Highlights

Targeted influence maximization in complex networks
Renquan Zhang, Xiaolin Wang, Sen Pei

- We developed a theoretical framework to analyze targeted influence maximization using a message passing process.
- We proposed a metric based on non-backtracking matrix to select influential spreaders.
- We validated the proposed metric in both synthetic and real-world networks.
Targeted influence maximization in complex networks

Renquan Zhang\textsuperscript{a}, Xiaolin Wang\textsuperscript{a} and Sen Pei\textsuperscript{b,\ast}

\textsuperscript{a}School of Mathematical Sciences, Dalian University of Technology, Dalian, 116024, China
\textsuperscript{b}Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, 10032, NY, USA

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Influential Spreaders
Targeted Influence Maximization
Spreading Dynamics
Targeted Spreading
Non-backtracking Matrix
Optimal Perturbation

\textbf{ABSTRACT}

Many real-world applications based on spreading processes in complex networks aim to deliver information to specific target nodes. However, it remains challenging to optimally select a set of spreaders to initiate the spreading process. In this paper, we study the targeted influence maximization problem using a susceptible-infected-recovered (SIR) model as an example. Formulated as a combinatorial optimization, the objective is to identify a given number of spreaders that can maximize the influence over target nodes while minimize the influence over non-target nodes. To find a practical solution to this optimization problem, we develop a theoretical framework based on a message passing process and perform a stability analysis on the equilibrium solution using non-backtracking (NB) matrices. We propose that the spreaders can be selected by imposing optimal perturbation on the equilibrium solution for the subgraph consisting of the target nodes and their multi-step nearest neighbors while avoiding such perturbation on the complement graph that excludes target nodes from the original network. We further introduce a metric, termed targeted collective influence, for each node to identify influential spreaders for targeted spreading processes. The proposed method, validated in both synthetic and real-world networks, outperforms other competing heuristic approaches. Our results provide a framework for analyzing the targeted influence maximization problem and a practical method to identify spreaders in real-world applications.

\section{1. Introduction}

Spreading processes in complex networks can describe a wide variety of real-world phenomena, ranging over epidemic outbreaks \cite{1,2,3}, online information diffusion \cite{4,5,6,7}, behavior adoption \cite{8,9,10}, and viral marketing\cite{11,12}. Due to the structural heterogeneity of networks, a small set of nodes play a disproportionate role in shaping the outcome of spreading dynamics. Identifying such pivotal nodes, or influencers, is a critical question in network science. Over the last decades, a plethora of studies developed methods to locate influential spreaders in networks \cite{13,14,15,16}, either a single node initiating the spreading process or multiple spreaders considering their collective influence \cite{17,18,19,20,21,22,23,24,25,26,27,28}. A review on recent advances in this area can be found in Ref \cite{29}.

Many real-world applications aim to deliver information to certain target nodes while avoiding reaching non-target nodes. For instance, in advertising, it is desirable to promote products to potential customers and minimize the coverage among other users for precise marketing; in an election campaign, it is more cost-effective to disseminate information to swing voters and save the resources spent on decided voters. Studies on targeted spreading and control have attracted much attention in recent years \cite{30,31,32,33,34,35,36,37,38}. A number of heuristic methods were developed to identify influencers for targeted spreading processes. For instance, a greedy algorithm was proposed to select seeds in a spreading process with several constrains \cite{32}; a heuristic method based on local path counting was used to identify single spreaders in a targeted spreading process \cite{38}; and real-time targeted online advertisement informed by key words was also tested \cite{33}. While these approaches were demonstrated effective in different settings, a general framework for analyzing the targeted influence maximization problem is lacking.

In this study, we focus on the influence maximization problem for a general targeted spreading process. Specifically, we aim to identify a given number of spreaders that maximize the influence over target nodes and minimize the influence over non-target nodes in a susceptible-infected-recovered (SIR) model. To solve this optimization problem, we first develop a mathematical framework to formulate the system as a message passing process, and then perform a stability analysis on the equilibrium solution using the non-backtracking (NB) matrix of the system \cite{39}. We argue that the spreaders can be selected by imposing optimal perturbation on the equilibrium solution for the subgraph consisting of

\ast Corresponding author
ORCID(s): 0000-0002-4927-1827 (R. Zhang); 0000-0002-7072-2995 (S. Pei)
the target nodes and their multi-step nearest neighbors while avoiding such perturbation on the complement graph that excludes target nodes from the original network. Our analysis leads to a theoretically based metric, termed targeted collective influence, to quantify the targeted influence of each node, which allows the selection of multiple influencers in the targeted spreading process. We validate the proposed method in both synthetic and real-world networks and demonstrate that it outperforms commonly used heuristic approaches. Our analysis provides a theoretical framework for analyzing the targeted influence maximization problem and a practical method to identify spreaders in real-world applications.

