Big Data
Volume 11, Number 2, 2023
© Mary Ann Liebert, Inc.
DOI: 10.1089/big.2021.0398

ORIGINAL ARTICLE

Identification of Potential Cooperation Relationships Among Scientists

Fuzhong Nian,* Yinuo Qian, and Yabing Yao

Abstract

In this article, the phenomenon of scientist cooperation in the scientist cooperation network is studied from the perspectives of information spread and link prediction. By mining the information in the scientist cooperation network, analyzing the cooperation has been generated and discovering potential cooperation opportunities. It helps to build a richer cooperation network with more content. Information spread can reflect the inner laws of network structure formation, and the link prediction method can retain the integrity of network information to the maximum extent. First, the real network is abstracted by analyzing its structure as well as node attributes into a simulated network. Second, the process of information spread in the cooperation network is simulated by improving the traditional SIS model. Some improvements are made to the link prediction algorithm for the impact brought to the network by information spread. Finally, the experimental results in the scientist cooperation network show that the hybrid weighted link prediction algorithm combining node attributes and spread factors can improve the accuracy of link prediction and provide suggestions for scientists to find partners. The comparative experiments on simulated and real networks not only validate the effectiveness of the propagation model in the scientist cooperation network, but also verify the accuracy of the hybrid weighted link prediction algorithm.

Keywords: information spread; link prediction; scientist cooperation network; node attributes

Introduction

In the era of rapid development of the internet, it is easier to find and meet more suitable partners. Internet records detailed personal information about scientists, and the possibility of cooperation between scientists who do not know each other is greatly increased. Scientists relying only on information exchanges in small circles to find partners are far from keeping up with the development of the era, and Karlovčec et al had done a detailed study of cross-disciplinary cooperation in the process of scientists' cooperation.[1] Therefore, information mining of scientist network can not only help scientists to seek more suitable partners, but also promote the development of cross-field achievements and create new sparks.

In previous studies, researchers have investigated the characteristics of network propagation and the structure of networks in many social networks. Tang et al proposed a dynamic friend network with clustering and community features and a propagation model based on interest attributes.[2] Based on the study of the propagation process of failures, Wang et al found the two factors affecting the propagation of cascading failures,[3] and proposed a preferential attachment strategy to improve the robustness of interdependent networks against cascading failures.[4] Ma and Zhu proposed a novel rumor propagation model by taking into account the subjective judgment and diverse characteristics of individuals.[5] Wang and Qin proposed a novel method of finding and expanding the core communities in bipartite networks.[6] Cui and Wang proposed an algorithm to detect one-mode community structures in bipartite networks, and to deduce which one-mode community structures are weighted.[7]

Zhu and Ma investigated how contact differences and individual similarity affect the rumor propagation process in complex heterogeneous networks.[8] Ren and Wang studied the epidemic spreading in time-varying community networks.[9] Yao and Gao proposed an

School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China.

*Address correspondence to: *Fuzhong Nian, School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China,* E-mail: gdnfz@lut.edu.cn

SE2IR rumor propagation model with hesitation mechanism based on the actual situation of investor networks.[10] Wang et al provided two new strategies for epidemic prevention and controlled by representing different fractional rates of incubation and invasion periods.[11] Hosni et al proposed an HISB model by analyzing personal and social behaviors in social networks.[12] Liu et al used the characteristic infected cluster size to investigate the inhomogeneity of the epidemic spreading in static and dynamic complex networks.[13] Gebhart and Funk studied the emergence of higher order structures in collaborative networks.[14] Hui et al studied the statistical properties of the network by constructing a weighted network of scientific cooperation.[15]

Zhou and Wang considered the difference in the sizes of the infected clusters in the dynamic complex networks, and the normalized entropy based on infected clusters was proposed to characterize the inhomogeneity of epidemic spreading.[16] Wang et al explored the propagation of multimessages by considering their correlation degree.[17] Ren and Wang studied cooperation in prisoner's dilemma game.[18] Fewer studies have been conducted to address the spread of information in collaborative networks of scientists. In fact, the publication of an article by a scientist and the resulting emergence of a collaborative authorship of scientists are similar to the process of information spread. When information about a certain research field appears in the limelight, the country needs talents to conduct in-depth research on a phenomenon or a technology, which contributes to the phenomenon of cooperation.

