

Node ranking based on graph curvature and PageRank

Hongbo Qu(曲鸿博), Yu-Rong Song(宋玉蓉), Ruqi Li(李汝琦), Min Li(李敏), and Guo-Ping Jiang(蒋国平)

Citation: Chin. Phys. B, 2025, 34 (**2**): 028901. DOI: 10.1088/1674-1056/ad9a9b
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Hyperbolic map unravels eight regions in temperature volatility regionalization of Mainland China

Yuxuan Song(宋雨轩), Changgui Gu(顾长贵), Muhua Zheng(郑木华), Aixia Feng(冯爱霞), Yufei Xi(席雨菲), Haiying Wang(王海英), and Huijie Yang(杨会杰)
Chin. Phys. B, 2024, 33 (**12**): 128902. DOI: 10.1088/1674-1056/ad84c2

Identifying influential spreaders in complex networks based on density entropy and community structure

Zhan Su(苏湛), Lei Chen(陈磊), Jun Ai(艾均), Yu-Yu Zheng(郑雨语), and Na Bie(别娜)
Chin. Phys. B, 2024, 33 (**5**): 058901. DOI: 10.1088/1674-1056/ad20d6

A multilayer network diffusion-based model for reviewer recommendation

Yiwei Huang(黄羿炜), Shuqi Xu(徐舒琪), Shimin Cai(蔡世民), and Linyuan Lü(吕琳媛)
Chin. Phys. B, 2024, 33 (**3**): 038901. DOI: 10.1088/1674-1056/ad181d

Source localization in signed networks with effective distance

Zhi-Wei Ma(马志伟), Lei Sun(孙蕾), Zhi-Guo Ding(丁智国), Yi-Zhen Huang(黄宜真), and Zhao-Long Hu(胡兆龙)
Chin. Phys. B, 2024, 33 (**2**): 028902. DOI: 10.1088/1674-1056/ad1482

Important edge identification in complex networks based on local and global features

Jia-Hui Song(宋家辉)
Chin. Phys. B, 2023, 32 (**9**): 098901. DOI: 10.1088/1674-1056/aca6d8

Node ranking based on graph curvature and PageRank

Hongbo Qu(曲鸿博)[1], Yu-Rong Song(宋玉蓉)[2,†], Ruqi Li(李汝琦)[1], Min Li(李敏)[2], and Guo-Ping Jiang(蒋国平)[2]

[1] *School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing* 210023, *China*
[2] *College of Automation & College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing* 210023, *China*

(Received 30 September 2024; revised manuscript received 16 November 2024; accepted manuscript online 5 December 2024)

Identifying key nodes in complex networks is crucial for understanding and controlling their dynamics. Traditional centrality measures often fall short in capturing the multifaceted roles of nodes within these networks. The PageRank algorithm, widely recognized for ranking web pages, offers a more nuanced approach by considering the importance of connected nodes. However, existing methods generally overlook the geometric properties of networks, which can provide additional insights into their structure and functionality. In this paper, we propose a novel method named Curv-PageRank (C-PR), which integrates network curvature and PageRank to identify influential nodes in complex networks. By leveraging the geometric insights provided by curvature alongside structural properties, C-PR offers a more comprehensive measure of a node's influence. Our approach is particularly effective in networks with community structures, where it excels at pinpointing bridge nodes critical for maintaining connectivity and facilitating information flow. We validate the effectiveness of C-PR through extensive experiments. The results demonstrate that C-PR outperforms traditional centrality-based and PageRank methods in identifying critical nodes. Our findings offer fresh insights into the structural importance of nodes across diverse network configurations, highlighting the potential of incorporating geometric properties into network analysis.

Keywords: important nodes, graph curvature, complex networks, network geometry

PACS: 89.75.–k, 89.75.Fb, 02.40.–k

DOI: 10.1088/1674-1056/ad9a9b **CSTR:** 32038.14.CPB.ad9a9b

1. Introduction

In many complex systems, whether natural or artificial, certain nodes or entities within a network play a disproportionately important role in the overall dynamics and behavior of the system.[1,2] These important nodes, often referred to as central or influential nodes, are crucial for understanding and controlling the network's functionality and evolution. Such nodes are characterized by their ability to significantly influence the flow of information, signals, or resources across the network. This influence can manifest in various forms, such as facilitating the rapid spread of information in social networks,[3] controlling the transmission of diseases in biological networks,[4] or stabilizing the flow of electricity in power grids.[5] Identifying these key nodes is essential, particularly in scenarios where interventions must be strategically targeted due to constraints such as limited resources or time.

The identification of important nodes often relies on traditional centrality measures, such as degree centrality,[6] betweenness centrality,[7] and closeness centrality.[8] These measures have provided valuable insights into the structure and dynamics of networks, but they may not fully capture the complex interplay of structural and functional roles that nodes can play within a network. This limitation has led researchers to explore more sophisticated methods that incorporate additional dimensions of network analysis.

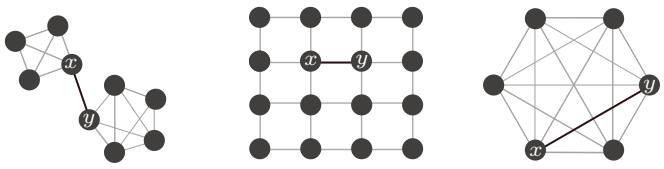
One of the most well-known methods for identifying important nodes is the PageRank algorithm, which was originally developed by Larry Page and Sergey Brin as a way to rank web pages based on their importance.[9] PageRank operates on the principle that a node is important if it is connected to other important nodes. In the context of web pages, this means that a page with many inbound links from other highly-ranked pages is considered more important. The algorithm iteratively calculates the rank of each node based on the ranks of its neighbors, leading to a measure that reflects both the quantity and quality of connections. Over time, various improvements to the original PageRank algorithm have been proposed to enhance its effectiveness and applicability. For instance, LeaderRank[10] is an algorithm designed to improve upon PageRank by addressing some of its limitations. Unlike PageRank, which can sometimes fail to converge or produce biased results due to dangling nodes and dead-ends, LeaderRank introduces a ground node connected to all other nodes, ensuring a more stable and unbiased ranking process.

Despite the widespread success of PageRank and its improved algorithms, these methods often overlook the geometric properties of the network, which can provide additional insights into the underlying structure and relationships between nodes. Network geometry provides a geometric perspective

[†]Corresponding author. E-mail: songyr@njupt.edu.cn

© 2025 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

on network structure and dynamics, offering tools to analyze networks using the principles of geometry and topology.[11,12] Central to the concept of network geometry is the notion of curvature, a measure that quantifies the deviation of a network from a flat, Euclidean space. Curvature can be understood in various ways, with different definitions capturing different aspects of network structure.[13] For instance, the Ollivier–Ricci curvature for a graph, inspired by concepts from Riemannian geometry, assesses the degree to which a network's local structure resembles that of a curved surface.[14–17] As shown in Fig. 1, the figure illustrates network structures characterized by different Ricci curvature values, including negative, zero, and positive curvature.



(a) $\kappa(x,y)<0$ (b) $\kappa(x,y)=0$ (c) $\kappa(x,y)>0$

Fig. 1. Illustration of network structures represented by different Ricci curvature values: (a) negative curvature, (b) zero curvature, (c) positive curvature.

The application of curvature in network analysis has opened new avenues for understanding complex systems. One significant application is in community detection,[17] where curvature-based methods have demonstrated significant potential in uncovering the modular structure of networks. In fact, the curvature of a network can implicitly encode information about its community structure. In networks with a modular structure, where nodes are organized into tightly-knit communities, the importance of a node may depend not only on its direct connections but also on its strategic position within and between these communities. This modular structure is commonly observed in social networks,[18] where groups of individuals with similar interests form distinct communities, and in biological networks, where proteins involved in related functions cluster together.[19] Nodes that reside at the intersections of multiple communities, or those that act as bridges between different regions of the network, are often critical for maintaining overall connectivity and facilitating the flow of information. Understanding the significance of these bridge nodes requires a nuanced approach that accounts for both their geometric properties and their broader network's structure features.