2. Model

We use a susceptible-infectious-recovered (SIR) agent-based model to simulate the spreading process in networks. Considering a network composed of $N$ nodes and $M$ undirected edges, denote $a_{ij}$ as the binary adjacency matrix ($a_{ij} = 1$ if node $i$ is connected to node $j$, and $a_{ij} = 0$ otherwise). The binary variables $S_i(t)$, $I_i(t)$ and $R_i(t)$ represent that node $i$’s state is susceptible, infectious and recovered at time $t$, respectively. We denote the probability that an infectious node will infect its susceptible neighbor as $\beta$, and define $\gamma$ as the infectious period (without loss of generality, $\gamma = 1$). At each time step $t$, a susceptible node $i$ ($S_i(t) = 1$) can be infected by its neighbor $j$ in the infectious state ($I_j(t) = 1$) with probability $\beta$. Meanwhile, nodes in the infectious state will transit to the recovered state after $\gamma$ steps and can never be infected again. The spreading process can be described as follows:

$$\frac{dS_i(t)}{dt} = -S_i(t)\left[1 - \prod_j (1 - \beta a_{ij} I_j(t))\right],$$  
$$\frac{dI_i(t)}{dt} = S_i(t)\left[1 - \prod_j (1 - \beta a_{ij} I_j(t))\right] - \frac{I_i(t)}{\gamma},$$  
$$\frac{dR_i(t)}{dt} = \frac{I_i(t)}{\gamma}.$$  

Here we focus on a combinatorial optimization problem - how to select a given number of initial infected nodes, or seeds, to maximize infection among a specific group of nodes and minimize infection among others? Specifically, denote $V = V^T \bigcup V^{NT}$ as the set of nodes, where $V^T$ represents the set of target nodes and $V^{NT}$ the set of non-target ones. The seeds can be only selected from $V^{NT}$ and the number of seeds is denoted as $n^*$. Based on the SIR dynamics, the number of nodes that have been infected is equal to the number of the recovered nodes at the end of the spreading process. For a given network, the influence over the target and non-target nodes are defined as

$$f(s_{n^*}) = \lim_{t \to \infty} \sum_{i \in V^T} R_i(t),$$  
$$g(s_{n^*}) = \lim_{t \to \infty} \sum_{i \in V^{NT}} R_i(t),$$

where $s_{n^*}$ denotes the set of seeds with the size $n^*$. As the topological structure of the network plays a significant role in the spreading process, different combinations of seeds with the same size $n^*$ could lead to contrasting outcomes. We aim to find the optimal set of seeds such that $g(s_{n^*})/f(s_{n^*})$ is minimized for $f(s_{n^*}) > 0$.

3. Method

3.1. Message passing equations

We first formulate the propagation as a message passing process. Compared to the master equations defined using the adjacency matrix, the message passing process can better represent the SIR dynamics. Specifically, in the SIR model, backtracking infections ($i \rightarrow j \rightarrow i$) are not allowed as the transmission is irreversible. The adjacency matrix allows backtracking infections, which can introduce excessive dynamical resonance between pairs of connected nodes. In contrast, the message passing process excludes backtracking spreading and was found superior in analyzing a number of dynamical models in complex networks [19, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50].

To study the impact of a node $j$ on its neighbor node $i$, we investigate the probability of node $i$ being infected if node $j$ is assumed to be absent from the network. For a link from $i$ to $j$ ($i \rightarrow j$, even if the link is undirected), suppose
node $j$ is "virtually" removed from the network (i.e., creating a "cavity" at node $j$) and calculate the probability of node $i$ being infected in the absence of node $j$ at time $t$, which is represented as $S_{i\rightarrow j}(t)$. We apply the same procedure for $I$ and $R$. For sparse networks without too many short loops, the message passing process can be described by

$$S_{i\rightarrow j}(t+1) = S_{i\rightarrow j}(t) \prod_{k\not= j} (1 - \beta a_{ik} I_{k\rightarrow i}(t)), \quad (6)$$

$$I_{i\rightarrow j}(t+1) = I_{i\rightarrow j}(t) \left[1 - \prod_{k\not= j} (1 - \beta a_{ik} I_{k\rightarrow i}(t)) \right] + I_{i\rightarrow j}(t)(1 - \frac{1}{\gamma}), \quad (7)$$

$$R_{i\rightarrow j}(t+1) = R_{i\rightarrow j}(t) + \frac{I_{i\rightarrow j}(t)}{\gamma}. \quad (8)$$

Here $k\not= j$ means $k$ runs over all nodes except $j$. Denote $\lim_{t\rightarrow \infty} S_{i\rightarrow j}(t) = S_{i\rightarrow j}$, $\lim_{t\rightarrow \infty} I_{i\rightarrow j}(t) = I_{i\rightarrow j}$ and $\lim_{t\rightarrow \infty} R_{i\rightarrow j}(t) = R_{i\rightarrow j}$. Note that Eq. (8) is redundant. The steady state of the nonlinear dynamical system can be obtained by solving the following self-satisfying equations:

$$S_{i\rightarrow j} = S_{i\rightarrow j} \prod_{k\not= j} (1 - \beta a_{ik} I_{k\rightarrow i}), \quad (9)$$

$$I_{i\rightarrow j} = \gamma S_{i\rightarrow j} \left[1 - \prod_{k\not= j} (1 - \beta a_{ik} I_{k\rightarrow i}) \right]. \quad (10)$$