We can view this process as the process of information spread and use the idea of spread to analyze the phenomenon of cooperation.

It takes a long time for a scientist to go from finding a coauthor to successfully copublishing an article. Therefore, in the context of information spread in cooperation network, we analyze this phenomenon in depth. We propose the conjecture that multiple attributes of nodes affect the spread of information, improve the traditional propagation model, and construct a new simulation model to simulate the phenomenon of scientists' cooperation, which paves the way for the proposed hybrid link prediction algorithm.

Among the researches in the field of link prediction algorithms, a large part of the research is based on the form of node similarity. In the network, each node pair obtains a value according to the link prediction algorithm. For instance, the similarity index is based on local information, specifically, common neighbors, Salton Index, Jaccard coefficient, local path index, and preferential attachment are similarity indices defined using local information in the network. In addition, there are many similarity indices based on global information, such as the Matrix Forest Index, Katz Index, and Leicht–Holme–Newman Index. These similarity indices are described in detail in Lü and Zhou.[19]

In recent years, the research enthusiasm for link prediction algorithms has continued to rise, but there are still many problems to be solved. Among them, there is less research on the attributes of nodes themselves and the influence of spread factors on the network. The influence of information spread data on network structure has always been a relatively active field in data mining. Researchers conduct research on information related to information spread and use mathematical formula derivation to infer network structure changes.[20–28] The use of information spread is not just limited to the study of the influence of network structure, but also applied to many other fields. In our previous research, through the study of spread data, we successfully carry out the division of communities.

Tran et al simultaneously detected the community structure and network contagion between individuals in studying the activities of individuals to infer the underlying network and found coherent communities.[29] Therefore, it is possible to improve the link prediction algorithm by using the data of information spread.

From the analysis of previous work, the attributes of individuals in the scientist cooperation network are the basis for discovering the cooperation of scientists, and the factors of information spread in the network are important factors affecting the cooperation of scientists. Therefore, in this article, a link prediction algorithm based on the attributes of nodes is first proposed by studying the attributes of nodes for the scientists' cooperation network. Second, the spread factors are added to the link prediction based on node attributes after the successful simulation of scientists' cooperation by the joining cooperation related psychology and scientists personal attributes of SIS propagation model, and the hybrid weighted link prediction algorithm based on information spread factors and node attributes is proposed. Finally, the two algorithms are applied to the real scientist cooperation network to complete the validation of the model.

In this article, we investigate the phenomenon of scientists' cooperation in detail from several aspects, reflect the phenomenon of scientist cooperation network in the form of propagation, and reveal the evolution of the intrinsic structure in the cooperation network by link prediction.

Problem Description and Research Framework

In this article, we introduce node attributes and information spread influencing factors to construct link prediction models to identify potential scientists' cooperation by studying a scientist cooperation network. In reality, in addition to the existing cooperative relationship between scientists, scientists can also build an invisible cooperative relationship by studying the spread of information in the network. Cooperation network is constructed by using scientists as nodes and constructing connected edges with relationships between scientists who collaborate on publications. In the simulated network, first, the cooperation phenomenon existing in the scientist cooperation network is simulated by an improved information propagation model, and second, a single similarity index based on node attributes and a similarity index based on information spread factors are proposed.

Finally, the two single similarity indices are coupled to propose the hybrid weighted similarity index, and the relative optimal algorithm is formed by evaluating the accuracy of the algorithm to determine the weights of the two metrics. In the end, the prediction of the cooperation relationship in the scientist cooperation network is successfully performed, which provides a reference for scientists to seek partners in the future.