In this paper, we introduce a novel method called Curv-PageRank (C-PR), which combines network curvature with PageRank to identify key nodes in complex networks. Our approach is based on the premise that integrating the geometric insights provided by curvature with the network's structural properties allows for a more effective inference of transfer probabilities between nodes through PageRank. This integration results in a more comprehensive measure of a node's influence. Additionally, we incorporate the degree ratio of neighboring nodes to enhance the accuracy and robustness of the PageRank computation. By unifying these perspectives, our method captures the multifaceted roles that nodes play, particularly in networks with modular structures. C-PR not only identifies nodes with high centrality but also excels at efficiently pinpointing bridge nodes, which are crucial for maintaining structural integrity and facilitating information flow within the network. The overview of Curv-PageRank is shown in Fig. 2.

The main contributions of this paper are summarized as follows:

(i) We propose the Curv-PageRank method, which incorporates network curvature and PageRank, providing a more comprehensive metric for identifying key nodes in networks.

(ii) Our method demonstrates superior performance in identifying critical nodes, particularly in networks with community structures, compared to other centrality-based and PageRank approaches.

(iii) Extensive experiments, including network dismantling and virus spreading on both synthetic and real-world networks, validate the effectiveness of Curv-PageRank, offering fresh insights into the structural importance of nodes across diverse network configurations.

This paper is structured as follows: Section 2 introduces the preliminaries of PageRank and Olliver's Ricci curvature. Section 3 details the Curv-PageRank method. Section 4 presents experimental results on both synthetic and real-world networks. Finally, Section 5 concludes the paper.

2. Preliminaries

This section introduces three aspects: PageRank method, optimal transportation theory, and Olliver's Ricci curvature.

2.1. PageRank

PageRank[9] is an algorithm developed by Larry Page and Sergey Brin, to rank web pages in their search engine. It operates on the principle that the importance of a webpage can be determined by the number and significance of links pointing to it. Mathematically, the PageRank of a page P is defined recursively and depends on the collective PageRank of all pages Q that link to P. The basic formula is given by

$$PR(P)=\frac{1-d}{N}+d\sum_{Q\in M(P)}\frac{PR(Q)}{L(Q)}, \tag{1}$$

where $PR(P)$ is the PageRank of page P, N is the total number of pages, $M(P)$ is the set of pages that link to P, $L(Q)$ is the number of outbound links on page Q, and d is a damping factor that represents the probability of a user continuing to click on links. The damping factor d is usually set to 0.85, which means there is an 85% chance that a user will follow links from the current page.

The PageRank algorithm involves several steps:

Step 1 Initialize PageRanks: Assign an initial PageRank value to each page, typically $1/N$ where N is the total number of pages.

Step 2 Iterative Computation: Update the PageRank values iteratively using the PageRank formula until convergence.

Step 3 Convergence: The algorithm converges when the PageRank values stabilize within a predefined threshold.

PageRank has been widely used not only in web search engines but also in various applications such as determining the most influential individuals based on their connections and interactions in social networks,[20] and it is utilized to rank items or users for better personalization in recommendation systems.[21] Furthermore, in biological networks, it aids in identifying critical proteins or genes based on their interactions.[22]

2.2. Optimal transportation theory

Before introducing Ollivier's Ricci curvature, we need to first introduce the concept of optimal transportation theory, also known as the Monge–Kantorovich problem,[23] which concerns the optimal way to move mass from one distribution to another while minimizing a cost function. This theory has its origins in a problem posed by Gaspard Monge in 1781 and was later rigorously formulated by Leonid Kantorovich in the 1940s.

Given two probability measures μ and ν on a metric space (X,d), the goal is to find a transportation plan π that minimizes the total transportation cost. Mathematically, this can be expressed as

$$\inf_{\pi\in\Pi(\mu,\nu)}\int_{X\times X} d(x,y)\,\mathrm{d}\pi(x,y), \tag{2}$$

where inf denotes the infimum, or the greatest lower bound, of the set of values attained by the integral. The set $\Pi(\mu,\nu)$ represents all possible couplings (transport plans) between the probability measures μ and ν. The goal is to find a coupling π that minimizes the integral of the distance $d(x,y)$ with respect to the measure π, thus achieving the minimum total transportation cost.

In this context, the transportation plan π is a measure on $X\times X$ that describes how mass is moved from the distribution μ to ν. The cost function $d(x,y)$ represents the cost of transporting a unit mass from point x to point y.

The optimal transport problem has numerous applications in various fields, such as economics,[24] machine learning.[25] The solution to this problem provides a way to measure the distance between probability distributions, leading to the Wasserstein distance or earth mover's distance, which is defined as

$$W_p(\mu,\nu)=\left(\inf_{\pi\in\Pi(\mu,\nu)}\int_{X\times X} d(x,y)^p\,\mathrm{d}\pi(x,y)\right)^{1/p} \tag{3}$$

for $p\geq 1$. The case $p=1$ corresponds to the classical Monge–Kantorovich distance.

The Wasserstein distance has important properties, such as being a true metric in the space of probability measures, and it metrizes the weak convergence of measures. This makes it a powerful tool in the analysis of probability distributions and their convergence properties.

2.3. Ollivier's Ricci curvature

Ollivier's Ricci curvature extends the concept of curvature to metric-measure spaces. Unlike traditional Riemannian curvature, which operates on smooth manifolds, Ollivier's approach considers more general spaces. The key idea is to characterize lower bounds on the Ricci curvature using convexity properties of entropy along optimal transport. Specifically, if the entropy remains convex along optimal transport paths, it implies nonnegative Ricci curvature.

In analogy with Riemannian geometry, Ollivier's Ricci curvature controls the overlap between neighborhoods of neighboring points. It quantifies how much the geometry of a space deviates from Euclidean space due to the presence of mass and optimal transport. This notion has been successfully applied to graphs, providing estimates for curvature and diameter.

Ollivier's Ricci curvature provides a way to extend the concept of Ricci curvature from Riemannian manifolds to more general metric spaces, including graphs and discrete spaces. Introduced by Yann Ollivier,[14] it is based on the idea of comparing the transportation of small balls around points in the space.

Formally, given a metric space (X,d) with a probability measure m, the Ollivier–Ricci curvature between two points x and y is defined using the Wasserstein distance W_1 between the probability distributions obtained by taking a small step from x and y respectively. Let m_x^α and m_y^α be the probability measures centered at x and y with parameter α. The Ollivier–Ricci curvature $\kappa(x,y)$ is defined as

$$\kappa(x,y)=1-\frac{W_1(m_x^\alpha,m_y^\alpha)}{d(x,y)}. \tag{4}$$

This curvature measures how the distance between points changes when mass is transported according to the probability measures centered at these points. Positive Ollivier–Ricci curvature implies that the neighborhood distributions are closer together in the Wasserstein sense, indicating a form of "geometric closeness" between the neighborhoods.

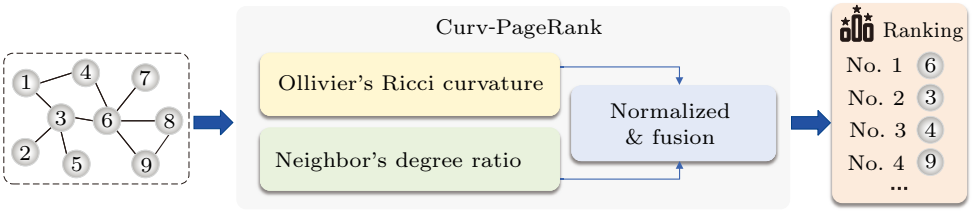


Fig. 2. Overview of Curv-PageRank.

