Hindawi
Complexity
Volume 2020, Article ID 8715619, 14 pages
https://doi.org/10.1155/2020/8715619

Research Article

Novel Node Centrality-Based Efficient Empirical Robustness Assessment for Directed Network

Xiaolong Deng,[1] **Hao Ding**,[1] **Yong Chen**,[2] **Cai Chen**,[3] **and Tiejun Lv**[1]

[1]*Key Lab of Trustworthy Distributed Computing and Service of Education Ministry, Beijing University of Post and Telecommunication, Beijing 100876, China*
[2]*North Automatic Control Technology Institute, Taiyuan, China*
[3]*China Academy of Information and Communications Technology (CAICT), Beijing 100037, China*

Correspondence should be addressed to Xiaolong Deng; shannondeng@bupt.edu.cn

Received 5 August 2020; Revised 11 October 2020; Accepted 28 October 2020; Published 21 November 2020

Academic Editor: Jia Wu

In recent years, while extensive researches on various networks properties have been proposed and accomplished, little has been proposed and done on network robustness and node vulnerability assessment under cascades in directed large-scale online community networks. In essential, an online directed social network is a group-centered and information spread-dominated online platform which is very different from the traditional undirected social network. Some further research studies have indicated that the online social network has high robustness to random removals of nodes but fails to the intentional attacks, particularly to those attacks based on node betweenness or node directed coefficient. To explore on the robustness of directed social network, in this article, we have proposed two novel node centralities of *ITG* (information transfer gain-based probability clustering coefficient) and $IM_p(v)$ (directed path-based node importance centrality). These two new centrality models are designed to capture this cascading effect in directed online social networks. Furthermore, we also propose a new and highly efficient computing method based on iterations for $IM_p(v)$. Then, with the abundant experiments on the synthetic signed network and real-life networks derived from directed online social media and directed human mobile phone calling network, it has been proved that our *ITG* and $IM_p(v)$ based on directed social network robustness and node vulnerability assessment method is more accurate, efficient, and faster than several traditional centrality methods such as degree and betweenness. And we also have proposed the solid reasoning and proof process of iteration times k in computation of $IM_p(v)$. To the best knowledge of us, our research has drawn some new light on the leading edge of robustness on the directed social network.

1. Introduction

With rapid increasing online social network, the network structure of online social networks has become more complicated than before. Analysis and explaining the dynamics and properties of social networks has become an interesting researching task with plenty of applications in social sciences and many other web application scripts. In some social networks, it is very common for some users who decide to leave the network or begin to stop being active in the activities of their community [1]. This phenomenon is also called as quitting or churn and has absorbed much research attention in social networks. And how to analyze and evaluate network robustness and resilience [2–4] after node departure or removal [5] has always been the hot research points [2, 6] in the last decade. And online social network has been classified as a scale-free network for demonstrating the power-law [7] distribution of degree by many famous complex network scientists [4, 8].

Some recent research results have also told us that network nodes which have a large betweenness [8] value are closely related to swift information and material dissemination in a graph [9, 10] which is useful for quick network robustness and node vulnerability assessment. Because the traditional network robustness and node vulnerability assessment theories are based on undirected and nonweighted networks, it is necessary to research on the relationship of resilience of directed social network after node departure

and the latest found rules of complex network which can be found in Figure 1 for Toy example. And it would be helpful to find some new discipline and cast new light on robustness and vulnerability assessment of directed social network.

In detail, our contributions are as follows:

(1) Basing on classic probability graph theory and clustering coefficient definition, we have proposed two new node centralities named *ITG* (information transfer gain) and $IM_p(v)$ (directed node importance). It can be used to measure the robustness and vulnerability in directed networks especially directed social networks which have been seldom tested before.

(2) We have proposed enough experiment results on undirected artificial networks and directed online social networks to make robustness assessment comparison of *ITG* and $IM_p(v)$, which was mentioned scarcely in former-related works.

(3) Our *ITG*- and $IM_p(v)$-based centrality has been proved to be more accurate, efficient, and faster than classical centrality methods such as degree and betweenness with sufficient experiments results for node robustness assessment in directed social networks.

(4) We have firstly proposed a new rigorous proving process of directed node importance centrality $IM_p(v)$ and its implementation method. We have found that, in the more densely connected directed social network, the $IM_p(v)$ node removal strategy is the most harmful to the network connecting structure. And to the best knowledge of us, we attain the varying trend of iteration times k to the marginal difference ε on directed social networks for the first time based on strict mathematical proof.

The outline of the paper is as follows: Section 2 introduces the related work on robustness and resilience of scale-free network which includes some traditional complex network and social network. Section 3 presents the definitions of network structure quantities which we used to evaluate the robustness and vulnerability of network datasets. And our novel *ITG*- and $IM_p(v)$-based directed node centrality will be introduced. Section 4 proposes the experiments results in synthetic signed network and real-life large networks derived from directed online social media and directed human mobile phone calling network with former undirected node centrality measures and our directed *ITG* centrality measure and $IM_p(v)$. Section 5 gives the final conclusion of this article and draws some new light on the future work.

2. Related Work on Network Robustness

There have been many important research studies on vulnerability assessment and robustness assessment in network structures [11, 12] after node departure or network attack and other significant research areas.

The robustness of ER random network and BA scale-free network was analyzed firstly by Albert [13] in 2000. He used the multiple correlation relationship data of l^{-1} (inverse geodesic length), S (size of the largest connected subgraph), and removed node ration under node attack to make vulnerability and robustness assessment. He discovered that collapse of scale-free network may reach a high price because of the selection and removal of a few nodes which play an important role in maintaining the connectivity of network.

Also, in 2000, Cohen and Callaway [10, 14] observed that real networks demonstrating power-law degree distribution are robust against random node removal but easy to crash in case of attacks to high-degree nodes.

Holme et al. [8] used four different removal strategies which will be introduced in Section 4 in 2002, and he found details of the response of networks according to these attacks on vertices and edges. Holme and his research team observed that the removals by the recalculated degrees and betweenness centralities are always more harmful than the attack strategies based on degrees and betweenness centralities of the initial network. But they only use one real communication network and did not propose some application points.

Then, some researchers have discovered a few valuable results on directed networks. Xu and Wang [6] have done some experiments on the cascading crash on weighted complex networks in 2008. Newman and Ghoshal [2] found that removal of some special single node in the network may cause the bicomponents in the graph to be disconnected.

Malliaros and Vazirgiannis [4] proposed a model to capture this cascading effect node departure of social network, based on engagement dynamics of social networks in 2015. Fragkiskos introduced a new concept of robustness assessment method under cascades triggered by the quitting of nodes based on their engagement level. His results indicated that social networks are very robust and strong under cascades triggered by randomly selected nodes but highly vulnerable in cascades caused by targeted departures of nodes with high engagement level.

In 2016, Gao et al. [10] developed a set of analytical tools with which to identify the natural control and state parameters of a multidimensional complex system for stimulation, and it can attain effective one-dimensional dynamics that can accurately predict the system's resilience. Their proposed analytical framework tool can systematically separate the roles of the system's dynamics and topology, collapsing the behavior of different networks onto a single universal resilience function.

But in the related research mentioned above, these researchers did not model on the information transfer process of directed social networks and cannot fulfill the network robustness and vulnerability assessment requirement of directed social networks nowadays, and it needs to construct some new model and methods to bridge the requirement gap in some efficient ways.

