Theories for Influencer Identification in Complex Networks

Sen Pei, Flaviano Morone, and Hernán A. Makse

1 Introduction

In spreading processes of information, it is well known that certain individuals are more influential than others. In the field of information diffusion, it has been accepted that the ability of influencers to initiate a large-scale spreading is attributed to their privileged locations in the underlying social networks [41, 59, 71, 92]. Due to the direct relevance of influencer identification in such phenomena as viral marketing [46], innovation diffusion [81], behavior adoption [17], and epidemic spreading [69], the research on searching for influential spreaders in different settings is becoming increasingly important in recent years [71].

In the relative simple case of locating individual influencers, given the rich structural information encoded in nodes' location in the network, it is straightforward to measure the influence of a single node using centrality-based heuristics. Over the years, a growing number of predictors have been developed and routinely employed to rank single node's influence in spreading processes, among which the most widely used ones include number of connection [1], k-core [85], betweenness centrality [25], and PageRank [13], just to name a few. Beyond this non-interacting problem, a more challenging task is to identify a set of influencers to achieve maximal collective influence. Originally formulated in the context of viral marketing [80], collective influence maximization is in fact a core optimization problem in an array of important applications in various domains, ranging from cost-effective

S. Pei (✉)

Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY, USA
e-mail: sp3449@cumc.columbia.edu

F. Morone · H. A. Makse
Levich Institute and Physics Department, City College of New York, New York, NY, USA
e-mail: hmakse@lev.ccny.cuny.edu

marketing in commercial promotion, optimal immunization in epidemic control, to strategic protection against targeted attacks on infrastructures. In addition to the topological complexity of network structure, collective influence maximization is further complicated by the entwined interactions between multiple spreaders, which renders the aforementioned centrality-based approaches invalid. As a result, it is required to treat the problem from a collective point of view to develop effective solutions [61].

2 Finding Individual Influencers

In reality, many spreading phenomena are typically initiated by a single spreader. For instance, an epidemic outbreak in a local area is usually caused by the first infected person. For such processes, ranking the spreading capability of individual spreaders is of great significance in both accelerating and confining the diffusion.

2.1 *Topological Measures*

Intuitively, the nodes with large numbers of connections should have more influence on their direct neighbors. The disproportionate effect of highly-connected nodes, or hubs, on dynamical processes has been revealed in the early works on the vulnerability of scale-free networks [1, 23]. The targeted attack on a very small number of high-degree nodes will rapidly collapse the giant component of networks with heavy-tailed degree distribution. Compared with other more complex centrality measures, the computational burden of degree is almost negligible. Due to this, the simple degree centrality has been playing an important role in influencer identification. In implementation, the performance of high-degree ranking can be further enhanced by a simple adaptive calculation procedure, that is, recalculating the degree of remaining nodes after the removal of previously selected nodes.

An obvious drawback of degree centrality is that it only considers the number of direct neighbors. However, as indicated by empirical studies, most spreading phenomena are proceeded in a cascading fashion. Therefore, the ultimate influence of a single spreader is also affected by the global network structure. In realistic complex networks, high-degree nodes can appear at either the core area or the periphery region. This implies, the number of connections may not be a reliable indicator of influencers in real-world systems. Recently, Kitsak et al. confirmed this speculation through extensive simulations of susceptible-infected-recovered (SIR) and susceptible-infected-susceptible (SIS) dynamics on diverse real-world social networks [41]. In SIR model, a susceptible individual will become infected with a probability β upon contact with his/her infected neighbors, and infected population will recover with a probability μ and become immune to the disease. In SIS model, the infection follows the same dynamics but infected persons will become
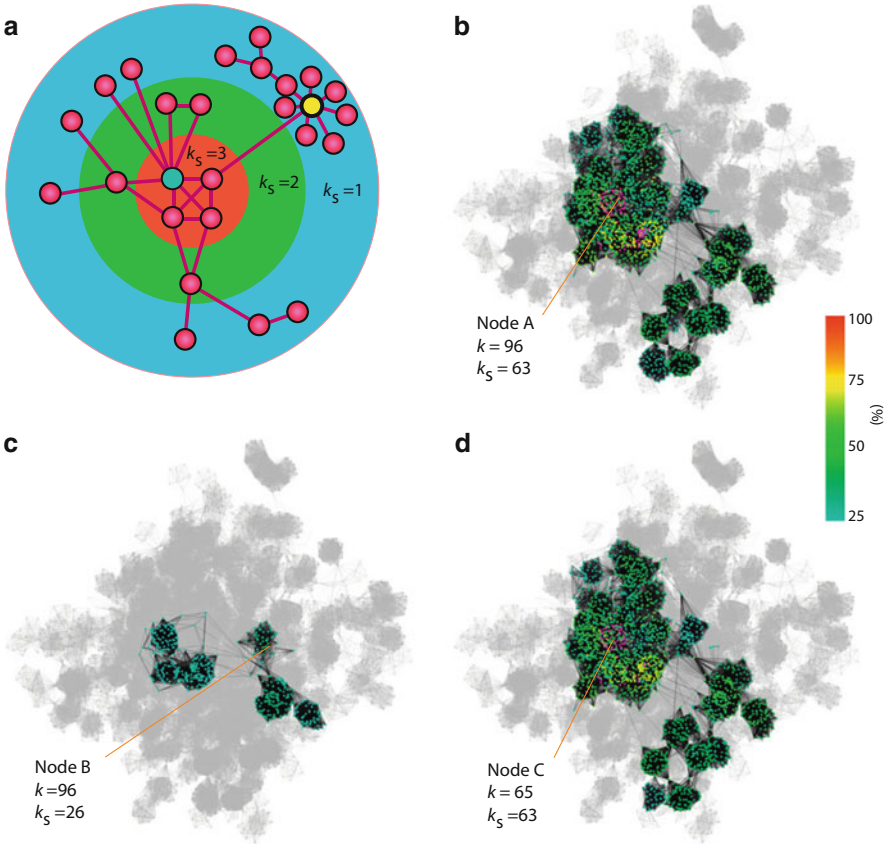
Fig. 1 (**a**), A schematic diagram of k-shell decomposition. The two highlighted nodes (blue and yellow), although both with degree $k = 8$, are in different k-shells. (**b–d**), Infections starting from single nodes with same degree $k = 96$ (A and B) can result in totally different outcomes. Whereas, infections originating from node C, locating in the same k-shell of node A ($k_S = 63$) but with a smaller degree, are quite similar to the spreading from node A. The colors indicate nodes' probability to be infected in SIR simulations with infection rate $\beta = 0.035$ and recovery rate $\mu = 1$. Results are averaged over 10,000 realizations. Figure is adapted from Kitsak et al. [41]

susceptible again with a probability μ. As shown in Fig. 1b–d, SIR spreading processes initiated by two hubs with the same degree could result in quite different infected population, depending on their global position in the network. In contrast, the k-core index, which distinguishes the network core and periphery, is a more reliable predictor of influence.

The k-core index is obtained by the k-shell decomposition in which nodes are iteratively pruned according to their remaining degree in the network (see Fig. 1a) [85]. Specifically, nodes with degree $k = 1$ are first removed successively until there is no node left with one link. The removed nodes are assigned with

k-core index $k_S = 1$. Then we remove nodes with degree $k = 2$ similarly and continue to prune higher k-shells until no node left in the network. In terms of computational complexity, the above decomposition process can be finished within $O(M)$ operations, where M is the number of links [7]. Thus k-core ranking is feasible for large-scale complex networks encountered in big-data analysis.

As illustrated in Fig. 1a, the classification of k-core can be very different from that of degree. A hub with low k-core index is usually surrounded by many low-degree neighbors that limit the influence of the hub. On the contrary, nodes located in the core region, although may have moderate degree, are capable of generating large-scale spreading facilitated by their well-connected neighbors. In the case where recovered individuals do not develop immunity, infections would persist in the high k-core area. These findings challenge the previous predominate focus on the number of connections. The simple yet effective measure k-core has inspired several generalizations in consideration of the detailed local environment in the vicinity of high k-core nodes [50, 51, 54, 95].

Although k-core was found effective in SIR and SIS spreading dynamics, some studies indicate that it may not be a good predictor of influence for other spreading models. For instance, in rumor spreading model, Borge-Holthoefer and Moreno [11] showed that the spreading capabilities of the nodes did not depend on their k-core values. These contradictory results relying on the choice of specific spreading model necessitate more extensive empirical validation with real information flow [72].