3. The proposed model

3.1. Ollivier's Ricci curvature for graph

For any edge e_{xy} linking nodes x and y on a metric graph $G=(V,E)$, we assign a mass distribution m_x to the neighboring nodes of x, reflecting its local structure. A discrete transport plan is represented by a map $A: V\times V\to[0,1]$, where $A(u,v)$ signifies the mass transferred from node v to node u. This plan must fulfill the conditions $\sum_{v'\in V}A(u,v')=m_x(u)$ and $\sum_{u'\in V}A(u',v)=m_y(v)$. The Wasserstein distance $W(m_x,m_y)$, which measures the minimal total weighted distance required to transport m_x to m_y, is expressed as

$$W(m_x,m_y)=\inf\left\{\sum_{u,v\in V}A(u,v)d(u,v)\right\}. \tag{5}$$

The discrete Ollivier's Ricci curvature for an edge e_{xy} can be formulated as

$$\kappa(x,y)=1-\frac{W(m_x,m_y)}{d(x,y)}, \tag{6}$$

where $d(x,y)$ represents the shortest path distance between nodes x and y. The Wasserstein distance $W(m_x,m_y)$ illustrates the extent of connectivity or overlap between the local structures of nodes x and y. A smaller $W(m_x,m_y)$ indicates a higher overlap in the local structures of x and y, and *vice versa*. In another view, when nodes x and y belong to different communities, their neighbors are less likely to share common nodes. Consequently, the optimal way to transfer m_x from x's neighbors to m_y in y's neighbors is along the edge e_{xy}, resulting in a larger Wasserstein distance and negative Ricci curvature. Conversely, if there is significant overlap, the curvature is positive. Figure 1 provides examples of edges with varying curvatures: positive, zero, and negative.

Next, we need to define the probability distribution m_x for each $x\in V$. Using the definition from Ref. [17], and different with other methods[13] that use probability distribution based on x's neighbors, we introduce a generalized family of probability distributions $m_x^{\alpha,p}$, characterized by parameters $\alpha\in[0,1]$ and $p\geq 0$:

$$m_x^{\alpha,p}(x_i)=\begin{cases}\alpha & \text{if } x_i=x,\\ (1-\alpha)\cdot\dfrac{\exp(-d(x,x_i)^p)}{\sum_{x_i\in\pi(x)}\exp(-d(x,x_i)^p)} & \text{if } x_i\in\pi(x),\\ 0 & \text{otherwise,}\end{cases} \tag{7}$$

where α sets the probability of remaining at node x, while the parameter p determines the depth to which neighbors x_i are discounted based on their distance $d(x,x_i)$. When $p=0$, the distribution is uniform across all neighbors of x, as suggested in Ref. [13]. For larger values of p, distant neighbors are increasingly discounted.

To ensure that $W(m_x,m_y)$ corresponds to the minimal transport cost, we verify that the transport plan A satisfies the marginal conditions $\sum_{v'\in V}A(u,v')=m_x(u)$ and $\sum_{u'\in V}A(u',v)=m_y(v)$. By minimizing $\sum_{u,v\in V}A(u,v)d(u,v)$, the Wasserstein distance optimally captures the distance between the distributions m_x and m_y. Given this minimal transport plan, the Ricci curvature $\kappa(x,y)$ reflects the degree to which local neighborhoods overlap, by normalizing the distance $W(m_x,m_y)$ with the shortest path $d(x,y)$ between x and y.

The parameter p modulates the influence of neighborhood distances, ensuring flexibility in capturing various network topologies and interaction strengths. For completeness, these choices of m_x ensure that $W(m_x,m_y)$ appropriately characterizes the neighborhood connectivity, yielding meaningful curvature values for community detection and network structure analysis.

3.2. Curvature transform module

In the PageRank method, the transition probability between two nodes should be between 0 and 1. However, since the curvature of an edge can range from -1 to 1, it is necessary to design a strategy to transform negative values into positive ones while preserving the original order of all curvatures.

To address this issue, we propose a curvature transformation module that converts all curvatures to positive values while maintaining their original order. For a graph G, let the set of curvatures for all edges be $K=\{\kappa_{0,1},\ldots,\kappa_{i,j}\}$, where $\kappa_{i,j}$ denotes the curvature of edge $e_{ij}\in E$. Since we focus on negative curvatures, which can represent bridge edges, we first invert all elements of the set K, formulating it as $-K$. The transformation process is defined as follows:

$$\kappa'_{i,j}=-\kappa_{i,j}-\min(-K)+\varepsilon, \tag{8}$$

where $\min(-K)$ is the minimum value in $-K$, and $\varepsilon\geq 0$ denotes the bias. Consequently, we obtain $K'=\{\kappa'_{0,1},\ldots,\kappa'_{i,j}\}$.

This transformation is both intuitive and efficient, ensuring that the differences in curvature between edges remain constant.

Next, we need to normalize K' to ensure the values range from 0 to 1. For a node i and its neighboring node j, the normalized curvature $c_{i,j}$ is defined as

$$c_{i,j} = \text{softmax}(\kappa_{i,j}) = \frac{\exp(\kappa_{i,j})}{\sum_{k\in\mathcal{N}(i)}\exp(\kappa_{i,k})}, \tag{9}$$

where $\kappa_{i,j}$ is the curvature of edge e_{ij}, and $\mathcal{N}(i)$ is the set of neighbors of node i. Thus, we have $C = \{c_{0,1},\ldots,c_{i,j}\}$ with all normalized curvatures.

3.3. Neighbor's degree ratio

For a network, the higher the degree of a node, the larger its range of contact with surrounding nodes. If a node needs to disseminate information, transmitting the information to neighbors with relatively high degrees can facilitate the spread of information throughout the network. As shown in Fig. 3, node 0 has four first-order neighbors, which are nodes 1, 2, 3, and 4. It is apparent that node 1 is more important than node 4 when allocating the PageRank value for node 0. This is because node 1 has a higher degree value and can spread more information. Therefore, we design the neighbor's degree ratio to ensure that neighbors with higher degree values are allocated more PageRank value, thereby making the information spread wider. The neighbor's degree ratio between node i and its neighbor node j is defined as

$$R_{(i,j)} = \frac{d_j}{\sum_{k\in N(i)} d_k}, \tag{10}$$

where $N(i)$ denotes the first-order neighbor nodes of node i, and d_j is the degree of node j.

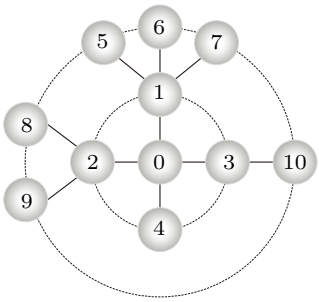


Fig. 3. An example of calculating Neighbor's degree ratio.