Cooperation Spread Model of Scientists Based on Multifactor Coupling

Modeling of the spread dynamic model

We select the classical SIS model as the base model.[30] The traditional model divides the population within the epidemic range of infectious diseases into the following three categories: state S (Susceptible), which refers to people who do not have the disease but lack immunity. These nodes are susceptible to infection after contact with infected persons. State I (Infective), which refers to a person who has contacted an infectious disease and can spread to members of state S.

In the process of information spread, there are some similarities between the spread of information and the spread of diseases. In the SIS model, the spread of information usually refers to the fact that after the information has spread for some time, the information spreads again in the network because of certain factors. In the scientist cooperation network, there is a high probability that scientists will continue to be involved in writing articles in the field after they have been involved in collaborating on articles in that field. Nowadays, people have more ways to access information independently. The spread of information is no longer limited to a small area, and two people who did not know each other may become friends by chance through the accident of information in the internet. The same phenomenon exists in the scientist cooperation network.

Theoretically, any two scientists in the world have a certain probability of cooperation, and it is easier to produce collisions of thinking. Therefore, it is very meaningful for us to predict the future cooperation of scientists. In the traditional SIS model, both the propagation rate β and the recovery rate γ are static, but information spread in the network is affected by numerous factors. So, we improve the traditional propagation model to make the process of propagation closer to the real scientists' cooperation spread effect. This paves the way for the subsequent validation of our weighted hybrid link prediction algorithm in a simulated network.

Dynamic rate of change. When information about a field of research appears in the public eye, each article is published with a ranking of its authors. This ranking is defined by the degree of contribution to the article, with the more contributing scientists being ranked higher. We define this as the scientist's degree of involvement $p(i,t)$. The activeness of the scientist is also related to the field of study of the information. The more relevant a scientist's field of study is to the content of that information field of study, the higher the activeness of the scientist. We define the scientist's field similarity as ω_i.

To characterize this variability in the degree of activity, we define the activeness of the scientist (the extent to which scientists have been affected by the field) as $E(i,t)$. Where i represents the node and the node represents the scientist. The activity of node i at time t is defined as Equation (1).

$$E(i,t)=p(i,t)\times\omega_i \tag{1}$$

$p(i,t)$ is activeness of the node i at time t. $p(i,t)$ is defined as shown in Equation (2).

$$p(i,t)=\left[\left(\frac{v_1}{\sum_{i=1}^{q}v_i}\right)\left(\frac{v_2}{\sum_{i=1}^{q}v_i}\right)\left(\frac{v_3}{\sum_{i=1}^{q}v_i}\right)\cdots\left(\frac{v_q}{\sum_{i=1}^{q}v_i}\right)\right]_{1\times q}\times\tau_i\times[1,1,1\cdots1]_{n\times1}^{T} \tag{2}$$

The weight value of the article published by the qth author is labeled as v_q in Equation (2), and its weight value increases as q increases. τ_i is a 0, 1 matrix of $(q\times n)$ to represent the articles published by node i. τ_i is defined as shown in Equation (3).

$$\tau_i = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{q,1} & \cdots & a_{q,n} \end{bmatrix}_{q\times n} \tag{3}$$

The number of $a_{q,n}$s with a value of 1 is determined by the number of published articles for that node i with the qth author, and n is the number of articles. Due to the spread of information, τ_i changes over time. Therefore, the activeness of each node is updated after each round of spread.

ω_i is the similarity between the research field of node i and the field where the information is located. It is defined as shown in Equation (4).

$$\omega_i = \frac{|\Omega(i) \cap R|}{|\Omega(i) \cup R|} \tag{4}$$

R in Equation (4) represents the set containing all the keywords in the research field where this information is located, and $\Omega(i)$ is the set of keywords in the field of node i. The relevance of the research content of node i to the field in which the information is located is measured by calculating the similarity ω_i between the keywords of the research field of node i and the set R using Equation (4). Bringing Equation (2), Equation (3), and Equation (4) into Equation (1), we get the activeness of node i as shown in Equation (5).