3. Definitions of Network Structure Quantities and Node Centrality Measures

In a scale-free network which includes undirected and directed ones, after some very important nodes are intentionally selected and removed, the network would suffer a serious collapse. It is necessary to research on the network

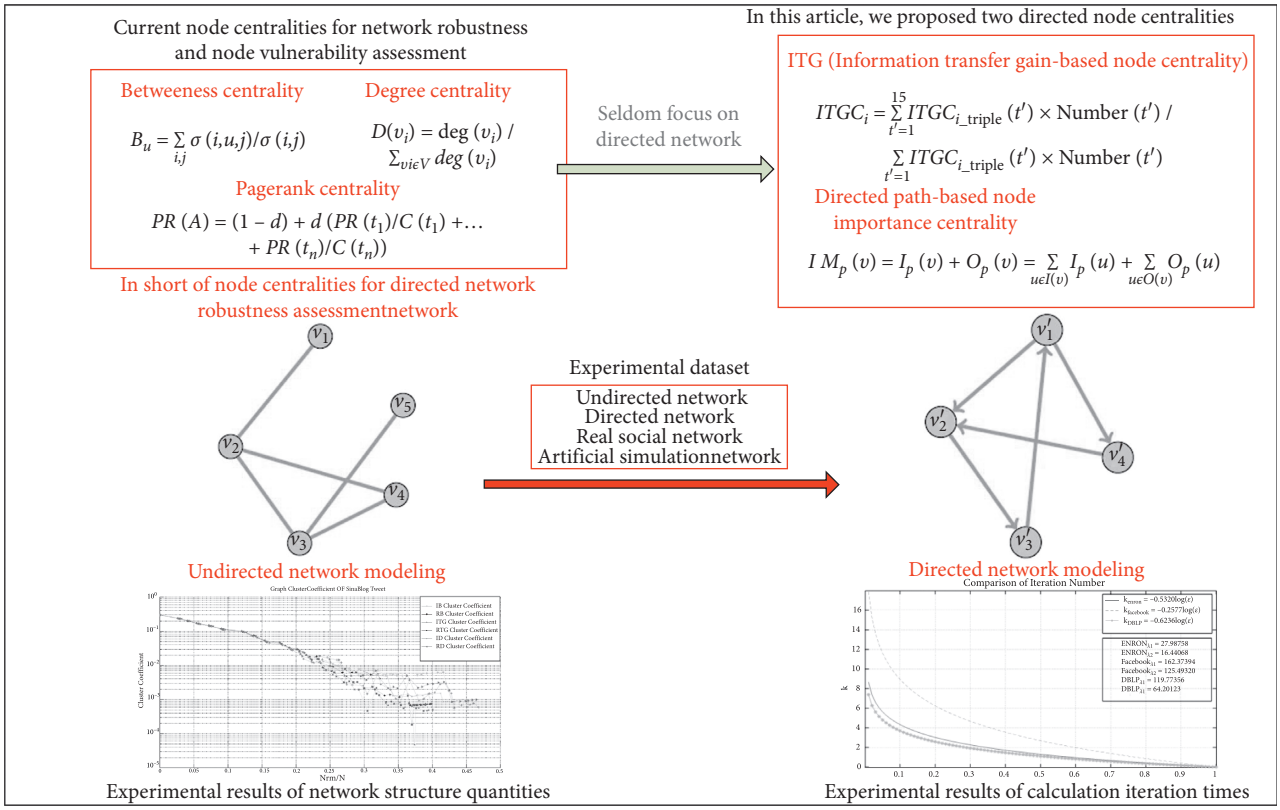

Figure 1: Toy example figure.

resilience in directed networks, so we adopted three key performance quantitative indicators and some famous centralities indicators to make robust assessment of simulated and real networks.

3.1. Network Structure Quantitative Indicator Definition. In this paper, we used two types of network which are undirected and directed networks. On the one hand, for the undirected networks, the network model is unweighted and undirected, which can be demonstrated as $G = (V, E)$. V is the set of nodes with number $N = |V|$, and E is the set of edges with number $L = |E|$. On the other hand, the directed network model is always defined as $\overrightarrow{G} = (V, E)$ with each edge having its own weight and direction to supply the directed vivid information and material spreading.

3.1.1. Average Inverse Geodesic Length. In an undirected network, average inverse geodesic length l is the important network structure quantities after node failure. And in the directed network especially in social networks, the average inverse geodesic length l can be calculated by removal of the direction of edges:

$$l \equiv \langle \mathrm{d}(v, w) \rangle \equiv \sum_{v \in V} \sum_{w \neq v \in V} \mathrm{d}(v, w). \tag{1}$$

In formula (1), $\mathrm{d}(v, w)$ stands for the geodesic path length between different nodes v and w. In traditional social networks of human relationship which is a small-world network, l is always around 6. In the new social networks such as Twitter and Facebook, it would decline to 4 which has been proved by Robert and Sebastiano in Laboratory for web algorithm [15] in 2015. After some nodes are removed, if there is no path between nodes v and w, $\mathrm{d}(v, w)$ would reach $+\infty$. And there is another length quantity l^{-1} instead of l:

$$l^{-1} \equiv \langle \mathrm{d}(v, w) \rangle \equiv \frac{1}{N(N-1)} \sum_{v \in V} \sum_{w \neq v \in V} \frac{1}{\mathrm{d}(v, w)}. \tag{2}$$

The value of $1/\mathrm{d}(v, w)$ is zero where there does not exist path from node v to node w.

3.1.2. Network Average Cluster Coefficient. In most network models, the node cluster coefficient C_i reflects the density of connection around some focus nodes. In the whole network, the network average cluster coefficient $C_{G=(V,E)}$ demonstrates the macroscopically cluster characteristics of the whole network to reveal the density of links among neighbor nodes and the cluster coefficient of node can be found in the following formula:

$$C_i = \frac{2E_i}{k_i(k_i - 1)}, \tag{3}$$

where k_i stands for stands for the neighbor node number of node i and E_i stands for the actual number of existing links among neighbor nodes of node i. Network average cluster coefficient can be found in the following formula:

$$C_{G=(V,E)} \equiv \frac{1}{N} \sum_{v_i \in V, i=1}^{N} \frac{2E_i}{k_i(k_i - 1)}. \quad (4)$$

3.1.3. Relative Size of Largest Connected Subgraph. In some disconnected large networks, there would exist some separated subgraphs and cannot connect to each other. Size of the largest connected subgraph is also called the node number of the largest connected subgraph in the whole graph which is important to reveal inner connectivity characteristics of graph. In the real network, it can assume that if the connected subgraphs set are $g_1, g_2, g_3, g_4, \ldots\ldots, g_n$ in $G = (V, E)$, then definition of the relative size of the largest connected subgraph can be found in the following formula:

$$S = \frac{\text{Max}\left\{\left|V_{g1}\right|, \left|V_{g2}\right|, \left|V_{g3}\right|, \ldots\ldots, \left|V_{gn}\right|\right\}}{|V|}. \quad (5)$$

3.2. Node Centrality Measures. In network, there are many node centrality measures to calculate the importance of node and each node has its own status and influence to its neighbors. Particularly, in the social relations network, it can demonstrate the relations among nodes. Based on node centrality measures, they can be used to analyze the closeness in the each other dependent relations between different nodes. Furthermore, node centrality measures can be used to calculate the role and graph position of every node for information dissemination influence analysis and other advancing applications [16–21]. Among these researches, node centrality is always the focus of analysis and research results of social network researchers [22] to find the role and status of node. Among these well-known node centralities, we have chosen three most common node centrality statistics and proposed two new node centrality measures which can be found as follows.

3.2.1. Betweenness Centrality. Betweenness is a very famous and vital node centrality measure in network statistics computing. Its detailed definition is as follows: in graph G, if there exists a path between every node v_i and v_j while $\forall v_i, v_j \in V, G$ is a connected graph and there must be a shortest graph to connect v_i and v_j which is always called the geodesic path. Thus, we can use the assuming shortest paths number between two vertices to quantify the importance of a vertex or an edge in terms of its betweenness centrality [23] which can be found in the following formula:

$$B_u = \sum_{i,j} \frac{\sigma(i, u, j)}{\sigma(i, j)}. \quad (6)$$

In formula (6), $\sigma(i, u, j)$ is the number of shortest paths between vertices v_i and v_j that pass through vertex u and $\sigma(i, j)$ is the total number of shortest paths between nodes v_i and v_j.

3.2.2. PageRank Centrality. PageRank is a typical link analysis centrality measure invented by Larry Page in 1997 with the purpose of "calculating" its relative importance within the linking data set. And PageRank also can be applied to calculate node centrality. The simple definition of PageRank can be found in formula (7). In our calculation, damping factor d equals 0.85 as a constant value. The calculation formula of PageRank is as follows.

When page number increases and the damping factor is used to recalculate the PageRank mark of every page, damping factor stands for the real mark when one page is linked to the other page with the range from zero to 1. In our calculation, the damping factor value equals 0.85 as a constant value. The calculation formula of PageRank is as follows:

$$PR(A) = (1 - d) + d\left(\frac{PR(t_1)}{C(t_1)} + \cdots + \frac{PR(t_n)}{C(t_n)}\right). \quad (7)$$

In formula (7), $PR(t_1)$ stands for PR (PageRank) marks in which t_1 brings to page A (t_1 is linked to A). In the same way, $C(t_1)$ stands for pages which are linked to page t_1.

3.2.3. Degree Centrality. In social network, the degree centrality is that central users have the most ties to other actors in the network. If the central user is more powerful and influential in the connecting network, it will have much more links to other users including in and out links. The number of adjacent edges a node has is defined as a degree. In graph G, if the degree of node v_i is $\deg(v_i)$ and the total degree of nodes in the graph is $\sum_{v_i \in V} \deg(v_i)$, the degree centrality of node v_i is $D(v_i)$ which defined in as follows [24]:

$$D(v_i) = \frac{\deg(v_i)}{\sum_{v_i \in V} \deg(v_i)}. \quad (8)$$

From the definition above in formula (8), it can be found that node degree centrality expressed the ratio of degree of the node u to the total degree in the whole graph and some nodes with a higher degree centrality reflect that this node may have some powerful role linking to other node in the network and it may be the most vital person living in the focus of attention.