As shown in Fig. 3, node 0 has four neighbors, which is $N_0 = \{1,2,3,4\}$, and their neighbors are $N_1 = \{0,5,6,7\}$, $N_2 = \{0,8,9\}$, $N_{(3)} = 0,\{10\}$, $N_4 = \{0,4\}$ respectively. Therefore, for node 0, its four neighbor's degree ratios are

$$\begin{aligned}
R_{(0,1)} &= \frac{d_1}{d_1+d_2+d_3+d_4} \\
&= \frac{|N_1|}{|N_1|+|N_2|+|N_3|+|N_4|} = \frac{4}{10}, \\
R_{(0,2)} &= \frac{d_2}{d_1+d_2+d_3+d_4} \\
&= \frac{|N_2|}{|N_1|+|N_2|+|N_3|+|N_4|} = \frac{3}{10}, \\
R_{(0,3)} &= \frac{d_3}{d_1+d_2+d_3+d_4} \\
&= \frac{|N_3|}{|N_1|+|N_2|+|N_3|+|N_4|} = \frac{2}{10}, \\
R_{(0,4)} &= \frac{d_4}{d_1+d_2+d_3+d_4} \\
&= \frac{|N_4|}{|N_1|+|N_2|+|N_3|+|N_4|} = \frac{1}{10}.
\end{aligned} \tag{11}$$

3.4. Fusion

Now, we have two methods to measure the connection between two nodes: graph curvature and the neighbor's degree ratio. Next, we design a linear accumulation (LA) strategy to fuse these two methods to better capture the network's structural properties.

Linear accumulation (LA): For an edge e_{ij} in G, which has graph curvature c_{ij} and neighbor's degree ratio r_{ij}, the linear accumulation is defined as

$$p'_{ij} = \alpha\cdot c_{ij} + \beta\cdot r_{ij}, \tag{12}$$

where α and β control the different ratios of graph curvature and neighbor's degree ratio, respectively, to form the composite transition probability. We impose the constraint $\alpha+\beta = 1$. The discussion on how the values of α and β affect the results will be introduced in Subsection 4.1.3. In the majority of the experiments discussed later, we set $\alpha = 0.5$ and $\beta = 0.5$.

Finally, we need to perform the softmax function again to normalize these values

$$p_{ij} = \text{softmax}(p'_{ij}) = \frac{\exp(p'_{ij})}{\sum_{k\in\mathcal{N}(i)}\exp(p'_{ik})}, \tag{13}$$

where p_{ij} is the transition probability for nodes i and j.

4. Experiment and discussion

In this section, we evaluate our model's effectiveness on both synthetic and real-world networks. We begin by outlining the experimental settings, including the baselines and a description of the datasets. Following that, we introduce the metrics used, such as the SIR infection experiment and network dismantling experiment. We then analyze the performance of our model under various values of α and β as specified in Eq. (12). Finally, we validate the proposed method through experiments conducted on both artificial and real-world networks, and provide a thorough analysis of the results.

4.1. Experimental settings

4.1.1. Baselines

We evaluate Curv-PageRank by comparing it with several established methods aimed at determining node importance. The methods used for comparisons include:

(i) Degree centrality (DC):[6] This approach quantifies the importance of a node by counting its edges, assigning higher significance to nodes with a greater number of connections.

(ii) Closeness centrality (CC):[8] This metric assesses a node's proximity to all other nodes in the network, prioritizing nodes that exhibit shorter average distances to others.

(iii) Betweenness centrality (BC):[7] This measure evaluates how often a node lies on the shortest paths between all pairs of nodes in the network, highlighting nodes that exert more control over information flow.

(iv) K-shell (KS):[26] The metric identifies core nodes in a network by progressively removing nodes with lower degrees. Nodes situated in higher k-shells are deemed more central and influential in maintaining network connectivity.

(v) FINDER:[27] This metric designed to identify influential nodes within a network through a combination of local and global topological features. It leverages machine learning techniques to analyze patterns in network structure, enabling the detection of critical nodes in networks.

(vi)PageRank (PR):[9] This metric measures the importance of a node based on the idea of random walks. A node is considered important if it is linked to by other important nodes, with each link contributing a share of the linking node's importance.

4.1.2. Datasets description

To assess the efficacy and distinguishability of the Curv-PageRank model, we employed both synthetic and real-world networks. For the synthetic networks, we utilized three types of generated networks:

Type 1 Barabási–Albert (BA) networks:[28]

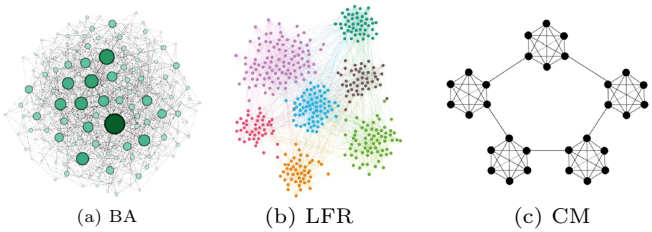
As shown in Fig. 4(a), this model generates scale-free networks through a preferential attachment mechanism. Nodes with higher degrees are more likely to attract new connections, resulting in a power-law degree distribution that is frequently observed in real-world networks.

Type 2 Lancichinetti–Fortunato–Radicchi (LFR) networks:[29]

As shown in Fig. 4(b), this model creates synthetic networks with known community structures and realistic features akin to those found in real-world networks. It produces graphs with heterogeneous distributions of node degrees and community sizes, making it ideal for studying complex network dynamics.

Type 3 Caveman (CM) networks:[30]

As shown in Fig. 4(c), this network model social structures with tightly-knit communities (cliques) that are sparsely interconnected. They provide a clear example of community structure, useful for evaluating algorithms related to community detection and network resilience.



(a) BA (b) LFR (c) CM

Fig. 4. Illustration of different types of networks: (a) Barabási–Albert networks (BA), (b) Lancichinetti–Fortunato–Radicchi networks (LFR), (c) Caveman networks (CM).

For real-world networks, we selected several datasets from different domains, encompassing various scales and types, to ensure the robustness and generalizability of our model:

(i) Karate:[31] This network represents social interactions among members of a karate club, including players, a coach, and a manager. Nodes represent individuals, while edges represent social ties between them. The network can be divided into two communities, one centered around the coach and the other around the manager.

(ii) Dolphins:[32] This network is constructed from the observation of 62 bottlenose dolphins. Nodes represent individual dolphins, and edges represent social relationships between them, such as frequent associations or interactions.

(iii) Footballs:[32] This network represents the games played between American College football teams. Each node represents a team, and an edge between two nodes indicates that at least one game has been played between those two teams.

(iv) Email:[33] This network captures the email communication within a research institution. Nodes represent anonymized members of the institution, and edges indicate that at least one email has been exchanged between two members.

(v) Cora:[34] This network constitutes a citation network of scientific papers. In this network, each node signifies a publication, and an edge between any two nodes denotes a citation relationship where one paper references another.

(vi) PGP:[35] This social network consists of nodes that represent individuals utilizing the pretty good privacy (PGP) encryption algorithm to share confidential information. Edges between nodes indicate that two individuals have exchanged encrypted communications.

The statistical characteristics of the aforementioned real-world networks are detailed in Table 1. In this table, $|N|$ represents the number of nodes, $|E|$ represents the number of edges, $\langle k \rangle$ stands for average degree, D indicates the diameter, $\rho(10^{-3})$ represents network density, c means the clustering coefficient, and Q indicates the modularity.

Modularity serves as a quality metric to gauge the robustness of the network's division into communities or clusters.[36] In other words, it assesses how closely the structure of a network resembles that of a community-based system. It quantifies the density of links within communities in comparison to links between different communities. A higher modularity value suggests a strong community structure, indicating that nodes within the same community are more densely interconnected than nodes belonging to different communities. Mathematically, modularity Q is defined as

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j), \tag{14}$$

where A_{ij} represents the adjacency matrix of the network, k_i and k_j stand for the degrees of nodes i and j, respectively. The variable m denotes the quantity of edges in the network. The terms c_i and c_j indicate the communities to which nodes i and j are assigned. The Kronecker delta function, δ, equals 1 if c_i is the same as c_j, and 0 otherwise.

Considering the community structure is not known a priori in many networks, we first use label propagation algorithms[37] to partition the network into communities and then the modularity score can be calculated based on the resulting partitions. This approach helps in determining whether the network exhibits community structure.

