TUDelft

Delft University of Technology

Predicting nodal influence via local iterative metrics

Zhang, Shilun; Hanjalic, Alan; Wang, Huijuan

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

scientific reports

OPEN

Predicting nodal influence via local iterative metrics

Shilun Zhang, Alan Hanjalic & Huijuan Wang✉

Nodal spreading influence is the capability of a node to activate the rest of the network when it is the seed of spreading. Combining nodal properties (centrality metrics) derived from local and global topological information respectively has been shown to better predict nodal influence than using a single metric. In this work, we investigate to what extent local and global topological information around a node contributes to the prediction of nodal influence and whether relatively local information is sufficient for the prediction. We show that by leveraging the iterative process used to derive a classical nodal centrality such as eigenvector centrality, we can define an iterative metric set that progressively incorporates more global information around the node. We propose to predict nodal influence using an iterative metric set that consists of an iterative metric from order 1 to K produced in an iterative process, encoding gradually more global information as K increases. Three iterative metrics are considered, which converge to three classical node centrality metrics, respectively. In various real-world networks and synthetic networks with community structures, we find that the prediction quality of each iterative based model converges to its optimal when the metric of relatively low orders ($K \sim 4$) are included and increases only marginally when further increasing K. This fast convergence of prediction quality with K is further explained by analyzing the correlation between the iterative metric and nodal influence, the convergence rate of each iterative process and network properties. The prediction quality of the best performing iterative metric set with $K = 4$ is comparable with the benchmark method that combines seven centrality metrics: their prediction quality ratio is within the range [91%, 106%] across all three quality measures and networks. In two spatially embedded networks with an extremely large diameter, however, iterative metric of higher orders, thus a large K, is needed to achieve comparable prediction quality with the benchmark.

Spreading processes are ubiquitous in various systems of nature and society. Examples include the spreading of epidemics, the propagation of information, and cascade of failures. Complex networks, usually considered as the underlying structure of such systems, provide the substrate upon which the spreading process unfolds via links connecting nodes. The spreading influence of a node represents the extent to which the node, where the spread originates, can eventually activate other nodes in the network. For a given spreading process, the spreading influence of a node is defined as the expected outbreak size when the spreading process starts from the node, also called the seed node. Due to the topological heterogeneity of nodes in many real networks[1], some nodes may have significantly higher spreading influence and are evidently more influential than the other nodes[2–4]. Identifying these influential nodes and predicting their spreading influence is crucial for controlling the spread of epidemics[5,6] or rumors[7,8], promoting strategic marketing[9–11], quantifying the impact of researchers and publications[12], and more[13–15].

Two generic influence prediction problems have been addressed in prior research. The first involves identifying the most influential nodes among all nodes based on the given network topology. To solve this problem, previous studies have proposed to rank nodes by a single nodal topological metric, so-called centrality metric[16–18], which encodes either local[19,20] or global[16,21] topological information around a given node. The highest-ranked nodes are then identified as the most influential ones. Nonetheless, these prior work suggests that no single centrality metric can outperform all other centralities for different epidemic parameters and in diverse types of networks, since a centrality metric only captures a certain topological feature of a node. It has been shown that nodal degree, i.e., number of 1-hop neighbors, is more (less) predictive than eigenvector centrality[22] when the spreading rate is small (large)[6,23]. The coreness better predicts the top spreaders than nodal degree in Susceptible-Infected-Recovered model below epidemic threshold. Further studies put forward methods to integrate local and global centralities or their rankings. Zhe Li et al.[24] used the sum of normalized degree, eigenvector centrality, and coreness as the mass of a node in a gravity model to derive a new nodal metric. Andrea Madotto et al.[25]

Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands. ✉email: H.Wang@tudelft.nl

1

aggregated the ranking lists by local and global node centralities to produce a new ranking list based on the correlations between the rankings. These methods usually exhibit better performance than merely using a local or global centrality.

In many practical scenarios, it is possible to observe or derive the spreading influences of a small fraction of nodes. For example, the average number of retweets of content posted by a node can be used as an approximation of the spreading influence of the node[6,26]. This motivates the second influence prediction problem: identify the most influential nodes given the network topology and the influence of a small fraction of nodes. Bucur[27] recently proposed to train a statistical model on the set of nodes whose spreading influences are known to classify the rest of nodes into binary classes, representing whether a node is among the top (e.g., top 10%) influential ones or not. The statistical model maps the relation between the class of a node in spreading influence and centrality metrics including both local centrality metrics like degree and global centrality metrics like betweenness[28] and eigenvector centrality. These centrality metrics were shown to be able to complement each other to achieve universally good performance in locating the most influential nodes across various real-world networks. However, global centrality metrics have a high computational complexity, which limits their application to large-scale networks. Moreover, the non-trivial correlation among different metrics makes it difficult to interpret to what extent global nodal properties are needed to estimate nodal spreading influence.

To bridge this gap, we will systematically explore two foundational questions: how local and global topological information around a node contribute to the prediction of the spreading influence of this node, and whether relatively local information, i.e., topological information derived from the neighborhood within a small hopcount from a target node, can predict its nodal spreading influence effectively. The general prediction task is considered: given the topology of a network and the spreading influences of a fraction of nodes, how to predict the spreading influences of the other nodes in the network, beyond their ranking. To solve the prediction task, a node-level regression model is trained on the set of nodes whose spreading influences are known and used to predict the influences of the remaining nodes. To understand how local and global topological information contribute to the prediction, we design the input of the regression model based on nodal properties as follows. We show that by leveraging the iterative process used to derive a classical node centrality such as eigenvector centrality, we can define an iterative metric that gradually encodes more global information as the order grows. Then, an iterative metric set that consists of an iterative metric from order 1 to order K is used as input features of the regression model. For example, the number of k-hop walks originate from a node, which is determined by the k-hop neighborhood of the node, can be derived in an iterative process starting from $k = 1$. The resultant iterative metric set is composed of the iterative metric (the number k-hop walks) with order $k \in [1, K]$ after K iterations. The benefits of using an iterative metric set to predict nodal influence are as following. Firstly, it allows us to explore to what extent global network information is needed to estimate the nodal influence, i.e., is K necessarily large for accurate prediction? Secondly, it enables us to identify the prediction method with low computational cost, i.e., the regression model with an iterative metric set of a small K. Moreover, in practical applications, one has the flexibility to choose an appropriate K to achieve a well-balanced trade-off between prediction accuracy and computational efficiency. The intuition is illustrated in Fig. 1, which shows a network example of 1000 nodes
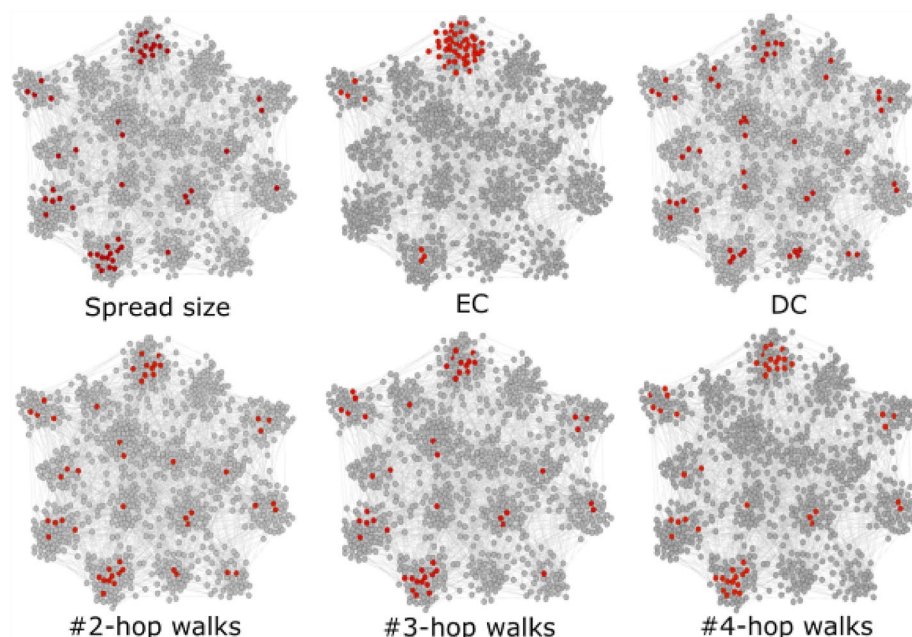

Figure 1. Location of top ranked nodes in a network generated by LFR model. The red-colored nodes are the top 10% nodes when nodes are ranked by spread size (top left), eigenvector centrality (EC, top middle), degree centrality (DC, top right), 2-hop walk counts (bottom left), 3-hop walk counts (bottom middle), and 4-hop walk counts (bottom right), respectively.

with community structure generated by Lancichinetti-Fortunato-Radicchi model[29]. The red-colored nodes are the top 10% nodes when nodes are ranked by spreading influence (top left), eigenvector centrality (EC, top middle, which corresponds the component of the eigenvector corresponds to the largest eigenvalue of the adjacency matrix), degree (DC, top right), number of 2-hop (bottom left), 3-hop (bottom middle) 4-hop (bottom right) walks originating from a node, respectively. The example suggests that the number of 2-, 3- and 4-hop walks possibly reflect nodal spreading influence better than the global metric (eigenvector centrality). Furthermore, it has been observed and partially proved in previous work that a centrality metric like betweenness with a high computational complexity is correlated with local metrics derived from a low order neighborhood[18,30]. Hence, global network information, i.e., large K, is not necessarily needed in nodal influence prediction.