### 3.2. Stability analysis

To maximize influence over the target nodes and minimize influence over the non-target nodes, we consider two subgraphs of the original network: 1) a subgraph $G_s$ consisting of the target nodes and their nearest neighbors within $L$ steps, and 2) a subgraph $G_c$ that excludes the target nodes from the original network. We create $G_s$ using a breadth-first-search algorithm starting from the target nodes. An illustration for $G_s$ ($L = 2$) and $G_c$ is shown in Fig. 1. As seeds can be only selected from the non-target nodes, we define the set of candidate seeds as the non-target nodes in $G_s$ (i.e., nodes in $G_s$ excluding the target nodes).

For both $G_s$ and $G_c$, a trivial equilibrium solution exists: $(S^*_{i\rightarrow j}, I^*_{i\rightarrow j})^T = (1, 0)^T$, corresponding to the state that all nodes are susceptible. The stability of the trivial solution is controlled by the largest eigenvalue of the Jacobian matrix ($J$) at this solution. Now we derive the Jacobian matrix at the solution $(1, 0)^T$ for $G_s$ and $G_c$. For a given network, we take the partial derivatives of Eq. (9). For the directed links $k \rightarrow l$ and $i \rightarrow j$, we have

$$\frac{\partial S_{i\rightarrow j}}{\partial S_{k\rightarrow l}} = \prod_{l\not= k} (1 - \beta a_{lk} I_{k\rightarrow l}(t)) \bigg|_{(1,0)} = \begin{cases} 1 & \text{if } k = i \text{ and } l = j \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

$$\frac{\partial S_{i\rightarrow j}}{\partial I_{k\rightarrow l}} = -\beta S_{i\rightarrow j}(t) \prod_{k'\not= j, k} (1 - \beta a_{ik'} I_{k'\rightarrow i}(t)) \bigg|_{(1,0)} = \begin{cases} -\beta & \text{if } l = i \text{ and } k \neq j \\ 0 & \text{otherwise} \end{cases} \quad (12)$$
Targeted influence maximization

The same analysis on Eq. (10) yields:

\[
\frac{\partial I_{l\rightarrow j}}{\partial S_{k\rightarrow l}} = \gamma \left[ 1 - \prod_{i\neq j} (1 - \beta d_{ik} I_{k\rightarrow l}(t)) \right]_{(1,0)} = 0
\]

(13)

\[
\frac{\partial I_{l\rightarrow j}}{\partial I_{k\rightarrow l}} = \beta \gamma S_{l\rightarrow j}(t) \prod_{k'\neq l,k} (1 - \beta d_{ik'} I_{k'\rightarrow l}(t)) \right]_{(1,0)} = \begin{cases} 
\beta \gamma & \text{if } l = i \text{ and } k \neq j \\
0 & \text{otherwise}
\end{cases}
\]

(14)

So the Jacobian matrix at the solution \((1, 0)^T\) is given by:

\[
J_{(1,0)} = \begin{pmatrix} I & \Lambda \\
0 & B \end{pmatrix}.
\]

(15)

Here \(I\) is the identity matrix, \(\Lambda = \left\{ \frac{\partial S_{l\rightarrow j}}{\partial I_{k\rightarrow l}} \right\}_{2M \times 2M}\) and \(B = \left\{ \frac{\partial I_{l\rightarrow j}}{\partial I_{k\rightarrow l}} \right\}_{2M \times 2M}\), where \(M\) is the number of links. The stability of \(I^*_{l\rightarrow j}\) is determined by the largest eigenvalue of the matrix \(B\), denoted by \(\lambda_B\). The trivial solution is stable if \(\lambda_B < 1\) and unstable if \(\lambda_B > 1\). The matrix \(B\) is a generalization of the non-backtracking (NB) matrix of the network \(W\), which was found important for a range of dynamical processes in complex networks. Precisely, \(B = \beta \gamma W\), where \(W_{k\rightarrow l,j\rightarrow j} = \begin{cases} 1 & \text{if } l = i \text{ and } k \neq j, \\
0 & \text{otherwise.}
\end{cases}\)

(16)

3.3. Optimal perturbation

Selecting seeds to initiate a spreading process acts as a perturbation on the equilibrium solution. Define \(L_*(0) = (\cdots, I_{l\rightarrow j}(0), \cdots)^T\) as the initial state. For the equilibrium solution, we have \(L_*(0) = 0^T\). If node \(i\) is chosen as a seed, we set the elements \(I_{l\rightarrow j}(0) = 1\) for all \(j \in \partial i\), where \(\partial i\) represents the set of neighbors of node \(i\). To select a given number of \(n^*\) seeds, there exist a total of \(C_n^N\) possible combinations. In order to maximize influence over the target nodes, we can impose the optimal perturbation along the leading eigenvector of the NB matrix \(W\) for \(G_s\) so that more nodes are infected in this subgraph. Meanwhile, we should avoid the perturbation in \(G_c\), along the leading eigenvector of its NB matrix to minimize infections among non-target nodes. Similar approaches have been used in numerical weather prediction [51, 52, 53] and infectious disease forecasting [54, 55, 56].