Table 1. Statistical properties of real-world networks.

Network	$\|N\|$	$\|E\|$	$\langle k\rangle$	D	ρ (10^{-3})	c	Q
Karate	34	78	4.59	5	139.03	0.57	0.32
Dolphins	62	159	5.13	8	84.08	0.26	0.49
Footballs	115	616	10.66	4	93.51	0.40	0.58
E-mail	1005	25571	31.97	7	31.84	0.40	0
Cora	2708	5429	3.90	19	1.44	0.24	0.65
PGP	10680	24316	4.55	24	0.43	0.27	0.72

4.1.3. Metrics

First, we use two metrics to verify the effectiveness of the ranking results: the SIR infection experiment and the network dismantling experiment.

(i) **SIR infection experiment:** The susceptible–infected–recovered (SIR) model is an extension of the basic SI model, used to simulate the spread of infectious diseases within a network where recovery is possible. In this experiment, nodes are initially classified as susceptible (S), infected (I), or recovered (R). A susceptible node may contract the infection through contact with an infected node. Subsequently, an infected node can transition to the recovered state after a certain period, at which point it ceases to contribute to the spread of the infection, akin to the progression of illnesses such as influenza or measles. The behavior of the SIR model can be mathematically captured through the following set of differential equations:

$$\begin{cases} \frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI, \\ \frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - \gamma I, \\ \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I, \end{cases} \tag{15}$$

where S, I, and R represent the quantities of susceptible, infected, and recovered nodes, respectively. The infection probability per contact is given by the parameter β, and the recovery rate is indicated by γ. The expressions $\mathrm{d}S/\mathrm{d}t$, $\mathrm{d}I/\mathrm{d}t$, and $\mathrm{d}R/\mathrm{d}t$ capture the temporal dynamics of these populations.

In the SIR experiment, we first select the top n nodes from the ranking lists generated by different methods as the initial infected nodes. Then, we conduct the SIR infection experiment, where the infection probability is chosen as $\beta = 1.5\langle k\rangle/\langle k^2\rangle$[38] and the recovery probability is set to $\gamma = 0.1$. Considering the inherent randomness in the SIR infection process, each network is simulated $t = 100$ times, and the average result is taken as the final outcome.

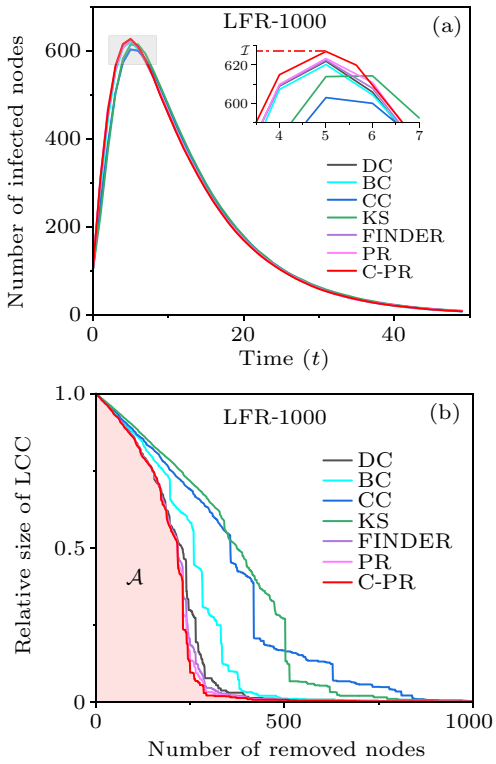


Fig. 5. (a) SIR infection experiment on LFR network with 1000 nodes. (b) Network dismantling experiment on LFR network with 1000 nodes.

As shown in Fig. 5(a), the SIR infection experiment is conducted on the LFR network with 1000 nodes. The curve represents the change in the number of infected nodes over time. The top 100 nodes from the ranking lists generated by different methods are selected as the initial infected nodes. At the same time t, a higher number of infected nodes indicates a better ranking result. We define $\mathcal{I}$ as the peak number of infected nodes in all nodes. It can be observed that our method (red curve) achieves the highest $\mathcal{I}$, enabling faster and wider

infection across the network, demonstrating the effectiveness of the proposed approach.

The SIR experiment evaluates the effectiveness of node ranking by observing how quickly and extensively an infection spreads through the network and how many nodes eventually recover when starting from highly ranked nodes. This helps to assess the influence, centrality, and resilience of nodes within the network.

(ii) **Network dismantling experiment:** The problem of network dismantling (ND) can be mathematically formulated as follows: Consider a network $G = (V, E)$ comprising N nodes and M edges. The task is to determine a subset S of nodes such that the removal of S from G yields a residual network $G' = (V \setminus S, E')$ where no connected component exceeds a predefined size T. The objective is to minimize the size of S. Formally, this can be written as

$$S = \underset{S \subseteq V}{\operatorname{argmin}} |S| \quad \text{subject to} \quad \max_{C \in C(G')} |C| \leq T, \tag{16}$$

where $C \in C(G')$ denotes the set of connected components in G', and $\max |C|$ indicates the number of nodes in the largest connected component (LCC) of G'. The size of the LCC reflects the degree of connectivity and robustness of the network, indicating how well the network functions as a whole.

In this experiment, nodes are removed one at a time according to the ranking lists generated by various methods, and after each removal, we calculate the relative size $\Theta = \max |C|/N$ of the LCC, where N is the total number of nodes in the network. As illustrated in Fig. 5(b), the ND experiment is conducted on an LFR network with 1000 nodes under different ranking methods. The curve represents the relative size of the LCC as nodes are progressively removed based on the rankings. A better ranking method should lead to a rapid reduction in the relative size of the LCC with minimal node removals. Therefore, the area under the curve (AUC), denoted as $\mathcal{A}$ and shown as the red-shaded region in Fig. 5(b), is used as an evaluation metric. Our method demonstrates faster and more efficient network dismantling, achieving the smallest $\mathcal{A}$ with fewer node removals.

This experiment involves progressively removing nodes from the network based on their ranking and observing the resultant fragmentation. The goal is to determine how the removal of top-ranked nodes impacts the network's connectivity and robustness. A more effective ranking method will lead to a quicker disintegration of the network, indicating that the most critical nodes are being accurately identified and prioritized for removal.

Notably, we did not use the Kendall coefficient to demonstrate the effectiveness of the ranking results. Instead, we relied on practical experiments, such as the SIR infection experiment and the network dismantling experiment, to evaluate the performance of our ranking method. The SIR infection experiment simulates the spread of an infection through the network, starting from highly ranked nodes, to observe the speed and extent of the infection spread. This helps to measure the influence and centrality of the nodes. On the other hand, the network dismantling experiment involves systematically removing nodes based on their ranking and assessing the resulting network fragmentation. This approach allows us to determine how quickly the network disintegrates, thereby validating the criticality and accuracy of the identified top-ranked nodes.

These practical experiments provide a more tangible and intuitive assessment of the ranking results, focusing on real-world implications and network behaviors, rather than relying solely on statistical measures. This approach ensures that the ranking method is robust and effective in identifying the most influential nodes within a network.

4.2. The impact of the parameters α and β

In this part, we explore the impact of the parameters α and β in Eq. (12) on the experimental results. As shown in Fig. 6(a), the infection results of the SIR model on an LFR network with 500 nodes are displayed for various values of α and β. The x-axis represents the values of α/β, while the y-axis shows the peak infection value for each experiment, i.e., the proportion of infected nodes at the peak of the infection $\mathcal{I}/N$. The larger this value, the better the effectiveness of the node ranking, indicating that the initially selected infected nodes can spread the infection more rapidly throughout the network. From Fig. 6(a), we observe a clear trend indicating that as the proportion of graph curvature, denoted as α, increases, and the proportion of another factor, β, decreases, the overall performance of our model exhibits significant improvement. This observation suggests that the incorporation of graph curvature plays a crucial role in enhancing the model's effectiveness. Specifically, when initiating the infection process with the same number of initial nodes, a model that accounts for a higher proportion of graph curvature is capable of infecting a greater number of nodes within the network.