In this work, we consider three iterative metrics, which converge, respectively to three global node centrality metrics: eigenvector centrality, PageRank centrality[31], and H index of a node[32]. The computation of each iterative metric set can be done in $\mathcal{O}\ (K \cdot |E|)$ time, where $|E|$ is the number of links in the network. Based on each iterative metric set, a statistical regression model is built and trained to predict nodal influence. We evaluate the prediction quality of the corresponding three regression models, in comparison with a benchmark[27], i.e., the regression model that uses 7 nodal centrality metrics, in both real-world networks and synthetic networks with community structure. We find that in almost all networks, an iterative metric set with $K \sim 4$ is able to accurately predict nodal spreading influence, and the prediction quality increases marginally when more global metrics are included as K grows. This suggests the low computational complexity of our iterative metric based prediction methods. Additionally, the best performing iterative metric based model with $K \sim 4$ performs as well as the benchmark model, which has higher computational cost due to the computation of global centrality metrics. An exception holds for two infrastructure networks, i.e., US power grid and Chicago regional road network, which are spatially embedded networks and have an extremely large diameter (> 40). In these two networks, nearly optimal prediction quality is achieved only when using the iterative metric set that includes metrics of large orders, thus when K is large. Hence, the proposed iterative metric method utilizing relatively local network information could predict nodal influence as well as the benchmark in networks with the small-world property and has a lower computational complexity.

This paper is organized as follows. In "Method" section, we introduce the definition of nodal spreading influence and iterative metrics, and regression models to predict nodal influence. "Results" section evaluates the performance of the proposed influence predication methods in both real-world networks and synthetic networks with community structure. "Discussion and future work" section summarizes our findings and discusses limitations and potential extensions of our work.

Method

In this section, we present the definition of nodal spreading influence ("Nodal spreading influence" subsection), followed by the definition of iterative metrics ("Iterative metrics" subsection). We then describe the regression model that uses an iterative metric set to predict nodal spreading influence ("Nodal influence prediction method" subsection).

Nodal spreading influence

We consider the continuous-time Susceptible-Infected-Recovered (SIR) spreading process on a static network[3,33]. At any time, each node can be in one of three possible states: susceptible, infected, or recovered. At the beginning, one seed node gets infected, while the rest are susceptible. A susceptible node gets infected by each of its infected neighbors at an infection rate β, and each infected node recovers at a recovery rate γ. Both the infection and recovery processes are independent Poisson processes. In the steady state, all nodes are either susceptible or recovered. The ratio $\lambda = \beta/\gamma$ is called the effective infection rate. Without loss of generality, we assume the recovery rate $\gamma = 1$, thus $\lambda = \beta$. For a given network, an epidemic threshold λ_c exists. When $\lambda > \lambda_c$, a nonzero fraction of recovered nodes exist in the stable state. When $\lambda < \lambda_c$, the epidemic dies out. The number of recovered nodes in the steady state, or equivalently, the number of nodes that have ever been infected is called the outbreak size.

The spreading influence of a node is defined as the average outbreak size when the node is chosen as the seed node. We derive the influence of a node as the average outbreak size over $r = 10^4$ realizations of the SIR spreading process on a given network. When the effective infection rate $\lambda \ll \lambda_c$ or when $\lambda \gg \lambda_c$, nodes tend to have similar influence. We focus on predicting influence when the effective infection rate is around the epidemic threshold, e.g., $\lambda = 0.5\lambda_c, \lambda_c, 1.5\lambda_c, 2\lambda_c$. This is when nodes differ evidently in influence, and influence prediction is crucial. We estimate the epidemic threshold λ_c using the numerical approach introduced in[34]. Specifically, referring to ρ as a random variable denoting the influence of a random node in the network, we consider the variability $\sqrt{\langle \rho^2 \rangle - \langle \rho \rangle^2}/\langle \rho \rangle$ as a function of λ. The epidemic threshold λ_c is then the value of λ that maximizes the variability.

Iterative metrics

Given an undirected network $G = (V, E)$, where V is the set of nodes and E is the set of links between nodes in V, the network can be represented by the adjacency matrix A, whose element $A_{ij} = 1$ if there is a link between node i and j, otherwise $A_{ij} = 0$. Various node centrality metrics have been proposed to measure the topological importance of a node, such as eigenvector centrality, PageRank, and coreness[32]. For a given centrality metric, the centralities of all nodes can be denoted by a vector $\mathscr{M}$, where the entry $\mathscr{M}_i$ represents the centrality of node i. The iterative process used to derive the corresponding iterative metric set starts with an initial metric vector $\mathscr{M}^{(0)}$ and

updates the metric vector based on a specific rule $\mathscr{M}^{(k)} = f(\mathscr{M}^{(k-1)})$. Eventually, this process converges to the target centrality metric $\mathscr{M}$. We refer to the derived metric vectors $\{\mathscr{M}^{(k)}, k = 1, 2, ...K\}$ as the iterative metric set.

In this paper, we consider three iterative processes that converge to three global centrality metrics: eigenvector centrality, PageRank centrality, and coreness of a node, respectively. Three different iterative metrics are derived using these processes.

- *Normalized Walk Count (NWC).* We adopt the power iteration process for the computation of eigenvector centrality to derive the NWC iterative metric. The centrality vector is initialized as the normalized all-one vector $w^{(0)} = u/\sqrt{N}$, where u is the all-one vector, and is updated iteratively following the updating equation $w^{(k)} = Aw^{(k-1)}/||Aw^{(k-1)}||$. The k-th order NWC follows $w^{(k)} = A^k u/||A^k u||$. Its element $w_i^{(k)}$ represents the normalized number of distinct k-hop walks starting from node i and can be derived from the neighborhood within k hops of the node i. As k increases, $w^{(k)}$ converges to the eigenvector centrality w. The rate of convergence is determined by the ratio of the largest eigenvalue $\lambda_1(A)$ and the second largest eigenvalue $\lambda_2(A)$ of the adjacency matrix A of the network. The convergence rate is higher when $\frac{|\lambda_2(A)|}{|\lambda_1(A)|}$ is smaller[35].

- *Visiting Probability (VP)* is derived using the iteration process for the computation of PageRank centrality[31]. The metric vector is initiated as the normalized all-one vector, $p^{(0)} = u/N$, and updated iteratively as $p_i^{(k)} = \alpha \sum_{j=1}^{N} A_{ji} p_j^{(k-1)}/d_j + (1-\alpha)/N$, where d_j is the degree of node j and the teleportation parameter α is set to 0.85, which is a common choice for calculating the PageRank centrality[36]. As k increases, $p_i^{(k)}$ converges to PageRank centrality. The updating equation can be formulated in matrix form: $p^{(k)} = Gp^{(k-1)}$, where $G = \alpha A^T D^{-1} + \frac{1-\alpha}{N} uu^T$, matrix D is a diagonal matrix with $D_{ii} = \sum_j A_{ij}$. Since matrix G is a stochastic matrix, the largest eigenvalue $\lambda_1(G) = 1$. The rate of convergence is determined by the second largest eigenvalue $\lambda_2(G)$ of the matrix G. The smaller $|\lambda_2(G)|$ is, the faster the convergence is[35]. The iterative process can be interpreted as a random walk: the walker starts at a randomly selected node. At each time step, with a probability α it moves to a random neighbor of the current visiting node, and with a probability $1-\alpha$ it jumps to a node that is randomly selected from the network. The k-th order iterative metric $p_i^{(k)}$ of a node i is the probability that node i is visited by the random walker at the k-th hop. Since the information of neighbors' degree is needed in each iteration step, $p_i^{(1)}$ actually encodes 2-hop neighbors' information. Similarly, the $(k+1)$-hop neighborhood information of a node i is needed to derive $p_i^{(k)}$.

