0.0027	0	0	0	0	0	0	0	0
0.7	0.0041	**0.8605**	0.0163	0	0	0	0	0	0	0
0.8	0	0.0038	**0.7904**	0	0	0	0	0	0	0
0.9	0	0	0.0113	0.0345	0	0	0	0	0	0
1.0	0	0	0	**0.9913**	0	0	0	0	0	0
λ	θ=5	θ=5.5	θ=6	θ=6.5	θ=7	θ=7.5	θ=8	θ=8.5	θ=9	θ=9.5
0	0.001	0.0109	**0.0626**	**0.2026**	**0.4421**	**0.6468**	**0.9066**	**0.9002**	**0.7454**	**0.637**
0.1	0	0.0001	0.0018	0.0113	0.0425	**0.0864**	**0.1689**	**0.2572**	**0.3513**	**0.4373**
0.2	0	0	0	0	0.0002	0.0006	0.0019	0.0042	0.0077	0.0121
0.3	0	0	0	0	0	0	0	0	0	0
0.4	0	0	0	0	0	0	0	0	0	0
0.5	0	0	0	0	0	0	0	0	0	0
0.6	0	0	0	0	0	0	0	0	0	0
0.7	0	0	0	0	0	0	0	0	0	0
0.8	0	0	0	0	0	0	0	0	0	0
0.9	0	0	0	0	0	0	0	0	0	0
1.0	0	0	0	0	0	0	0	0	0	0

Table 5
The p-value of Kolmogorov–Smirnov test for the difference of two empirical distribution under distinct infection rate β, i.e. $RA(\theta = 2.5)$ and other baseline methods. The full name of methods in the table is respectively degree(*K*), k-shell(*KS*), betweenness (*BTN*), closeness (*CLO*), eigenvector (*EIG*), Katz (*KZ*), subgraph (*SG*), non-backtracking (*NB*), and collective influence(*CI*). The methods/k denotes the ratio between methods and degree.

β	K	KS/k	BTN/k	CLO/k	EIG/k	KZ/k	SG/k	NB/k	CI(3)/k
0.02	0	0	0	0	0	0	0	0	0
0.04	0	0	0	0	0	0	0	0	0
0.06	0	0	0	0	0	0	0	0	0
0.08	0	0	0	0	0	0	0	0	0
0.10	0	0	0	0	0	0	0	0	0
0.12	0	0	0	0	0	0	0	0	0
0.14	0	0	0	0	0	0	0	0	0
0.16	0	0	0	0	0	0	0	0	0
0.18	0	0	0	0	0	0	0	0	0
0.20	0	0	0	0	0	0	0	0	0

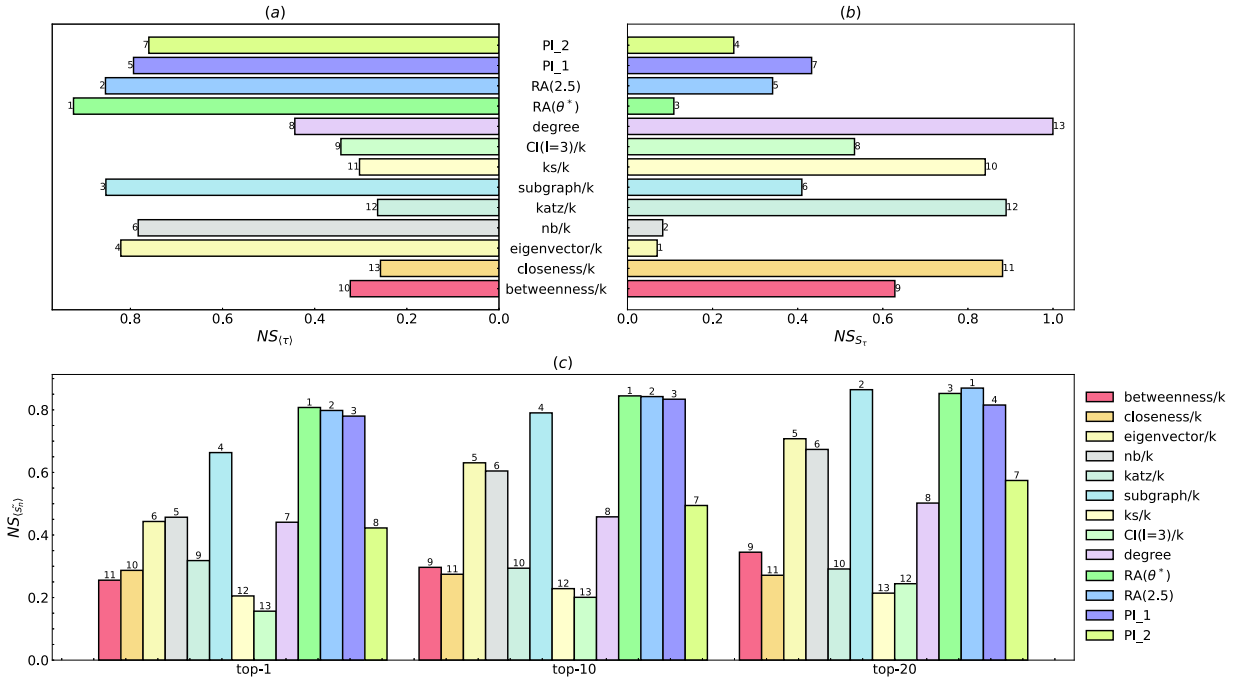

Fig. 5. *NS* **of evaluation metrics for different methods in 40 networks**. The numbers on the bar denote the ranking among all methods. (**a**) *NS* based on the average Kendall rank correlation $\langle\tau\rangle$. (**b**) *NS* based on the standard deviation of the Kendall rank correlation S_τ. (**c**) *NS* based on the average effective spreading coverage $\langle\tilde{s_n}\rangle$. Results shown for $n = 1, 10$, and 20. The simulation results are for a critical infection rate of β_c.

fix the parameter to make *RA* achieve the highest possible τ under different λ. When $\theta = 2.5$, the risk-aware metric performs relatively well under different λ, i.e., $RA(\theta = 2.5)$ has the lowest standard deviation $\left(S_\tau(\theta) = \sum_{\lambda=0}^{0.9} \frac{(\tau_\lambda(\theta) - \langle\tau(\theta)\rangle)^2}{9}\right)$. Thus, we further compare $RA(\theta = 2.5)$ with other baseline methods for various λ in Fig. 4(b). The τ value of $RA(\theta = 2.5)$ is lower than that of the subgraph, Katz, degree, and *KS* methods with smaller values of λ, but $RA(\theta = 2.5)$ outperforms the other methods significantly at higher values of λ, indicating that our method is more accurate than other methods in ranking the effective spreading of nodes. Additionally, the τ value of *PI*_1 and *PI*_2 is lower than that of $RA(\theta = 2.5)$ when $\lambda < 0.6$, because high-degree nodes are penalized more heavily than in the *RA* metric. This could explain why τ is higher in *PI*_1 and *PI*_2 than in *RA* for larger values of λ. Some statistical significance tests can be applied to verify the Kendall rank correlation between the centrality value and the effective spreading of nodes (see Table 3 in [sec: Appendix D] Appendix D). In fact, we cannot exactly determine the value of λ for practical problems. To systematically study the performance of $RA(\theta = 2.5)$, we calculate $\langle\tau\rangle$ by averaging all τ over different λ, i.e., $\langle\tau\rangle = \sum_{\lambda=0}^{0.9}\tau(\lambda)/10$. This quantifies the performance of the different methods in identifying the effective spreader under various conditions. We further compare $RA(2.5)$ with the other methods under different infection rates β in Fig. 4(c). Visual observation indicates that, although $\langle\tau\rangle$ of $RA(\theta = 2.5)$ for larger values of β is the same as for the subgraph benchmark centrality, the Kolmogorov–Smirnov(K-S) test confirms that the Kendall rank correlation distribution under various conditions is significantly different (see Table 5 in [sec: Appendix D] Appendix D). This result shows that $RA(\theta = 2.5)$ outperforms most baseline methods when the infection rate β is relatively large. Moreover, the standard deviation of τ for $RA(2.5)$ is rather low compared with that of other methods (see the error bar in Fig. 4(c)), further revealing that our method is robust under various conditions. Finally, we investigate how λ and θ together affect the value of τ in Fig. 4(d). The black dashed line shows that the optimal λ corresponding to the maximum τ varies as θ increases, suggesting that *RA* achieves better performance with larger λ as θ increases. Similarly, the gray dashed line indicates that the optimal θ corresponding to the maximum τ varies with λ, meaning that a larger θ value should be selected to identify the effective spreaders as λ increases. The two dashed lines together reveal that there is a positive correlation between λ and θ. When there is greater difference in activation risks, *RA* with a larger θ value might perform better in terms of identifying the effective spreaders. The test of Kendall rank correlation between *RA* and effective spreading confirms that the correlation for most pairs of parameters (θ and λ) is significant and the result can be accepted (see Table 4,5 in [sec: Appendix D] Appendix D).