Apart from the k-core index, another measure that takes into account the global network structure is eigenvector centrality [10, 79]. The reasoning behind the eigenvector centrality is that the influence of an individual is determined by the spreading capability of his/her neighbors. Starting from a uniform score assigned to each node, the scores propagate along the links until a steady state is reached. In calculation, each step of score propagation corresponds to a left multiplication of the adjacency matrix to the current score vector. This procedure is actually the power method to compute the principal eigenvalue of the adjacency matrix. As a result, the steady score vector is in fact proportional to the right eigenvector corresponding to the largest eigenvalue. Notice that, supposing the initial score of each node is one, the first step of iteration will recover the degree centrality.

Despite the wide application of eigenvector centrality, it was recently found that the scores could be localized at a few high degree nodes due to the repeated reflection of scores from their neighbors during the iteration. Martin et al. solved this problem by using the leading eigenvector of the Hashimoto Non-Backtracking (NB) matrix [56]. In NB matrix, the immediate backtracking paths $i \rightarrow j$ and $j \rightarrow i$ are not permissible [34], thus avoiding the heavy score accumulation caused by the recurrent one-step reflection. Recently, by mapping the SIR spreading process to bond percolation, Radicchi and Castellano proved that the NB centrality was an optimized predictor for single influencers in SIR model at criticality [76]. In the next section, we will see the important role of NB matrix in collective influence maximization and optimal percolation [61].

2.2 Dynamics-Based Measures

Beyond the above pure topological measures, a number of centralities are developed on the basis of specific assumptions on the spreading dynamics. In some classical centralities proposed in the field of social networks, much emphasis is put on the shortest path. Along this way, several renowned centralities were developed and widely accepted in social network ranking. For instance, the closeness centrality quantifies the shortest distance from a given node to all other reachable nodes in the network [84], while betweenness centrality measures the fraction of shortest paths cross through a certain individual between all node pairs [25]. A useful generalization of closeness centrality is the Katz centrality [39], which considers all possible paths in the network, but assigns a larger weight to shorter paths using a tunable parameter. In application, the applicability of these shortest-path-based centralities is limited by the high computational complexity of calculating the shortest paths between all pairs of nodes. As a result, they are more suitable for small or medium scale networks.

Another group of metrics are designed based on random walks. A famous random walk based centrality is PageRank [13]. As a revolutionary webpage ranking algorithm, PageRank mimics a random walk process along the directed hyperlinks. To avoid the random walker trapped in the dangled nodes, a jumping probability α is introduced to allow the walker jump to a randomly chosen node. The PageRank score is the stationary probability of each node to be visited by the random walker, which can be calculated through iteration. In applications, the PageRank of a node i in a network can be calculated from $p_t(i) = \frac{1-\alpha}{N} + \alpha \sum_j \frac{A_{ij} p_{t-1}(j)}{k_{\text{out}}(j)}$, where $k_{\text{out}}(j)$ is the number of outgoing links from node j and α is the jumping probability. In a generalization called LeaderRank [53], a ground node is connected to all other nodes by additional bidirectional links. This procedure ensures the network to be strongly connected so that the convergence becomes faster.

In addition to the aforementioned centralities designed for general spreading processes, several measures are proposed aimed at specific dynamics, depending explicitly on model parameters. In these approaches, the development of measures is based on the equations depicting the dynamical process. Usually, the analysis of equations will naturally lead to the procedure of path counting in which the number of possible spreading paths is assessed. For instance, Klemm et al. developed a general framework to evaluate the dynamical importance (DI) of nodes in a series of dynamical processes [43]. The iterative calculation of DI centrality essentially counts the total number of arbitrarily long walks departing from each node. Another metric relying on possible spreading paths is the expected force (ExF) proposed by Lawyer [45]. To compute the expected force, all possible clusters of infected nodes after n transmission events starting from a given node are enumerated. Then the entropy of their cluster degree (i.e., number of outgoing links of the cluster, or infected-susceptible edges) is calculated as the expected force for each node.

The approaches introduced here are far from complete. A growing number of metrics and methods are continuously proposed in the active area of finding single influencers [52]. In designing effective methods for more complex spreading models, the basic principles behind these measures should be universal.

3   Finding Multiple Influencers

In spite of the great value of estimating individual nodes' influence with centralities, in a realistic situation, it is more relevant to understand spreading processes initiated by several spreaders. In applications such as viral marketing, it is expected that the spreaders can be coordinated in an optimal manner so that the final collective influence will be maximized. Although it sounds similar to the problem of locating single influencers, the collective influence maximization is in fact a fundamentally different and more difficult problem. In the seminal work of Kempe et al. [40], the influence maximization problems in both Independent Cascade Model (ICM) and Linear Threshold Model (LTM) were mapped to the NP-complete Vertex Cover problem. This implies, the influence maximization problem cannot be solved exactly within a polynomial time, leaving us the only choice of heuristic approach.

A straightforward idea to find multiple influencers is to select the top-ranked spreaders as individual seeds using centrality measures. However, this approach neglects the interactions and collective effect among spreaders. As demonstrated in SIR simulations, the selected spreaders have significant overlap in their influenced population [41]. Therefore, the set of influencers identified with centrality metrics are usually far from optimal. To solve this conundrum, it needs to be treated from a collective point of view [61].

3.1   Optimal Percolation

We start our discussion from the percolation model point of view. As a well-studied dynamical process, percolation was shown to be closely related to spreading and immunization [16, 67, 70]. Percolation is a classical physical process in which nodes or links are randomly removed from a graph [86]. The critical quantity that is of particular interest is the fraction of nodes or links whose removal will collapse the giant component. It is well known that the size of giant component decreases continuously to zero as the number of removed nodes or links increases. In the pioneering works of Newman [67, 68], the class of SIR models were mapped to the percolation process for which the critical point of the continuous transition could be solved exactly.

In contrast to the studies focused on random removal, the problem of optimal percolation aims to find the minimal set of nodes that could guarantee the global connectivity of the network, or equivalently, dismantle the network if removed.

Morone and Makse showed that, mathematically, the optimization of spreading process following *exactly* the Linear Threshold Model with threshold $k - 1$ (k is the degree of each node) can be mapped to the optimal percolation problem [61]. For this specific spreading model, finding the minimum number of seeds so that the information percolates the entire network is essentially equivalent to locating the optimal set of nodes in the optimal percolation problem. Similarly, the optimal immunization problem, dual of optimal spreading, can also be mapped to optimal percolation [61]. The relation between the cohesion of a network and influence spreading indicates that the most influential spreaders are the nodes that maintain the integrity of the network.

The collective influence theory for optimal percolation is developed based on the message passing equations of the percolation process. For a network with N nodes and M edges, suppose $\mathbf{n} = (n_i, \cdots, n_N)$ indicates whether node i is removed ($n_i = 0$) or left ($n_i = 1$) in the network. The total fraction of removed nodes is therefore $q = 1 - \sum_{i=1}^{N} n_i / N$. For a directed link from i to j ($i \rightarrow j$), let $v_{i \rightarrow j}$ denote the probability of node i belonging to the giant component G in the absence of node j. The evolution of $v_{i \rightarrow j}$ satisfies the following self-consistent equation:

$$v_{i \rightarrow j} = n_i \left[1 - \prod_{k \in \partial i \setminus j} (1 - v_{k \rightarrow i}) \right], \tag{1}$$

where $\partial i \setminus j$ denotes the nearest neighbors of i excluding j. The final probability v_i of node i belonging to the giant component is then determined by $v_{k \rightarrow i}$ ($k \in \partial i$) through

$$v_i = n_i \left[1 - \prod_{k \in \partial i} (1 - v_{k \rightarrow i}) \right]. \tag{2}$$

The fraction of nodes in the giant component is then given by $G(q) = \sum_{i=1}^{N} v_i / N$.

For the continuous phase transition in percolation process, the stability of the zero solution $G = 0$ is determined by the largest eigenvalue $\lambda(\mathbf{n}; q)$ of the coupling matrix $\mathscr{M}$ for the linearized Eq. (1) evaluated at $\{v_{i \rightarrow j} = 0\}$ (see Fig. 2a). Concretely, $\mathscr{M}$ is defined on the $2M \times 2M$ directed links as $\mathscr{M}_{k \rightarrow \ell, i \rightarrow j} \equiv \frac{\partial v_{i \rightarrow j}}{\partial v_{k \rightarrow \ell}}|_{\{v_{i \rightarrow j} = 0\}}$. A simple calculation reveals that for locally-tree like random networks, $\mathscr{M}$ is given in terms of the Non-Backtracking (NB) matrix $\mathscr{B}$ [34] via $\mathscr{M}_{k \rightarrow \ell, i \rightarrow j} = n_i \mathscr{B}_{k \rightarrow \ell, i \rightarrow j}$ in which $\mathscr{B}_{k \rightarrow \ell, i \rightarrow j} = 1$ if $\ell = i$ and $j \neq k$, and 0 otherwise.