3.2.4. Directed Path-Based Node Importance Centrality. With the rapid development of information technology, our life is becoming more and more networked. The graph information is closely related to our daily life, such as the directed WeChat relation network and directed Facebook relation network with such complex structures. And most important of all is that the scale of these graphs is enormous and it is hard to handle these so complex and massive graphs. In a directed graph network, how to quantitatively analyze and calculate the importance of each vertex has become an important problem to be solved urgently for those directed graph networks, while they all have some common features [25, 26]. Among these statistical indicators for directed graph networks, directed path-based directed

node importance centrality has become more prominent for its easy and fast calculation.

In the directed graph network G, if there exists a path $e_i = (v_i, v_j)$ between nodes v_i and v_j, directed path-based directed node importance centrality of node v_j can be expressed by the paths across node v_j. The more the across paths are, the more important the node v_j is. Directed path-based node importance centrality $IM_p(v)$ of node v has been posed in the famous book with the name "Computer Science Theory for the Information Age" by John Hopcroft and Kannan [27], but this book did not demonstrate some fast and efficient computing methods of $IM_p(v)$ which we have found in formula (9). And definition of directed path-based node importance centrality $IM_p(v)$ of node v can be found in formula as follows:

$$IM_p(v) = I_p(v) + O_p(v) = \sum_{u \in I(v)} I_p(u) + \sum_{u \in O(v)} O_p(u). \tag{9}$$

In formula (9), $I_p(v)$ stands for the directed paths number which finally ends in node v where $I(v) = \{v | (u, v) \in E\}$ and $O_p(v)$ stands for the directed paths number which starts from node v where $O(v) = v | (u, v) \in E$. Basing on formula (9), we can calculate the $IM_p(v)$ of nodes in directed networks when we solve two critical problems first. One problem is that, in complex structured graphs with huge number nodes, the time complexity of the iterative calculation will be very high and the other problem is that the convergence of the calculation process of formula (9) cannot be reached if we do not use some approximate calculation methods.

For $I_p(v) = \sum_{u \in I(v)} I_p(u)$ in formula (9), it can be converted to iterative computation in the following formula:

$$I_p(v)_k = \sum_{u \in I(v)} I_p(u)_{k-1}. \tag{10}$$

In formula (10), $I_p(v)_k$ stands for the k_{th} iteration calculation result, and for any node v, the initial value of $I_p(v)_k$ is $I_p(v)_0 = 1$. For the reason that computation of formula (10) cannot converge by its diverging computation process, the iterative results of formula (10) need to be standardized in formula (11a) in which α is the damping factor with the value range of $0 < \alpha < 1$ just like the damping factor in the PageRank computation model:

$$I_p(v)_k = (1 - \alpha) + \alpha \cdot \frac{\sum_{u \in I(v)} I_p(u)_{k-1}}{\sqrt{\sum_{u \in I(v)} \left(I_p(u)_{k-1}\right)^2}}. \tag{11a}$$

For the similar computation principle, the standardized form of $O_p(v)$ in formula (9) can be written as follows:

$$O_p(v)_k = (1 - \alpha) + \alpha \cdot \frac{\sum_{u \in I(v)} O_p(u)_{k-1}}{\sqrt{\sum_{u \in I(v)} \left(O_p(u)_{k-1}\right)^2}}. \tag{11b}$$

In computation of formulas (11a) and (11b), we can use M to stand for the adjacency matrix of directed graph network G and I_p to stand for the vector composed by all the $I_p(v)$ values of the node set $v \in V$ in G while $|v| = n$. And I_p can be expressed by $I_p = [I_p(v_1), I_p(v_2), \ldots, I_p(v_n)]^{\mathrm{T}}$, while formula (11a) can be changed as follows:

$$I_p = (1 - \alpha) \cdot I + \alpha \cdot \frac{M \cdot I_p}{\left\| M \cdot I_p \right\|}. \tag{12a}$$

In formula (12a), $I = [1, 1, \ldots, 1]^{\mathrm{T}}$ and I_p can be initialized to $I_p = I$. For the similar computation principle, formula (11b) can be changed as follows:

$$O_p = (1 - \alpha) \cdot I + \alpha \cdot \frac{M \cdot O_p}{\left\| M \cdot O_p \right\|}. \tag{12b}$$

By using formulas (12a) and (12b), the computation of directed path-based node importance centrality $IM_p(v)$ of node v can be transferred to the computation of adjacency matrix M of directed graph network G with great promotion. Furthermore, we will demonstrate the whole proof process which using the Power Iteration Theory [28] to prove that formulas (12a) and (12b) can converge to a definite vector by finite times of computations. For the adjacency matrix $M = (m_{ij}) \in R^{n*n}$ of directed graph network G, it has the eigenvalue set of $\lambda_1, \ldots, \lambda_n (\lambda_i \in R)$, while $\lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$ and $\lambda_1 = 1$. At the same time, $Mx_i = \lambda_i x_i$ and eigenvectors set $\longrightarrow_{x_i} (i = 1, \ldots, n)$ can satisfy that $\longrightarrow_{x_i} \cdot \longrightarrow_{x_j} = 0 \ (i \neq j)$.

We can assure that $y^{(0)} = \sum_{i=0}^{n} c_i x_i$ and $y^{(k)}$ can be expressed as $y^{(k)} = \sum_{i=0}^{n} c_i \lambda_i^k x_i$.

Here, $[y^{(0)}, y^{(1)}, \ldots, y^{(k)}]$ can be used to replace $I_p = [I_p(v_1), I_p(v_2), \ldots, I_p(v_n)]^{\mathrm{T}}$ or $O_p = [O_p(v_1), O_p(v_2), \ldots, O_p(v_n)]$ for formulas (11a) and (11b).

For $y^{(i)} \in \{y^{(0)}, y^{(1)}, \ldots, y^{(k)}\}$, $0 \leq i < k$,

$$\begin{aligned} \left\| y^{(k+1)} - y^{(k)} \right\| &= \left\| \sum_{i=1}^{n} c_i \lambda_i^{k+1} x_i - \sum_{i=1}^{n} c_i \lambda_i^k x_i \right\| = \left\| \sum_{i=2}^{n} c_i \lambda_i^k (\lambda_i - 1) x_i \right\| \\ &\leq \sum_{i=2}^{n} \left\| c_i \lambda_i^k (\lambda_i - 1) x_i \right\| \leq (n-1) |c|_{\max} |\lambda_2 - 1| |\lambda_2|^k \|x_i\|_{\max} \leq a |\lambda_2|^k, \end{aligned} \tag{13a}$$

where $a > 0, a = (n-1)|c|_{\max}|\lambda_2 - 1|\|x_i\|_{\max}$.

We can also prove that the iteration times k ($k = [1, N]$ of computation for $\|y^{k+1} - y^k\|$) are around a constant level value in the following formula:

$$\forall\left\|y^{(k+1)} - y^{(k)}\right\| < \varepsilon \Longrightarrow a|\lambda_2|^k < \varepsilon (0 < \varepsilon < 1) \Longrightarrow k\log|\lambda_2| < \log\varepsilon - \log a \Longrightarrow k > \frac{\log\varepsilon - \log a}{\log|\lambda_2|}. \tag{13b}$$

And the constant level value of k is

$$k > \eta\log\varepsilon + \beta\left(\eta < 0, \eta = \frac{1}{\log|\lambda_2|}, \beta = -\frac{\log a}{\log|\lambda_2|}\right). \tag{13c}$$

3.2.5. ITG-Based Directed Node Centrality. In most actual social networks such as Twitter, Facebook, and WeChat, the two friends of someone can be friends with each other. And this phenomenon can be found in both undirected networks and directed networks. So this attribute of social networks sometimes is called the clustering characteristic of network [7]. Network average clustering coefficient reflects the microscopically clustering characteristic of network and has become the very important measure of adjacent nodes which are connected closely. The definition of network node clustering coefficient can be found in Section 3.1.2. Based on the classic probabilistic graphical model (PGM) theory from Turing Award Owner Pearl [29], we have made a detailed research on directed node influence clustering coefficient [30] to propose a new vector influence clustering coefficient model with both information propagation direction and information propagation probability. The new vector influence clustering coefficient model starts from the two basic directed triple forms of vertex i in the directed graph which can be found in Figure 2.