- *H index (HI)*[32]. The 1-st order H index is defined as the degree of a node, i.e. $h_i^{(1)} = d_i$. The k-th order H index of node i can be derived as $h_i^{(k)} = \mathscr{H}[h_{j_1}^{(k-1)}, h_{j_2}^{(k-1)}, ..., h_{j_{d_i}}^{(k-1)}]$, where $j_1, ..., j_{d_i}$ are neighbors of node i and $\mathscr{H}$ is an operator that returns an integer. Specifically, $h_i^{(k)}$ is the maximum integer such that at least $h_i^{(k)}$ elements of $[h_{j_1}^{(k-1)}, h_{j_2}^{(k-1)}, ..., h_{j_{d_i}}^{(k-1)}]$ are no less than $HI_i^{(k)}$. It has been proved that $h^{(k)}$ will converge to the coreness[16,37] as k increases.

The iterative rules f in the three iterative processes only involve operations among a node's 1-hop neighbors. As a result, the metric vector $\mathscr{M}^{(k)}$ after one step iteration encodes information about the neighborhood one hop further than $\mathscr{M}^{(k-1)}$ (see Sect. S1 in Supplementary Information for a more detailed explanation). Given an iterative process, the obtained metric set $\{\mathscr{M}_i^{(1)}, \mathscr{M}_i^{(2)}, ..., \mathscr{M}_i^{(K)}\}$ will be used to predict the influence of node i using the regression model described in "Nodal influence prediction method" subsection. The parameter K controls the scope of information around a node encoded in the iterative metric set $\{\mathscr{M}_i^{(1)}, \mathscr{M}_i^{(2)}, ..., \mathscr{M}_i^{(K)}\}$.

Nodal influence prediction method

We assume two key types of information are given to predict nodal influence. Firstly, the network topology is known. Secondly, the influences of a small fraction of nodes are available. In practical scenarios, these influences can often be estimated from real-world diffusion data within social media networks. Our objective is to predict the influences of the remaining nodes in the network. We approach the prediction of nodal influence as a node-level regression problem. Specifically, given a static network $G = (V, E)$ represented by its adjacency matrix A and the spreading influences of a fraction q of nodes, which is randomly selected and denoted as S_q, we aim to predict spreading influences of the remaining $1-q$ nodes, referred to as S_{1-q}.

We choose $q = 10\%$ assuming only the influences of a small fraction of nodes are known. We train a statistical regression model, which maps the nodal features into the influence of a node, on the training node set S_q, and evaluate it on the remaining test node set S_{1-q}. For each of the three iterative metrics, the iterative metric set $\{\mathscr{M}_i^{(1)}, \mathscr{M}_i^{(2)}, ..., \mathscr{M}_i^{(K)}\}$ is used as nodal features in the regression model to predict nodal influence. As a benchmark model, we consider a regression model that uses the same set of 7 classic centrality metrics as in Bucur's classification model[27] as nodal features. These 7 centrality metrics include both local and global centrality metrics and are able to complement each other in improving the performance in the node classification task. Finally, we evaluate the prediction quality of the regression models based on 50 realizations of the random sampling of the training node set S_q and the training of the regression model.

We choose the Random Forest Regression model (RFR), a classic model that captures the nonlinear relationship between input features and the outcome variable, i.e., nodal influence, in our case. We also considered the

Ridge regression, a linear regression model with L2 regularization, and obtained qualitatively similar observations (in Supplementary Information) as the Random Forest Regression.

Results

We evaluate the performance of the regression models based on each of the three iterative metrics and the benchmark model based on classic centrality metrics, first in real-world networks in "Performance analysis in real-world networks" subsection, and afterwards in synthetic networks with community structures in "Prediction on networks with communities" subsection. Finally, we explore the performance of these models in relation to parameters of the spreading process in "Prediction of nodal spreading influence near epidemic threshold" subsection.

Networks and measures to evaluate prediction quality

We consider 9 real-world networks that differ in network properties such as size and and diameter (i.e. the largest shortest path length between a node pair among all possible node pairs), including four online social networks (advogato, facebook, deezerEU, github), a scientific collaboration networks (Arxiv Astro), a file sharing network (Gnutella04), two infrastructure networks (US power grid, ChicagoRegional road network), and an email communication network (Email Enron). All the datasets are obtained from the repository of KONECT project[38,39]. We treat all networks as simple, undirected and unweighted. Basic properties of these networks are listed in Table 1. Notably, the two infrastructure networks, US powergrid and ChicagoRegional, have significantly larger diameters, higher modularity, and lower average degree than the other networks.

We evaluate the prediction quality of the proposed regression models using the following 3 classic measures:

r^2 measures the proportion of the variance in the dependent variable (nodal influence) that is predictable from the input features in the regression model. r^2 is defined as:

$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{1}$$

Here, y_i and $\hat{y}_i$ are the ground truth and the predicted nodal influence of node i given by the regression model, respectively. $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the mean value of y_i.

Kendall's correlation coefficient $\tau(\hat{s}, s)$ measures the similarity of the two ranking lists of nodes based on the predicted nodal influence $\hat{s}$ and the ranking based on the actual nodal influence obtained by SIR simulation. A value of 1 for $\tau(\hat{s}, s)$ indicates that the predicted nodal influence gives the same node ranking as the ground truth, while a value of -1 indicates that the two rankings are reverse. Kendall's correlation coefficient[40] $\tau(\hat{s}, s)$ is defined as follows:

$$\tau(\hat{s}, s) = \frac{n_c - n_d}{\sqrt{(n_c + n_d + T) * (n_c + n_d + U)}} \tag{2}$$

where n_c and n_d are the total number of node pairs that are concordant and discordant respectively, based on the influence s and the predicted influence $\hat{s}$. For example, node pair (i, j) is concordant if $(\hat{s}_i - \hat{s}_j)(s_i - s_j) > 0$, and is discordant if $(\hat{s}_i - \hat{s}_j)(s_i - s_j) < 0$. T is the number of node pairs that have the same influence but different predicted influence, i.e., $s_i = s_j, \hat{s}_i \neq \hat{s}_j$ and U is the number of node pairs that have the same predicted influence but different influence, i.e., $\hat{s}_i = \hat{s}_j, s_i \neq s_j$.

Recognition rate of top-f% measures the performance of a regression model in identifying the most influential f% nodes in the test set S_{1-q}. It is calculated as the fraction of nodes that are present in the top f% of both the ranking by predicted nodal influence $\hat{s}$ and the ranking by actual nodal influence s. A higher recognition rate of top-f% implies better performance of the regression model in identifying the most influential nodes.

Dataset	$|N|$	$|E|$	$\langle d \rangle$	Diameter	Q	λ_c
Advogato	5042	41791	16.577	9	0.408	0.020
Arxiv-astrophics (astroph)	17903	196972	22.004	14	0.626	0.015
Enron	33696	180811	10.732	13	0.608	0.013
Facebook	63392	816886	25.773	15	0.632	0.010
Gnutella04 (gnu04)	10876	39994	7.355	10	0.386	0.080
Github	37700	289003	15.332	11	0.453	0.011
Deezer EU (deezereu)	28281	92752	6.559	21	0.683	0.070
US power grid (uspower)	4941	6594	2.669	46	0.935	0.870
ChicagoRegional (Chicago)	12979	20627	3.179	106	0.931	1.230