Denote the leading eigenvector of \(W\) as \(v\) such that \(Wv = \lambda v\), where \(\lambda\) is the largest eigenvalue of \(W\). Computing the leading eigenvector for \(W\) can be challenging for large-scale networks due to the high dimensionality of the NB matrix. For a network with \(M\) edges, \(W\) has a dimension of \(2M \times 2M\). An effective method to compute the largest eigenvalue and the leading eigenvector is the power iteration. Starting from an initial vector \(y_0 = 1^T\), we multiply \(W\) from the left (\(y_{t+1} = Wy_t\)) repeatedly until the ratio \(\|y_{t+1}\|/\|y_t\|\) is stabilized. The largest eigenvalue is \(\lambda = \lim_{t \to \infty} \|y_{t+1}\|/\|y_t\|\) and the normalized leading eigenvector is \(v = \lim_{t \to \infty} y_t/\|y_t\|\).

To approximate the leading eigenvector using the power iteration, we multiply \(W\) from the left on \(y_0 = 1^T\) for \(\ell\) times. We denote the approximated influence of node \(i\) for \(\ell\) as \(CI_{\ell}(i) = \sum_{j \in \partial i} y_{\ell,j \rightarrow j}^2\), where \(y_{\ell,j \rightarrow j}\) is the entry of \(y_{\ell}\) corresponding to the link \(i \rightarrow j\). Following the method in Ref. [19], we derive that \(CI_{\ell}(i)\) for node \(i\) in a network is given by

\[
CI_{\ell}(i) = \left( k_i - 1 \right) \sum_{j \in \partial Ball(i, 2\ell - 1)} \left( k_j - 1 \right),
\]

(17)

where \(\ell\) is the iteration time and \(\partial Ball(i, 2\ell - 1)\) is the set of nodes whose shortest distance to node \(i\) is \(2\ell - 1\). In order to find spreaders for the targeted spreading process, we define the targeted collective influence for node \(i\) at level \(\ell\) as

\[
\Delta CI_{\ell}(i) = CI_{\ell}^{G_s}(i) - CI_{\ell}^{G_c}(i).
\]

(18)

Here \(CI_{\ell}^{G_s}(i)\) and \(CI_{\ell}^{G_c}(i)\) are calculated on the subgraph \(G_s\) and \(G_c\). For the targeted influence maximization problem, we select top \(n^*\) nodes from the candidates with the highest \(\Delta CI_{\ell}\) score as the seed set \(s_{n^*}\). Nodes with higher \(\Delta CI_{\ell}\) scores tend to have higher \(CI_{\ell}^{G_s}\) and lower \(CI_{\ell}^{G_c}\).
4. Numerical validation

In order to test the performance of the proposed targeted collective influence, we first run numerical simulations on synthetic networks. We consider the case that target nodes are connected in a local cluster. In experiments, we first randomly select a target node and then apply a breadth-first search to assign other target nodes until the predefined number of target nodes is reached. Without loss of generality, here we set 10 target nodes in each cluster. Once the target node set \( V_T \) is assigned, we select spreaders using different methods. Starting from the selected seeds, we perform 100 independent realizations of the SIR model. Results are evaluated using the average of the 100 simulations.

To compute the targeted collective influence \( \Delta CI_\ell \), we define \( G_s \) as the subgraph consisting of the target nodes and their nearest neighbors within \( L = 7 \) steps. Other values of \( L \) were tested. We find that \( L = 7 \) is enough to capture potential optimal spreaders and increasing \( L \) does not improve the performance. In model simulations, we consider \( \ell = 1 \) and \( \ell = 2 \) to calculate \( \Delta CI_\ell \). We compare the targeted collective influence \( \Delta CI_1 \) and \( \Delta CI_2 \) with several other competing methods, including (1) dynamical importance defined based on the eigenvector of the adjacency matrix of the original network \( G \) (Eig) \cite{57}; (2) the dynamical importance defined based on the eigenvector of the adjacency matrix of the subgraph \( G_s \) (Eigs); (3) the degree centrality of \( G \) (HD); (4) the degree centrality of \( G_s \) (HDs); (5) the collective influence \( CI_\ell \) of \( G \) for \( \ell = 1 \) in Eq. (17) (\( CI_1 \)); and (6) the collective influence \( CI_\ell \) of \( G \) for \( \ell = 2 \) in Eq. (17) (\( CI_2 \)). More details of the competing methods are provided in Appendix A. For each metric, we select the top \( n^* \) nodes with the highest values as the initial seeds. For reference, we also test a random selection method (Rand) that chooses seeds randomly from \( G \).