Basing on deducting from Figures 2(a) and 2(b), we can get the related edge information propagation probability and direction of different two types including 36 subgraphs of all different triples [30]. And then we can calculate all numerical values of directed *ITG* node centrality in directed social networks in all kinds of situations by the following formula:

$$\sum_{i\longleftrightarrow j} ITG = \underset{i\longleftrightarrow j}{ITG} + \underset{i\longleftrightarrow k\longleftrightarrow j}{ITG} = \underset{i\longleftrightarrow j}{ITG} + \underset{i\longleftrightarrow k}{ITG} + \underset{k\longleftrightarrow j}{ITG}, \tag{14}$$

$$\sum_{i\longrightarrow k} ITG = \underset{i\longrightarrow k}{ITG} + \underset{i\longleftrightarrow j}{ITG} + \underset{j\longrightarrow k}{ITG}. \tag{15}$$

Because there are three possible directional statuses for each edge in Figures 2(a) and 2(b), the adjacent edge of node i has three different definitions, which are friends relationship ($i \longleftrightarrow j$), following relationship ($i \longrightarrow j$), and fan relationship ($i \longleftarrow j$). At the same time, the opposite edge $i \longleftrightarrow k$ of node i also has three definitions. Furthermore, the edge $j \longleftrightarrow k$ has three types of relationships, in which node j and node k are friends, node j follows node k, and node k follows node j. We can use 0, 1, and 2 to stand for the relationships and substitute the three different definitions, and we obtain the following 27 arrangements in Table 1 [30].

In the above 27 arrangement cases, because node i is the source node, we can find some symmetry results, and finally, we can get 15 independent results. And the *ITG*-based directed node centrality value from information transfer gain clustering coefficient (*ITGC*) of node i in a directed network can be finally summed by the 15 different independent results in formula (16) as follows:

$$ITGC_i = \frac{\sum_{t=1}^{15} ITG_{i\_\text{triangle}}(t) \times \text{Number}(t)}{\sum_{t'=1}^{6} ITG_{i\_\text{triple}}(t') \times \text{Number}(t')}. \tag{16}$$

$ITGC_i$ is the *ITG* value of node i in a directed network. $\sum_{t=1}^{15} ITG_{i\_\text{triangle}}(t) \times \text{Number}(t)$ is the weighted number of triangles which use node i as the top vertex (i.e., the information transfer source node), and its weight is the *ITG* (information transfer gain) contribution $ITG_{i\_\text{triangle}}(t)$ from the 15 different types of weighted triangles multiplied by its counted number $\text{Number}(t)$. $\sum_{t'=1}^{6} ITG_{i\_\text{triple}}(t') \times \text{Number}(t')$ is the weighted number of the triples using node i as the top vertex; its weight is the weighted sum of the six $ITG_{i\_\text{triple}}(t')$ values of different types of triples multiplied by its counted number $\text{Number}(t')$. Similar to undirected clustering coefficients, *ITGC* has the same characteristic, measuring the tightness of the graph to form tight communities.

4. Experiments on Network Robustness Assessment

4.1. Simulation of Node Failure Generation Method. Basing on the importance and connectivity of different nodes in the network, there are eight node attack strategies [8] always chosen by researchers to evaluate network robustness such as follows:

(1) The ID removal strategy: the attack starting from the node with the highest degree and node attack strategy uses the initial node degree distribution.

(2) The IB removal strategy: the attack starting from the node with the highest betweenness and node attack strategy uses the initial node betweenness distribution.

(3) The *ITG* removal strategy: the attack starting from the node with the highest *ITG* centrality and node attack strategy uses the initial node *ITG* centrality distribution.

(4) The *IMP* removal strategy: the attack starting from the node with the highest $IM_p(v)$ centrality and

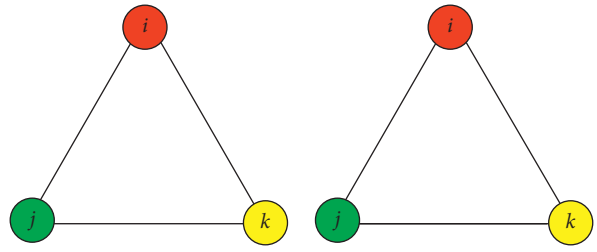

Figure 2: Two triple forms of vertex i in undirected graph.

node attack strategy uses the initial node $IM_p(v)$ centrality distribution.

(5) The RD removal strategy: using the recalculated node degree distribution at every removal step.

(6) The RB removal strategy: using the recalculated node betweenness at every step.

(7) The RTG removal strategy: using the recalculated node ITG centrality distribution at every removal step.

(8) The RIMP removal strategy: using the recalculated node $IM_p(v)$ centrality distribution at every removal step.

4.2. Experiment Result. Selection of typical and persuasive dataset is very important to experiment results, and we selected some classical undirected network and directed network dataset including synthetic signed network and real-life large networks used by Albert et al. [13] and Holme et al. [8]. And these typical datasets can be used to validate the network robustness and node vulnerability assessment in undirected and directed networks.

From Figures 3 to 21, all the prominence in experiment result which was created by *ITG*, *RTG*, *IMP*, and *RIMP* strategies have be pointed out by red ellipses and text tag. Based on these marks, we can easily find the more strong effects by our proposed *ITG* and $IM_p(v)$ centralities.

4.2.1. Undirected Dataset and Experiment Results. In Table 2, the network dataset includes the classical BA scale-free network (a) proposed by Albert et al. [13] (generation parameter is $m_0 = 5$, $m = 4$, $p_t = 0.8$, and $n = 490$) and the undirected call community [31] graph (b) from cellphone calling records in one month in China of a southern city, the LFR (Lancichinetti Fortunato Radicchi) [32] benchmark network with generation parameter of $N = 1000$, $k_{\text{degree}} = 2$, $C_{\min} = 20$, $C_{\max} = 100$, $u = 0.3$, and $C_{\text{degree}} = 1$. For the reason that the LFR benchmark has presented a much solid testing dataset for algorithms and having good performance from other dataset, we used LFR benchmark to generate testing dataset having typical attributes compared with real networks, such as real node degree distribution and heterogeneous distribution of community size. Because the above networks are important undirected networks, the three network datasets are chosen for our research.

Table 1: Permutation and combination of 27 conditions.

000	001	002	010	011	012	020	021	022
100	101	102	110	111	112	120	121	122
200	201	202	210	211	212	220	221	222

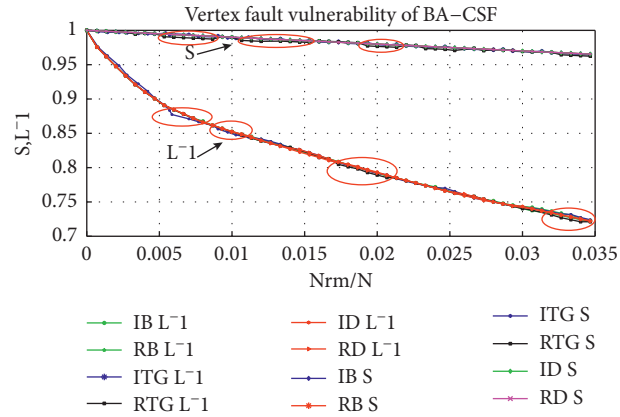

Figure 3: l^{-1}&S result of BACSF.

Now we will give the brief introduction of vertical and horizontal coordinates of the graph, in which the x-axis in Figures 3 to 8 stands for the removed node ration (Nrm: the number of removed nodes) to initial graph. The y-axis stands for the relative value of S and l^{-1} (the ration of S and l^{-1} after every step of node removal to initial S and l^{-1} in the network). Figures 7 and 11 consist of 400 calculated results each in 200 networks generated by the four simulated node generation mechanisms above to avoid random deviation (every mechanism generated 50 fault nodes).

In calculation of *ITG*-based node coefficient in undirected networks, all the information transfer gain probability a between nodes i and j for undirected edges will be equal to 0.5 and we would calculate *ITG* node centrality coefficient of each node and the whole graph.

4.2.2. Experiment Results Analysis for Undirected Dataset. After processing and analyzing the experimental data above, it can be clearly found in Figures 3, 5, and 7 that the influence to l^{-1} is as follows:

(1) When the number of removed nodes is very small that is to say when $0.0 < N_{rm}/N < 0.005$, it is the very

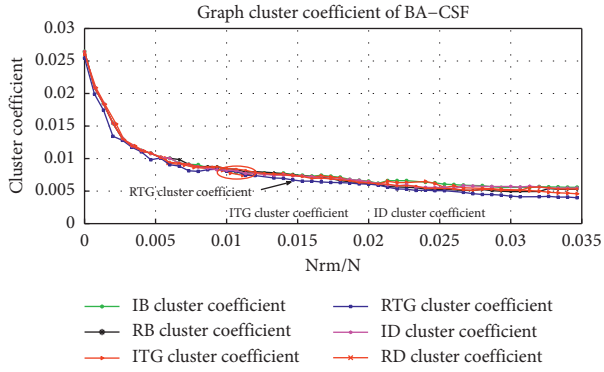

Figure 4: $C_{G=(V,E)}$ result of BACSF.