We now compare the *RA* metric with other baseline centralities on 40 disparate real networks according to the average Kendall rank correlation $\langle\tau\rangle$, standard deviation of the Kendall rank correlation S_τ, and average effective spreading coverage $\langle\tilde{s_n}\rangle$. Here, we employ *NS* to summarize the overall performance of different methods, providing a comprehensive under-

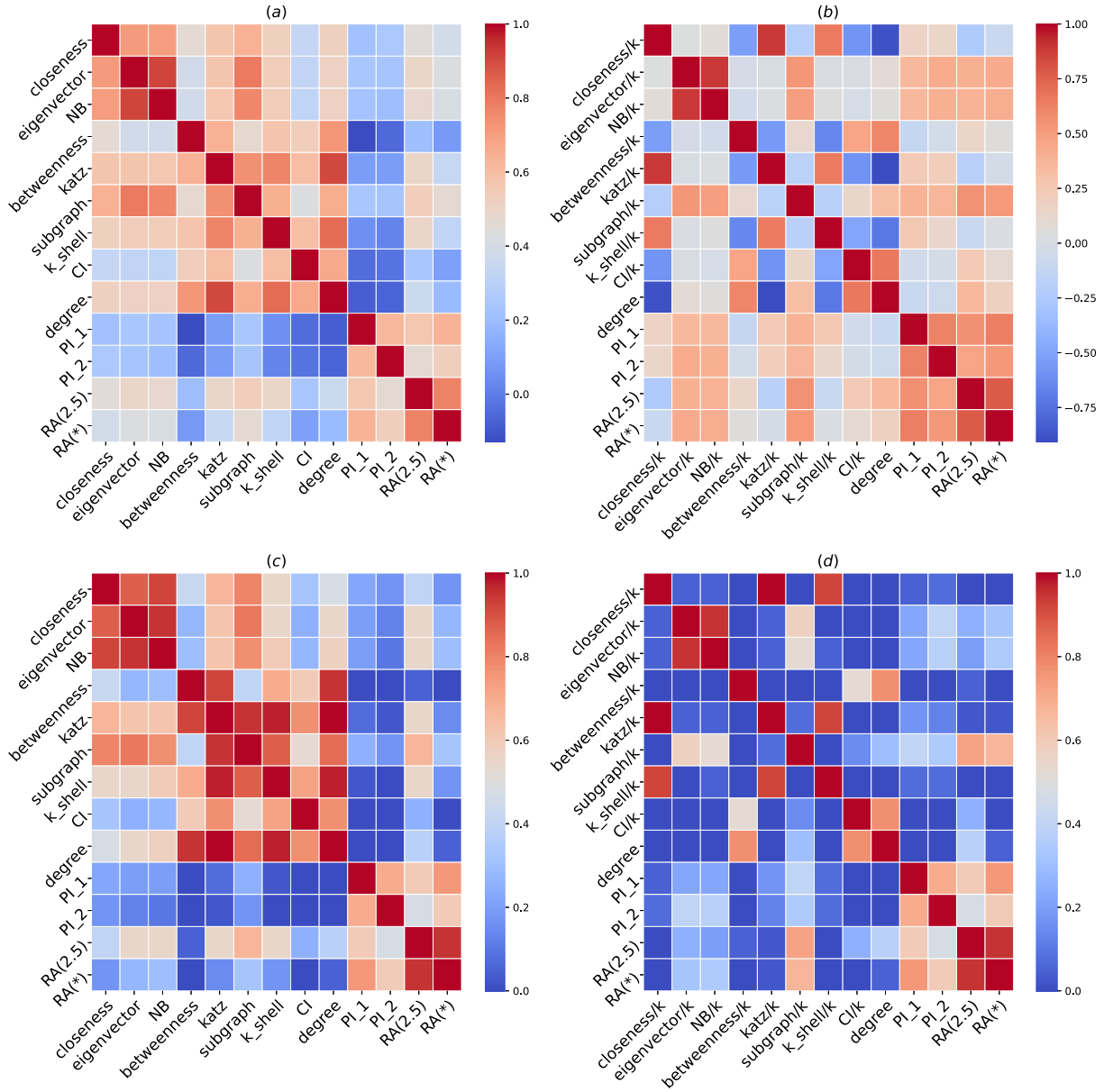

Fig. 6. The Kendall rank correlation between two centrality metrics on 40 networks. The $RA(*)$ denotes that parameter θ is set as θ^* in each network where RA(θ^*) has the smallest volatility over different λ. (**a**) The average of Kendall rank correlation between any two centralities. (**b**) The average of Kendall rank correlation between two benchmark metrics. (**c**) The proportion of network where Kendall rank correlation between two centralities is larger than 0.5. (**d**) The proportion of network where Kendall rank correlation between two benchmark centralities is larger than 0.5.

standing of each metric's performance. The value of θ^* in each network is determined by the lowest standard deviation $S_\tau(\theta)$ instead of the largest $\langle\tau\rangle$. Therefore, $RA(\theta^*)$ does not guarantee the best performance among all parameters, although it is more robust than other parameters under different λ. In Fig. 5,6(a), we present $NS_{\langle\tau\rangle}$ for different methods. Overall, $RA(\theta^*)$ and $RA(\theta = 2.5)$ outperform the other methods over the 40 real networks (normalized value of $\langle\tau\rangle$ for different methods in each network are provided in [sec: Appendix E] Appendix E). The *PI*_1 and *PI*_2 metrics perform worse than $RA(\theta = 2.5)$, although they are based on the same idea. The possible reason for this is that the extent to which high-degree nodes are penalized can be controlled by θ in *RA*. Therefore, *RA* with $\theta = 2.5$ is a good metric in terms of identifying effective spreaders. The value of NS_{S_τ} for the various centralities over the 40 networks is shown in Fig. 5(b) (normalized values of S_τ for different methods in each network can be found in [sec: Appendix E] Appendix E). Smaller values of NS_{S_τ} correspond to smaller values of the standard deviation of $\langle\tau\rangle$. The eigenvector/k method has the lowest volatility among all

methods under different λ. $RA(\theta = 2.5)$ remains relatively stable under the different risk cases, although it is outperformed by the eigenvector/k method. Finally, the performance of RA is discussed in terms of $\langle \tilde{s_n} \rangle$. Here, we identify the top-n effective spreaders given by the different methods for $n = 1, 10$ and 20 and calculate $\langle \tilde{s_n} \rangle$. $NS_{\langle \tilde{s_n} \rangle}$ for the different methods is shown in Fig. 5(c). $RA(\theta^*)$ and $RA(2.5)$ give higher NS values for identifying the top 1 and 10 nodes compared with the other baseline methods. Overall, $RA(2.5)$ outperforms all other methods except $RA(\theta^*)$ over the 40 networks, which again confirms that $RA(2.5)$ is a good predictor for identifying effective spreaders.

To demonstrate the advantages of the risk-aware metric in terms of accuracy, we present the improvement in the accuracy rate compared with the other baseline centralities in Table 1. Clearly, RA(*) and RA(2.5) have positive improvement rates (i.e., in $NS_{\langle \tau \rangle}, NS_{\langle \tilde{s_1} \rangle}, NS_{\langle \tilde{s_{10}} \rangle}, NS_{\langle \tilde{s_{20}} \rangle}$) when compared with the subgraph centrality, which is the state-of-the-art centrality measure among those considered here. The improvement rate in the overall accuracy $NS_{\langle \tau \rangle}$ is around 8%, and the spreading coverage initialized from the most highly-ranked node can be over 20% greater than that using the subgraph centrality. For the other baseline centralities, our method produces even more significant improvements. The results suggest that risk-aware metrics could be used to identify effective spreaders more accurately than other baseline centralities. Additionally, our method has a computational complexity of $O(m + n\langle k \rangle)$. The risk-aware metric employs local structural information to estimate the potential influence of nodes, which significantly reduces the runtime in large-scale networks. The subgraph, eigenvector, and nonbacktracking centralities have a computational complexity of $O(n^2)$. Obviously, our method produces a significant advantage in dealing with large-scale networks. In summary, the risk-aware metric is a more efficient means of maximizing the spreading than other baseline centralities.