4.1. Random networks

We first test on homogeneous Erdős-Rényi (ER) random networks with 10 target nodes. We generate undirected ER networks with size \( N \) and average degree \( \langle k \rangle \) by randomly connecting any possible pairs of nodes with a probability \( p = \langle k \rangle / N \). We use networks with \( N = 10,000 \) or \( 20,000 \) and \( \langle k \rangle = 3 \) or \( 4 \) in simulations shown in Figs. 2 and 3. To ensure the connectivity of the graph, all simulations are only applied on the giant connected component. We vary the
transmission rate $\beta \in [0.1, 0.4]$ for $\langle k \rangle = 3$ and $\beta \in [0.1, 0.3]$ for $\langle k \rangle = 4$. We don’t consider higher transmission rate $\beta$ as it will lead to large-scale outbreaks that infect almost the entire network. We test the number of seeds $n^* = 2, \ldots, 6$ and set $\gamma = 1$.

In Fig. 2, we show the probability that at least one target node is infected, $P_I$. As all target nodes are locally connected (as shown in Fig. 1a), $P_I$ measures the chance that the spreading process reaches the small cluster of target nodes in a large-scale network. The targeted collective influence $\Delta CI_1$ consistently outperforms other competing methods. However, $\Delta CI_1$ and $\Delta CI_2$ have similar results. As the transmission rate $\beta$ increases, $P_I$ increases for all methods. For $\beta < \langle k \rangle / (\langle k^2 \rangle - \langle k \rangle)$, seeds selected by Rand, Eig, and HD can hardly infect any target nodes. For $\beta > \langle k \rangle / (\langle k^2 \rangle - \langle k \rangle)$, all methods except Rand can almost always reach target nodes. Metrics defined on the subgraph $G_s$ performs better than their counterparts defined on the original network $G$. We examine the fraction of infected target nodes for all methods and find that the targeted collective influence performs best as well (see Appendix B).

We evaluate the ratio of infected non-target node to infected target node ($g(s_{nr}) / f(s_{nr})$) in Fig. 3. A lower ratio indicates a better performance of the method. Here we only compare Eigs, HDs, CI and $\Delta CI$ as other methods rarely infect target nodes. Again, we find that the targeted collective influence outperforms other competing approaches and $\Delta CI_2$ performs better than $\Delta CI_1$.

4.2. Scale-free networks

We perform the same analysis on scale-free (SF) networks with 10 target nodes. The SF networks have power-law degree distributions. We generate undirected SF networks with $N = 10,000$ or $20,000$ using the preferential attachment model [58]. Simulation results are shown in Figs. (4) and (5). We find that the targeted collective influence performs better than competing methods. Interestingly, for SF networks, $\Delta CI_1$ outperforms $\Delta CI_2$. This is possibly due to the existence of highly connected hubs. $\Delta CI_2$ may select global hubs that have both high $CI^{G_s}$ and $CI^{G_c}$ but are far from target nodes. In contrast, $\Delta CI_1$ can potentially select local hubs that are close to target nodes. Further analyses are needed to test this hypothesis in future works.
Figure 4: The probability that at least one target node is infected in scale-free (SF) networks with size $N = 10,000$ or 20,000 and degree distribution $P(k) \sim k^{-3}$. We use different methods (including Rand, Eig, Eigs, HD, HDs, CI and $\Delta CI$ with $\ell = 1$ and $\ell = 2$) to select the seeds for different size $n^* = 2, \ldots, 6$. The results are averaged over 100 independent realizations.

4.3. Real-world networks

We finally validate the proposed method in several real-world networks [59]. Four real networks of distinct types are selected, including a co-authorship network ($N = 379$ and $M = 914$), the US power grid network ($N = 4.9K$ and $M = 6.6K$), a web graph network (links between webpages) ($N = 16.1K$ and $M = 25.6K$), and a recommendation network of Amazon ($N = 91.8K$ and $M = 125.7K$). Data sources and statistics of the networks are reported in Appendix C.

In the first set of experiments, we assign 10 target nodes in one cluster and aim to find $n^* = 4$ seeds. Experiment results are shown in Fig. (6). Consistent with simulations on synthetic networks, the targeted collective influence performs best. A same method can have different performance in the four real-world networks depending on the network structure. For instance, HD performs much worse in the US power grid network and the Amazon recommendation network. However, the good performance of $\Delta CI_1$ and $\Delta CI_2$ is robust across all tested networks. $\Delta CI_1$ is generally better than $\Delta CI_2$ in the four networks. For the more heterogeneous web graph network (the maximum degree is 1.7K and the average degree is 3), the advantage of $\Delta CI_1$ over $\Delta CI_2$ is more prominent, which agrees with the results in SF networks.