$$E(i,t) = \left[\left(\frac{v_1}{\sum_{i=1}^{q} v_i}\right)\left(\frac{v_2}{\sum_{i=1}^{q} v_i}\right)\left(\frac{v_3}{\sum_{i=1}^{q} v_i}\right)\cdots\left(\frac{v_q}{\sum_{i=1}^{q} v_i}\right)\right]_{1\times q} \times \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{q,1} & \cdots & a_{q,n} \end{bmatrix}_{q\times n} \times [1,1,1\cdots 1]^T_{n\times 1} \times \frac{|\Omega(i)\cap R|}{|\Omega(i)\cup R|} \tag{5}$$

To better match the actual situation of information spread in the scientist cooperation network, we analyze the effects of scientists' own attributes and group effects on information spread. The effects of scientists' own attributes include the average influence of scientists' neighboring nodes. The influence of group effects includes the huddle effect and field heat effect. In our previous experiments, we studied the influence of incentive mechanism on information spread behavior and established a model of spread based on incentive mechanism in Nian et al.[31] Based on the analysis of the human search phenomenon, a new propagation model was proposed in Nian and Diao.[32] Also, in this article, to be more consistent with the actual process of cooperation situation in the scientist cooperation network, the dynamic propagation rate is redefined as a better response to the spread of information in the scientist cooperation network. As shown in Equation (6).

$$\beta(i,t) = \beta_0 + \delta(t) + \varphi(t)_{\mu,\sigma} + \theta(E(i,t) + F(i,t)) \tag{6}$$

In Equation (6), θ is a constant and β_0 is the initial infection rate. $\delta(t)$ corresponds to the huddle effect. $\varphi(t)_{\mu,\sigma}$ corresponds to the field heat effect. $E(i,t)$ corresponds to the average influence of the scientist's neighbor nodes. $F(i,t)$ corresponds to the activeness of the scientist. When scientists spread the information of cooperation, the influence on the depth and breadth of information spread differs due to the different influence of scientists themselves in the research field. The greater the average influence of neighboring nodes of node i, the greater the number of infected nodes, and the greater the possibility of node i to participate in the research. In the dynamic propagation rate, the average influence of the neighbor nodes of node i on node i is defined as Equation (7).

$$F(i,t) = \frac{\sum_{j\in\Gamma(i)} f_i}{|\Gamma(i)|} \times ln(hI(t)) \tag{7}$$

In Equation (7), $\Gamma(i)$ represents the set of neighboring nodes of node i, $I(t)$ represents the number of infected nodes in the neighboring nodes of node i, and h is a constant. f_i represents the scientist influence of node i itself, and it is defined as shown in Equation (8).

$$f_i = \sum_{z=1}^{k} \alpha_z \times \frac{c_{(i,z)} - c_{(i,z)\mathrm{min}}}{c_{(i,z)\mathrm{max}} - c_{(i,z)\mathrm{min}}} (0 < \alpha_z < 1) \tag{8}$$

In Equation (8), α_z represents the different weights of different attributes that measure the influence of scientists. k is the number of attributes that can measure the influence of scientists. The second term represents the normalized value of an attribute that can measure the influence of scientists in the set of attributes of node i, $c_{(i,z)}$ represents the value of an attribute that can measure the influence of scientists, $c_{(i,z)_{\mathrm{min}}}$ represents the minimum value of the attribute in the sample, and $c_{(i,z)_{\mathrm{max}}}$ denotes the maximum value of the attribute in the sample. Bringing Equation (8) into Equation (7), we get the activeness of node i as shown in Equation (9).

$$F(i,t) = \frac{\sum_{j\in\Gamma(i)} \sum_{z=1}^{k} \alpha_z \times \frac{c_{(j,z)} - c_{(j,z)\mathrm{min}}}{c_{(j,z)\mathrm{max}} - c_{(j,z)\mathrm{min}}}}{|\Gamma(i)|} \times ln(hI(t)), (0 < \alpha_z < 1) \tag{9}$$