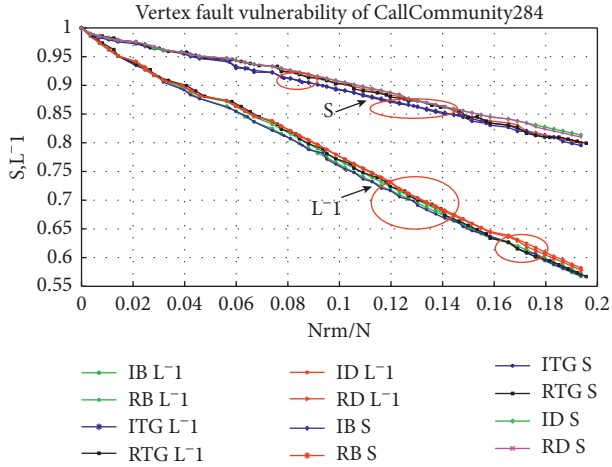

Figure 5: l^{-1}&S result of CallCommunity.

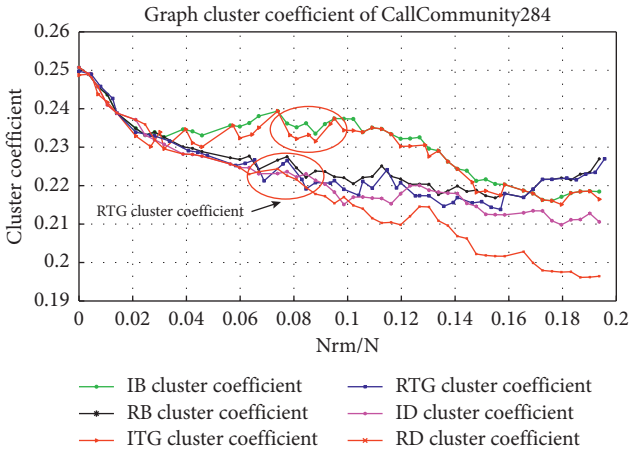

Figure 6: $C_{G=(V,E)}$ result of CallCommunity.

early stage of collapse time by node removal of the six different removal strategies and the l^{-1} curve falls quickly to make the whole graph shrink with a high speed.

(2) During the fall process of l^{-1}, the *RTG* curve, *RB* curve, and *RD* curve look like falling faster than other three removal strategies.

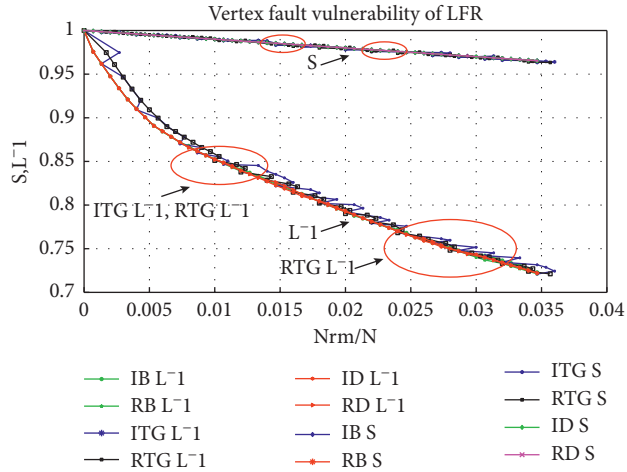

Figure 7: l^{-1}&S result of LFR.

(3) Among the six different curves of S, the curves of *ITG*, *IB*, and *ID* are more likely a linear curve than other ones, and the experimental result has shown the correctness of *ITG* in its almost linear curve with *IB* and *ID*.

In the common circumstances, the network will be more denser while having the larger value. But in generally, the value of network average cluster coefficient having a sudden increase after the continuous decreasing trend would tell that the node removal can make the network suddenly into disconnected subnetworks. Besides that, Albert et al. [13] and other researchers have also found that when the node having a very high betweenness value leaves the network, it may trigger the huge collapse in the whole network on the system level with a sudden unpredicted speed. Figures 4, 6, and 8 demonstrate the corresponding variation trend of $C_{G=(V,E)}$:

(1) It can be also clearly found in Figures 4, 6, and 8 that, among the six curves of S, the *RTG, ITG, RB, and IB* removal strategies are more harmful than *RD* and *ID* where in some datasets, the *RTG* removal strategy is much more harmful than others such as in S and $C_{G=(V,E)}$ curve of LFR in Figures 7 and 8.

(2) There have been much more rises in *ITG* and *IB* curves than other curves for the reason that it can break the whole network into disconnected subnetworks with much higher numbers. It has also proved that our *ITG* node centrality is a very good node centrality by the performance of curve S, $C_{G=(V,E)}$, and l^{-1}.

4.2.3. Directed Dataset and Experiment Results. Most social networks are directed networks with directed edges such as e-mail networks, twitter network, and calling record. We selected some representative social network dataset in Table 3. The OSLOM (a) dataset was provided by the open source algorithm OSLOM [33] as an example dataset of directed social networks. The subject reference (b) dataset

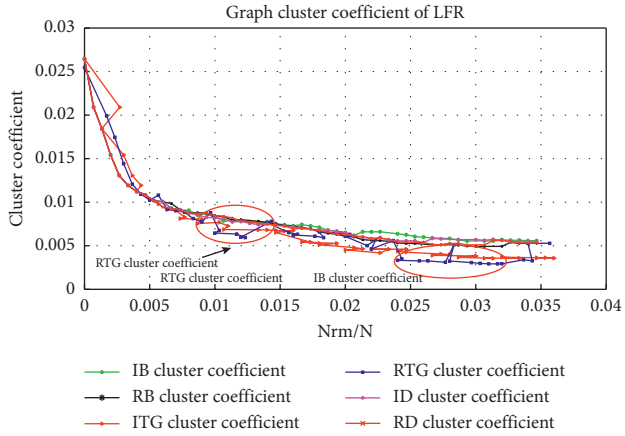

Figure 8: $C_{G=(V,E)}$ result of LFR.

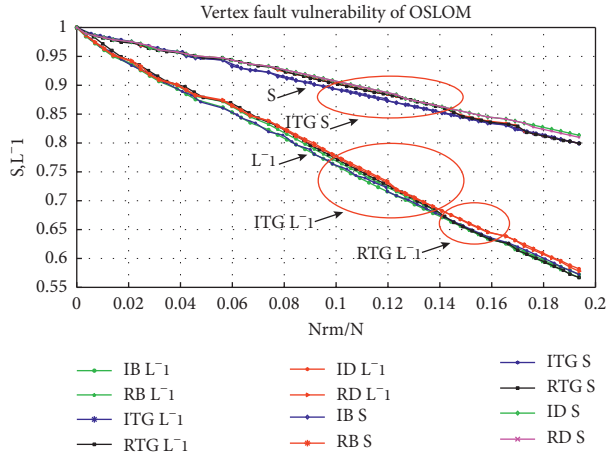

Figure 9: l^{-1}&S result of OSLOM.

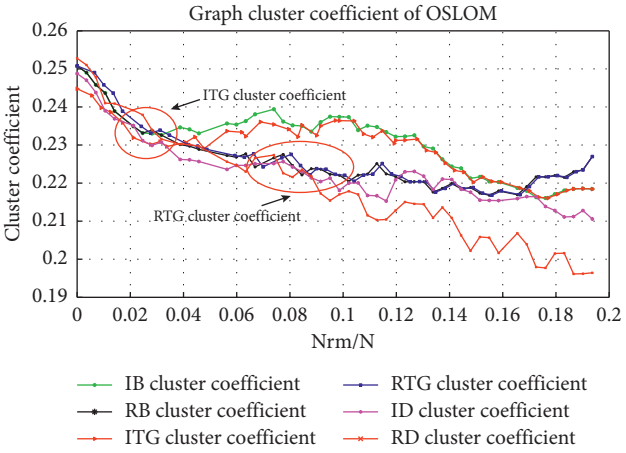

Figure 10: $C_{G=(V,E)}$ result of OSLOM.