We further consider the case that target nodes are located in several clusters that spread across the network. Specifically, we select two clusters of target nodes, each cluster with 10 target nodes. This optimization problem is more challenging as target nodes are not located in one place. Results shown in Fig. (9) indicate that $\Delta CI_1$ and $\Delta CI_2$ still outperform competing methods. The advantage of $\Delta CI_1$ and $\Delta CI_2$ is more prominent in sparse networks with lower average degrees (e.g., the recommendation network of Amazon and the US power grid). We additionally test the case with 30 target nodes in three clusters. Results in Fig. (10) demonstrate the consistent better performance of $\Delta CI_1$ and $\Delta CI_2$.

5. Conclusion

Targeted influence maximization has broad applications in real-world problems. In this study, we formulated the SIR model using a message passing process, which can better represent the transmission dynamics, and further
Targeted influence maximization

Figure 5: The ratio of \( g(s_n^*) \) to \( f(s_n^*) \) for scale-free (SF) networks with size \( N = 10,000 \) or \( 20,000 \). We compare the performances of different methods including Eigs, HDs, CI and \( \Delta CI \) with \( \ell = 1 \) and \( \ell = 2 \). The results are averaged over 100 independent realizations.

developed a theoretical framework to analyze the targeted influence maximization problem based on stability analysis and optimal perturbation. Our analysis led to a metric, termed targeted collective influence, that was used to identify influential spreaders in targeted spreading process. We validated the proposed method in both synthetic and real-world networks, demonstrating its robust performance that out-competes commonly used approaches. Our study provides a theoretically based metric that was shown effective in a range of network structures.

A. Competing Methods

We compare the targeted collective influence with several heuristic metrics that are widely used to rank the spreading capability of nodes. In numerical simulations, all nodes are ranked by each method and then the top \( n^* \) nodes with highest scores are selected as the set of seeds \( S_{n^*} \).

- Eigenvector-based ranking. The dynamical importance of nodes can be quantified using the eigenvector corresponding to the largest eigenvalue of the adjacency matrix [57]. Using a perturbation analysis on the largest eigenvalue, the dynamical importance of a node \( i \) is calculated as

\[
I_i = \frac{v_i u_i}{v^T u},
\]

where \( v \) and \( u \) denote the right and left eigenvectors of the adjacency matrix \( \{a_{ij}\}_{N \times N} \). In simulations, we use two versions of this method - Eig (for the original network) and Eigs (for the subgraph \( G_s \)).

- Degree-based ranking. In high degree (HD) ranking, the score of each node \( i \) is determined by the number of its connections: \( K^H D_i = \sum_{j \in \partial i} a_{ij} \). We also compare with the HD ranking in the subgraph \( G_s \) (HDs).

- Collective influence. The collective influence (CI) of each node is computed using power iteration that aims to estimate the largest eigenvalue of the NB matrix of the network [19]. Specifically, the CI score of node \( i \) at level \( \ell \) is \( CI_{\ell}(i) = (k_i - 1) \sum_{j \in \partial B_{2\ell}(i)} (k_j - 1) \).
Figure 6: The probability that at least one target node is infected in four real networks for 10 target nodes. The inset in each panel shows the ratio of $g(s_n^*)$ to $f(s_n^*)$. We use different methods (including Rand, Eig, Eigs, HD, HDs, CI and $\Delta CI$ with $\ell = 1$ and $\ell = 2$) to select the seeds for size $n^* = 4$. a. The co-authorship network of scientists with $N = 379$ and $M = 914$. b. The US power Grid network with $N = 4.9K$ and $M = 6.6K$. c. The web graph network with $N = 16.1K$ and $M = 25.6K$. d. The recommendation network of Amazon with $N = 91.8K$ and $M = 125.7K$.

B. Additional experiments

Figures (7) and (8) show the fraction of infected target nodes for all methods in Erdös-Rényi networks and scale-free networks respectively. Figures (9) and (10) show the results when target nodes are located in two and three clusters.

C. Network data

Network data are downloaded from the following websites. (1) The co-authorship network of scientists (https://networkrepository.com/ca-netscience.php). (2) The US power Grid network (https://networkrepository.com/USpowerGrid.php). (3) The web graph network (https://networkrepository.com/web-webbase-2001.php). (4) The recommendation networks of Amazon (https://networkrepository.com/rec-amazon.php).

Table 1
The properties of real-world networks used in this paper

| Network              | Size   | Links    | $\langle k \rangle$ | Maximum Degree | Maximum k-core |
|----------------------|--------|----------|---------------------|----------------|----------------|
| Co-authorship        | 379    | 914      | 4                   | 34             | 9              |
| US power Grid        | 4.9K   | 6.6K     | 2                   | 19             | 6              |
| Web-web graph        | 16.1K  | 25.6K    | 3                   | 1.7K           | 33             |
| Rec-amazon           | 91.8K  | 125.7K   | 2                   | 5              | 5              |

CRediT authorship contribution statement

Renquan Zhang: Designed this study, Wrote the code and first draft of the manuscript. Xiaolin Wang: Wrote the code, ran simulations and performed the analysis. Sen Pei: Designed this study, reviewed and edited the manuscript.
Targeted influence maximization

Figure 7: The fraction of infected target nodes in Erdős-Rényi (ER) networks with size $N = 10,000$ or $20,000$ and mean degree $\langle k \rangle = 3$ or $4$. We use different methods (including Rand, Eig, Eigs, HD, HDs, CI and $\Delta CI$ with $\ell = 1$ and $\ell = 2$) to select seeds for different size $n^* = 2, \cdots, 6$. The results are averaged over 100 independent realizations.