As shown in Fig. 6(b), the network dismantling results on an LFR network with 500 nodes are displayed under various values of α and β. The x-axis represents the values of α/β, while the y-axis shows the AUC $\mathcal{A}$ for each experiment under its corresponding α/β. As discussed previously, a smaller value of $\mathcal{A}$ indicates a better node ranking. We observe that as the proportion of graph curvature in the model increases, the $\mathcal{A}$ value decreases. At the point where curvature is fully considered, i.e., 1.0/0.0, the model achieves optimal performance. This is because LFR networks are community-based, and accurately identifying bridge nodes within these communities is

crucial. By considering curvature, the model can better capture the heterogeneous structure of the network and accurately identify bridge nodes.

In summary, our model, after considered with graph curvature, is better able to capture the structural information of networks, particularly in community-based networks, thereby identifying their important nodes more accurately. However, most networks may not possess strong community attributes. Therefore, to ensure the model's adaptability to various types of networks, we set $\alpha = 0.5$ and $\beta = 0.5$.

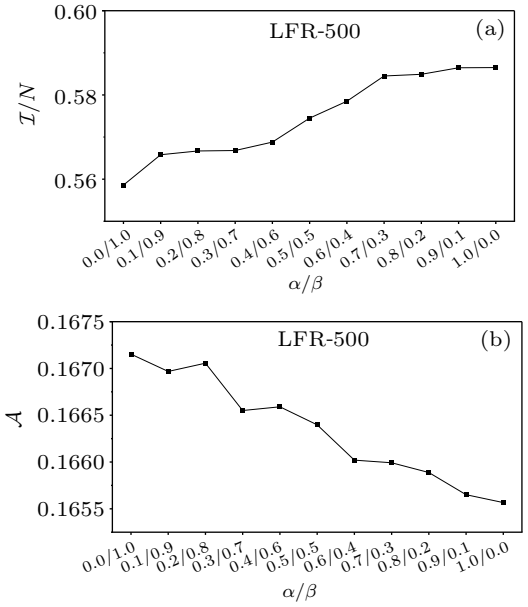


Fig. 6. (a) The SIR experiment on LFR-500 under different α and β, (b) the ND experiment on LFR-500 under different α and β.

4.3. Experiments on synthetic network

In this part, we perform SIR and network dismantling experiments on three types of synthetic networks: Barabási–Albert (BA) networks, LFR networks, and Caveman (CM) networks. For each network type, we generate networks with sizes of 500, 1000, and 2000 nodes.

For the BA networks, we set the parameter $m = 4$, which defines the number of edges a new node will link to existing nodes. In the case of the LFR networks, we configure the average degree to $a = 5$. Additionally, we specify the minimum community size as $m = N/10$ and the maximum community size as $m = N/5$, where N denotes the total number of nodes in the network. For the CM networks, each clique is composed of $l = 20$ nodes, and the entire network is structured by arranging N/l such cliques.

4.3.1. SIR experiments

Table 2 presents the results of the peak proportion $\mathcal{I}/N$ of infected nodes in the SIR experiment, which refers to the highest proportion of infected nodes during the infection process relative to the total number of nodes. A higher value indicates that, given the same initial number of infected nodes, more nodes can be infected, implying a more effective node ranking. The best result in each table is **bolded**, while the second-best result is underlined.

As we can see, our method, C-PR, consistently achieves the highest peak proportion of infected nodes across most network types and sizes, except for the BA-1000 networks. The second-best performing centrality measures vary depending on the network type and size. In BA networks, the second-best measures are FINDER and PR. The BA networks exhibit a distinctive scale-free property without a community structure, which results in C-PR's slight underperformance in the BA-1000 network. In contrast, FINDER and PR effectively capture this feature. This is because FINDER is trained using BA networks as part of its training set, enabling it to perform better on networks with scale-free properties. Consequently, PR ranks first in BA-1000 networks, while FINDER and PR rank second in the other two network types.

In LFR networks, C-PR consistently achieves the best results, while BC and PR come in second. LFR networks are characterized by heterogeneity and community structure, so finding both bridge nodes and highly influential nodes requires a robust algorithm. C-PR excels at this task. The BC algorithm is also effective at identifying bridges and bridge nodes, thus ranking second in the LFR-500 network. In contrast, FINDER performs poorly in networks with community structures due to its lack of feature learning related to community structures during the training process.

Table 2. The peak proportion $\mathcal{I}/N$ of infected nodes in the SIR experiment.

Type	Nodes	DC	BC	CC	KS	FINDER	PR	C-PR
BA	500	0.5645	0.565	0.5651	0.565	0.5667	0.5655	**0.5782**
	1000	0.5096	0.5101	0.5138	0.5034	0.515	**0.5158**	0.5151
	2000	0.5055	0.5024	0.5062	0.505	0.5077	0.5019	**0.5089**
LFR	500	0.5690	0.5737	0.5322	0.5448	0.5736	0.5719	**0.5745**
	1000	0.6219	0.6201	0.6029	0.6142	0.6239	0.6243	**0.6270**
	2000	0.4439	0.4407	0.3905	0.4154	0.4447	0.4505	**0.4520**
CM	500	0.7140	0.7367	0.7356	0.1521	0.7228	0.7124	**0.7459**
	1000	0.7138	0.7436	0.7449	0.1615	0.7439	0.7217	**0.7477**
	2000	0.6921	0.7275	0.7264	0.1588	0.7261	0.6950	**0.7299**

In CM networks, C-PR again achieves the best results across all cases. Unlike LFR networks, CM networks are homogeneous, with nodes within each clique being fully connected and having the same degree. C-PR effectively identifies key nodes within these homogeneous structures, thus infecting more nodes. BC and CC rank second in this network type, as they are also capable of accurately identifying bridge edges. It is worth noting that the KS method completely fails in this network. This is because most nodes have the same degree, making it difficult for the method to identify the key nodes, leading to its failure.

4.3.2. Network dismantling experiments

Table 3 presents the area under the curve (AUC) results for the network dismantling experiment across various types and sizes of networks. The AUC metric evaluates the effectiveness of different measures in terms of network dismantling, where lower AUC values indicate more effective node rankings. In each row, the lowest value is **bolded**, and the second-lowest value is underlined. Unlike the SIR experiment, the ND experiment focuses on the overall accuracy of the ranking list rather than just observing the top-n scenario.

In BA networks, C-PR achieves the lowest AUC scores for networks with 500 and 2000 nodes, with values of 0.1971 and 0.1884, respectively. However, in the BA-1000 network, PR slightly outperforms C-PR with an AUC of 0.1911, while C-PR closely follows with 0.1913. The scale-free nature and lack of community structure in BA networks may explain C-PR's marginal underperformance in the BA-1000 case. For LFR networks, C-PR consistently delivers the best performance across all network sizes, with PR coming in second. The heterogeneity and community structure in LFR networks require robust algorithms to identify bridge nodes and influential nodes, a task at which C-PR excels. FINDER, struggles in networks with pronounced community structures, such as LFR networks, due to its limited ability to learn community-related features during the training process. This limitation results in suboptimal performance compared to C-PR and PR in such environments.