To guarantee the stability of the solution $\{v_{i \rightarrow j} = 0\}$, it is required $\lambda(\mathbf{n}; q) \leq 1$. The optimal influence problem for a given q can be rephrased as finding the optimal configuration $\mathbf{n}$ that minimizes the largest eigenvalue $\lambda(\mathbf{n}; q)$. As q approaches the optimal threshold q_c, there exist a decreasing number of configurations that satisfy $\lambda(\mathbf{n}; q) \leq 1$. At q_c, only one configuration $\mathbf{n}^*$ exists such that $\lambda(\mathbf{n}^*; q_c) = 1$, and
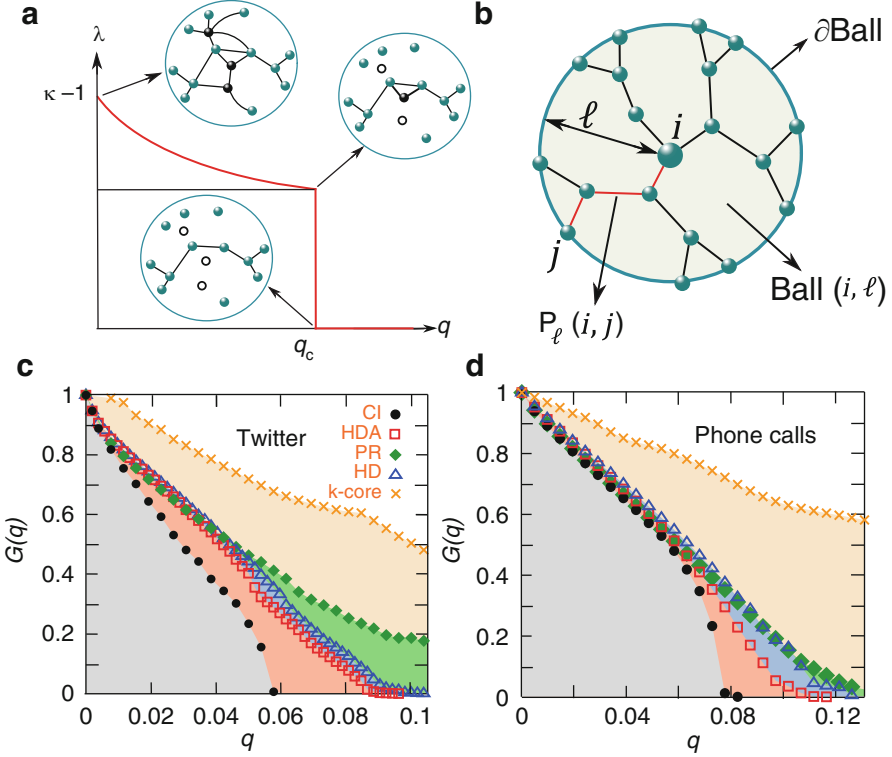
Fig. 2 (**a**), For $q \geq q_c$, the global minimum of the largest eigenvalue λ of the NB matrix over **n** is 0. In this case, $G = 0$ is stable, although there exist non-optimal configurations with $\lambda > 1$ for which $G > 0$. For $q < q_c$, the minimum of the largest eigenvalue is always $\lambda > 1$. Therefore the solution $G = 0$ is unstable and $G > 0$. At the optimal percolation transition, the minimum is at $\mathbf{n}^*$ such that $\lambda(\mathbf{n}^*, q_c) = 1$. At $q = 0$, $\lambda = \kappa - 1$ where $\kappa = \langle k^2 \rangle / \langle k \rangle$. At $\lambda = 1$, the giant component is reduced to a tree plus one single loop. This loop is destroyed at the transition q_c, and λ abruptly falls to 0. (**b**), Ball(i, ℓ) of radius ℓ around node i is shown. ∂Ball is the set of nodes on the boundary. The highlighted route is the shortest path from i to j. (**c** and **d**), Giant component $G(q)$ of Twitter ($N = 469,014$) and Mobile phone network in Mexico ($N = 1.4 \times 10^7$) computed using CI, high degree adaptive (HDA), PageRank (PR), high degree (HD), and k-core strategies. Figure is adapted from Morone et al. [61]

all other configurations will give $\lambda(\mathbf{n}; q) > 1$. The optimal configuration of $N q_c$ influencers $\mathbf{n}^*$ is therefore obtained when the *minimum* of the largest eigenvalue satisfies $\lambda(\mathbf{n}^*; q_c) = 1$. In practice, the largest eigenvalue can be calculated by the power method (we leave out q in $\lambda(\mathbf{n}; q)$):

$$\lambda(\mathbf{n}) = \lim_{\ell \to \infty} \left[\frac{|\mathbf{w}_\ell(\mathbf{n})|}{|\mathbf{w}_0|} \right]^{1/\ell}. \tag{3}$$

Here $|\mathbf{w}_\ell(\mathbf{n})|$ is the ℓ iterations of $\mathcal{M}$ on initial vector $\mathbf{w}_0$: $|\mathbf{w}_\ell(\mathbf{n})| = |\mathcal{M}^\ell \mathbf{w}_0|$. To find the best configuration of $\mathbf{n}$, we need to minimize the cost function $|\mathbf{w}_\ell(\mathbf{n})|$ for a finite ℓ. Through a proper simplification, we have an approximation of $|\mathbf{w}_\ell(\mathbf{n})|^2$ of order $1/N$ as

$$|\mathbf{w}_\ell(\mathbf{n})|^2 = \sum_{i=1}^N (k_i - 1) \sum_{j \in \partial \mathrm{Ball}(i, 2\ell-1)} \left(\prod_{k \in \mathscr{P}_{2\ell-1}(i,j)} n_k \right) (k_j - 1), \qquad (4)$$

in which $\partial \mathrm{Ball}(i, \ell)$ is the frontier of the ball of radius ℓ in terms of shortest path centered around node i, $\mathscr{P}_\ell(i, j)$ is the shortest path of length ℓ connecting i and j, and k_i is the degree of node i. See an example in Fig. 2b.

Based on the form of Eq. (4), an energy function for each configuration $\mathbf{n}$ can be defined as follows:

$$E_\ell(\mathbf{n}) = \sum_{i=1}^N (k_i - 1) \sum_{j \in \partial \mathrm{Ball}(i, \ell)} \left(\prod_{k \in \mathscr{P}_\ell(i,j)} n_k \right) (k_j - 1), \qquad (5)$$

where $E_\ell(\mathbf{n}) = |\mathbf{w}_{(\ell+1)/2}|^2$ for ℓ odd and $E_\ell(\mathbf{n}) = \langle \mathbf{w}_{\ell/2} | \mathcal{M} | \mathbf{w}_{\ell/2} \rangle$ for ℓ even. For $\ell = 1$, $E_\ell(\mathbf{n})$ is exactly the energy function of an Ising model which can be optimized using the cavity method [57]. For $\ell \geq 2$, it becomes a hard optimization problem involving many-body interactions. To develop a scalable algorithm for big-data analysis, an adaptive method is proposed, which is essentially a greedy algorithm for minimizing the largest eigenvalue of the stability matrix $\mathcal{M}$ for a given ℓ in the form of Eq. (4). In fact, Eq. (5) can be rewritten as the sum of collective influence from single nodes:

$$E_\ell(\mathbf{n}) = \sum_{i=1}^N \mathrm{CI}(i), \qquad (6)$$

in which the collective influence (CI) of node i at length ℓ is defined as:

$$\mathrm{CI}_\ell(i) = (k_i - 1) \sum_{j \in \partial \mathrm{Ball}(i, \ell)} (k_j - 1). \qquad (7)$$

The main idea behind the CI algorithm is to remove the nodes that can cause largest decrease of energy function in Eq. (4). In each iteration of CI algorithm, the node with the largest CI value is deleted, after which the CI values for remaining nodes are recalculated. The adaptive removal continues until the giant component is fragmented, i.e. $G(q) = 0$. Notice that the procedure minimizes q_c but does not guarantee the minimization of G in the percolation phase $G > 0$. If we want to optimize the configuration for $G(q) > 0$, a reinsertion procedure is applied from the configuration at $G(q) = 0$. In practice, if we use a heap structure to find the node

with the largest CI and only update the nodes inside the $(\ell+1)$-radius ball around the removed node, the computational complexity of CI algorithm can achieve $N \log(N)$ [62]. As a result, the CI algorithm is scalable for massively large-scale networks in modern social network analysis. For a Twitter network with 469,013 users (Fig. 2c) and a social network of 1.4×10^7 mobile phone users in Mexico (Fig. 2d), CI algorithm finds a smaller set of influencers than simple scalable heuristics including high degree adaptive (HDA), PageRank (PR), high degree (HD), and k-core [61]. To apply CI algorithm to real-time influencer ranking, a Twitter search engine was developed at http://www.kcore-analytics.com. Notice that, for $\ell = 0$, CI algorithm degenerates to high-degree ranking. So degree can be interpreted as the zero-order approximation of CI in Eq. (7).