Table 1. Basic properties of each real-world network considered: Number of nodes $|N|$, number of links $|E|$, average node degree $\langle d \rangle$, network diameter, the modularity Q[1], and epidemic threshold λ_C of the SIR process on the network.
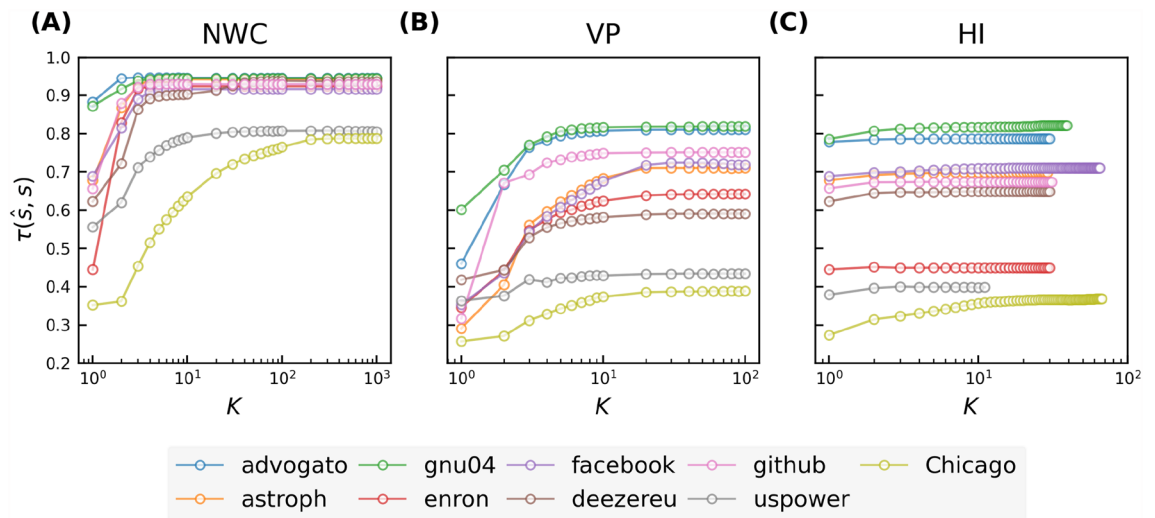
Figure 2. Kendall correlation between the actual nodal spreading influence s and the influence $\hat{s}$ predicted by a regression model based on NWC (panel (**A**)), VP (panel (**B**)), and H index (panel (**C**)) respectively. Results are averaged over 50 realizations of training set sampling and model training.
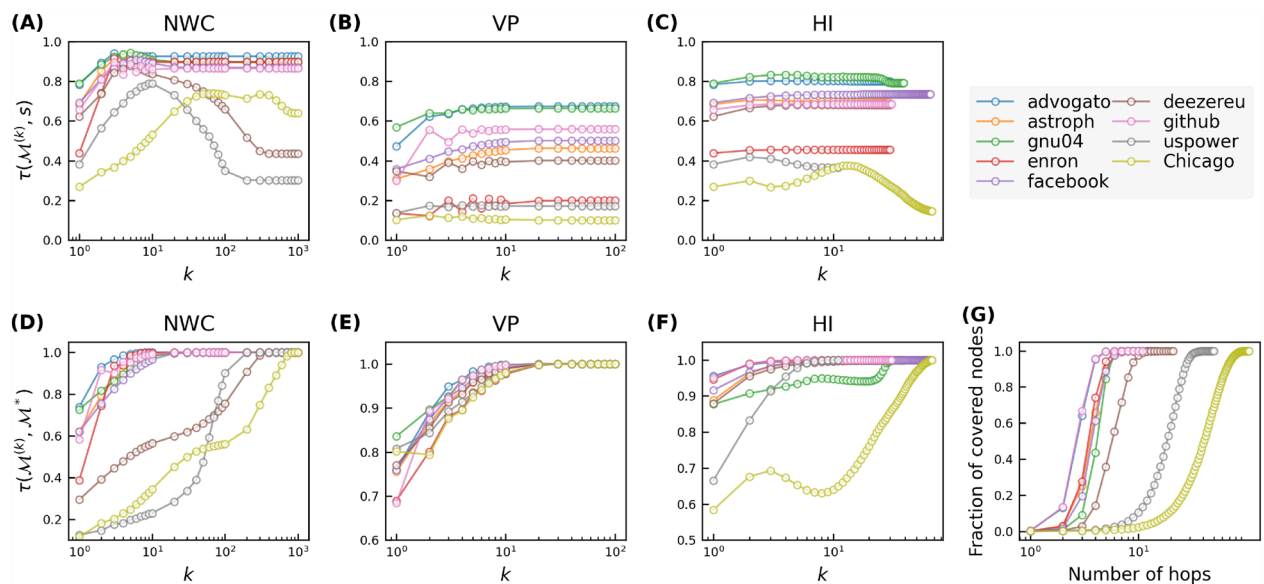

Figure 3. Kendall correlation between nodal spreading influence s and different orders of NWC ($w^{(k)}$, panel (**A**)), VP ($p^{(k)}$, panel (**B**)), and H index ($h^{(k)}$, panel (**C**)), and the convergence of NWC (**D**), VP (**E**), HI (**F**), measured by the Kendall's correlation between the iterative metric after k iterations and the corresponding global centrality metrics, as a function of iteration number k in 9 real-world networks. (**G**) shows the coverage, i.e. the average fraction of nodes covered by hopping step out from a node, as a function of the number of hops.

Performance analysis in real-world networks

We focus on the prediction of spreading influence when the effective infection rate of the SIR spreading process is $\lambda = \lambda_c$, where the epidemic threshold λ_c of each network is identified using the method described in "Method" section. The values of λ_c of each real-world network are shown in Table 1. Later in this section, we will discuss how the choice of the effective infection rate around the epidemic threshold impacts the performance of influence prediction methods.

We predict nodal influence in real-world networks using the iterative metric based regression models. Each model uses an iterative metric set $\{\mathcal{M}_i^{(1)}, \mathcal{M}_i^{(2)}, ..., \mathcal{M}_i^{(K)}\}$ as input features. Thus, topological information of the K-hop ($K + 1$-hop for VP) neighborhood of each node is used by the regression model for influence prediction. These regression models are evaluated using the evaluation metrics introduced in "Method" section. In Fig. 2, we show the Kendall correlations $\tau(\hat{s}, s)$ between the actual nodal influence s and the influence $\hat{s}$ predicted by a regression model as a function of K in real-world networks. As K grows, higher order iterative metrics are
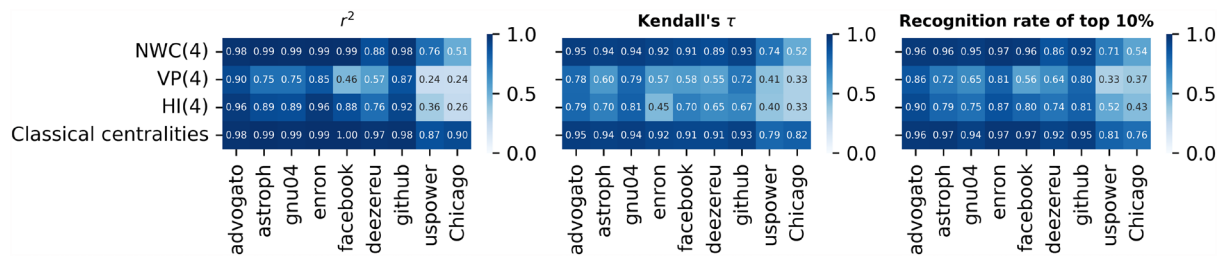
Figure 4. Comparison of prediction quality across different empirical networks (horizontal axis) of four prediction models based on different metrics (vertical axis): Normalized Walk Count when $K = 4$ (NWC(4)), Visiting Probability when $K = 4$ (VP(4)), H index when $K = 4$ (HI(4)), and classic centrality metrics[27]. Three panels correspond to different evaluation measures of prediction quality: r^2 (left panel), Kendall's τ (middle panel), and recognition rate of top 10% nodes (right panel), respectively. Results are averaged over 50 realizations of Random Forest Model training process.

included, and the prediction quality increases. For all three iterative metrics, the prediction quality converges relatively fast as K increases. As shown in Fig. 2A, the prediction quality of the NWC based model is already close to the highest at a small K ($K \sim 4$) and only increases marginally by choosing a $K > 4$. For example, the prediction quality when $K = 4$ reaches at least 95% of the highest prediction quality of the NWC based model. This suggests that a regression model using relatively local topological information could already achieve comparably good prediction quality as the one using more global information. This finding does not hold for the two infrastructure networks with an extremely large diameter, for which an iterative metric of higher orders (i.e., $K > 4$) is needed to achieve optimal prediction quality.

To understand why an iterative metric method achieves nearly its optimal prediction quality with a small $K \sim 4$ in all networks except for the two networks without the small-world property, we first explore the correlation $\tau(\mathscr{M}^{(k)}, s)$ between the k-th order iterative metric $\mathscr{M}^{(k)}$ and the spreading influence s. As shown in Fig. 3A–C, each iterative metric $\mathscr{M}^{(k)}$ exhibits positive correlation with spreading influence for any order k, indicating that each iterative metric has certain predictive power. As k increases, the correlation $\tau(\mathscr{M}^{(k)}, s)$ increases when k is small and achieves nearly the highest correlation around $k \sim 4$, implying the high predictive power of iterative metrics of up to order 4 in those small-world networks.