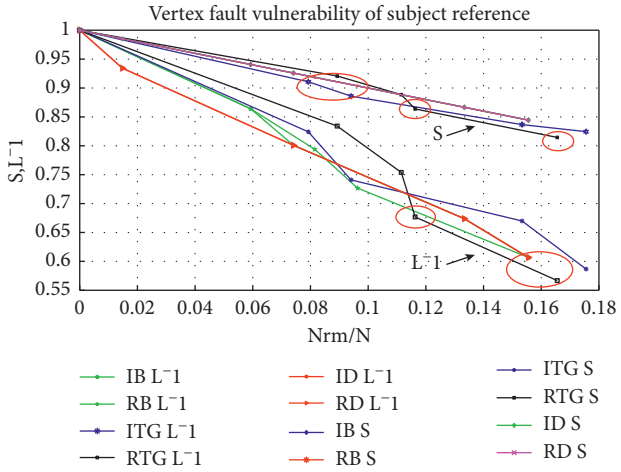

Figure 11: l^{-1}&S result of subject reference.

was provided by the INFOMAP algorithm [34] which was a subject reference network from the research of Physics, Chemistry, Biology, and Ecology in 6,128 journals connected by 6,434,916 citations. The SinaBlog Tweet (c) dataset came from the tweeting message reposting chain of one famous Chinese scholar in SinaBlog. The Calling Record (d) dataset was provided by cellphone calling records from another city in one month in China. All the above datasets are typical directed social networks.

The x-axis in Figures 9 to 16 stands for N_{rm}. The y-axis stands for the relative value of S and l^{-1}. Specially, all the network datasets in Table 3 would be processed as undirected graph in calculation of node betweenness and degree for *IB, RB, ID,* and *RD* strategy besides *ITG* and $IM_p(v)$ strategy. But, all the network datasets in Table 3 would be processed as directed graph in calculation of *ITG* node centrality coefficient. For the reason that there are only 40 nodes in subject reference (b) dataset, we only calculated the top five nodes.

(1) Experiment Results *of ITG with* $IM_p(v)$ *to IB, RB, ID, and RD.*

(2) Experiment Results Analysis for Directed Dataset of ITG with $IM_p(v)$ *to IB, RB, ID, and RD.* Figures 9 to 16 display the experiment results on directed social network, and we can find that the *ITG*, IMP, RTG, and RIMP strategies have amplified their harm in directed networks:

(1) l^{-1} and S: *ITG, IMP, RTG,* and *RIMP* are most harmful strategies, and the *RIMP* strategy is more powerful than *RTG* especially in the late stage of Figures 11, 13, and 15.

(2) $C_{G=(V,E)}$: among the six $C_{G=(V,E)}$ curves, the sudden rises caused by *IMP*, *ITG*, and IB are much more than the rises caused by other strategies where *IMP* is the most, where in Figure 14, we used the Y-axis logarithmic to show details more clear. The influence to $C_{G=(V,E)}$ approximately has shown that *RIMP* > *IMP* > *ITG* > *RTG* > *IB* > *RB* > *ID* > *RD*.

(3) Especially in the sparse directed network of subject reference, *IMDP* L^{-1}, *ITG* L^{-1}, *RTG* l^{-1}, and *RTG CC* curves in Figures 11 and 12 finally reached the finest experiment result and caused the biggest cascade. In Figures 13 and 14 of the SinaBlog Tweet network, the *RIMP* L^{-1}, *RIMP CC*, *RTG* l^{-1}, *RTG S*, and *RTG CC*

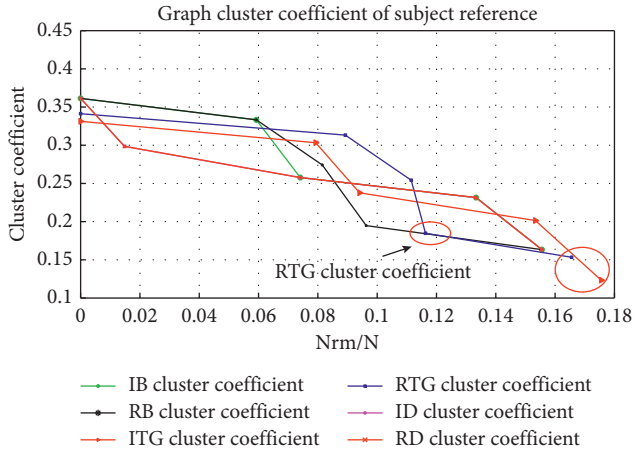

Figure 12: $C_{G=(V,E)}$ result of subject reference.

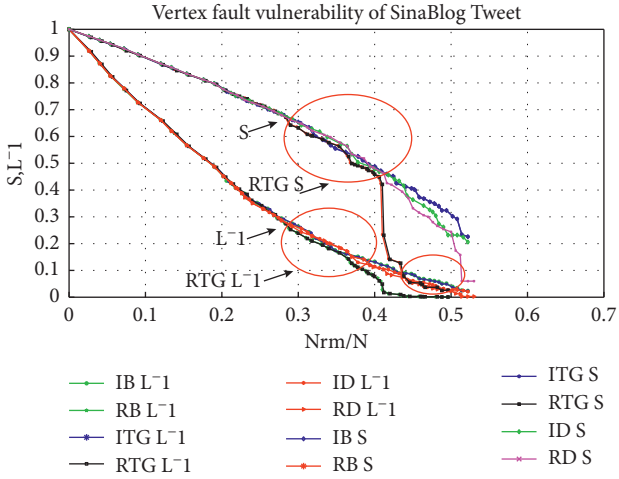

Figure 13: l^{-1}&S result of SinaBlog Tweet.

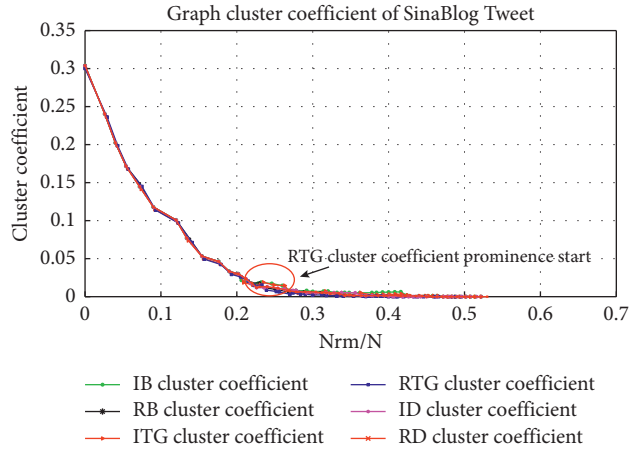

Figure 14: $C_{G=(V,E)}$ result of SinaBlog Tweet.

curves are the best result of experiment to other curves. In dense network of Calling Record, the *RIMP S*, *RIMP CC*, *RTG S*, and *RTG CC* also showed the power of $IM_p(v)$ and *ITG* node centrality by the leading experiment values.

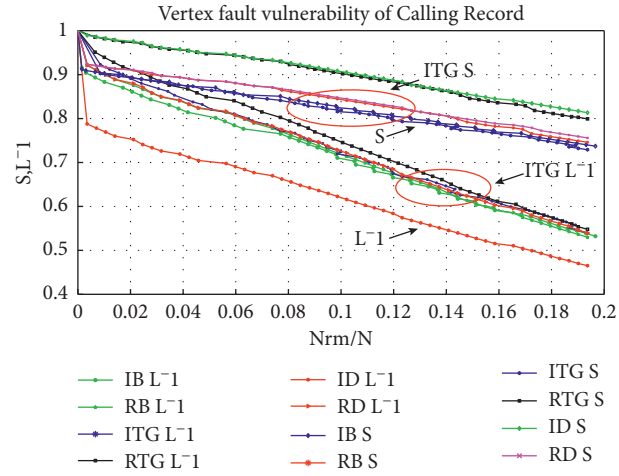

Figure 15: l^{-1}&S result of Calling Record.

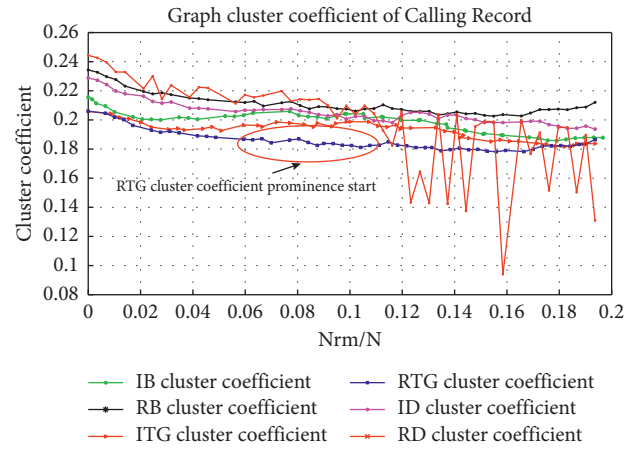

Figure 16: $C_{G=(V,E)}$ result of Calling Record.