5. Conclusion

As we all know, a product promoted by a celebrity on a social network will rapidly spread to millions of users, while a similar product posted by a less-well-connected individual will not reach as many people. One of the most important factors determining the fate of the spreading process is where the initial spreader is located within a social network with a given connectivity pattern. To date, many methods have attempted to identify influential spreaders using only node topology features. However, when dealing with the real application of influence maximization, most existing studies have ignored the fact that it is more costly and difficult to convince influential nodes to act as initial spreaders, resulting in a higher risk in terms of maximizing the spreading. Therefore, we have introduced the activation risk of initial spreaders into the problem. In this paper, we assumed that the probability of nodes agreeing to act as initial spreaders depends on their degree, with large-degree nodes having a lower activation probability than small-degree nodes. For simplicity, we used the exponential decay function to map the degree of the node into the activated probability and introduced a risk parameter λ to determine the difference in activation risk over various nodes. Through the theoretical results obtained from the percolation model on random networks, we found that the degree of the optimal initial spreader depends on λ and the infection rate β, rather than the node degree. Moreover, we confirmed our findings through numerical simulations conducted with the ER network. In a real-world network, we analyzed the correlation between existing centralities and the degree. It was found that simply identifying the effective spreaders by discounting for the degree in existing centralities is insufficient. Thus, the risk-aware metric was proposed to identify effective spreaders. Experimental results for the normalized score of the Kendall rank correlation $\langle \tau \rangle$ and the average effective spreading coverage $\langle \tilde{s_n} \rangle$ on 40 disparate real networks have shown that our method outperforms several existing benchmark centralities.

In actual problems, it is difficult to quantify the activation risk of influencers in marketing. The possibility of activating an influencer depends on many factors, such as cost, brand awareness, and quality of content. Among these factors, the cost of the influencer can be regarded as the main determinant of activation. Therefore, the activation risk is very relevant to the cost of the influencer. As the cost of an influencer is an unsolved empirical problem, we introduced the activation risk to characterize the problem, and assumed that large-degree nodes had a lower probability of being activated. To a certain extent, the degree-decaying effect in effective spreading can be interpreted as a higher cost. In addition, we employed the exponential decay function to describe the negative relation between the node degree and the activation risk. This functional form offers analytic tractability and produces reasonable analytic results that agree with our intuition. In fact, the conclusions obtained from our results rely on the specific form of the activation function, as proved in [sec:Appendix F] Appendix F. Although the analysis is not robust against the form of the activation function, this study has revealed that many existing centralities might not be able to identify the effective spreaders once we consider the real factors determining node activation. For practical problems, our findings suggest that it is critical to propose new methods of identifying effective spreaders, namely those that have a strong spreading ability but low degree, because this will allow advertising and immunization strategies to be designed at a lower cost. The risk-aware metric not only identifies the effective spreaders, but also evaluates the potential importance of a node in the network. Moreover, there is an additional "risk" in identifying the most influential spreaders, because the proposed metric does not guarantee that the outbreak size of detected nodes will be maximized. The topic covered in this paper is a general research problem, and many related issues could be studied in the near future. For example, different functional dependencies of effective spreading on the degree could be considered in follow-up studies. One could design the effective spreading coverage on the basis of the middle-status conformity theory [49] (i.e., individuals who are most likely to adopt an innovation or be susceptible to social contagion are those people in the middle strata in terms of social status).

CRediT authorship contribution statement

Leyang Xue: Methodology, Investigation, Validation, Writing-original-draft. **Peng Zhang:** Methodology, Writing-review-editing. **An Zeng:** Conceptualization, Investigation, Writing-review-editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 71731002), Fundamental Research Funds for the Central Universities (Contract No. 2019XD-A10) and National Key R&D Program of China (Contract No. 2020YFF0305300). Declaration of Competing Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. The mean size of outbreak initiated from nodes with degree k

The mean size of the outbreak initiated from a randomly chosen node has been given by the bond percolation model and generation function[25]. Here, we wish to get the mean outbreak size of the disease $\langle s_k \rangle$ triggered from the single node with the degree k on a random network. The ultimate size of the outbreak starting with a single infective node would be precisely the size of the cluster of nodes that can be reached from the initial node. In fact, the SIR model is equivalent to a bond percolation with bond occupation probability T. We employ the percolation model and generation function to give the exact mean size of the outbreak initiated from the single node with the degree k.

Firstly, we need to define some generation functions because it could generate the probability distribution and easily work than probability distribution itself (The crucial properties of generation function could see the work [50]). For instance, a generating function of degree distribution is as follow,

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k, \tag{14}$$

the mean degree $\langle k \rangle$ of the node in the network is given by

$$\langle k \rangle = \sum_k k p_k = G_0'(1). \tag{15}$$

If we follow an edge to the node at one of its ends, the probability of the reached nodes with degree k is $q_k = \frac{k p_k}{\langle k \rangle}$. In general, we will concern with the number of ways of leaving such a node excluding the edge we arrived along, which is the degree minus 1. The distribution of degrees of the nodes reached by following a randomly chosen edge is generated by

$$G_1(x) = \sum_{k=1}^{\infty} \frac{k p_k}{\langle k \rangle} x^{k-1} = \frac{G_0'(x)}{G_0'(1)}. \tag{16}$$

For a node with degree k, the probability of node having exactly m edges occupied from the k edges follows the binomial distribution $\binom{k}{m} T^m (1-T)^{k-m}$, hence the probability distribution of occupied edges for a node with degree k is generated by Eqs. 17,

$$G_0^k(x,T) = \sum_{m=0}^{\infty} \binom{k}{m} T^m (1-T)^{k-m} x^m = \sum_{m=0}^{k} \binom{k}{m} T^m (1-T)^{k-m} x^m = (1-T+xT)^k = (1-T(1-x))^k. \tag{17}$$

Likewise, the probability distribution of occupied edges leaving the node arrived by following a randomly chosen edge is generated by Eqs. 18

$$G_1(x,T) = \sum_{m=0}^{\infty} \sum_{k=m}^{\infty} q_k \binom{k}{m} T^m (1-T)^{k-m} x^m = G_1(1-T(1-x)). \tag{18}$$

We define the $\rho_s(T)$ as the distribution of cluster size of nodes reached by following a randomly chosen edge. Let $H_1(x,T)$ be the generating function for this distribution as is shown in the Eqs. 19,

$$H_1(x,T) = \sum_{s=0}^{\infty} \rho_s(T) x^s. \tag{19}$$

The H_1 can be broken down into an additive set of contributions as follows. We follow an edge to reach the cluster, which might be consisted by the following two parts: (1) a single node with no occupated edges connected to it other than the one along which we passed to reach it; (2) a single node with any number m ($m > 1$) of occupated edge attached to it excluding the one along which we have reached it. Each occupated edge leading to another cluster whose size distribution is also generated by the H_1. The chance that a finite cluster containing a closed loop of edges goes as N^{-1}, and is zero in the limit of $N \to \infty$. Using these results, the $H_1(x,T)$ can be expressed in a Dyson-equation-like self-consistent form by Eqs. 20,

$$H_1(x,T) = xG_1(H_1(x,T),T). \tag{20}$$

The self-consistent process mentioned-above can be a better understanding by the following analysis. By analogy with the preceding part, then we represent the $p_s^k(T)$ as the distribution of the size of cluster reachable from a starting node with the degree k. We also give the generating function for this distribution in the Eqs. 21. The $p_s^k(T)$ can be expressed by the $P(s-1|k)$, which means the probability of cluster size $s-1$ reached from the starting node with degree k, other than the starting node. The size of cluster s can be broken down into different contributions from the m occupied edges for a node with a degree k, which is described in the second row of Eqs. 21. The δ is the Dirac delta function. If $s = \sum_{r=1}^{m} j_r$, then $\delta = 1$, otherwise $\delta = 0$. The $\delta\left(s, \sum_{r=1}^{m} j_r\right)$ aims to hold that the sum of clustered size reached from the occupied edge is the same as s. The ρ_{j_r} refers to the probability of cluster size j_r of nodes reached by following a randomly chosen edge. By simplifying the equation, we obtain the form $h_0^k(x,T) = x(1-T(1-H_1(x,T)))^k$.