To guarantee the scalability of the algorithm, CI essentially takes an adaptive greedy approach. The performance of CI algorithm can be further improved by a simple extension of CI using the message passing framework for $\ell \to \infty$—the CI propagation algorithm (CI$_P$) [62]. Remarkably, the CI propagation algorithm can reproduce the exact analytical threshold of optimal percolation for cubic random regular graphs [8]. Another belief-propagation variant of CI algorithm based on optimal immunization (CI$_{BP}$) also has similar performance of CI$_P$ [62]. However, the improvement over CI algorithm is at the price of higher computational complexity $O(N^2 \log(N))$, which makes both CI$_P$ and CI$_{BP}$ unscalable.

Recent studies have shown that the optimal percolation problem is closely related to the optimal decycling problem, or minimum feedback vertex set (FVS) problem [38]. Using belief-propagation (BP) algorithms, the optimal percolation problem was solved in recent works [12, 65]. The result of BP algorithms was found better than CI algorithm. Another approach to the optimal destruction of networks makes use of the explosive percolation theory [22].

3.2 Independent Cascade Model

The percolation process is deterministic on a given network with a given seed set. An important class of spreading model with stochasticity is the independent cascade model (ICM) [42]. In these models, a node is infected or activated by its neighbors with a predefined probability independently. Frequently used independent cascade models include susceptible-infected (SI) model, susceptible-infected-susceptible (SIS) model, and susceptible-infected-removed (SIR) model. These models are widely adopted in modeling infectious disease outbreaks and information spreading in social networks [35, 41, 74, 87, 93, 94]. Therefore, it is of particular interest in relevant applications.

In the pioneering work of Kempe et al. [40], influence maximization was first formalized as a discrete optimization problem: For a given spreading process on a network and an integer k, how to find the optimal set of k seeds that could generate the largest influence. For a large class of ICM and LTM, the influence

maximization problem can be well approximated by a simple greedy strategy, with a provable approximation guarantee [40]. In the basic greedy algorithm, the seed set is obtained by repeatedly selecting the node that provides the largest marginal increase of influence at each time step. The performance guarantee is built on the submodular property of the influence function $\sigma(S)$ [66], which is defined as the expected number of active nodes if the initial seed set is S. The influence function $\sigma(\cdot)$ is submodular if the incremental influence of selecting a node u into a seed set S is no smaller than the incremental influence of selecting the same node into a larger set V containing S. That is, $\sigma(S \cup \{u\}) - \sigma(S) \geq \sigma(V \cup \{u\}) - \sigma(V)$ for all nodes u and any sets $S \subseteq V$. Leveraging on the result of submodular function [66], the greedy algorithm is guaranteed to approximate the true optimal influence within a factor of $1 - 1/e \approx 63\%$, i.e., $\sigma(S) \geq (1 - 1/e)\sigma(S^*)$, where S is the seed set obtained by the greedy algorithm and S^* is the true optimal seed set. Although the basic greedy algorithm is simple to implement and performance-guaranteed, it requires massive Monte Carlo simulations to estimate the marginal gain of each candidate node. Several works were proposed to improve the efficiency of greedy algorithm [19, 20, 30, 47].

While performance guaranteed, from an optimization point of view, the greedy algorithm may be stuck into local optimum. This drawback can be solved by a more sophisticated message passing approach. Altarelli et al. developed the message passing algorithms (both belief-propagation (BP) and max-sum (MS)) for the problem of optimal immunization for SIR and SIS model [4], which can be applied to general ICMs. From another point of view, the independent cascade model can be naturally mapped to a bond percolation. Hu et al. found that in a series of real-world networks, most SIR spreading would be restrained to a local area while global-scale spreading rarely occurs [37]. Using the bond percolation theory, a characteristic local length termed influence radius was revealed. They argue that the global spreading optimization problem in fact can be solved locally, with the knowledge of the local environment within the influence radius.

3.3 Linear Threshold Model

Compared with independent cascade model, linear threshold model is more complex in the sense that a node's state is collectively determined by its neighbors' state. In a typical instance of LTM, each node v is assigned with a threshold value θ_v and each link (u, v) is assigned with a weight $w(u, v)$. During the cascade, a node is activated only if the sum of weights of its activated neighbors reaches the threshold value, i.e. $\sum_{u \in \partial v} w(u, v) \geq \theta_v$. In the case where the weights and thresholds are drawn uniformly from the interval $[0, 1]$, LTM was proven to be submodular [40]. Therefore, the influence maximization in this class of LTM can be well approximated by the greedy strategy, as we introduced in the above section. However, even with the lazy forward update [47], the algorithm is still unscalable

for large networks. Chen et al. found a way to approximate the influence of a node in a local subgraph [21], and developed a scalable greedy algorithm. Goyal et al. [31] further improved this algorithm by considering more choices of paths.

The above greedy approach and its variants are applicable to LTM with submodular property. However, for the general class of LTM with fixed weight and threshold, it is not guaranteed to be submodular [40]. An important class of LTM that may not be submodular is defined as follows: A node i is activated only after a certain number m_i of its neighbors are activated. The choice of different threshold m_i can generate two qualitatively different cascade regimes with continuous and discontinuous phase transitions. For instance, in the special case of $m_i = k_i - 1$ (k_i is the degree of node i), a continuous phase transition of influence occurs as the seed set grows [61]. However, there also exist a wide class of LTM exhibiting a first-order, or discontinuous phase transition. In the case that seeds are selected randomly, the transition between these two regimes is explored in detail in the context of bootstrap percolation [9, 29] and a simple cascade model [91]. But these results are based on the typical dynamical properties starting from random initial conditions. For influence maximization with a special initial condition, the dynamical behavior should be deviated from the average ones. Altarelli et al. proposed a BP algorithm that could estimate statistical properties of nontypical trajectories and found the initial conditions that lead to cascading with desired properties [2]. To obtain the exact set of seeds, MS equations were derived by setting the inverse temperature $\beta \to \infty$ in the energy function [3]. Extending the work under the assumption of replica symmetry, the theoretical limit of the minimal contagious set (the minimal seed set that can activate the entire graph) in random regular graphs is obtained using the cavity method with the effect of replica symmetry breaking [33].

In big-data analysis, an efficient and scalable algorithm designed for general LTM is needed. Starting from the message passing equations of LTM, generalized from Eq. (1) of percolation, a scalable algorithm named collective influence for threshold model (CI-TM) can be developed [75]. By iteratively solving the linearized message passing equations, the cascading process can be decomposed to separate components, each of which corresponds to the contribution made by a single seed. Interestingly, it is found the contribution of a seed is determined by the subcritical paths along which cascade propagates. In order to design a scalable algorithm, the node with the largest number of subcritical paths is recursively selected into the seed set. After each selection, the selected node and the subcritical paths attached to it are removed, and the status of the remaining nodes is recalculated. Making use of the heap structure, CI-TM algorithm can achieve the complexity of $O(N \log N)$. On one hand, computing $\text{CI} - \text{TM}_\ell$ value for a given length ℓ is equivalent to iteratively visiting subcritical neighbors of each node layer by layer within ℓ radius. Because of the finite search radius, computing $\text{CI} - \text{TM}_\ell$ for each node takes $O(1)$ time. Initially, we have to calculate $\text{CI} - \text{TM}_\ell$ for all nodes. However, during later adaptive calculation, there is no need to update $\text{CI} - \text{TM}_\ell$ for all nodes. We only have to recalculate for nodes within $\ell + 1$ steps from the removed vertices, which scales as $O(1)$ compared to the network size as $N \to \infty$ as shown in [62]. On the other hand, selecting the node with maximal CI-TM can

be realized by making use of the data structure of heap that takes $O(\log N)$ time [62]. Therefore, the overall complexity of ranking N nodes is $O(N \log N)$ even when we remove the top CI-TM nodes one by one. In both homogeneous and scale-free random networks, CI-TM achieves larger collective influence given the same number of seeds compared with other scalable approaches. This provides a practical method that can be applied to massively large-scale networks.

4 Applications of Influencer Identification

The problem of influencer identification is ubiquitous in a wide class of applications. So far, the theory of influencer identification has been applied to a number of important problems. In this section, we will introduce the application of influencer identification in three different areas: information diffusion, brain networks, and socioeconomic systems.