In CM networks, C-PR outperforms all other methods across all cases, although the numerical differences between the methods are not substantial. This is due to the homogeneous nature of CM networks, which are characterized by fully connected cliques and uniform node degrees. After the removal of key nodes, the size of the LCC rapidly decreases. Therefore, quickly identifying bridge nodes is crucial for effectively dismantling the network. Our method, C-PR, is particularly effective at identifying these key nodes. Notably, similar to the results from the SIR experiments, the KS method fails in these networks due to the uniformity in node degrees, making it difficult to distinguish key nodes.

Table 3. Area under the curve (AUC) of the network dismantling experiment.

Type	Nodes	DC	BC	CC	KS	FINDER	PR	C-PR
BA	500	0.2069	0.2072	0.2923	0.2776	0.1998	0.2081	**0.1971**
	1000	0.1982	0.2168	0.3632	0.3026	0.1913	**0.1911**	0.1913
	2000	0.1948	0.2123	0.3459	0.2913	0.1889	0.1898	**0.1884**
LFR	500	0.1856	0.2168	0.3401	0.3417	0.1692	0.1681	**0.1663**
	1000	0.2074	0.2443	0.3496	0.3546	0.1927	0.1929	**0.1919**
	2000	0.1686	0.2325	0.3656	0.3367	0.1663	0.1661	**0.1658**
CM	500	0.1027	0.101	0.1083	0.4829	0.1029	0.101	**0.0996**
	1000	0.0574	0.0564	0.0651	0.4814	0.0571	0.0565	**0.0563**
	2000	0.0397	0.0394	0.0462	0.4807	0.0409	0.0393	**0.039**

4.4. Experiments on real-world network

In this part, we conduct SIR experiments and network dismantling experiments on real-world networks to demonstrate the effectiveness of our method. To make the results easier to observe, we visualize both the infection process and the dismantling process, with different methods represented by different colored curves. The red curve represents our method, C-PR. Additionally, for the SIR experiments, we include a zoomed-in view in the upper right corner of each subplot to facilitate better comparison.

4.4.1. SIR experiments

As shown in Fig. 7, the curves represent the number of infected nodes over time during the SIR infection process across different networks. The x-axis represents the infection time t, while the y-axis shows the number of infected nodes. A higher peak in the curve, along with a smaller corresponding value of t, indicates that the ranking method has more accurately identified the importance of nodes in the network.

From Fig. 7, it can be observed that our method achieves superior performance in most networks. For instance, in the E-mail network, despite its modularity value of $Q = 0$, which indicates the challenge label propagation faces in identifying a distinct community structure, C-PR effectively identifies critical nodes and achieves a higher infection rate. FINDER closely follows in performance, benefiting from its training process that leverages scale-free properties to enhance its ef-

fectiveness on such networks. In the PGP network, the BC method demonstrates the best performance, likely due to the network's distinct community structure with numerous sub-communities. The BC method excels at identifying key nodes within these communities, making it particularly effective in this scenario. Although our method performs well, it slightly trails behind BC, highlighting potential areas for refinement to further boost its effectiveness.

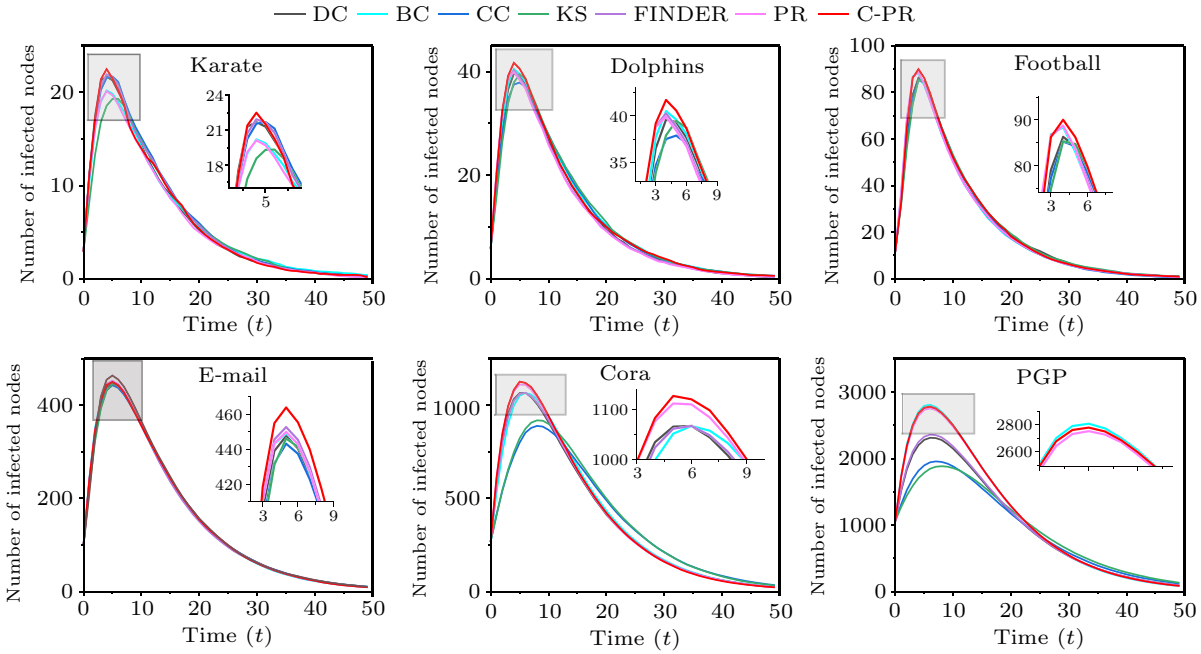


Fig. 7. The SIR experiments on six real-world datasets.

4.4.2. Network dismantling experiments

As illustrated in Fig. 8, the curves depict the variation in the relative size of the largest connected component (LCC) as a function of the number of nodes removed. To effectively showcase the experimental results, we choose not to remove all the nodes from each network, but rather a fraction of them. A steeper decline in the curve indicates a faster disintegration of the network, which in turn reflects a more accurate ranking of node importance by the corresponding method.

Upon examining the results, it is clear that our method, C-PR, generally surpasses other approaches in terms of efficiently dismantling network structures while minimizing the number of nodes removed. In the Football network, both the BC method and FINDER initially demonstrate strong performance in disrupting the network's connectivity.

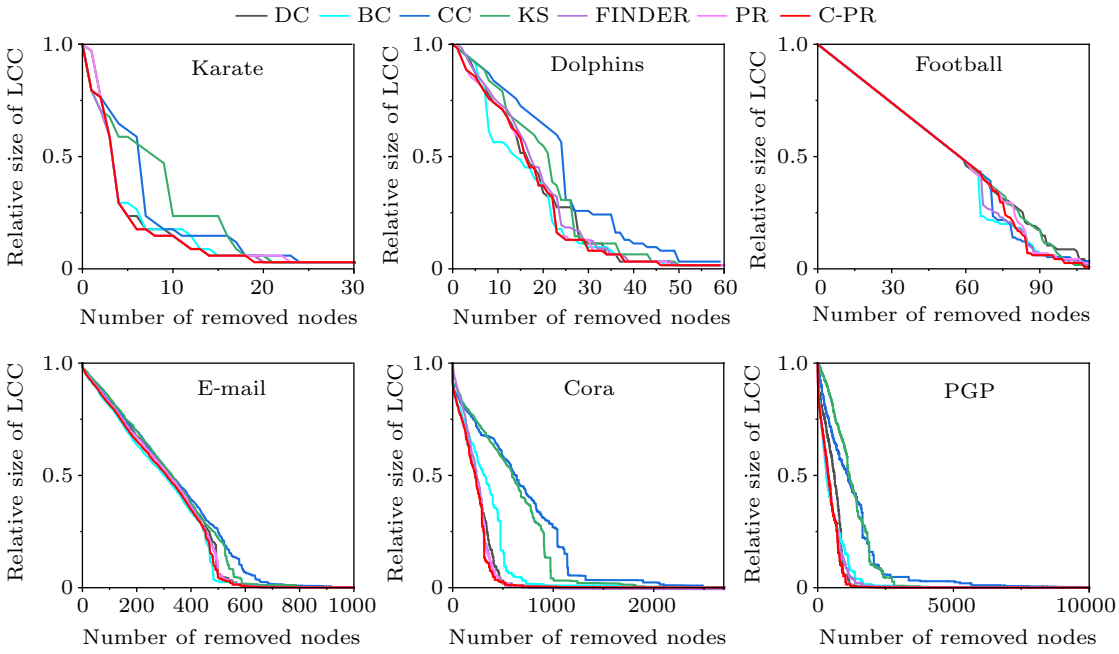


Fig. 8. The ND experiments on six real-world datasets.