In the scientist cooperation network, information about the cooperation spreads through the network, and people's interest and state of mind change over time. The huddle phenomenon arises when information about a research field first appears among the networks and many scientists are curious about the emergence of a new field. We define this as the huddle effect. An online survey showed that the majority of respondents believed that the phenomenon of online huddle was widespread and indicated that they had participated in online huddles. However, with the passage of time and the spread of news, this effect will slowly fade over time and will no longer have a large fluctuating effect on the spread of information. In the dynamic propagation rate, we define it as Equation (10), where a and b are constants.

$$\delta(t) = e^{\frac{1}{at+1}+b} \tag{10}$$

When a research field first appears in the public view, the field is still at a low level of enthusiasm. More and more people are involved in the research in the field when national policies are gradually improved, and the field will have a peak of enthusiasm after a period of time. When the field of research is hot, scientists will also want to use this opportunity to understand the development prospects of the field of research, and will be more willing to participate in research related to the field of information.

After a period of time, after the field heat reaches its peak, the field may reach a saturation of relevant research in the field, and the field heat returns to a plateau, we refer to this as the topic effect.[33] In the dynamic propagation rate, we define it as Equation (11). μ and σ are constants to control the time when the field heat effect reaches its peak.

$$\varphi(t)_{\mu,\sigma} = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{11}$$

In the process of spread, the recovery rate of nodes is also affected by group effects as well as individual behavior. In the process of scientists collaborating on an article, some of the scientists who have already collaborated on an article will be involved again in collaborating on an article in the field. Over time, some scientists will have a stronger desire to go deeper into cooperation in the field and they will be more interested in this field.

However, with the influence of the topic effect, some scientists will be less enthusiastic in this field, and the novelty level brought by the huddle effect will gradually decrease. The scientists may participate in cooperation about other fields and focus on other new fields. Therefore, the recovery rate of a node is also dynamic and consists of four components, including the huddle effect, the field heat effect, the average influence of the scientist's neighbor nodes, and the activeness of the scientist. Therefore, we define the dynamic recovery rate as shown in Equation (12), where α and k are defined as constants and γ_0 is an initial value of the propagation rate.

$$\gamma(i,t) = \gamma_0 - \alpha(\delta(t) + \varphi(t)_{\mu,\sigma}) - F(i,t) + kE(i,t) \tag{12}$$

We bring Equation (5), Equation (9), Equation (10), and Equation (11) into Equation (6) and Equation (12) to obtain the dynamic propagation rate [Eq. (13)] and dynamic recovery rate [Eq. (14)].

$$\beta(i,t) = \beta_0 + e^{\frac{1}{at+1}+b} + \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} + \theta \times \left\{ \begin{array}{l} \frac{\sum_{j\in\Gamma(i)} \sum_{z=1}^{k} \alpha_z \times \frac{c_{(j,z)} - c_{(j,z)_{\min}}}{c_{(j,z)_{\max}} - c_{(j,z)_{\min}}}}{|\Gamma(i)|} \times ln(hI(t)) + \\ \left[\left(\frac{v_1}{\sum_{i=1}^{q} v_i}\right) \left(\frac{v_2}{\sum_{i=1}^{q} v_i}\right) \left(\frac{v_3}{\sum_{i=1}^{q} v_i}\right) \cdots \left(\frac{v_q}{\sum_{i=1}^{q} v_i}\right) \right]_{1\times q} \\ \times \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{q,1} & \cdots & a_{q,n} \end{bmatrix}_{q\times n} \times [1,1,1\cdots 1]_{n\times 1}^{T} \times \frac{|\Omega(i)\cap R|}{|\Omega(i)\cup R|} \end{array} \right\} \tag{13}$$