Secondly, we study the convergence of the iterative metric $\mathscr{M}^{(k)}$ as the order k grows. As k increases, each centrality metric $\mathscr{M}^{(k)}$ converges to the global centrality metric $\mathscr{M}^*$. The three iterative metrics converge to three global metrics: eigenvector centrality, PageRank centrality, coreness, respectively. Figure 3D–F shows the Kendall's correlation $\tau(\mathscr{M}^{(k)}, \mathscr{M}^*)$ between the k-th order metric $\mathscr{M}^{(k)}$ and the global metric $\mathscr{M}^*$ as a function of k for each iterative metric. For each iterative metric, $\mathscr{M}^{(k)}$ converges to $\mathscr{M}^*$ with different convergence rates in different networks. Importantly, $\mathscr{M}^{(k)}$ exhibits relatively high correlation with $\mathscr{M}^*$ at $k \sim 4$ in most networks. Hence, the predictive power of an high-order iterative metric could be inherited by a low-order iterative metric. This explains why the corresponding regression model improves in prediction quality only marginally as K increases when $K \geq 4$. Furthermore, the large correlation $\tau(h^{(k)}, h^*)$ for any k, as shown in Fig. 3F, explains why the prediction quality of the regression model based on HI hardly improves when K grows, as observed in Fig. 2C.

In the two infrastructure networks with a large diameter and strong community structure, iterative metrics converge relatively slowly, indicating the possibility that a large K or high-order iterative metric is needed for better prediction quality. Still, the convergence of the prediction quality $\tau(\mathscr{M}^{(k)}, s)$ is faster than that of the metric NWC $\tau(\mathscr{M}^{(k)}, \mathscr{M}^*)$. This is likely because the higher-order metric is less predictive, thus possibly less needed for the prediction, as shown in the decreasing trend of the correlation $\tau(\mathscr{M}^{(k)}, s)$ with an increasing k when k is large. The different performance of the iterative metric based model in the two infrastructure networks from the other networks as well as the weakness of using a single classical centrality to predict influence precisely in networks with community structure[41,42] motivate us to investigate the impact of the strength of community structure on nodal influence prediction in the next section.

To gain insight into why each iterative metric $\mathscr{M}^{(k)}$ exhibits relatively high correlation with $\mathscr{M}^*$ at $k \sim 4$ in most networks, we investigate the average size of the k-hop neighborhood, i.e., the fraction of nodes that is reachable (covered) from a random node in k hops. This indicates the proportion of nodes whose information is considered in the metric $\mathscr{M}^{(k)}$. Figure 3G shows that in most real-world networks, more than half of nodes are reachable from a random node within 4 hops. Hence, the 4-th order iterative metric possibly captures the topological information of a significant amount of nodes, supporting why $\tau(\mathscr{M}^{(k)}, \mathscr{M}^*)$ is high when $k \sim 4$. The 4-hop coverage of network *deezer EU* and the two infrastructure networks is lower than in the other networks, which is likely due to their community structure or large diameter. Correspondingly, $\tau(\mathscr{M}^{(k)}, \mathscr{M}^*)$ when $k \sim 4$ for NMC is relatively lower in these three networks.

Among all three iterative metrics, NWC achieves evidently the highest prediction quality when $K \sim 4$. This is supported by the higher correlation $\tau(s, w^{(k)})$ between the NWC centrality $w^{(k)}$ and the spreading influence s at each order k, as shown in Fig. 3A–C.

Iterative metric	Degree, neighborhood, two-hop neighborhood	Coreness	Eigenvector and PageRank	Closeness
$\mathcal{O}\ (K \cdot |E|)$	$\mathcal{O}\ (|E|)$	$\mathcal{O}\ (|E|)$	$\mathcal{O}\ (K^* \cdot |E|)$	$\mathcal{O}\ (|V||E|)$

Table 2. Comparison of the computational complexity of different nodal metrics for all nodes in a network: an iterative metric set $\{\mathcal{M}^{(1)}, ..., \mathcal{M}^{(K)}\}$ and classical centrality metrics used in the benchmark model. Neighborhood stands for the sum of degrees of direct neighbors, and two-hop neighborhood are the sum of degrees of nodes that are two hops away. K^* is the number of iterations at which the iterative process used to compute the centrality metric converges.

μ	Diameter	Q	λ_c
0.02	10	0.924	0.090
0.05	6	0.872	0.080
0.1	5	0.608	0.070
0.2	5	0.632	0.070
0.3	5	0.386	0.070
0.4	5	0.453	0.070

Table 3. Basic properties of networks generated by LFR model with different mixing parameter μ and $N = 1000$: network diameter, the modularity Q, epidemic threshold λ_C of the SIR process on the network.

It has been found that combining local and global node centrality metrics can more accurately identify top influencers than using either local or global centralities alone[27]. Hence, we build a benchmark regression model that uses the same 7 centrality metrics (local ones like degree and global ones like closeness) as in the classification model in[27] as input features. Now, we compare the prediction quality of the proposed iterative metric based models with the benchmark model. We choose $K = 4$ for iterative metric based models. The choice of $K = 4$ corresponds the case where the iterative metric based model only uses relatively local information, which ensures the computational efficiency and reasonably good prediction quality in most networks.

Figure 4 shows three evaluation measures of the regression models: r^2 (left panel), Kendall correlation between the actual nodal spreading influence s and the predicted influence $\hat{s}$ of the node by a regression model (middle panel), and the recognition rate of top 10% nodes (right panel). Across all real-world networks, the prediction quality of the NWC based model is evidently better than the other two iterative metric based models. In all networks except for the two infrastructure networks, the NWC based model achieves prediction quality comparable to the benchmark model. The prediction quality ratio between the NWC based model and the benchmark model is within the range $[91\%, 101\%]$ for any of the three evaluation measures. In those two infrastructure networks uspower and Chicago, the NWC based model with $K = 4$ performs worse than the benchmark, whereas the NWC based model with a large K performs as well as the benchmark, achieving 96% to 105% of the prediction quality of the benchmark model.

Moreover, the computational complexity of the NWC based model with $K = 4$ is lower than that of the benchmark model, which requires the computation of global centrality metrics. We summarize in Table 2 the computational complexity of an iterative metric of orders up to K and the 7 classical centrality metrics used in the benchmark model for all nodes. In each iteration of an iterative process, the iterative metric of each node is updated via aggregating the metrics of its 1-hop neighbors derived in the previous iteration. Thus, updating the metric for all nodes in each iteration requires $2|E|$ basic operations. The computational complexity of an iterative metric set $\{\mathcal{M}_i^{(1)}, \mathcal{M}_i^{(2)}, ..., \mathcal{M}_i^{(K)}\}$ for all nodes equals that of $\mathcal{M}_i^{(K)}$ for all nodes, which is $\mathcal{O}\ (K \cdot |E|)$. Hence, a relatively small K facilitates the application of iterative metric based method in large networks. In contrast, the global metrics used in the benchmark model, such as closeness centrality, have a higher complexity.