4.1 Information Diffusion in Social Networks

The most direct application of influencer identification is to maximize the information diffusion in social networks. In recent years, a huge number of research works have been performed aiming to relate users' spreading power to their locations, or personal features [58, 72, 89]. These works, mainly focusing on various types of online social networks including email communication [49], Facebook [60, 90], Twitter [6, 18, 44], and blogs sharing communities [5, 77], enrich our understanding of information diffusion in social networks.

A great challenge of developing effective predictors of influencers comes from the validation. In most of the previous works, the validation of proposed measures depends on modeling of information spreading in a given network. This approach, however, has led to several contradictory results on the best predictor of influence depending on the particular models [11, 41]. These models are built on simplified assumptions on human behavior [36] that neglect some of the most important features in real information diffusion [27], such as activity frequency [64, 83], behavior pattern [48, 73, 88], etc. Therefore, it is required to validate the various proposed predictors using empirical diffusion records in real-world social media.

We first compare the performance of different predictors for single influencers [72]. Realistic information diffusion instances as well as the underlying social networks are collected in four dissimilar social platforms: a blog-sharing community LiveJournal, scientific journals of American Physical Society, an online social network Facebook, and microblog service Twitter. To determine the real influence of each node, a directed diffusion graph is first constructed for each system by combining all directed diffusion links together. Then starting from a source node
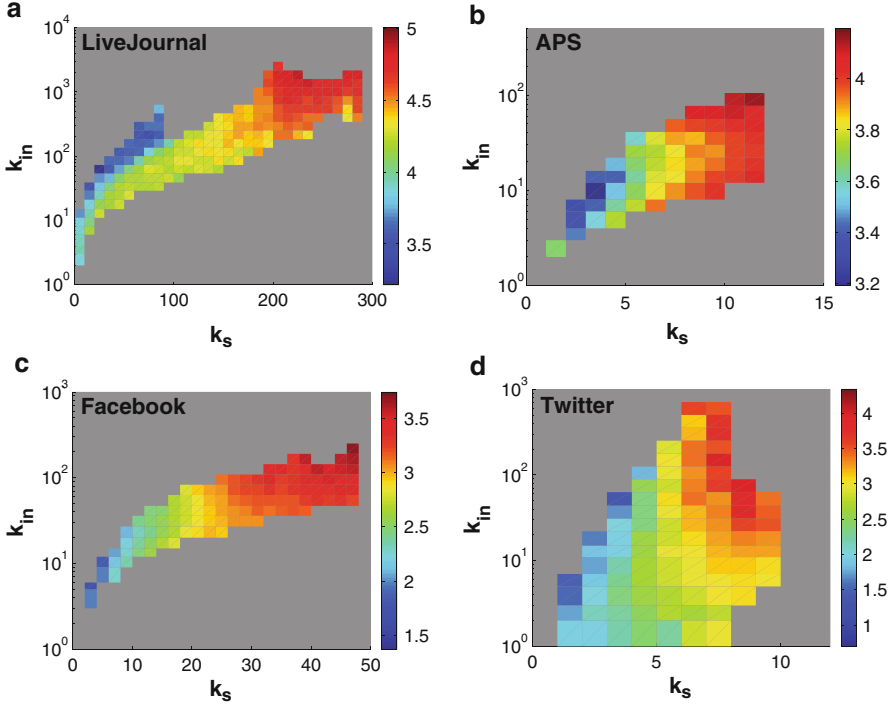
Fig. 3 K-core predicts the average influence of spreading more reliably than in-degree. Logarithmic values (base 10) of the average size of influence region $M(k_S, k_{in})$ when spreading originates from nodes with (k_S, k_{in}) for LiveJournal (**a**), APS journals (**b**), Facebook (**c**), and Twitter (**d**) are shown. Figure is adapted from Pei et al. [72]

i, the total influence M_i of node i is computed by tracking the diffusion links layer by layer in a breadth-first-search (BFS) fashion. Once we get the realistic influence, it is convenient to compare the performance of different predictors, including degree, k-core, and PageRank. Specifically, we can calculate the average influence $M(k_S, k_{in})$ for nodes with a given combination of k-core value k_S and in-degree k_{in}: $M(k_S, k_{in}) = \sum_{i \in \Upsilon(k_S, k_{in})} M_i / N(k_S, k_{in})$, where $\Upsilon(k_S, k_{in})$ is the collection of users in the (k_S, k_{in}) bin, and $N(k_S, k_{in})$ is the size of this collection. In all the systems, it is consistently observed that nodes with fixed degree can have either large or small influence, while nodes located in the same k-core have similar influence (see Fig. 3). Thus the influence of nodes is more related to their global location in the network, indicated by their k-core values. The same conclusion is also obtained in the comparison with PageRank. K-core does not only predict the average influence better, but also recognize influencers more accurately. Although k-core is effective, it is too coarse to distinguish different nodes within same shells. In some cases, there may be millions of nodes in one shell.

We further investigate the identification of multiple influencers [89]. Again, we use the realistic diffusion instances in the above four platforms. However, the empirical data cannot be directly mapped to ideal multi-source spreading. Such ideal multi-source spreading instances in which spreaders send out the same piece of message at the same time rarely exist in reality. Even though we can find such instances, the initial spreaders are hardly the same as the set of nodes selected by CI or other heuristic strategies. To circumvent this difficulty, we can construct virtual multi-source spreading processes by leveraging the behavior patterns of users extracted from the data. Suppose n spreaders $S = \{s_i | i = 1, 2, \cdots, n, n = qN\}$ are activated at the beginning of the virtual process. The influence strength $I_{g_1}(s)$ from seed s to its neighbor g_1 depends on the tendency of g_1 to receive information from s. Assume during the observation time, s has sent out $r(s)$ pieces of messages and g_1 has accepted $r(s, g_1)$ of them. Then the influence strength can be approximated by $I_{g_1}(s) = r(s, g_1)/r(s)$. In subsequent spreading, g_1 may affect its neighbor $g_2 \neq s$ in the same manner. Following the spreading paths, we can acquire the influence strength s enforcing on its ℓ-step neighbor g_ℓ: $I_{g_\ell}(s) = \prod_{k=1}^{\ell} r(g_{k-1}, g_k)/r(g_{k-1})$, where $g_0 = s$. The collective influence I_u for node u imposed by the seed set S is therefore $I_u = \max_{i=1}^{n} I_u(s_i)$. See Fig. 4 for an example. Finally, summing up all the N nodes in the network, the collective influence of the spreaders imposed on the entire system is $Q(q) = \sum_{u=1}^{N} I_u/N$. Based on this virtual spreading process, we can evaluate the collective influence of the spreaders selected by different methods. In particular, we compare the influencers selected by collective influence algorithm (CI), adaptive high degree (HDA), high degree (HD), PageRank (PR), and k-core. In all the systems, CI consistently outperforms other ranking methods.
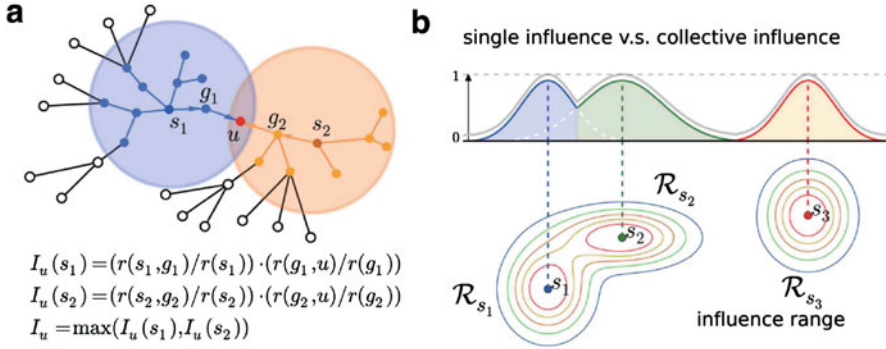

$$I_u(s_1) = (r(s_1, g_1)/r(s_1)) \cdot (r(g_1, u)/r(g_1))$$
$$I_u(s_2) = (r(s_2, g_2)/r(s_2)) \cdot (r(g_2, u)/r(g_2))$$
$$I_u = \max(I_u(s_1), I_u(s_2))$$

Fig. 4 (**a**) Calculation of influence strength to node u. Suppose the maximum spreading layer is set as $L = 2$ for two distinct seeds s_1 and s_2. The collective influence enforcing to u is selected as the largest value of the strength $I_u(s_1)$ and $I_u(s_2)$. (**b**) An illustration of single influence and collective influence. The three circle-like areas represent influence range $\mathcal{R}_{s_1}$, $\mathcal{R}_{s_2}$ and $\mathcal{R}_{s_3}$, for different spreaders s_1, s_2, and s_3. The contour lines show the levels of influence strength. The collective influence (grey curve) is obtained by combining single influence strengths of all spreaders. Figure is adapted from Teng et al. [89]

4.2   Collective Influence in Brain Networks

The human brain is a robust modular system interconnected as a Network of Networks (NoN) [15, 28, 78]. How this robustness emerges in a modular structure is an important question in many disciplines. Previous interdependent NoN models inspired by power grid are extremely fragile [14], thus cannot explain the observed robustness in brain networks. To reveal the mechanism beneath this robustness, a NoN model is proposed which can afford inter-link functionality and remain robust at the same time [63, 82].