Acknowledgements

R. Zhang is supported by National Key Research and Development Program of China (Grant No.2021ZD0112400 and No.2020YFA0713702), National Natural Science Foundation of China (Grant No.11801058) and High-level Talents program of Dalian City (Grant No.2020RQ061).

References

[1] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Reviews of modern physics 87 (2015) 925.
[2] M. E. Newman, Spread of epidemic disease on networks, Physical review E 66 (2002) 016128.
[3] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Physical review letters 86 (2001) 3200.
[4] Z.-K. Zhang, C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, Y.-C. Zhang, Dynamics of information diffusion and its applications on complex networks, Physics Reports 651 (2016) 1–34.
[5] D. J. Watts, P. S. Dodds, Influentials, networks, and public opinion formation, Journal of consumer research 34 (2007) 441–458.
[6] S. Goel, A. Anderson, J. Hofman, D. J. Watts, The structural virality of online diffusion, Management Science 62 (2016) 180–196.
[7] B. Zhou, S. Pei, L. Muchnik, X. Meng, X. Xu, A. Sela, S. Havlin, H. E. Stanley, Realistic modelling of information spread using peer-to-peer diffusion patterns, Nature Human Behaviour 4 (2020) 1198–1207.
[8] D. Centola, The spread of behavior in an online social network experiment, science 329 (2010) 1194–1197.
[9] M. Granovetter, Threshold models of collective behavior, American journal of sociology 83 (1978) 1420–1443.
[10] S. Aral, D. Walker, Creating social contagion through viral product design: A randomized trial of peer influence in networks, Management science 57 (2011) 1623–1639.
[11] P. Domingos, M. Richardson, Mining the network value of customers, in: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001, pp. 57–66.
[12] J. Leskovec, L. A. Adamic, B. A. Huberman, The dynamics of viral marketing, ACM Transactions on the Web (TWEB) 1 (2007) 5–es.
[13] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, H. A. Makse, Identification of influential spreaders in complex networks, Nature physics 6 (2010) 888–893.
[14] S. Aral, D. Walker, Identifying influential and susceptible members of social networks, Science 337 (2012) 337–341.
Figure 8: The fraction of infected target nodes in scale-free (SF) networks with size $N = 10,000$ or $20,000$ and degree distribution $P(k) \sim k^{-3}$. We use different methods (including Rand, Eig, Eigs, HD, HDs, CI and $\Delta CI$ with $\ell = 1$ and $\ell = 2$) to select the seeds for different size $n^* = 2, \ldots, 6$. The results are averaged over 100 independent realizations.

[15] S. Pei, F. Morone, H. A. Makse, Theories for influencer identification in complex networks, in: Complex spreading phenomena in social systems, Springer, 2018, pp. 125–148.

[16] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, T. Zhou, Vital nodes identification in complex networks, Physics Reports 650 (2016) 1–63.

[17] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, pp. 137–146.

[18] W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 199–208.

[19] F. Morone, H. A. Makse, Influence maximization in complex networks through optimal percolation, Nature 524 (2015) 65–68.

[20] S. Pei, L. Muchnik, J. S. Andrade Jr, Z. Zheng, H. A. Makse, Searching for superspreaders of information in real-world social media, Scientific reports 4 (2014) 1–12.

[21] S. Aral, P. S. Dhillon, Social influence maximization under empirical influence models, Nature human behaviour 2 (2018) 375–382.

[22] X. Teng, S. Pei, F. Morone, H. A. Makse, Collective influence of multiple spreaders evaluated by tracing real information flow in large-scale social networks, Scientific reports 6 (2016) 1–11.

[23] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, Cost-effective outbreak detection in networks, in: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007, pp. 420–429.

[24] A. Braunstein, L. Dall’Asta, G. Semerjian, L. Zdeborová, Network dismantling, Proceedings of the National Academy of Sciences 113 (2016) 12368–12373.

[25] S. Aral, P. S. Dhillon, Social influence maximization under empirical influence models, Nature human behaviour 2 (2018) 375–382.
Figure 9: The probability that at least one target node is infected in four real networks with 20 target nodes. The inset in each panel shows the ratio of $g(s_{n_2})$ to $f(s_{n_0})$. We use different methods (including Rand, Eig, Eigs, HD, HDs, CI and $\Delta CI$ with $\ell = 1$ and $\ell = 2$) to select the seeds for size $n_0 = 4$. a, The co-authorship network of scientists with $N = 379$ and $M = 914$. b, The US power Grid network with $N = 4.9K$ and $M = 6.6K$. c, The web graph network with $N = 16.1K$ and $M = 25.6K$. d, The recommendation network of Amazon with $N = 91.8K$ and $M = 125.7K$.