$$\begin{aligned}
h_0^k(x,T) &= \sum_{s=1}^{\infty} p_s^k(T)x^s = \sum_{s=1}^{\infty} P(s-1|k)(T)x^s = x\sum_{s=0}^{\infty} P(s|k)(T)x^s \\
&= x\sum_{s=0}^{\infty}\sum_{m=0}^{k}\binom{k}{m}T^m(1-T)^{k-m}\sum_{j_1}^{\infty}\cdots\sum_{j_m}^{\infty}\delta\left(s,\sum_{r=1}^{m} j_r\right)\prod_{r=1}^{m}\rho_{j_r}(T)x^s \\
&= x\sum_{m=0}^{k}\binom{k}{m}T^m(1-T)^{k-m}\sum_{s=0}^{\infty}\sum_{j_1}^{\infty}\cdots\sum_{j_m}^{\infty}\prod_{r=1}^{m}\rho_{j_r}(T)x^{j_r} \\
&= x\sum_{m=0}^{k}\binom{k}{m}T^m(1-T)^{k-m}\left(\sum_{j_r}^{\infty}\rho_{j_r}(T)x^{j_r}\right)^m \\
&= x\sum_{m=0}^{k}\binom{k}{m}(1-T)^{k-m}(TH_1(x,T))^m = x(1-T(1-H_1(x,T)))^k.
\end{aligned} \tag{21}$$

Then, we make full use of the important properties of generating function. The mean of the probability distribution is given by the first derivative of the generating function, evaluated at 1. Using the Eqs. 21, we obtained the mean outbreak size of disease initiated from the node with degree k by differentiating for x. At $x = 1$, we have

$$\langle s_k\rangle = \frac{\partial h_0^k}{\partial x}|_{x=1} = \left[(1-T(1-H_1(x,T)))^k + xk(1-T(1-H_1(x,T)))^{k-1}TH_1'(x,T)\right]|_{x=1}. \tag{22}$$

By differentiating the Eqs. 20, we have

$$H_1'(x,T) = G_1(H_1(x,T),T) + xG_1'(H_1(x,T),T)H_1'(x,T). \tag{23}$$

By simplifying the Eqs. 23, $H_1'(x,T)$ would be

$$H_1'(x,T) = \frac{G_1(H_1(x,T),T)}{1 - xG_1'[H_1(x,T),T]}. \tag{24}$$

Due to the fact that the generating functions are 1 at $x = 1$ if the distributions generated by generating functions are normalized, hence $H_1(1,T) = 1, G_1(1,T) = 1$. $G_1'(x,T) = TG_1'(1-T(1-x))$. Thus, we have

$$H_1'(1,T) = \frac{1}{1-G_1'(1,T)} = \frac{1}{1-TG_1'(1)}. \tag{25}$$

Substuting the Eqs. 25, $H_1(1,T) = 1$ and $G_1(1,T) = 1$ into the Eqs. 22, we obtain the exact value of mean outbreak size of disease triggered from the node with degree k,

$$\langle s_k\rangle = 1 + \frac{kT}{1-TG_1'(1)}. \tag{26}$$

We take the derivative with respect to $G_1(x)$, it would be $G_1'(x) = \sum_{k=2}^{\infty}\frac{k(k-1)p_k}{\langle k\rangle}x^{k-2}$, hence $G_1'(1) = \frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}$. We get the mean outbreak size of disease triggered from the node of degree k in closed form when $T_c < \frac{\langle k\rangle}{\langle k^2\rangle - \langle k\rangle}$.

Appendix B. Network

The risk-aware metric does not aim at specific datasets, so we examine the performance of our method on 40 datasets from different domains. Most of the datasets are downloaded from the network repository [1], other datasets are downloaded from the website [2]. In the network repository, we randomly select 3 to 5 networks from different categories to form the corpus. We consider the largest connected component in each original network and analyze the statistical characteristic of networks. Detailed information see the Table 2.

Appendix C. Correlation analysis

The effective spreading $\langle \tilde{s}_k \rangle$ is defined as $\langle s_k \rangle * p_k$ where the activation risk p_k is quantified by the degree-decay function. In this way, some existing centrality metrics might underperform because they are correlated with the degree. To confirm the guess, we analyze Kendall rank correlation between degree and other centrality metrics. In fact, the correlation between two metrics depends on the network structure and varies with real networks. Here, we use two distinct ways to analyze the correlation in order to obtain a relative objective and comprehensive result on 40 networks. On one hand, we average the correlation between any two centrality metrics over 40 networks, which might have some bias due to the existence of extreme value. On the other hand, we count the ratio of the network in which Kendall rank correlation between two centralities is higher than a threshold (threshold = 0.5). If the results obtained from different method show the same phenomenon, a general conclusion could be drawn.

The Kendall rank correlation between any two centrality metrics is analyzed. The results in Fig. 6(a)(c) show that centrality metrics could be roughly classified into three groups. The first group contains betweenness, Katz, subgraph, *KS*, and *CI*. The Kendall rank correlation between degree and them are very strong. The second group includes closeness, eigenvector, and *NB*. There is a weak correlation between them and degree. The third group includes $PI\_1, PI\_2, RA(2.5)$, and $RA(*)$, they are no correlated with a degree. Besides, the results show that centrality metrics within each group are correlated with each other, further confirming the reasonability of classification. Through the correlation analysis, we verify that it is not competitive to employ existing metrics as baseline methods to compare with the *RA* method, because the activation probability in effective spreading has already considered the degree-decay effect and there is a strong correlation between degree and centralities in existing metrics, which naturally weakens the performance of existing metrics.

To eliminate the degree-decay effect, we calculate the ratio between centralities and degree (e.g. Katz score/degree, betweenness score/degree) as benchmark centralities. In Fig. 6(b)(d), the results show that there is no correlation between most benchmark centralities and degree. As for the benchmark centrality of closeness, Katz, and *KS*, they have a negative correlation with the degree, which suggests that making a comparison between them and *RA* is more competitive when λ is very large. As a result, it is relatively fair to compare *RA* with other benchmark centralities by the effective spreading as a target function.

Appendix D. Statistical tests

Some rigorous statistical tests are provided to validate the result in the main text. We firstly make the significance tests of Kendall rank correlation between different methods and effective spreading coverage $\tilde{s}$. For various λ, the p-value for Kendall rank correlation coefficient is shown in Table 3. The result corresponds to Fig. 4(b). When $\lambda = 0.5$ and 0.6, the Kendall rank correlation is not significant for some baseline methods, so the null hypothesis is supported, which suggests that there is no correlation between methods (BTN/k, CLO/k, Katz/k, KS/k, Degree) and effective spreading. In most cases, the baseline methods are correlated to effective spreading. Besides, we also test the Kendall rank correlation between *RA* and effective spreading under different pair of parameters (λ and β) in Table 4, which corresponds to Fig. 4(d). For most pairs of parameters, the p-value is lower than 0.05, which means the correlation between *RA* and effective spreading coverage could be accepted. But for larger λ and smaller θ, or larger θ and smaller λ, there is no correlation between *RA* and effective spreading coverage.

For the Fig. 4 (c), some points among different methods might overlap due to the visual observation, we thus employ the Kolmogorov–Smirnov test (K-S test) to measure the difference of Kendall rank correlation distribution between *RA* and other baseline methods under various λ, further verify that performance of our method is distinct with others. The K-S test is one of the most useful nonparametric methods to quantify a distance between the empirical distribution function of two samples. For a given infection rate β, two samples in the K-S test are respectively from Kendall rank correlation of both baseline method and *RA* under different conditions. The p-value of the K-S test under different infection rates β is shown in Table 5. One could see that the p-value of the K-S test is equal to 0 between *RA* and any baseline methods, suggesting that the distribution of Kendall correlation of *RA* is statistically significantly different with other baseline methods.

[1] http://networkrepository.com/
[2] http://networksciencebook.com/translations/en/resources/data.html

Appendix E. Normalized score

According to different evaluation metrics, we normalize the performance of different methods on each network, further obtaining the overall performance on all networks. The normalized value of Kendall rank correlation between benchmark centralities and effective spreading could be seen in Table 6. Besides, we also show the normalized value of the standard deviation of Kendall rank correaltion between centralities and effective spreading coverage in Table 7.