In NoN system, the links are classified into two types: inter-modular links that represent the mutual dependencies between modules and intra-modular links that do not involve in the inter-modular dependencies. Denote $\mathscr{S}(i)$ and $\mathscr{F}(i)$ as the set of nodes connected to node i via intra-modular and inter-modular links, respectively. Suppose the variable state of node i is $\sigma_i \in \{0, 1\}$ (inactive or active), and the external input to node i is $n_i \in \{0, 1\}$ (no input or input). In the general activation model, the variable state is related to the input through $\sigma_i = n_i \left[1 - \prod_{j \in \mathscr{F}(i)} (1 - n_j) \right]$. That is, the node i is activated only if i receives the input ($n_i = 1$) and at least one of its neighbors connected with inter-modular links receives the input. In a robust brain network, for typical input configuration $\mathbf{n} = (n_1, \cdots, n_N)$, the giant (largest) component of the active nodes G with $\sigma_i = 1$ should be globally connected. Therefore, the robustness of the brain network can be characterized by the critical value $q_{rand} = 1 - \langle \mathbf{n} \rangle$ of zero inputs such that $G(q_{rand}) = 0$. Here the input configuration $\mathbf{n}$ is sampled from a flat distribution. Ideally, the robust NoN should have no disconnected phase, with a large value of q_{rand} close to 1.

To explain both robustness and inter-link functionality of brain networks, a robust NoN (R-NoN) model is proposed [63]. Define $\rho_{i \to j} \in \{0, 1\}$ as the message running along an intra-modular link $i \to j$, $\varphi_{i \to j} \in \{0, 1\}$ as the message running along an inter-modular link $i \to j$. The information flow follows the self-consistent equations

$$\rho_{i \to j} = \sigma_i \left[1 - \prod_{k \in \mathscr{S}(i) \setminus j} (1 - \rho_{k \to i}) \prod_{\ell \in \mathscr{F}(i)} (1 - \varphi_{\ell \to i}) \right], \tag{8}$$

$$\varphi_{i \to j} = \sigma_i \left[1 - \prod_{k \in \mathscr{S}(i)} (1 - \rho_{k \to i}) \prod_{\ell \in \mathscr{F}(i) \setminus j} (1 - \varphi_{\ell \to i}) \right]. \tag{9}$$

The physical meaning of the above equations is easy to be interpreted. For instance, in Eq. (8), a positive message $\rho_{i \to j}$ is transmitted from i to j in the same module if node i is active $\sigma_i = 1$ and if it receives at least one positive message from either a node k in the same module $\rho_{k \to i} = 1$ *or* a node ℓ in the other module $\varphi_{\ell \to i} = 1$.

Notice that, the logical OR is important since it is the basis of the robustness of R-NoN. The final probability of node i belonging to the largest active component G is

$$\rho_i = \sigma_i \left[1 - \prod_{k \in \mathscr{S}(i)} (1 - \rho_{k \to i}) \prod_{\ell \in \mathscr{F}(i)} (1 - \varphi_{\ell \to i}) \right]. \tag{10}$$

The size of G is therefore $G = \langle \rho_i \rangle$. In the R-NoN model, the system is robust since a node can be active $\sigma_i = 1$ even it does not belong to G. This prevents catastrophic cascading effects in the catastrophic C-NoN model inspired by power grid failure [14]. In the C-NoN model, a node remains functional only if it belongs to the giant component in *both* networks. This implies the status of a node in one network is interdependent on its status in the other network. The fundamental difference between C-NoN and R-NoN is that, in C-NoN model, the size of G is computed through

$$\rho_i = \sigma_i \left[1 - \prod_{k \in \mathscr{S}(i)} (1 - \rho_{k \to i}) \right] \left[1 - \prod_{\ell \in \mathscr{F}(i)} (1 - \varphi_{\ell \to i}) \right]. \tag{11}$$

So the logical OR in Eq. (10) is replaced by the logical AND in C-NoN. This stricter condition makes the system extremely sensitive to small perturbations. In synthetic NoN made of ER and SF random graphs, it is found the percolation threshold q_{rand} of R-NoN model is close to 1. On the contrary, the C-NoN model has threshold q_{rand} close to 0. This indicates that the two models indeed capture two different phenomena.

After exploring the behavior of R-NoN model under typical inputs, it is required to study the response to rare events targeting the influencers in the brain networks. Rare malfunction of nodes in the brain network that targets influencers may interrupt the global communication in the brain, which have been conjectured be responsible for certain neurological disorders. Or conversely, activating the influencers would optimally broadcast information to the entire network. Therefore, it is important to predict the location of the most influential nodes involved in information processing in the brain. To find the minimal fraction of nodes q_{infl} in the brain network whose removal would optimally fragment the giant component, the R-NoN model is mapped to the optimal percolation. The collective influence of nodes is calculated by minimizing the largest eigenvalue of the modified NB matrix. Particularly, the collective influence of node i is given by

$$\mathrm{CI}_\ell(i) = z_i \sum_{j \in \partial \mathrm{Ball}(i,\ell)} z_j + \sum_{j \in \mathscr{F}(i):k_j^{\mathrm{out}}=1} z_j \sum_{m \in \partial \mathrm{Ball}(j,\ell)} z_m, \tag{12}$$
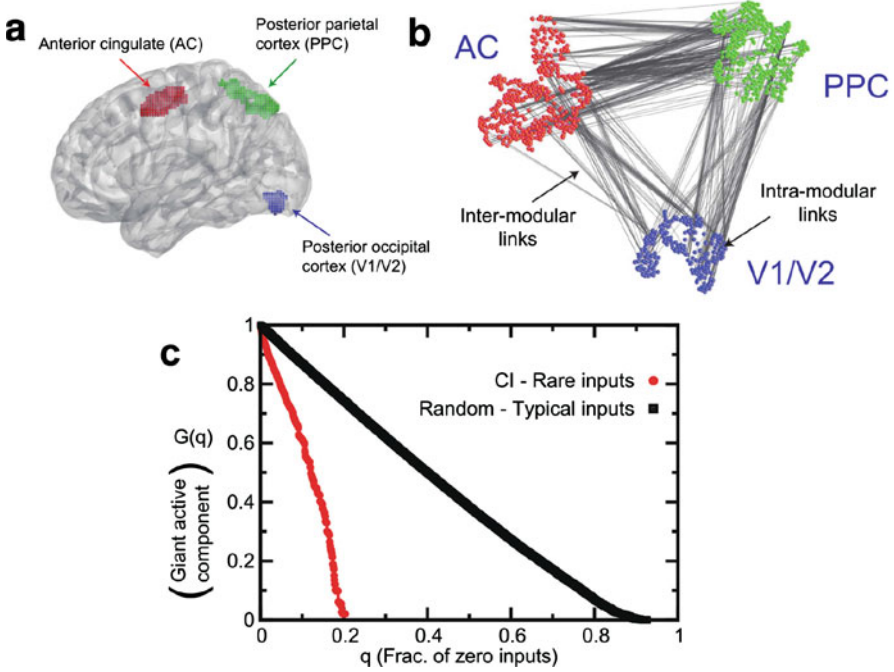
Fig. 5 (**a**) Spatial location of the three main modules (AC, PPC, and V1/V2) in the 3NoN. (**b**) Topology of the 3NoN. Inter-links and intra-links are displayed. (**c**) Size of the largest active cluster $G(q)$ as a function q of the nodes with $n_i = 0$ following CI optimization (red curve, $\ell=3$) and random states (black curve, random percolation). Figure is adapted from Morone et al. [63]

where $z_i \equiv k_i^{\text{in}} + k_i^{\text{out}} - 1$. The first term is the node-centric contribution, which presents in the single network case of optimal percolation, while the second term is the node-eccentric contribution, which is a new feature of the brain NoN.

Applying the R-NoN model and collective influence theory to real brain networks, it is possible to obtain the collective influence map of brain NoN. The brain network is constructed from the functional magnetic resonance imaging (fMRI) data of the experiment of stimulus driven attention [26, 28, 63]. In the experiment, each subject performs a dual visual-auditory task when receiving a visual stimulus and an auditory pitch simultaneously. This experiment requires the deployment of high level control modules in the brain, thus captures the role of dependency inter-modular connections. In the obtained brain network (see Fig. 5a, b), it is observed that the system is robust with large threshold $q_{rand} \approx 0.9$. While the minimal set of influencers only requires $q_{infl} \approx 0.2$ fraction of nodes (see Fig. 5c). Using the CI-map of the brain network, it is confirmed that control is deployed from the higher level module (Anterior cingulate) towards certain strategic locations in the lower ones (posterior parietal cortex, posterior occipital cortex). Moreover, the coarse-grain of the NoN to top CI nodes can predict the strategic areas in the brain.