Similarly, in the Dolphins and E-mail networks, the BC method shows notable initial effectiveness in breaking down the network structure. These methods quickly identify and remove nodes that are crucial to maintaining network cohesion, causing an immediate reduction in connectivity.

However, as the process of node removal progresses into the latter stages, C-PR proves to be superior. It achieves a more dispersed largest connected component (LCC) with fewer nodes removed compared to the other methods. This indicates that C-PR is not only effective in its initial impact but also maintains its efficacy throughout the dismantling process. The method's ability to strategically identify and target the most critical nodes ensures a sustained disruption of the network, ultimately leading to a more effective breakdown.

5. Conclusion

In this study, we introduce Curv-PageRank, a novel method for identifying influential nodes in complex networks by incorporating both geometric and structural properties. Unlike traditional centrality measures, Curv-PageRank leverages the concept of network curvature alongside the degree ratio of neighboring nodes, offering a more nuanced and effective approach to assessing node importance, especially in networks with community structures.

Our experimental results on both synthetic and real-world networks demonstrate that Curv-PageRank outperforms existing centrality-based and PageRank-based methods in identifying critical nodes for network robustness and information flow. In particular, Curv-PageRank excels in identifying bridge nodes that are crucial for maintaining network connectivity, which is often overlooked by standard methods.

The findings of this research open new avenues for incorporating geometric properties in network analysis, providing deeper insights into the complex interplay between network topology and functionality. Future work could explore extending this approach to dynamic networks or further refining the curvature-based model for specific applications such as epidemic modeling or social influence analysis.

By bridging the gap between network geometry and centrality measures, Curv-PageRank offers a valuable tool for researchers and practitioners seeking to better understand complex networks.

Data availability statement

All networks can be downloaded from Refs. [33,39].

Acknowledgements

Project partially supported by the National Natural Science Foundation of China (Grant Nos. 61672298 and 62373197), the Major Project of Philosophy and Social Science Research in Colleges and Universities in Jiangsu Province, China (Grant No. 2018SJZDI142), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province, China (Grant No. KYCX23_1045).

References

[1] Artime O, Grassia M, DeDomenico M, Gleeson J P, Makse H A, Mangioni G, Perc M and Radicchi F 2024 *Nat. Rev. Phys.* **6** 114
[2] Liu X M, Li D Q, Ma M Q, Szymanski B K, Stanley H E and Gao J X 2022 *Phys. Rep.* **971** 1
[3] Watts D J and Strogatz S H 1998 *Nature* **393** 440
[4] Newman M E 2002 *Phys. Rev. E* **66** 016128
[5] Albert R, Albert I and Nakarado G L 2004 *Phys. Rev. E* **69** 025103
[6] Bonacich P 1972 *J. Math. Sociol.* **2** 113
[7] Brandes U 2001 *J. Math. Sociol.* **25** 163
[8] Freeman L C, *et al.* 2002 *Soc. Netw.* **1** 215
[9] Page L, Brin S, Motwani R and Winograd T *Tech. Rep.* **1999** 66
[10] Lu L Y, Zhang Y C, Yeung C H and Zhou T 2011 *PLoS One* **6** e21202
[11] Boguna M, Bonamassa I, DeDomenico M, Havlin S, Krioukov D and Serrano M Á 2021 *Nat. Rev. Phys.* **3** 114
[12] Mulder D and Bianconi G 2018 *J. Stat. Phys.* **173** 783
[13] Lin Y, Lu L Y and Yau S T 2011 *Tohoku Math. J.* **63** 605
[14] Ollivier Y 2009 *J. Funct. Anal.* **256** 810
[15] DoCarmo M P and FlahertyFrancis J 1992 *Riemannian geometry* (Boston: Birkhäuser) p. 10
[16] Ollivier Y 2010 *Adv. Stud. Pure Math.* **57** 343
[17] Wandelt S, Shi X, Sun X Q and Zanin M 2020 *IEEE Access* **8** 111954
[18] Bedi P and Sharma C 2016 *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **6** 115
[19] Barrio-Hernandez I, Schwartzentruber J, Shrivastava A, Del-Toro N, Gonzalez A, Zhang Q, Mountjoy E, Suveges D, Ochoa D, Ghoussaini M, Bradley G, Hermjakob H, Orchard S, Dunham I, Anderson C, Porras P and Beltrao P 2023 *Nat. Genet.* **55** 389
[20] Heidemann J, Klier M and Probst F 2010 *Proceedings of the International Conference on Information Systems*, December 12–15, 2010, St. Louis, Missouri, USA, p. 79
[21] Nguyen P, Tomeo P, DiNoia T and DiSciascio E 2015 *Proceedingsof the 24th International Conference on World Wide Web*, May 18–22, 2015, Florence, Italy, p. 1477
[22] Morrison J L, Breitling R, Higham D J and Gilbert D R 2005 *BMC Bioinformatics* **6** 233
[23] Monge G 1781 *Mem. Math. Phys. Acad. Royale Sci.* **1** 666
[24] Galichon A 2018 *Optimal transport methods in economics* (Princeton: Princeton University Press) p.96
[25] Peyré G and Cuturi M 2019 *Found. Trends Mach. Learn.* **11** 355
[26] Carmi S, Havlin S, Kirkpatrick S, Shavitt Y and Shir E 2007 *Proc. Natl. Acad. Sci. USA* **104** 11150
[27] Fan C J, Zeng L, Sun Y Z and Liu Y Y 2020 *Nat. Mach. Intell.* **2** 317
[28] Barabási A L and Albert R 1999 *Science* **286** 509
[29] Lancichinetti A, Fortunato S and Radicchi F 2008 *Phys. Rev. E* **78** 046110
[30] Watts D J 1999 *Am. J. Sociol.* **105** 493
[31] Zachary W W 1977 *J. Anthropol. Res.* **33** 452
[32] Lusseau D and Newman M E 2004 *Proc. R. Soc. Lond. B Biol. Sci.* **271** S477
[33] Leskovec J and Sosič R 2016 *ACM Trans. Intell. Syst. Technol.* **8** 1
[34] Yang Z L, Cohen W and Salakhudinov R 2016 *Proceedings of the 33rd International Conference on Machine Learning*, June 19–24, New York, USA, p. 40
[35] Boguná M, Pastor-Satorras R, Díaz-Guilera A and Arenas A 2004 *Phys. Rev. E* **70** 056122
[36] Clauset A, Newman M E and Moore C 2004 *Phys. Rev. E* **70** 066111
[37] Cordasco G and Gargano L 2010 *Proceedings of the IEEE International Workshop on Business Applications of Social Network Analysis*, December 15, 2010, Bangalore, India, p. 1
[38] Pastor-Satorras R and Vespignani A 2001 *Phys. Rev. Lett.* **86** 3200
[39] Rossi R A and Ahmed N K 2015 *Proceedings of the AAAI Conference on Artificial Intelligence*, January 25–30, 2015, Austin, Texas USA, p. 4292