Appendix F. Activation function

In defined problem, the choice of activation function is important to analytic tractability, and the optimal initial spreaders about the problem also depends on the slection of activation function. To provide evidence to support the claim, we choose the $\frac{1}{k^\gamma}$ as activation function to make a further analysis. $p_k = \frac{1}{k^\gamma}, \gamma$ is a parameter to control the risk difference among nodes with different degree. Firstly, we substitute $\frac{1}{k^\gamma}$ into the $\langle \tilde{s_k} \rangle = p_k * s_k$, see the Eqs. 27.

$$\langle \tilde{s_k} \rangle = \frac{1}{k^\gamma}\left(1 + \frac{k\beta}{1-\frac{\beta}{\beta_c}}\right) = k^{-\gamma} + k^{1-\gamma}\frac{\beta\beta_c}{\beta_c - \beta}. \tag{27}$$

Then, we derive analytic solution by setting $\frac{\partial\langle \tilde{s_k} \rangle}{\partial k} = 0, k^* = \left(\frac{1}{\beta} - \frac{1}{\beta_c}\right)\left(\frac{\gamma}{1-\gamma}\right), \beta < \beta_c$. (1) When $-1 \leqslant \gamma < 0, k^*$ is less than 0 and is the minimal point of $\langle \tilde{s_k} \rangle$, which suggests that the degree of the optimal initial spreaders in the problem should be the largest

Table 6
Normalized value of Kendall rank correlation between methods and effective spreading in all networks. The name of methods in the table is the same as the Table 3. The $RA(*)$ denotes that θ is determined by the smallest standard deviation of Kendall rank correlation over different λ in each network. The number in bold denotes the method performs well than other methods in the current network.

Networks	BTN/k	CLO/k	EIG/k	NB/k	Katz/k	SG/k	KS/k	CI(3)/k	Degree	PI_1	PI_2	RA(2.5)	RA(*)
email-Univ	0.69	0.00	0.85	0.84	0.00	0.99	0.11	0.85	0.88	0.74	0.80	**1.00**	0.94
ani-Aves-Songbird	0.57	0.00	0.87	0.86	0.00	0.98	0.11	0.12	0.82	0.59	0.59	**1.00**	**1.00**
ani-Dolphins	0.60	0.03	0.91	0.89	0.00	0.86	0.02	0.61	0.77	0.55	0.43	**1.00**	**1.00**
ani-Reptilia	0.42	0.03	0.79	0.79	0.00	0.90	0.13	0.87	0.70	0.68	0.90	**1.00**	0.94
ani-Mammalia	0.47	0.00	0.76	0.72	0.01	**1.00**	0.15	0.94	0.68	0.73	0.83	0.95	0.93
bio-Celegans	0.18	0.57	**1.00**	**1.00**	0.50	0.78	0.63	0.00	0.22	0.95	0.84	0.64	0.95
bio-Grid-Plant	0.00	0.31	0.81	0.62	0.35	0.99	0.40	0.32	0.15	**1.00**	0.95	0.87	0.98
bio-Grid-Worm	0.07	0.99	0.69	0.67	**1.00**	0.61	0.89	0.07	0.00	0.69	0.68	0.43	0.69
bio-Yeast	0.07	0.51	0.86	0.87	0.51	0.86	0.34	0.23	0.00	**1.00**	0.97	0.93	0.93
bn-Mouse-Kasthuri	0.17	0.72	0.83	0.89	0.73	0.86	0.71	0.28	0.00	0.95	0.98	0.77	**1.00**
ca-CSphd	0.06	0.62	0.70	0.44	0.88	0.93	0.61	0.07	0.00	**1.00**	0.94	0.95	0.91
ca-Erdos992	0.00	0.68	0.88	0.87	0.67	**1.00**	0.48	0.15	0.07	0.94	0.89	0.78	0.96
ca-GrQc	0.06	0.02	0.97	0.93	0.00	**1.00**	0.11	0.71	0.26	0.91	0.95	0.83	0.94
ca-Netscience	0.08	0.00	0.66	0.61	0.01	0.99	0.11	0.66	0.31	0.94	0.94	**1.00**	**1.00**
econ-Mahindas	0.56	0.00	0.97	0.97	0.01	**1.00**	0.32	0.65	0.89	0.98	0.89	0.99	0.90
econ-Poli	0.07	0.71	0.55	0.58	0.95	0.96	0.69	0.08	0.00	**1.00**	0.94	0.89	0.89
econ-Wm1	0.35	0.02	0.99	0.99	0.00	0.94	0.25	0.24	0.78	0.90	0.39	**1.00**	0.92
email-Dnc	0.09	0.99	0.70	0.68	**1.00**	0.50	0.87	0.10	0.00	0.55	0.63	0.30	0.49
email- Corecipient	0.27	0.02	0.57	0.57	0.18	0.85	0.07	0.00	0.30	**1.00**	0.69	0.60	0.94
email-Enron-Only	0.48	0.01	0.90	0.89	0.00	0.98	0.21	0.46	0.81	0.67	0.76	**1.00**	**1.00**
hs-Arenas-Jazz	0.53	0.01	0.90	0.89	0.00	**1.00**	0.23	0.11	0.88	0.55	0.60	0.98	0.98
hs-Physical	0.56	0.00	0.81	0.77	0.01	0.96	0.07	0.10	0.72	0.47	0.50	0.87	**1.00**
hs-Zachary	0.50	0.28	0.65	0.60	0.31	0.72	0.34	0.00	0.45	0.93	0.65	0.99	**1.00**
ia-Crime-Moreno	0.64	0.02	0.97	0.94	0.00	0.51	0.09	0.84	0.58	0.81	0.86	**1.00**	**1.00**
ia-Fb-Messages	0.75	0.00	0.80	0.79	0.00	0.96	0.13	0.71	0.87	0.99	0.91	**1.00**	0.99
ia-Infect-Dublin	0.51	0.04	0.86	0.86	0.00	**1.00**	0.23	0.80	0.87	0.53	0.59	**1.00**	0.95
inf-Euroroad	0.06	0.16	0.99	**1.00**	0.00	0.31	0.08	0.95	0.37	0.34	0.59	0.73	0.81
inf-Openflights	0.16	0.00	0.87	0.87	0.00	0.87	0.29	0.40	0.34	0.99	0.86	0.71	**1.00**
inf-Power	0.21	0.01	**1.00**	0.25	0.09	0.52	0.00	0.75	0.32	0.59	0.76	0.87	0.90
inf-Usair97	0.30	0.00	0.70	0.70	0.00	0.87	0.22	0.05	0.50	**1.00**	0.78	0.79	0.98
Metabolic	0.26	**1.00**	0.72	0.71	0.98	0.46	0.88	0.54	0.00	0.44	0.66	0.24	0.44
Protein	0.00	0.38	0.85	0.88	0.39	0.96	0.33	0.22	0.01	**1.00**	0.93	0.86	0.97
rt-Retweet	0.00	0.49	**1.00**	0.98	0.47	0.75	0.13	0.17	0.08	0.98	**1.00**	**1.00**	**1.00**
rt-Twitter	0.00	0.49	**1.00**	**1.00**	0.46	0.95	0.24	0.17	0.00	**1.00**	0.94	0.87	0.95
soc-Karate	0.20	0.14	0.46	0.44	0.16	0.62	0.28	0.00	0.33	0.87	0.50	0.96	**1.00**
socfb-Caltech36	0.74	0.00	0.75	0.75	0.04	**1.00**	0.11	0.16	0.97	0.70	0.62	**1.00**	**1.00**
socfb-Haverford76	0.72	0.02	0.88	0.88	0.01	**1.00**	0.11	0.00	0.93	0.58	0.49	0.98	0.98
socfb-Reed98	0.60	0.00	0.83	0.82	0.07	0.97	0.17	0.06	0.90	0.73	0.72	**1.00**	0.97
socfb-Simmons81	0.80	0.00	0.74	0.74	0.03	**1.00**	0.06	0.16	0.96	0.58	0.55	0.94	0.87
web-EPA	0.11	0.96	0.81	0.77	**1.00**	0.73	0.88	0.11	0.00	0.85	0.86	0.43	0.83

Table 7
Normalized value of the standard deviation of Kendall rank correlation between methods and effective spreading. The name of methods in the table is the same as the Table 3. The $RA(*)$ denotes that θ is determined by the smallest standard deviation of Kendall rank correlation among all parameters θ ($\theta \in [0, 9.5]$ with a step of 0.5). The number in bold denotes that Kendall rank correlation of the method has the smallest volatility under different λ.