Prediction on networks with communities

Community structure has been observed in many real-world networks[43], where nodes within a community are densely connected while nodes from different communities have fewer connections. The existence of communities affects significantly the spreading process unfolding on a network[44,45] and has been ignored in most centrality metrics used to predict nodal influence[42,46]. Here we evaluate the performance of our influence prediction methods in networks with communities and investigate how community structure affects the prediction quality. To this end, we adopt the Lancichinetti-Fortunato-Radicchi (LFR) model[29] to generate networks with power-law degree distribution and community size distribution, as observed in real-world networks. One advantage of LFR model is that the strength of the community structure in the generated networks can be changed via tuning its parameters. We use LFR model to generate networks with the following properties: network size $N = 1000$ and $N = 10000$ respectively, the exponent of the power-law degree distribution $\tau_1 = 2$, and exponent of the power-law community size distribution $\tau_2 = 3$, the average degree $\langle k \rangle = 10$, the maximum degree $d_{max} = \sqrt{10N}/2$, the range of community sizes $[50, \sqrt{10N}]$. The mixing parameter μ represents the fraction of inter-community links of a node. When $\mu = 0$, the generated networks have the strongest community structure, with communities being disjoint from each other. The model with $\mu = 1$ generates networks where
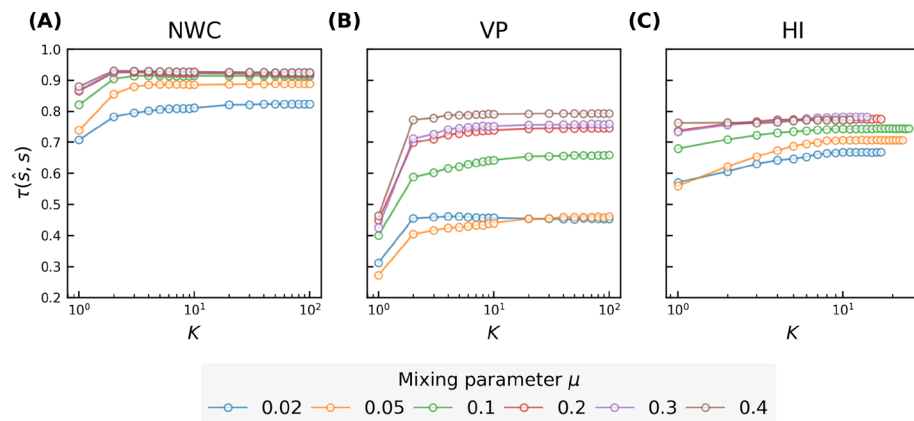
Figure 5. Kendall correlation between nodal spreading influence $\hat{s}$ predicted by different numbers of iterative metrics as features and nodal spreading influence given by SIR simulations of NWC (**A**), VP (**B**), and H index (**C**). Results are averaged over 50 realizations of training set sampling and model training.
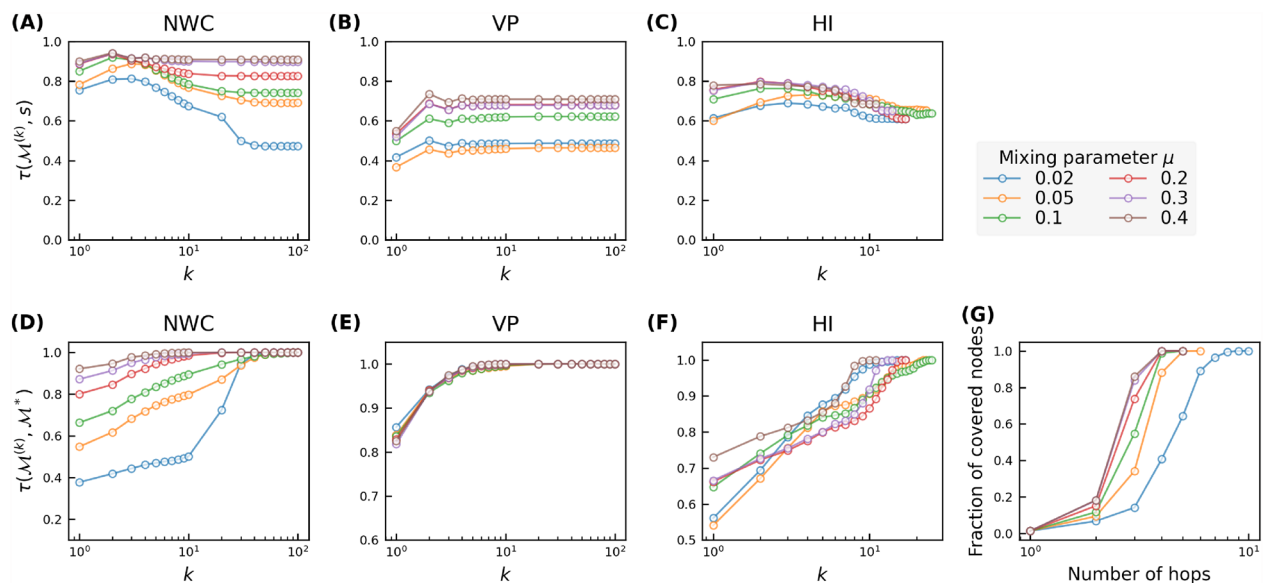

Figure 6. Kendall correlation between nodal spreading influence s and different orders of NWC ($w^{(k)}$, panel (**A**)), VP ($p^{(k)}$, panel (**B**)), and H index ($h^{(j)}$, panel (**C**)), and the convergence of NWC (**D**), VP (**E**), HI (**F**), measured by the Kendall's correlation between the iterative metric after k iterations and the corresponding global centrality metrics, as a function of iteration number k in Lancichinetti-Fortunato-Radicchi (LFR) networks with different $\mu = 0.02, 0.05, 0.1, 0.2, 0.3, 0.4$. (**G**) shows the coverage, i.e. the average fraction of nodes covered by hopping step out from a node, as a function of the number of hops.

all links fall between different clusters. When $\mu > 0.5$, the community structure is not evident anymore[29]. We set $\mu = [0.02, 0.05, 0.1, 0.2, 0.3, 0.4]$, thus six networks with different strength of communities are generated. We will focus on the results for $N = 1000$, since results for $N = 10000$ (as shown in the Supplementary Information) lead to the same observation. The generated networks vary in network properties such as diameter and modularity, as shown in Table 3 and Table S1.

We first evaluate our iterative metric based models in predicting nodal influence in LFR networks when the effective infection rate of the SIR model is around epidemic threshold, i.e., $\lambda = \lambda_c$. Figure 5A–C show Kendall correlations $\tau(\hat{s}, s)$ between the nodal spreading influence s and the prediction $\hat{s}$ by a regression model based on an iterative metric set $\{\mathcal{M}^{(1)}, \mathcal{M}^{(2)}, ..., \mathcal{M}^{(K)}\}$, as a function of K in LFR networks. Like what we observed in real-world networks, the prediction quality increases as K increases. Notably, the prediction quality only improves marginally when choosing a $K > 4$. This can be understood by the correlation $\tau(\mathcal{M}^{(k)}, s)$ between $\mathcal{M}^{(k)}$ and nodal influence s, which is shown in Fig. 6A–C. As k increases up to $k \sim 4$, the correlation $\tau(\mathcal{M}^{(k)}, s)$ increases. As k increases further, the correlation tends to decrease. This decreasing trend is more evident in networks with more evident community structure, but not observed in real-world networks that have a relatively small diameter and
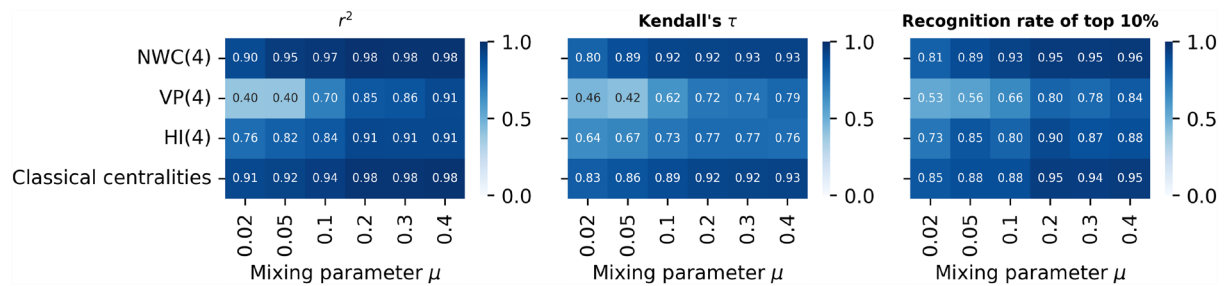
Figure 7. Prediction performance on model networks generated with LFR model with varying mixing parameter μ (horizontal axis) of five sets of metrics (vertical axis): Normalized Walk Count when $K = 4$ (NWC(4)), Visiting Probability when $K = 4$ (VP(4)), H index when $K = 4$ (HI(4)), and classical centralities. Three panels correspond to different evaluation measures of predictive models: r^2 (left panel), Kendall's τ (middle panel), and recognition rate of top 10% nodes (right panel), respectively. Results are averaged over 50 realizations of training process of Random Forest Model.

modularity as shown in Fig. 3. This suggests that high-order ($k > 4$) iterative metrics are less predictive than an iterative metric of an order around $k \sim 4$, thus less needed to predict nodal influence in networks with a higher modularity. Furthermore, we explore the convergence of an iterative metric $\mathcal{M}^{(k)}$ as k increases. Figure 6D–F show the Kendall's correlation $\tau(\mathcal{M}^{(k)}, \mathcal{M}^*)$ as a function of k for the three iterative metrics, respectively. For NWC, the correlation tends to be lower when $k \sim 4$ as the mixing parameter μ gets smaller or equivalently in network with more evident community structure. In networks with strong community structure, NWC converges relatively slowly. Still, the prediction quality of the regression models in these networks is close to optimal when $K \sim 4$, since the higher order metric is less predictive. This is also in line with the intuition that in networks with strong community structure and when the infection rate is around the critical epidemic threshold, nodal influence is supposed to be mainly determined by nodal property derived within or around the community that the node belongs to.