[36] X. Ke, A. Khan, G. Cong, Finding seeds and relevant tags jointly: For targeted influence maximization in social networks, in: Proceedings of the 2018 International Conference on Management of Data, 2018, pp. 1097–1111.

[37] S. Su, X. Li, X. Cheng, C. Sun, Location-aware targeted influence maximization in social networks, Journal of the Association for Information Science and Technology 69 (2018) 229–241.

[38] Y. Sun, L. Ma, A. Zeng, W.-X. Wang, Spreading to localized targets in complex networks, Scientific reports 6 (2016) 1–10.

[39] K.-i. Hashimoto, Zeta functions of finite graphs and representations of p-adic groups, in: Automorphic forms and geometry of arithmetic varieties, Elsevier, 1989, pp. 211–280.

[40] S. Pei, X. Teng, J. Shaman, F. Morone, H. A. Makse, Efficient collective influence maximization in cascading processes with first-order transitions, Scientific reports 7 (2017) 1–13.

[41] B. Karrer, M. E. Newman, L. Zdeborová, Percolation on sparse networks, Physical review letters 113 (2014) 208702.

[42] K. E. Hamilton, L. P. Pryadko, Tight lower bound for percolation threshold on an infinite graph, Physical review letters 113 (2014) 208701.

[43] J. Wang, S. Pei, W. Wei, X. Feng, Z. Zheng, Optimal stabilization of boolean networks through collective influence, Physical Review E 97 (2018) 032305.

[44] D. Aleja, R. Criado, A. J. G. del Amo, Á. Pérez, M. Romance, Non-backtracking pagerank: From the classic model to hashimoto matrices, Chaos, Solitons & Fractals 126 (2019) 283–291.

[45] R. Zhang, S. Pei, Dynamic range maximization in excitable networks, Chaos: An Interdisciplinary Journal of Nonlinear Science 28 (2018) 013103.

[46] J. Wang, R. Zhang, W. Wei, S. Pei, Z. Zheng, On the stability of multilayer boolean networks under targeted immunization, Chaos: An Interdisciplinary Journal of Nonlinear Science 29 (2019) 013133.

[47] T. Martin, X. Zhang, M. E. Newman, Localization and centrality in networks, Physical review E 90 (2014) 052808.

[48] R. Zhang, G. Quan, J. Wang, S. Pei, Backtracking activation impacts the criticality of excitable networks, New Journal of Physics 22 (2020) 013038.

[49] T. Kawamoto, Localized eigenvectors of the non-backtracking matrix, Journal of Statistical Mechanics: Theory and Experiment 2016 (2016) 023404.

[50] C. Bordeneuve, M. Lelarge, L. Massoulié, Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs, in: 2015 IEEE 56th Annual Symposium on Foundations of Computer Science, IEEE, 2015, pp. 1347–1357.

[51] T. N. Palmer, Predicting uncertainty in forecasts of weather and climate, Reports on Progress in Physics 63 (2000) 71–116.

[52] Z. Toth, E. Kalnay, Ensemble forecasting at nmc: The generation of perturbations, Bulletin of the american meteorological society 74 (1993) 2317–2330.

[53] Z. Toth, E. Kalnay, Ensemble forecasting at ncep and the breeding method, Monthly Weather Review 125 (1997) 3297–3319.

[54] S. Pei, M. A. Cane, J. Shaman, Predictability in process-based ensemble forecast of influenza, PLoS computational biology 15 (2019)
Targeted influence maximization

Figure 10: The probability that at least one target node is infected in four real networks with 30 target nodes. The inset in each panel shows the ratio of $g(s_n^*)$ to $f(s_n^*)$. We use different methods (including Rand, Eig, Eigs, HD, HDs, CI and $\Delta CI$ with $\ell = 1$ and $\ell = 2$) to select the seeds for size $n^* = 4$. a, The co-authorship network of scientists with $N = 379$ and $M = 914$. b, The US power Grid network with $N = 4.9K$ and $M = 6.6K$. c, The web graph network with $N = 16.1K$ and $M = 25.6K$. d, The recommendation network of Amazon with $N = 91.8K$ and $M = 125.7K$.

[55] S. Pei, J. Shaman, Counteracting structural errors in ensemble forecast of influenza outbreaks, Nature communications 8 (2017) 1–10.
[56] S. Pei, X. Teng, P. Lewis, J. Shaman, Optimizing respiratory virus surveillance networks using uncertainty propagation, Nature communications 12 (2021) 1–10.
[57] J. G. Restrepo, E. Ott, B. R. Hunt, Characterizing the dynamical importance of network nodes and links, Physical review letters 97 (2006) 094102.
[58] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[59] R. Rossi, N. Ahmed, The network data repository with interactive graph analytics and visualization, in: Twenty-ninth AAAI conference on artificial intelligence, 2015.