Networks	BTN/k	CLO/k	EIG/k	NB/k	Katz/k	SG/k	KS/k	CI(3)/k	Degree	PI_1	PI_2	RA(2.5)	RA(*)
email-Univ	0.68	0.96	0.03	**0.00**	0.97	0.56	0.98	0.34	1.00	0.52	0.13	0.45	0.07
ani-Aves-Songbird	0.55	0.98	**0.00**	0.02	0.98	0.57	0.99	0.53	1.00	0.80	0.60	0.07	0.07
ani-Dolphins	0.46	0.92	0.04	0.13	0.95	0.60	0.96	**0.00**	1.00	0.78	0.81	0.17	0.17
ani-Reptilia	0.42	0.91	0.02	**0.00**	0.92	0.33	0.64	0.54	1.00	0.59	0.29	0.29	0.06
ani-Mammalia	0.68	0.93	**0.00**	**0.00**	0.95	0.53	0.95	0.71	1.00	0.63	0.29	0.17	0.12
bio-Celegans	0.69	0.92	0.12	0.11	0.94	0.55	0.76	0.30	1.00	0.36	**0.00**	0.65	0.19
bio-Grid-Plant	0.59	0.85	0.14	**0.00**	0.87	0.26	0.81	0.66	1.00	0.30	0.10	0.35	0.14
bio-Grid-Worm	0.73	0.77	**0.00**	0.02	0.77	0.19	0.59	0.88	1.00	0.15	0.09	0.48	0.11
bio-Yeast	0.74	0.79	0.01	**0.00**	0.79	0.05	0.88	0.88	1.00	0.31	0.08	0.06	0.06
bn-Mouse-Kasthuri	0.80	0.79	0.06	**0.00**	0.79	0.09	0.76	0.80	1.00	0.14	0.09	0.22	0.10
ca-CSphd	0.88	0.69	0.01	**0.00**	0.69	0.46	0.95	0.89	1.00	0.33	0.12	0.19	0.13
ca-Erdos992	0.86	0.63	0.01	**0.00**	0.63	0.18	0.78	0.95	1.00	0.19	0.05	0.37	0.13
ca-GrQc	0.42	0.92	0.06	**0.00**	0.93	0.23	0.50	0.57	1.00	0.08	0.03	0.51	0.01
ca-Netscience	0.77	0.90	**0.00**	0.01	0.93	0.54	0.84	0.46	1.00	0.67	0.32	0.20	0.20
econ-Mahindas	0.16	0.95	0.41	0.41	0.95	0.55	0.44	0.26	1.00	0.14	0.08	0.68	**0.00**
econ-Poli	0.85	0.68	**0.00**	0.01	0.69	0.14	0.89	0.90	1.00	0.23	0.13	0.16	0.16
econ-Wm1	**0.00**	0.92	0.46	0.45	0.93	0.72	0.80	0.16	1.00	0.19	0.45	0.62	0.07
email-Dnc	0.73	0.74	**0.00**	0.01	0.74	0.38	0.62	0.72	1.00	0.36	0.14	0.67	0.42
email- Corecipient	0.48	0.93	0.05	0.04	0.73	0.45	0.63	0.10	1.00	0.07	0.02	0.78	**0.00**
email-Enron-Only	0.53	0.94	0.11	0.09	0.97	0.62	0.87	**0.00**	1.00	0.70	0.28	0.17	0.17
hs-Arenas-Jazz	0.55	0.98	**0.00**	**0.00**	0.99	0.62	0.85	0.42	1.00	0.77	0.59	0.09	0.09
hs-Physical	0.70	0.95	**0.00**	0.19	0.95	0.44	0.99	0.34	1.00	0.86	0.35	0.42	0.03
hs-Zachary	0.73	0.89	0.41	0.54	0.91	0.31	0.95	**0.00**	1.00	0.51	0.73	0.27	0.12
ia-Crime-Moreno	0.77	0.90	**0.00**	0.23	0.90	**0.00**	0.99	0.87	1.00	0.60	0.24	0.11	0.11
ia-Fb-Messages	0.74	0.96	**0.00**	0.04	0.96	0.75	0.99	0.08	1.00	0.19	0.11	0.66	0.12
ia-Infect-Dublin	0.31	0.96	**0.00**	**0.00**	0.98	0.28	0.80	0.06	1.00	0.69	0.38	0.24	0.07
inf-Euroroad	0.45	0.77	0.01	**0.00**	0.79	0.28	0.82	0.70	1.00	0.69	0.43	0.17	0.06
inf-Openflights	0.61	0.92	0.11	0.11	0.92	0.47	0.80	0.34	1.00	0.11	0.34	0.70	**0.00**
inf-Power	0.62	0.86	0.12	**0.00**	0.87	0.30	0.91	0.80	1.00	0.73	0.46	0.25	0.14
inf-Usair97	0.74	0.96	**0.00**	0.01	0.96	0.60	0.83	0.03	1.00	0.28	0.04	0.80	0.22
Metabolic	0.32	0.91	0.08	0.08	0.93	0.51	0.65	0.52	1.00	0.58	**0.00**	0.75	0.52
Protein	0.73	0.82	0.12	0.08	0.82	0.09	0.84	0.87	1.00	0.03	0.07	0.20	**0.00**
rt-Retweet	0.84	0.78	0.02	0.19	0.78	**0.00**	0.93	0.78	1.00	0.41	0.07	0.05	0.05
rt-Twitter	0.82	0.78	0.06	0.02	0.78	**0.00**	0.93	0.91	1.00	0.28	0.09	0.22	0.05
soc-Karate	0.75	0.89	0.29	0.49	0.89	0.26	1.00	0.02	0.99	0.47	0.70	0.22	**0.00**
socfb-Caltech36	0.64	0.99	**0.00**	0.03	0.93	0.87	0.97	0.82	1.00	0.61	0.38	0.12	0.12
socfb-Haverford76	0.62	1.00	0.02	**0.00**	1.00	0.76	0.97	0.79	1.00	0.67	0.61	0.01	0.01
socfb-Reed98	0.70	0.99	0.01	**0.00**	0.94	0.84	0.96	0.80	1.00	0.58	0.07	0.34	0.10
socfb-Simmons81	0.66	0.99	0.02	**0.00**	0.93	0.79	0.98	0.65	1.00	0.60	0.23	0.16	0.09
web-EPA	0.80	0.83	**0.00**	0.01	0.83	0.21	0.82	0.91	1.00	0.09	0.01	0.60	0.08

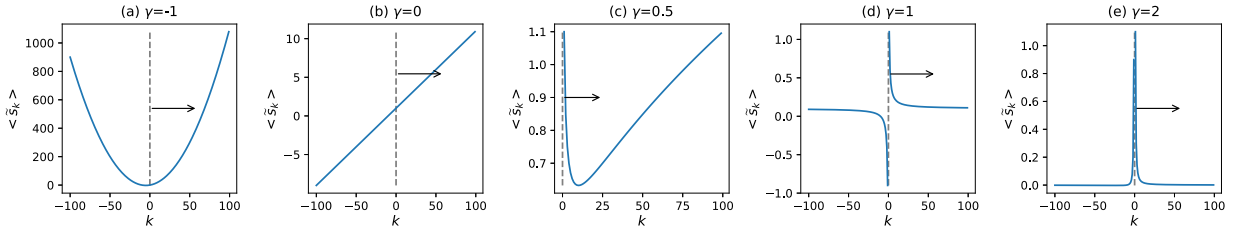

Fig. 7. The effective spreading $\langle \tilde{s_k} \rangle$ for various γ, $\langle \tilde{s_k} \rangle = k^{-\gamma} + k^{1-\gamma} q, q = \frac{\beta\beta_c}{\beta_c - \beta}$. Here q is set as 0.1. One could clearly see that the change of function form as the increase of γ. The k value corresponding to largest effective spreading coverage on the right side of dashed line is the optimal degree of node in the problem.

degree among all nodes in the network (see the Fig. 7(a)). (2) When $\gamma = 0$, then $\langle \tilde{s_k} \rangle = 1 + \frac{k\beta}{1-\frac{\beta}{\beta_c}}$, the problem degrenerates into the original problem (see the Fig. 7(b)) and the degree of the optimal initial spreaders is the largest degree among all nodes in the network. (3) When $0 < \gamma < 1, k^*$ is greater than 0 and corresponds to the minimal point of $\langle \tilde{s_k} \rangle$ (see the Fig. 7(c)). The degree of the optimal initial spreaders in the problem is difficult to be determined. By calculating partial derivatives for