Figure 6G shows the average fraction of nodes that are reachable (covered) from a randomly chosen node within k hops, i.e., the so called coverage, as a function of k. In networks with strong community structure (small μ), the coverage and $\tau(\mathcal{M}^{(k)}, \mathcal{M}^*)$ when $k \sim 4$ tend to be small. In such networks, an order $k \sim 4$ iterative metric encodes topological information of a small fraction of nodes, which explains partially the weak correlation $\tau(\mathcal{M}^{(k)}, \mathcal{M}^*)$ when $k \sim 4$.

Now we compare the prediction quality of iterative metric based models (when $K = 4$) with the benchmark model in LFR networks via the same three evaluation measures as in real-world networks. Figure 7 shows that the NWC based model with $K = 4$ performs comparably as (mostly slightly better than) the benchmark model, the prediction quality ratio between the NWC based model and the benchmark model ranges from 95% to 106%. Among the three iterative metric based models, the NWC based model performs the best whereas VP based model performs the worst. As the strength of community structure grows, all models perform worse. This can be explained by the small (large) correlation $\tau(\mathcal{M}^{(k)}, s)$ in networks with a strong (weak) community structure, as shown in Fig. 6A–C. The same has also been observed in real-world networks. As shown in Fig. 4, both the NWC based model and the benchmark model perform the worst in the two infrastructure networks that have the stronger community structure than the other considered real-world networks. In the two infrastructure networks, the correlation $\tau(\mathcal{M}^{(k)}, s)$ is also weaker (see Fig. 3).

Prediction of nodal spreading influence near epidemic threshold

So far, we have focused on the influence prediction problem, where the influence is defined for the SIR epidemic spreading process with $\lambda = \lambda_c$. It has been shown that the change of parameters in the epidemic spreading can lead to different rankings of nodes according to their influences[23,47,48]. Hence, we evaluate the average prediction quality of a regression model over all the networks except for the two infrastructure networks, at various effective infection rates around the epidemic threshold λ_c. Figure 8 (top panel) shows that NWC outperforms VP and HI, as λ varies from $0.5 \cdot \lambda_c$ to $2.0 \cdot \lambda_c$. The NWC based model with $K = 4$ and the benchmark model show comparable prediction quality. Their prediction quality is less sensitive to the effective infection rate λ. In the two infrastructure networks (Fig. 8 bottom panel), the NWC based model with $K = 4$ exhibits lower prediction quality than the benchmark at different effective infection rates except that they perform similarly at $\lambda = 0.5 \cdot \lambda_c$, when the SIR spreading is relatively local.

Discussion and future work

In summary, we explore to what extent local and global topological information of a node is needed for the prediction of nodal spreading influence and whether relatively local topological information around a node is sufficient for the prediction. We propose to predict nodal influence by an iterative metric set derived from an iterative process. Three iterative metrics are considered: Normalized Walk Counts (NWC), Visiting Probability (VP), and H index (HI), which converge to eigenvector centrality, PageRank, and H index, respectively. The regression model using an iterative metric set as input features is trained on a fraction of nodes whose influence is known and is used to predict the nodal influence of the remaining nodes. We evaluate and interpret the performance of
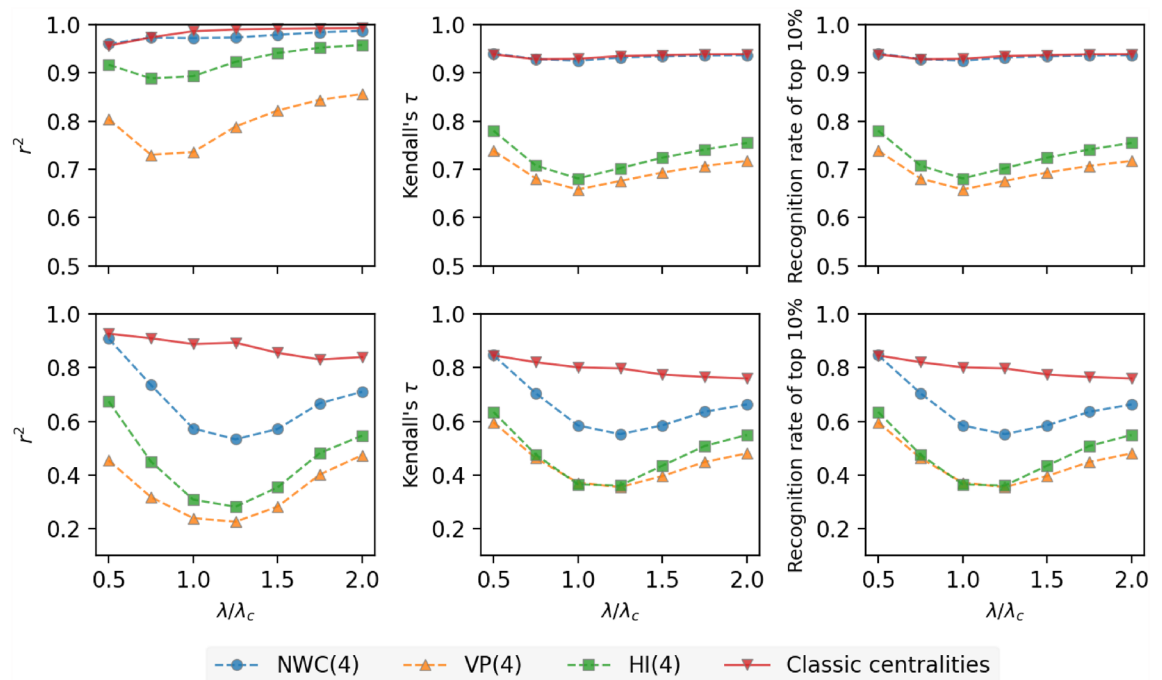
Figure 8. Average prediction quality over all considered real-world networks (shown in Table 1) excluding the two spatially embedded networks (top panels) and over the two spatially embedded networks (bottom panels) as a function of λ/λ_c of 4 different metric sets: Normalized Walk Counts (NWC), Visiting Probability (VP), H index (HI), and 7 node centralities[27]. Three columns correspond to different evaluation measures of predictive models: r^2 (left panel), Kendall's τ (middle panel), and recognition rate of top 10% nodes (right panel), respectively. Results are averaged over 50 realizations of training process of Random Forest Model.

these three iterative metric based models in predicting nodal influence in SIR spreading processes with diverse effective infection rates around the epidemic threshold, on both real-world networks and synthetic networks with different strength of community structure. We find that the prediction quality of each iterative metric based model converges to its optimal when the iterative metric set of relatively low orders (up to order 4) are included and increases only marginally when further increasing K. This is explained via the correlation between an iterative metric of order k and nodal influence and the fast convergence of each iterative metric. The prediction quality of the best performing iterative metric set (NWC) with $K = 4$ is comparable with the benchmark method that combines seven centrality metrics. In two spatially embedded networks with an extremely large diameter and modularity, however, iterative metric of higher orders, thus a large K, is needed to achieve comparable predict quality as the benchmark. These findings suggest that the NWC metric of relatively low orders contain sufficient information to predict nodal influence reasonably well in networks with the small-world property, whereas its computation complexity is lower than that of the global centrality metrics needed by the benchmark model. In these networks, the NWC metric has almost the highest correlation with nodal influence when $k \approx 4$ in most networks, indicating that a node with more distinct 4-hop walks starting from the node tends to be more influential. However, the interpretability of the iterative metric-based regression model is limited by the strong correlation among the iterative metric of different orders. Nodes with what kind of combination of low order the iterative metrics are more influential remains an interesting question.