Then, we can summarize from Figures 3 to 16 that the $IM_p(v)$- and *ITG*-based strategies have good experiment result in undirected and directed networks, and it has been proved for its correctness.

(3) Directed Experiment Dataset of ITG to $IM_p(v)$. In this part, we added two typical directed networks with much nodes inside which can be found in Table 4. The first one is the DBLP directed network dataset which is a famous directed heterogeneous information network that contains a dataset of author-centric English literature in the field of computer science with 14736 papers and 14475 authors [28]. We selected the data related to the field of computer, including database, data mining, artificial intelligence, and information retrieval including titles of papers published in various fields, authors who published more than five papers, abstracts of papers, and conferences titles. The another directed network dataset is the well-known ENRON e-mail network among employees of ENRON company from May 11, 1999, to May 21, 2002 (http://www.cs.cmu.edu/~/enron/). This ENRON e-mail network is divided into small parts by time stamp in each seven days, and the whole network is composed by 150 persons and 1526 emails among them [35, 36]. Its network

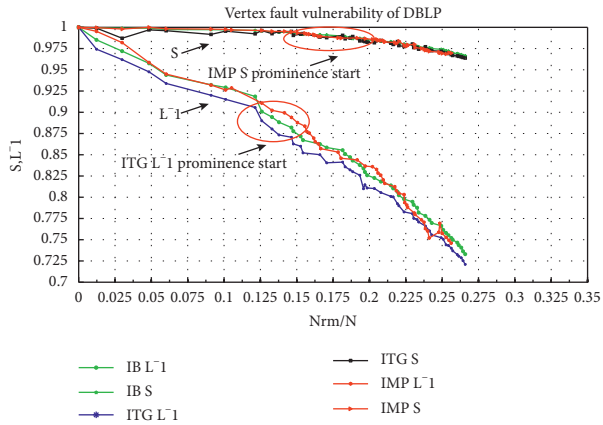

Figure 17: l^{-1}&S result of DBLP.

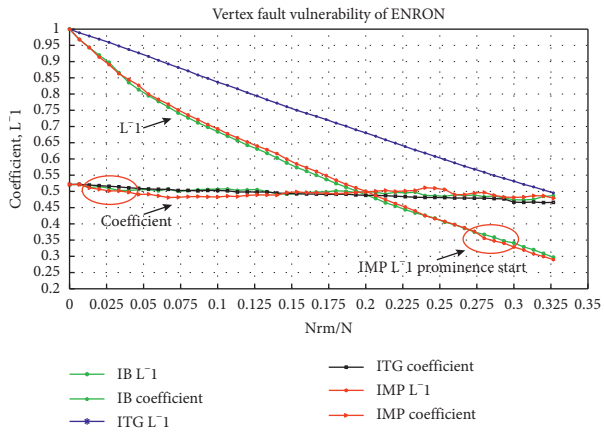

Figure 18: l^{-1}&$C_{G=(V,E)}$ result of ENRON.

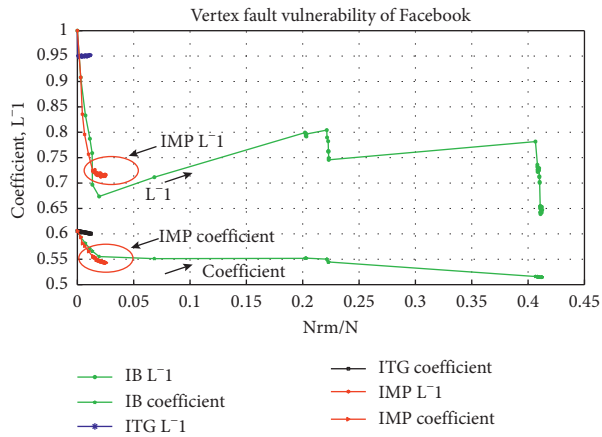

Figure 19: l^{-1}&$C_{G=(V,E)}$ result of Facebook.

structure clearly shows some important moments in the company's development, such as the company's collapse and the suicide of its former CEO. Dataset C is a Facebook dataset which was collected from survey participants using Facebook app. The dataset includes node features (profiles), circles, and ego networks [37]. The Facebook dataset has been

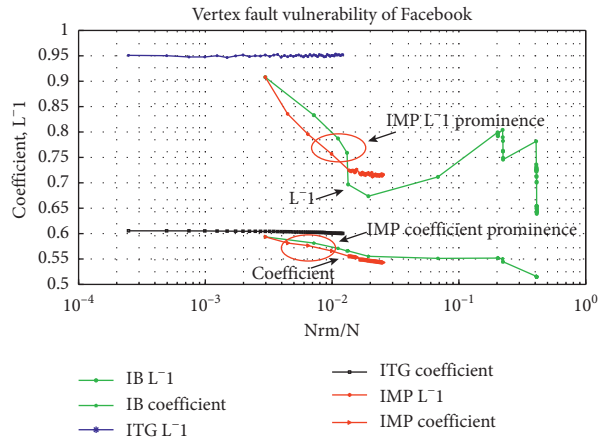

Figure 20: l^{-1}&$C_{G=(V,E)}$ result of Facebook (X-axis logarithmic coordinate).

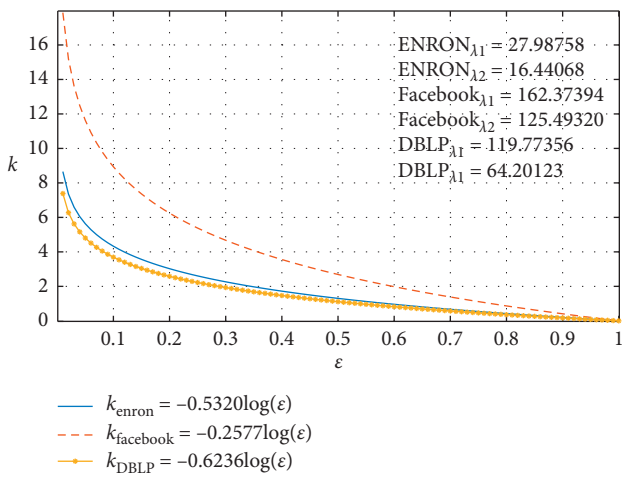

Figure 21: l^{-1} comparison result of iteration times.

anonymously processed by replacing the Facebook-internal ids for each user with a new value.

(4) Experiment Results for Directed Dataset of ITG to $IM_p(v)$.

Figures 17 and 18 display the experiment results of l^{-1} and coefficient on directed social network DBLP and ENRON. For the reason of the small scale of ENRON network dataset and very little change in when deleting the top 50 nodes, we only offered the l^{-1} and coefficient result of it.

It is obviously that, in these directed networks, the influence of IB, ITG, and $IM_p(v)$ has demonstrated a different varying trend in the undirected networks from Table 1. The computation of average inverse geodesic length l always cause high computing costs, while it costs us about 2.4 billion times shortest path search in the dataset of Facebook which has 4039 nodes and 88234 edges ($4000 * 4000 * 3 * 50 = 2.4 * 10^9$) and its computation result only can make up the three curves of $IB\,l^{-1}$, $ITG\,l^{-1}$, and $IMP\,l^{-1}$.

Table 2: Undirected network dataset.

	BACSF (a)	Call Community (b)	LFR (c)
Number of node N	490	284	1000
Number of link L	1892	3030	15384
Edge density (L/N)	3.8612	10.6690	15.3840

N: number of nodes; K_{degree}: degree power-law distribution index; $C_{\min}$: minimum number of nodes; $C_{\max}$: maximum number of nodes; u: mix hybrid parameter; C_{degree}: community size power-law distribution index.

Table 3: Directed network dataset.

	OSLOM (a)	SubjectReference (b)	SinaBlogTweet (c)	CallingRecord (d)
Number of node N	301	40	1492	286
Number of link L	6234	306	1490	3934
Edge Density (L/N)	20.7110	7.6500	0.9987	13.7552

Table 4: Another directed network dataset.

	DBLP (a)	ENRON (b)	Facebook (c)
Number of node N	14376	150	4039
Number of link L	14475	1526	88234
Edge Density (L/N)	1.0069	10.1733	21.8455

Figures 19 and 20 demonstrate the experiment results of l^{-1} and coefficient on directed social network Facebook. In order to demonstrate more details, the x-axis in Figure 20 used the logarithmic coordinates.