4.3 Financial Status in Socioeconomic Systems

It has long been recognized that the pattern of individuals' social connection in society can affect people's financial status [32]. However, how to quantify the relationship between the location of an individual in social network and his/her economic wellness remains an open question. Despite that the effect of network diversity on economic development has been tested in the community level [24], inference of people's financial status from social network centralities or metrics in individual level is still needed. The difficulty of such investigation comes from the lack of empirical data containing both individual's financial information and pattern of social ties.

To find a reliable social network predictor of people's financial status, a massively large social network of the mobile and residential communication in Mexico containing 1.10×10^8 users together with financial banking data are analyzed [55]. With this dataset, it is possible to precisely cross-correlate the financial information of a person with his/her location in the communication network at the country level. Particularly, the financial status of individuals is reflected by their credit limit. In the analysis of the 5.02×10^5 bank clients identified in the phone call network, the top 10% and bottom 10% individuals present completely different communication pattern (see Fig. 6). Richer people maintain more active and diverse links, some connecting to remote locations and forming tightly linked "rich clubs."

To characterize the affluent people with network metrics, several centralities that are feasible for large-scale networks are compared, including degree, PageRank, k-core, and collective influence (CI). In the communication network, these four metrics are correlated. Therefore, they all show correlations with financial status when age is controlled. Among them, both k-core and CI capture the strong correlation with credit line with a R^2 value of 0.96 and 0.93, respectively. However, CI is more preferable since it satisfies both, a strong correlation and a high resolution. According to the definition of CI, top CI nodes are surrounded by hubs hierarchically. This is exactly the structure of ego-centric network of the top 1% wealthy people.
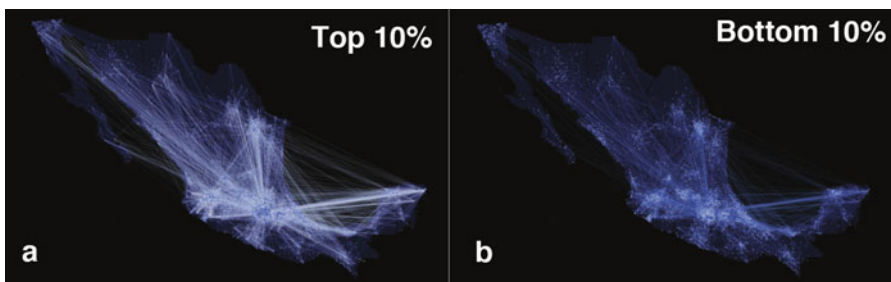

Fig. 6 (**a** and **b**) Visualization of communication activity of population in the top 10% and bottom 10% total credit limit classes. Figure is adapted from Luo et al. [55]

The performance of predictions can be further enhanced by considering the factor of age. An age-network combined metric $\text{ANC} = \alpha\text{Age} + (1-\alpha)\text{CI}$ with $\alpha = 0.5$ can achieve a correlation with $R^2 = 0.99$. Moreover, it is able to identify 70% high credit individuals at the highest earner level. To validate the effectiveness, a real social marketing campaign was performed. Specifically, text messages inviting new credit card clients were sent to 656,944 people selected by their high CI values in the social network. Meanwhile, the same message was sent to a control group of 48,000 individuals selected randomly. The response rate, measured by the fraction of recipients who requested the product, is augmented by threefold in the top influencers identified by CI compared with the random control group.

The same analysis was also applied to individuals' diversity of links [24]. The diversity of an individual can be measured by the diversity ratio $\text{DR} = W_{\text{out}}/W_{\text{in}}$, i.e., the ratio of total communication events with people in other communities W_{out} and within the same community W_{in}. The correlation between DR and CI is weak so they should reflect different aspects of network structure. In comparison with financial data, the age-diversity composite $\text{ADC} = \alpha\text{Age} + (1-\alpha)\text{DR}$ ($\alpha = 0.5$) well correlates with people's financial status. These evidences indicate that both CI and DR are effective predictors of people's financial situation in an individual level. This finding has a great practical value in relevant applications, for instance, social marketing campaigns.

References

1. Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. Nature 406(6794):378–382
2. Altarelli F, Braunstein A, Dall'Asta L, Zecchina R (2013) Large deviations of cascade processes on graphs. Phys Rev E 87(6):062115
3. Altarelli F, Braunstein A, Dall'Asta L, Zecchina R (2013) Optimizing spread dynamics on graphs by message passing. J Stat Mech: Theory Exp 2013(09):P09011
4. Altarelli F, Braunstein A, Dall'Asta L, Wakeling JR, Zecchina R (2014) Containing epidemic outbreaks by message-passing techniques. Phys Rev X 4(2):021024
5. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, pp 44–54
6. Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on twitter. In: Proceeding of the 4th ACM international conference on web search and data mining. Association for Computing Machinery, New York, pp 65–74
7. Batagelj V, Zaversnik M (2003) An o (m) algorithm for cores decomposition of networks. arXiv preprint cs/0310049
8. Bau S, Wormald NC, Zhou S (2002) Decycling numbers of random regular graphs. Random Struct Algoritm 21(3–4):397–413

9. Baxter GJ, Dorogovtsev SN, Goltsev AV, Mendes JF (2010) Bootstrap percolation on complex networks. Phys Rev E 82(1):011103
10. Bonacich P (1972) Factoring and weighting approaches to status scores and clique identification. J Math Socio 2(1):113–120
11. Borge-Holthoefer J, Moreno Y (2012) Absence of influential spreaders in rumor dynamics. Phys Rev E 85(2):026116
12. Braunstein A, Dall'Asta L, Semerjian G, Zdeborová L (2016) Network dismantling. Proc Natl Acad Sci U S A 113(44):12368–12373
13. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1):107–117
14. Buldyrev SV, Parshani R, Paul G, Stanley HE, Havlin S (2010) Catastrophic cascade of failures in interdependent networks. Nature 464(7291):1025–1028
15. Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci 10(3):186–198
16. Callaway DS, Newman ME, Strogatz SH, Watts DJ (2000) Network robustness and fragility: percolation on random graphs. Phys Rev Lett 85(25):5468
17. Centola D (2010) The spread of behavior in an online social network experiment. Science 329(5996):1194–1197
18. Cha M, Haddadi H, Benevenuto F, Gummadi PK (2010) Measuring user influence in twitter: the million follower fallacy. In: Proceeding of the 4th international AAAI conference on weblogs and social media 10(10–17):30
19. Chen W, Wang Y, Yang S (2009) Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, pp 199–208
20. Chen W, Wang C, Wang Y (2010) Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, pp 1029–1038
21. Chen W, Yuan Y, Zhang L (2010) Scalable influence maximization in social networks under the linear threshold model. In: 2010 IEEE 10th international conference on data mining (ICDM). IEEE, Los Alamitos, CA, pp 88–97
22. Clusella P, Grassberger P, Pérez-Reche FJ, Politi A (2016) Immunization and targeted destruction of networks using explosive percolation. Phys Rev Lett 117(20):208301
23. Cohen R, Erez K, Ben-Avraham D, Havlin S (2001) Breakdown of the internet under intentional attack. Phys Rev Lett 86(16):3682
24. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328(5981):1029–1031
25. Freeman LC (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239
26. Gallos LK, Sigman M, Makse HA (2007) The conundrum of functional brain networks: small-world efficiency or fractal modularity? Front Psychol 3:123
27. Gallos LK, Song C, Makse HA (2008) Scaling of degree correlations and its influence on diffusion in scale-free networks. Phys Rev Lett 100(24):248701
28. Gallos LK, Makse HA, Sigman M (2012) A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. Proc Natl Acad Sci U S A 109(8):2825–2830
29. Goltsev AV, Dorogovtsev SN, Mendes JFF (2006) k-core (bootstrap) percolation on complex networks: critical phenomena and nonlocal effects. Phys Rev E 73(5):056101
30. Goyal A, Lu W, Lakshmanan LV (2011) Celf++: optimizing the greedy algorithm for influence maximization in social networks. In: Proceedings of the 20th international conference on world wide web. Association for Computing Machinery, New York, pp 47–48
31. Goyal A, Lu W, Lakshmanan LV (2011) Simpath: an efficient algorithm for influence maximization under the linear threshold model. In: 2011 IEEE 11th international conference on data mining (ICDM). IEEE, Los Alamitos, CA, pp 211–220

32. Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380
33. Guggiola A, Semerjian G (2015) Minimal contagious sets in random regular graphs. J Stat Phys 158(2):300–358
34. Hashimoto KI (1989) Zeta functions of finite graphs and representations of p-adic groups. Adv Stud Pure Math 15:211–280
35. Hethcote HW (2000) The mathematics of infectious diseases. SIAM Rev 42(4):599–653
36. Hu Y, Havlin S, Makse HA (2014) Conditions for viral influence spreading through multiplex correlated social networks. Phys Rev X 4(2):021031
37. Hu Y, Ji S, Feng L, Havlin S, Jin Y (2015) Optimizing locally the spread of influence in large scale online social networks. arXiv preprint arXiv:1509.03484
38. Karp RM (1972) Reducibility among combinatorial problems. In: Complexity of computer computations. Springer, Berlin, pp 85–103
39. Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1): 39–43
40. Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, pp 137–146
41. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. Nat Phys 6(11):888–893
42. Kleinberg J (2007) Cascading behavior in networks: algorithmic and economic issues. Algorithmic Game Theory 24:613–632
43. Klemm K, Serrano M, Eguiluz VM, Miguel MS (2012) A measure of individual role in collective dynamics. Sci Rep 2:292
44. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceeding of the 19th ACM international conference on world wide web. Association for Computing Machinery, New York, pp 591–600
45. Lawyer G (2015) Understanding the influence of all nodes in a network. Sci Rep 5:8665
46. Leskovec J, Adamic LA, Huberman BA (2007) The dynamics of viral marketing. ACM Trans Web 1(1):5
47. Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N (2007) Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, pp 420–429
48. Li W, Tang S, Pei S, Yan S, Jiang S, Teng X, Zheng Z (2014) The rumor diffusion process with emerging independent spreaders in complex networks. Physica A 397:121–128
49. Liben-Nowell D, Kleinberg J (2008) Tracing information flow on a global scale using internet chain-letter data. Proc Natl Acad Sci U S A 105(12):4633–4638
50. Liu Y, Tang M, Zhou T, Do Y (2015) Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition. Sci Rep 5:9602
51. Liu Y, Tang M, Zhou T, Do Y (2015) Improving the accuracy of the k-shell method by removing redundant links-from a perspective of spreading dynamics. Sci Rep 5:13172
52. Lü L, Chen D, Ren XL, Zhang QM, Zhang YC, Zhou T (2016) Vital nodes identification in complex networks. Phys Rep 650:1–63
53. Lü L, Zhang YC, Yeung CH, Zhou T (2011) Leaders in social networks, the delicious case. PLoS One 6(6):e21202
54. Lü L, Zhou T, Zhang QM, Stanley HE (2016) The h-index of a network node and its relation to degree and coreness. Nat Commun 7:10168
55. Luo S, Morone F, Sarraute C, Makse HA (2017) Inferring personal financial status from social network location. Nat Commun 8:15227
56. Martin T, Zhang X, Newman M (2014) Localization and centrality in networks. Phys Rev E 90(5):052808
57. Mézard M, Parisi G (2003) The cavity method at zero temperature. J Stat Phys 111(1):1–34
58. Min B, Liljeros F, Makse HA (2015) Finding influential spreaders from human activity beyond network location. PLoS One 10(8):e0136831

59. Min B, Morone F, Makse HA (2016) Searching for influencers in big-data complex networks. In: Bunde A, Caro J, Karger J, Vogl G (eds) Diffusive spreading in nature, technology and society. Springer, Cham
60. Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on internet measurement. Association for Computing Machinery, New York, pp 29–42
61. Morone F, Makse HA (2015) Influence maximization in complex networks through optimal percolation. Nature 524:65–68
62. Morone F, Min B, Bo L, Mari R, Makse HA (2016) Collective influence algorithm to find influencers via optimal percolation in massively large social media. Sci Rep 6:30062
63. Morone F, Roth K, Min B, Stanley HE, Makse HA (2017) A model of brain activation predicts the neural collective influence map of the human brain. Proc Natl Acad Sci U S A 114(15):3849–3854
64. Muchnik L, Pei S, Parra LC, Reis SD, Andrade Jr, JS, Havlin S, Makse HA (2013) Origins of power-law degree distribution in the heterogeneity of human activity in social networks. Sci Rep 3:1783
65. Mugisha S, Zhou HJ (2016) Identifying optimal targets of network attack by belief propagation. Phys Rev E 94(1):012305
66. Nemhauser GL, Wolsey LA, Fisher ML (1978) An analysis of approximations for maximizing submodular set functions–I. Math Program 14(1):265–294
67. Newman ME (2002) Spread of epidemic disease on networks. Phys Rev E 66(1):016128
68. Newman ME, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 64(2):026118
69. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86(14):3200
70. Pastor-Satorras R, Vespignani A (2002) Immunization of complex networks. Phys Rev E 65(3):036104
71. Pei S, Makse HA (2013) Spreading dynamics in complex networks. J Stat Mech: Theory Exp 2013(12):P12002
72. Pei S, Muchnik L, Andrade Jr JS, Zheng Z, Makse HA (2014) Searching for superspreaders of information in real-world social media. Sci Rep 4:5547
73. Pei S, Muchnik L, Tang S, Zheng Z, Makse HA (2015) Exploring the complex pattern of information spreading in online blog communities. PLoS One 10(5):e0126894
74. Pei S, Tang S, Zheng Z (2015) Detecting the influence of spreading in social networks with excitable sensor networks. PLoS One 10(5):e0124848
75. Pei S, Teng X, Shaman J, Morone F, Makse HA (2017) Efficient collective influence maximization in threshold models of behavior cascading with first-order transitions. Sci Rep 7:45240
76. Radicchi F, Castellano C (2016) Leveraging percolation theory to single out influential spreaders in networks. Phys Rev E 93(6):062314
77. Ramos M, Shao J, Reis SD, Anteneodo C, Andrade Jr JS, Havlin S, Makse HA (2015) How does public opinion become extreme? Sci Rep 5:10032
78. Reis SD, Hu Y, Babino A, Andrade Jr JS, Canals S, Sigman M, Makse HA (2014) Avoiding catastrophic failure in correlated networks of networks. Nat Phys 10(10):762–767
79. Restrepo JG, Ott E, Hunt BR (2006) Characterizing the dynamical importance of network nodes and links. Phys Rev Lett 97(9):094102
80. Richardson M, Domingos P (2002) Mining knowledge-sharing sites for viral marketing. In: Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, pp 61–70
81. Rogers EM (2010) Diffusion of innovations. Simon and Schuster, London
82. Roth K, Morone F, Min B, Makse HA (2017) Emergence of robustness in networks of networks. Phys Rev E 95(6):062308

83. Rybski D, Buldyrev SV, Havlin S, Liljeros F, Makse HA (2012) Communication activity in a social network: relation between long-term correlations and inter-event clustering. Sci Rep 2:560
84. Sabidussi G (1966) The centrality index of a graph. Psychometrika 31(4):581–603
85. Seidman SB (1983) Network structure and minimum degree. Soc Netw 5(3):269–287
86. Stauffer D, Aharony A (1994) Introduction to percolation theory. CRC press, Boca Raton
87. Tang S, Teng X, Pei S, Yan S, Zheng Z (2015) Identification of highly susceptible individuals in complex networks. Physica A 432:363–372
88. Teng X, Yan S, Tang S, Pei S, Li W, Zheng Z (2014) Individual behavior and social wealth in the spatial public goods game. Physica A 402:141–149
89. Teng X, Pei S, Morone F, Makse HA (2016) Collective influence of multiple spreaders evaluated by tracing real information flow in large-scale social networks. Sci Rep 6:36043
90. Viswanath B, Mislove A, Cha M, Gummadi KP (2009) On the evolution of user interaction in facebook. In: Proceedings of the 2nd ACM workshop on online social networks. Association for Computing Machinery, New York, pp 37–42
91. Watts DJ (2002) A simple model of global cascades on random networks. Proc Natl Acad Sci U S A 99(9):5766–5771
92. Watts DJ, Dodds PS (2007) Influentials, networks, and public opinion formation. J Constr Res 34(4):441–458
93. Yan S, Tang S, Pei S, Jiang S, Zheng Z (2014) Dynamical immunization strategy for seasonal epidemics. Phys Rev E 90(2):022808
94. Yan S, Tang S, Fang W, Pei S, Zheng Z (2015) Global and local targeted immunization in networks with community structure. J Stat Mech: Theory Exp 2015(8):P08010
95. Zeng A, Zhang CJ (2013) Ranking spreaders by decomposing complex networks. Phys Lett A 377(14):1031–1035