This study has several limitations that call for further exploration. Firstly, we observe the trend that a larger K is needed for the iterative metric based method to perform close to its optimal in networks with a significant large diameter. It is interesting to explore the minimal K needed for the NWC based model to perform at least, for example, 95% of the optimal performance of the model in relation to the diameter of the network. Secondly, the diameter and strength of community structure are possibly correlated in real-world networks and network models. We have observed the influence of community structure or diameter on the prediction quality of the NWC based model and the benchmark model. An open question is how the diameter influences the prediction quality while the community strength is fixed. For both objectives, network models with a controllable diameter and more real-world networks, especially those without the small-world property are needed. Thirdly, we confine ourselves to the SIR spreading process on a static network. However, in many scenarios, both the spreading process and the underlying topology can be more complicated. Our proposed method can be extended to explore its capability of predicting nodal influence defined in such more complex context using local network information.

Data availibility

The data sets used are publicly available. More information can be found in the corresponding references.

References

1. Newman, M. *Networks* (Oxford University Press, 2018).
2. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
3. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925 (2015).
4. Hu, Y. *et al.* Local structure can identify and quantify influential global spreaders in large scale social networks. *Proc. Natl. Acad. Sci.* **115**, 7468–7472 (2018).
5. Woolhouse, M. E. *et al.* Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proc. Natl. Acad. Sci.* **94**, 338–342 (1997).
6. Pei, S. & Makse, H. A. Spreading dynamics in complex networks. *J. Stat. Mech. Theory Exp.* **2013**, P12002 (2013).
7. Chen, X. & Wang, N. Rumor spreading model considering rumor credibility, correlation and crowd classification based on personality. *Sci. Rep.* **10**, 1–15 (2020).
8. Bovet, A. & Makse, H. A. Influence of fake news in twitter during the 2016 us presidential election. *Nat. Commun.* **10**, 1–14 (2019).
9. Watts, D. J. & Dodds, P. S. Influentials, networks, and public opinion formation. *J. Consumer Res.* **34**, 441–458 (2007).
10. Leskovec, J., Adamic, L. A. & Huberman, B. A. The dynamics of viral marketing. *ACM Trans. Web (TWEB)* **1**, 5 (2007).
11. Kempe, D., Kleinberg, J. & Tardos, É. Influential nodes in a diffusion model for social networks. In *International Colloquium on Automata, Languages, and Programming* (eds Kempe, D. *et al.*) 1127–1138 (Springer, 2005).
12. Zhou, Y.-B., Lü, L. & Li, M. Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity. *New J. Phys.* **14**, 033033 (2012).
13. Zhan, X.-X., Li, Z., Masuda, N., Holme, P. & Wang, H. Susceptible-infected-spreading-based network embedding in static and temporal networks. *EPJ Data Sci.* **9**, 30 (2020).
14. Wang, J., Xu, S., Mariani, M. S. & Lü, L. The local structure of citation networks uncovers expert-selected milestone papers. *J. Informet.* **15**, 101220 (2021).
15. Zhang, S., Medo, M., Lü, L. & Mariani, M. S. The long-term impact of ranking algorithms in growing networks. *Inf. Sci.* **488**, 257–271 (2019).
16. Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nat. Phys.* **6**, 888–893 (2010).
17. Lü, L. *et al.* Vital nodes identification in complex networks. *Phys. Rep.* **650**, 1–63 (2016).
18. Li, C., Li, Q., Van Mieghem, P., Stanley, H. E. & Wang, H. Correlation between centrality metrics and their application to the opinion model. *Eur. Phys. J. B* **88**, 1–13 (2015).
19. Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C. & Zhou, T. Identifying influential nodes in complex networks. *Physica A* **391**, 1777–1787 (2012).
20. Lawyer, G. Understanding the influence of all nodes in a network. *Sci. Rep.* **5**, 1–9 (2015).
21. Klemm, K., Serrano, M., Eguíluz, V. M. & Miguel, M. S. A measure of individual role in collective dynamics. *Sci. Rep.* **2**, 1–8 (2012).
22. Maharani, W. *et al.* Degree centrality and eigenvector centrality in twitter. In *2014 8th International Conference on Telecommunication Systems Services and Applications (TSSA)* (eds Maharani, W., Gozali, A. A. *et al.*) 1–5 (IEEE, 2014).
23. Liu, J.-G., Lin, J.-H., Guo, Q. & Zhou, T. Locating influential nodes via dynamics-sensitive centrality. *Sci. Rep.* **6**, 1–8 (2016).
24. Li, Z. & Huang, X. Identifying influential spreaders by gravity model considering multi-characteristics of nodes. *Sci. Rep.* **12**, 9879 (2022).
25. Madotto, A. & Liu, J. Super-spreader identification using meta-centrality. *Sci. Rep.* **6**, 38994 (2016).
26. Pei, S., Muchnik, L., Andrade, J. S. Jr., Zheng, Z. & Makse, H. A. Searching for superspreaders of information in real-world social media. *Sci. Rep.* **4**, 1–12 (2014).
27. Bucur, D. Top influencers can be identified universally by combining classical centralities. *Sci. Rep.* **10**, 1–14 (2020).
28. Wang, H., Hernandez, J. M. & Van Mieghem, P. Betweenness centrality in a weighted network. *Phys. Rev. E* **77**, 046105 (2008).
29. Lancichinetti, A., Fortunato, S. & Radicchi, F. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* **78**, 046110 (2008).
30. Bartolucci, S., Caccioli, F., Caravelli, F. & Vivo, P. Ranking influential nodes in networks from aggregate local information. *Phys. Rev. Res.* **5**, 033123 (2023).
31. Page, L., Brin, S., Motwani, R. & Winograd, T. *The Pagerank Citation Ranking: Bringing Order to the Web* (Tech. Rep, Stanford InfoLab, 1999).
32. Lü, L., Zhou, T., Zhang, Q.-M. & Stanley, H. E. The h-index of a network node and its relation to degree and coreness. *Nat. Commun.* **7**, 1–7 (2016).
33. Kiss, I. Z. *et al.* *Mathematics of Epidemics on Networks* Vol. 598 (Springer, 2017).
34. Shu, P., Wang, W., Tang, M. & Do, Y. Numerical identification of epidemic thresholds for susceptible-infected-recovered model on finite-size networks. *Chaos Interdiscipl. J. Nonlinear Sci.* **25**, 063104 (2015).
35. Björck, Å. *et al.* *Numerical Methods in Matrix Computations* Vol. 59 (Springer, 2015).
36. Gleich, D. F. Pagerank beyond the web. *SIAM Rev.* **57**, 321–363 (2015).
37. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. K-core organization of complex networks. *Phys. Rev. Lett.* **96**, 040601 (2006).
38. Kunegis, J. The koblenz network collection. *URL: http://konect.uni-koblenz.de/* Accessed April 16 2023 (2020).
39. Kunegis, J. Konect: the koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, 1343–1350 (2013).
40. Kendall, M. G. The treatment of ties in ranking problems. *Biometrika* **33**, 239–251 (1945).
41. Ghalmane, Z., Cherifi, C., Cherifi, H. & Hassouni, M. E. Centrality in complex networks with overlapping community structure. *Sci. Rep.* **9**, 1–29 (2019).
42. Rajeh, S., Savonnet, M., Leclercq, E. & Cherifi, H. Characterizing the interactions between classical and community-aware centrality measures in complex networks. *Sci. Rep.* **11**, 10088 (2021).
43. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Phys. Rep.* **659**, 1–44 (2016).
44. Saxena, R., Kaur, S. & Bhatnagar, V. Social centrality using network hierarchy and community structure. *Data Min. Knowl. Disc.* **32**, 1421–1443 (2018).
45. Kumar, M., Singh, A. & Cherifi, H. An efficient immunization strategy using overlapping nodes and its neighborhoods. *Companion Proc. Web Conf.* **2018**, 1269–1275 (2018).
46. Costantini, L., Sciarra, C., Ridolfi, L. & Laio, F. Measuring node centrality when local and global measures overlap. *Phys. Rev. E* **105**, 044317 (2022).
47. Šikić, M., Lančić, A., Antulov-Fantulin, N. & Štefančić, H. Epidemic centrality-is there an underestimated epidemic impact of network peripheral nodes?. *Eur. Phys. J. B* **86**, 1–13 (2013).
48. Qu, B., Li, C., Van Mieghem, P. & Wang, H. Ranking of nodal infection probability in susceptible-infected-susceptible epidemic. *Sci. Rep.* **7**, 9233 (2017).

Author contributions

S.Z. conceived the idea, S.Z. and H.W. designed research, S.Z. performed the numerical simulations, all authors analyzed and discussed the results. S.Z. and H.W. wrote the manuscript. All authors reviewd the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-55547-y.

Correspondence and requests for materials should be addressed to H.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.