Figure 21 demonstrates the comparison result of iteration times in different datasets of Table 4 when computing the directed path-based node importance centrality $IM_p(v)$. By calculating formula $\Longrightarrow k > ((\log\varepsilon - \log a)/\log|\lambda_2|)$ (13b), we attain the varying curve of iteration times k to the marginal difference ε, while ε can be calculated by $\varepsilon \geq \|y^{(k+1)} - y^{(k)}\|$. Furthermore, each largest eigenvalue value λ_1 and second largest eigenvalue value λ_2 of dataset DBLP, ENRON, and Facebook also can be computed.

(5) Experiment Results Analysis for Directed Dataset of ITG to $IM_p(v)$. Figures 16 to 21 display the compared experiment results of *IB*, *ITG*, and $IM_p(v)$ node removal strategies on directed social networks. In addition, we explore the varying trend of iteration times k to the marginal difference ε.

(1) Generally after analyzing the demonstrated data in Figures 17 and 18, we can find that the influence of node with high *ITG* values have the most harmful effect to the l^{-1} in directed networks with low edge density such as *DBLP*. But in directed networks with high edge density such as ENRON, *IB* strategy and *ITG* strategy are just like doing the same effect.

(2) But in the directed network Facebook with a more higher edge density, we can find that the *ITG* node removal strategy suddenly lost its magic and the $IM_p(v)$ removal strategy does the best harmful effect to the Facebook network, while the *IB* removal strategy followed. And it maybe needs to be varied in more densely connected directed social networks such as Twitter and WeChat.

(3) In Figure 21, we can clearly find that the relationship of k and ε. The more densely connected directed network Facebook has the larger iteration times number k. And in dataset Facebook, when ε is more close to zero, iteration times rises to around twenty which has shown high efficiency of our node $IM_p(v)$ computing algorithm.

5. Conclusion

In this paper, we have proved new information transfer gain-(*ITG*-) based probability clustering coefficient and directed node importance centrality $IM_p(v)$ for measuring directed graph. Our comparisons in the variation trend of some key performance quantities of network robustness and node vulnerability assessment are useful and helpful. Comparison could help us to capture this cascading effect in directed online social networks. Experiments results showed that node *RIMP* and *RTG* strategies are more harmful than node betweenness-based strategies such as *RB* and *IB* in directed social networks including real Sina Blogging and Calling Record network. With sufficient experiments in synthetic signed networks and real networks derived from directed online social media and directed human mobile phone calling network, it has been proved that our *ITG*- and $IM_p(v)$-based directed social network robustness and node vulnerability assessment method is more accurate, efficient, and faster than several classical traditional centrality methods such as degree and betweenness. Furthermore, we will carry out our *ITG* centrality on more types and more large-scale directed social networks.

In addition, we propose a new proving process of directed node importance centrality $IM_p(v)$ in Section 3.2.4. By rigorous mathematical derivation and approximate calculation, to the best knowledge of us, we attain the varying trend of iteration times k to the marginal difference ε on directed social networks for the first time to our best knowledge.

Data Availability

These data used to support the findings of this study are included within the supplementary information file.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Key Research and Development Program of China (Grant no. 2017YFC0820603) and Advance Research Project of Shared Information System in 2019 (Grant no. 31511040103).

Supplementary Materials

The supplementary files contain the data which are used to support the findings of this study. (*Supplementary Materials*)

References

[1] F. Liu, S. Xue, J. Wu et al., "Deep learning for community detection: progress, challenges and opportunities," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence*, pp. 4981–4987, Yokohama, Japan, July 2020.

[2] M. Newman and G. Ghoshal, "Bicomponents and the robustness of networks to failure," *Physical Review Letters*, vol. 100, no. 13, p. 138701, 2008.

[3] T. Zhou, W. Bai, B. Wang, Z. Liu, and G. Yan, "A brief review of complex networks," *Physics*, vol. 34, no. 1, pp. 31–36, 2005.

[4] F. D. Malliaros and M. Vazirgiannis, "Vulnerability assessment in social networks under cascade-based node departures," *EPL (Europhysics Letters)*, vol. 110, no. 6, p. 68006, 2015.

[5] M. Subramanian, *Network Management: Principles and Practice*, Pearson Education India, Chennai, India, 2010.

[6] J. Xu and X. F. Wang, "Cascading failures in scale-free coupled map lattices," in *Proceedings of the 2005 IEEE International Symposium on Circuits and Systems*, pp. 3395–3398, Kobe, Japan, June 2005.

[7] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 3, p. 036104, 2006.

[8] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, "Attack vulnerability of complex networks," *Physical Review E*, vol. 65, no. 5, p. 056109, 2002.

[9] A. A. Nanavati, R. Singh, D. Chakraborty et al., "Analyzing the structure and evolution of massive telecom graphs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 5, pp. 703–718, 2008.

[10] J. Gao, B. Barzel, and A.-L. Barabási, "Universal resilience patterns in complex networks," *Nature*, vol. 530, no. 7590, pp. 307–312, 2016.

[11] J. Wu, X. Zhu, C. Zhang, and P. S. Yu, "Bag constrained structure pattern mining for multi-graph classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2382–2396, 2014.

[12] J. Wu, S. Pan, X. Zhu, and C. Zhihua, "Boosting for multi-graph classification," *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 416–429, 2015.

[13] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[14] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin, "Resilience of the internet to random breakdowns," *Physical Review Letters*, vol. 85, no. 21, p. 4626, 2000.

[15] R. Meusel, S. Vigna, O. Lehmberg, and C. Bizer, "The graph structure in the web - analyzed on different aggregation levels," *Journal of Web Science*, vol. 1, no. 1, pp. 33–47, 2015.

[16] R. Cohen and S. Havlin, *Complex Networks: Structure, Robustness and Function*, Cambridge University Press, Cambridge, UK, 2010.

[17] D. Garcia, P. Mavrodiev, and F. Schweitzer, "Social resilience in online communities: the autopsy of friendster," in *Proceedings of the First ACM Conference on Online Social Networks*, pp. 39–50, Boston, MA, USA, October 2013.

[18] S. Wu, A. Das Sarma, A. Fabrikant, S. Lattanzi, and A. Tomkins, "Arrival and departure dynamics in social networks," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pp. 233–242, Rome, Italy, February 2013.

[19] M. Kitsak, L. K. Gallos, S. Havlin et al., "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, 2010.

[20] M. Scheffer, S. R. Carpenter, T. M. Lenton et al., "Anticipating critical transitions," *Science*, vol. 338, no. 6105, pp. 344–348, 2012.

[21] T. Nepusz and T. Vicsek, "Controlling edge dynamics in complex networks," *Nature Physics*, vol. 8, no. 7, pp. 568–573, 2012.

[22] J. S. Carrington, J. Peter, and S. Wasserman, *Models and Methods in Social Network Analysis*, Cambridge University Press, Cambridge, UK, 2005.

[23] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, "Characterization of complex networks: a survey of measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167–242, 2007.

[24] C. Kiss and M. Bichler, "Identification of influencers-measuring influence in customer networks," *Decision Support Systems*, vol. 46, no. 1, pp. 233–253, 2008.

[25] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web," Technical Report, Stanford InfoLab, Stanford, CA, USA, 1999.

[26] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[27] J. Hopcroft and R. Kannan, *Computer Science Theory for the Information Age*, Shanghai Jiao Tong University Press, Shanghai, China, 2013.

[28] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao, "Graph regularized transductive classification on heterogeneous information networks," in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 570–586, Springer, Barcelona, Spain, September 2010.

[29] J. Pearl, *Probabilistic Reasoning in Intelligent Systems, Representation and Reasoning*, Morgan Kaufmann, San Mateo, CA, USA, 1988.

[30] X. Deng, J. Zhai, T. Lv, and L. Yin, "Efficient vector influence clustering coefficient based directed community detection method," *IEEE Access*, vol. 5, pp. 17106–17116, 2017.

[31] X. Deng, Y. Wen, and Y. Chen, "Highly efficient epidemic spreading model based lpa threshold community detection method," *Neurocomputing*, vol. 210, pp. 3–12, 2016.

[32] S. Fortunato and A. Lancichinetti, "Community detection algorithms: a comparative analysis: invited presentation, extended abstract," in *Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools*, pp. 1-2, Pisa, Italy, October 2009.

[33] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks," *PLoS One*, vol. 6, no. 4, 2011.

[34] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[35] S. Chakrabarti, *Mining the Web: Analysis*, Morgan Kaufmann, Burlington, MA, USA, 2002.

[36] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowledge and Information Systems*, vol. 42, no. 1, pp. 181–213, 2015.

[37] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in Neural Information Processing Systems*, pp. 539–547, Springer, Berlin, Germany, 2012.