$\gamma, \frac{\partial k^*}{\partial \gamma} = \left(\frac{1}{\beta} - \frac{1}{\beta_c}\right)\left(\frac{1}{1-\gamma}\right)^2 > 0$, we find that k^* increases as γ, which suggests that the minimal point of $\langle \tilde{s_k} \rangle$ tends to be large as the increase of γ, further inferring that the degree of the optimal initial spreaders should be the largest degree among all nodes when γ is close to 0 and the smallest degree among all nodes when γ is close to 1. (4) When $\gamma = 1, \langle \tilde{s_k} \rangle = k^{-1} + \frac{\beta\beta_c}{\beta_c - \beta}$ (see the Fig.7(d)), which suggests the degree of the optimal initial spreaders should be the smallest degree among all nodes in the network. In summary, the degree of the optimal initial spreaders to maximize the effective spreading $\langle \tilde{s_k} \rangle$ corresponds to the largest degree among all nodes in network when $\gamma < \gamma_c$, and corresponds to the smallest degree among all nodes in network when $\gamma > \gamma_c$ (γ_c is a threshold). The results here are qualitatively similar to the exponential decay function, namely the degree of the optimal initial spreaders have a negative relation with γ.

We conduct the experiment to test the performance of risk-aware metric according to the activation function ($\frac{1}{k^\gamma}$). γ is set from 0 to 1 with a step of 0.1. The setting of γ is the same as the risk parameter λ. The experimental result is shown in Fig. 8. One could see that the performance of $RA(\theta = 2.5)$ for $\langle \tau \rangle, S_\tau$ and $\langle \tilde{s} \rangle$ is poorer than other baseline methods. The $RA(\theta^*)$ is used to make a comparison, it lacks an advantage. The possible reason behind the results is that the optimal initial spreaders for most of γ in interval $\gamma \in [0, 1]$ favor the largest-degree nodes in the network, and the optimal initial spreaders for less γ favor the smallest-degree nodes. In other words, the degree of the optimal spreaders is not continuous when we consider $\frac{1}{k^\gamma}$ as an activation function, which could be confirmed by the simulation result. In the Eqs. 27, the effective spreading $\langle \tilde{s_k} \rangle = k^{-\gamma} + k^{1-\gamma} q, q = \frac{\beta\beta_c}{\beta_c - \beta}$. When q is given, the $\langle \tilde{s_k} \rangle$ plotted as a function of k for different γ could be seen in the Fig. 9(a). One could clearly observe the change of effective spreading as the increase of γ. In Fig. 9(b), the degree of the optimal initial spreaders for various γ shows the discontinuous transition. The condition where the degree of the optimal spreaders is in the intermediate degree does not exist, which is different from the analytical result of exponential decay activation. When we select $\frac{1}{k^\gamma}$ as an activation function, the largest effective spreading coverage among all nodes does not correspond to the nodes linked to many hubs but the largest-degree node or smallest-degree node in the real networks. This could explain why the *RA* has poorer performance than other metrics, and degree and subgraph benchmark centrality have better performance.

Through the above analysis, the inversely proportional function might not be a good form as activation function because it is hard to analyze the degree of the optimal initial spreaders and there is no exact analytical solution for the optimal degree value in the defined problem. Intuitively, the optimal initial spreaders obtained from the $\frac{1}{k^\gamma}$ is not agreed well with the real condition, although it could characterize the negative relation between the degree and activation probability. Therefore, the form of the activation function is important to the quantification of the problem. The conclusion obtained from the result depends on the activation function.

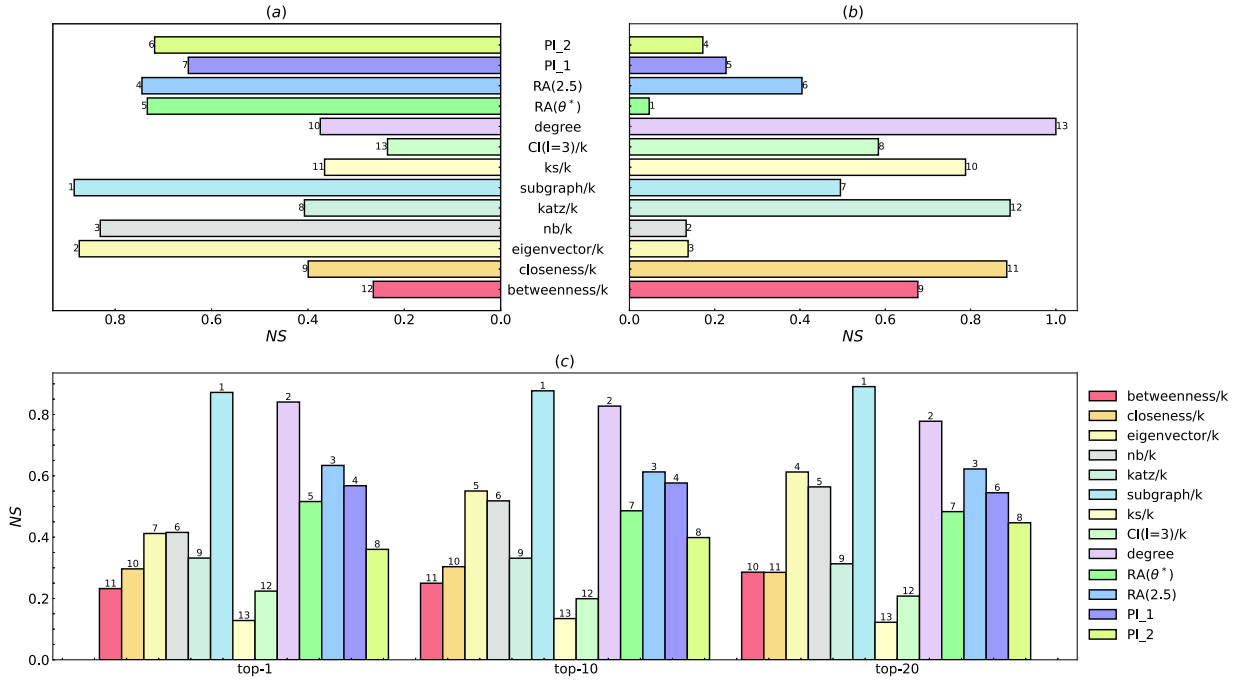

Fig. 8. Based on the activation function of $\frac{1}{k^\gamma}$, the normalized score of evaluation metrics for different methods in all networks. (**a**) The normalized score of average of Kendall τ. (**b**) The normalized score of standard deviation of Kendall τ. (**c**) The normalized score of average of effective spreading coverage $\langle \tilde{s_n} \rangle$.

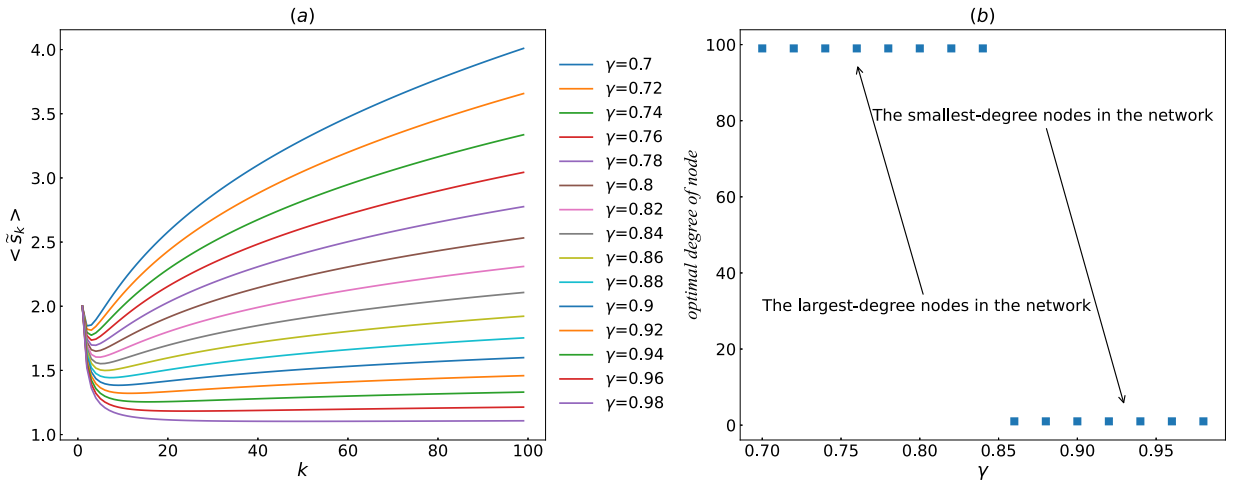

Fig. 9. (**a**) The $\langle \tilde{s}_k \rangle$ plotted as a function of k for different $\gamma, \gamma \in [0.7, 0.98]$ and $q = 1$. (**b**) The optimal initial spreaders to maximize the effective spreading for different γ. Here, we assume that the network size is 100.

References

[1] D. Centola, The spread of behavior in an online social network experiment, Science 329 (5996) (2010) 1194–1197.
[2] A. Montanari, A. Saberi, The spread of innovations in social networks, Proc. Nat. Acad. Sci. 107 (47) (2010) 20196–20201.
[3] L. Yang, Z. Li, A. Giua, Containment of rumor spread in complex social networks, Inf. Sci. 506 (2020) 113–130.
[4] R. Pastor-Satorras, A. Vespignani, Immunization of complex networks, Phys. Rev. E 65 (3) (2002) 036104.
[5] M. Opuszko, J. Ruhland, Effects of the network structure on the dynamics of viral marketing, Wirtschaftsinformatik 94 (2013).
[6] M. Kimura, K. Saito, R. Nakano, Extracting influential nodes for information diffusion on a social network, in: AAAI, Vol. 7, 2007, pp. 1371–1376..
[7] T.R. Frieden, C.T. Lee, Identifying and interrupting superspreading events-implications for control of severe acute respiratory syndrome coronavirus 2, Emerg. Infect. Dis. 26 (6) (2020).
[8] F. Morone, H.A. Makse, Influence maximization in complex networks through optimal percolation, Nature 524 (7563) (2015) 65–68.
[9] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, T. Zhou, Vital nodes identification in complex networks, Phys. Rep. 650 (2016) 1–63.
[10] S. Aral, P.S. Dhillon, Social influence maximization under empirical influence models, Nat. Human Behav. 2 (6) (2018) 375–382.
[11] G. Sabidussi, The centrality index of a graph, Psychometrika 31 (4) (1966) 581–603.
[12] L.C. Freeman, Centrality in social networks conceptual clarification, Social Networks 1 (3) (1978) 215–239.
[13] P. Bonacich, Some unique properties of eigenvector centrality, Social Networks 29 (4) (2007) 555–564.
[14] J. Zhan, S. Gurung, S.P.K. Parsa, Identification of top-k nodes in large networks using katz centrality, J. Big Data 4 (1) (2017) 1–19.
[15] E. Estrada, J.A. Rodriguez-Velazquez, Subgraph centrality in complex networks, Phys. Rev. E 71 (5) (2005) 056103..
[16] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks, Nat. Phys. 6 (11) (2010) 888–893.
[17] Ş. Erkol, C. Castellano, F. Radicchi, Systematic comparison between methods for the detection of influential spreaders in complex networks, Sci. Rep. 9 (1) (2019) 1–11.
[18] A. Zeng, C.-J. Zhang, Ranking spreaders by decomposing complex networks, Phys. Lett. A 377 (14) (2013) 1031–1035.
[19] L. Lü, T. Zhou, Q.-M. Zhang, H.E. Stanley, The h-index of a network node and its relation to degree and coreness, Nat. Commun. 7 (2016) 10168.
[20] A. Zareie, A. Sheikhahmadi, M. Jalili, M.S.K. Fasaei, Finding influential nodes in social networks based on neighborhood correlation coefficient, Knowl.-Based Syst. (2020), 105580.
[21] E. Bakshy, J.M. Hofman, W.A. Mason, D.J. Watts, Everyone's an influencer: quantifying influence on twitter, in, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 2011, pp. 65–74.
[22] A. Lanz, J. Goldenberg, D. Shapira, F. Stahl, Climb or jump: status-based seeding in user-generated content networks, J. Mark. Res. 56 (3) (2019) 361–378.
[23] D.J. Daley, D.G. Kendall, Epidemics and rumours, Nature 204 (4963) (1964) 1118.
[24] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Mod. Phys. 87 (3) (2015) 925.
[25] M.E. Newman, Spread of epidemic disease on networks, Phys. Rev. E 66 (1) (2002) 016128.
[26] H. Liao, M.S. Mariani, M. Medo, Y.-C. Zhang, M.-Y. Zhou, Ranking in evolving complex networks, Phys. Rep. 689 (2017) 1–54.
[27] F. Zhou, L. Lü, M.S. Mariani, Fast influencers in complex networks, Commun. Nonlinear Sci. Numer. Simul. 74 (2019) 69–83.
[28] W. Khaouid, M. Barsky, V. Srinivasan, A. Thomo, K-core decomposition of large networks on a single pc, Proc. VLDB Endowment 9 (1) (2015) 13–23.
[29] K. Das, S. Samanta, M. Pal, Study on centrality measures in social networks: a survey, Social Network Anal. Min. 8 (1) (2018) 1–11.
[30] U. Brandes, A faster algorithm for betweenness centrality, J. Math. Sociol. 25 (2) (2001) 163–177.
[31] M. Lin, W. Li, L.J. Song, C.-T. Nguyen, X. Wang, S. Lu, Sake: estimating katz centrality based on sampling for large-scale social networks, ACM Trans. Knowl. Discovery Data (TKDD) 15 (4) (2021) 1–21.
[32] E. Nathan, G. Sanders, J. Fairbanks, D.A. Bader, et al, Graph ranking guarantees for numerical approximations to katz centrality, Proc. Comput. Sci. 108 (2017) 68–78.
[33] F. Morone, B. Min, L. Bo, R. Mari, H.A. Makse, Collective influence algorithm to find influencers via optimal percolation in massively large social media, Sci. Rep. 6 (1) (2016) 1–11.
[34] F. Radicchi, C. Castellano, Leveraging percolation theory to single out influential spreaders in networks, Phys. Rev. E 93 (2016) 062314.
[35] T. Martin, X. Zhang, M.E.J. Newman, Localization and centrality in networks, Phys. Rev. E 90 (2014) 052808.
[36] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, P. Zhang, Spectral redemption in clustering sparse networks, Proc. Nat. Acad. Sci. 110 (52) (2013) 20935–20940.

[37] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93.
[38] H. Jeong, S.P. Mason, A.-L. Barabási, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (6833) (2001) 41–42.
[39] H. Yu, P. Braun, M.A. Yıldırım, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R.R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M.E. Hudson, J. Park, X. Xin, M.E. Cusick, T. Moore, C. Boone, M. Snyder, F.P. Roth, A.-L. Barabási, J. Tavernier, D.E. Hill, M. Vidal, High-quality binary protein interaction map of the yeast interactome network, Science 322 (5898) (2008) 104–110.
[40] J. Schellenberger, J.O. Park, T.M. Conrad, B.Ø. Palsson, Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions, BMC Bioinf. 11 (1) (2010) 1–10.
[41] M.E. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (3) (2006) 036104.
[42] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol. 54 (2003) 396–405.
[43] D.J. Watts, S.H. Strogatz, Collective dynamics of small-world networks, Nature 393 (6684) (1998) 440–442.
[44] D.A. Bader, H. Meyerhenke, P. Sanders, D. Wagner, Graph partitioning and graph clustering, in: 10th DIMACS Implementation Challenge Workshop, 2012..
[45] P.M. Gleiser, L. Danon, Community structure in jazz, Adv. Complex Syst. 6 (4) (2003) 565–573.
[46] J. Coleman, E. Katz, H. Menzel, The diffusion of an innovation among physicians, Sociometry (1957) 253–270.
[47] W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. 33 (1977) 452–473.
[48] R.A. Rossi, N.K. Ahmed, The network data repository with interactive graph analytics and visualization, in: AAAI, 2015.
[49] Y. Hu, C. Van den Bulte, Nonmonotonic status effects in new product adoption, Market. Sci. 33 (4) (2014) 509–533.
[50] M.E.J. Newman, S.H. Strogatz, D.J. Watts, Random graphs with arbitrary degree distributions and their applications, Phys. Rev. E 64 (2001) 026118.