$$\begin{array}{l} \gamma(i,t) = \gamma_0 - \alpha\left(e^{\frac{1}{at+1}+b} + \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) - \\ \frac{\sum_{j\in\Gamma(i)} \sum_{z=1}^{k} \alpha_z \times \frac{c_{(j,z)} - c_{(j,z)_{\min}}}{c_{(j,z)_{\max}} - c_{(j,z)_{\min}}}}{|\Gamma(i)|} \times ln(hI(t)) + \\ k \times \left[\left(\frac{v_1}{\sum_{i=1}^{q} v_i}\right) \left(\frac{v_2}{\sum_{i=1}^{q} v_i}\right) \left(\frac{v_3}{\sum_{i=1}^{q} v_i}\right) \cdots \left(\frac{v_q}{\sum_{i=1}^{q} v_i}\right) \right]_{1\times q} \\ \times \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{q,1} & \cdots & a_{q,n} \end{bmatrix}_{q\times n} \times [1,1,1\cdots 1]_{n\times 1}^{T} \times \frac{|\Omega(i)\cap R|}{|\Omega(i)\cup R|} \end{array} \tag{14}$$

Construction of the improved spread dynamic model. In the scientist cooperation network, a node that has published an article in the field of information will influence a node that has not published a related article with a dynamic propagation rate β, so that the node knows about the information of cooperation and has a certain probability of being influenced to publish articles in the related field. After successfully publishing an article, the node changes from the S state to one of the q I states (Different I states indicate that articles are published with different ranks of authors, and q represents the number of coauthors in an article). When some node wants to publish again, the node changes to S state again and reengages in the process of publishing a new article. Therefore, we choose the SIS model to simulate the process of scientists' cooperation. The traditional propagation dynamic formulation of SIS is shown in Equation (15).

$$\begin{cases} \frac{ds(t)}{dt} = \gamma i(t) - \beta s(t)i(t) \\ \frac{di(t)}{dt} = \beta s(t)i(t) - \gamma i(t) \end{cases} \tag{15}$$

The description of the differential formula based on the improved SIS model is shown in Equation (16).

$$\begin{cases} \frac{ds(t)}{dt} = \left[\gamma_0 - \alpha\left(\delta(t) + \varphi(t)_{\mu,\sigma}\right) - F(i,t) + kE(i,t)\right] i(t) - \\ \qquad \left[\beta_0 + \delta(t) + \varphi(t)_{\mu,\sigma} + \theta(E(i,t) + F(i,t))\right] s(t)i(t) \\ \frac{di(t)}{dt} = \left[\beta_0 + \delta(t) + \varphi(t)_{\mu,\sigma} + \theta(E(i,t) + F(i,t))\right] s(t)i(t) - \\ \qquad \left[\gamma_0 - \alpha(\delta(t) + \varphi(t)_{\mu,\sigma}) - F(i,t) + kE(i,t)\right] i(t) \end{cases} \tag{16}$$

We bring the dynamic propagation rate of Equation (13) and the dynamic recovery rate of Equation (14) into Equation (16) to obtain the final spread dynamic model [Eq. (17)].

$$\begin{cases} \frac{ds(t)}{dt} = \left[\begin{matrix} \gamma_0 - \alpha\left(e^{\frac{1}{at+1}+b} + \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) - \\ \frac{\sum_{j\in\Gamma(i)} \times \sum_{z=1}^{k} \alpha_z \times \frac{c_{(j,z)} - c_{(j,z)_{min}}}{c_{(j,z)_{max}} - c_{(j,z)_{min}}}}{|\Gamma(i)|} \times l\,n(hI(t)) + \\ k \times \left[\left(\frac{v_1}{\sum_{i=1}^{q} v_i}\right)\left(\frac{v_2}{\sum_{i=1}^{q} v_i}\right)\left(\frac{v_3}{\sum_{i=1}^{q} v_i}\right)\cdots\left(\frac{v_q}{\sum_{i=1}^{q} v_i}\right)\right]_{1\times q} \times \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{q,1} & \cdots & a_{q,n} \end{bmatrix}_{q\times n} \times \\ [1,1,\cdots,1]_{n\times 1}^{T} \times \frac{|\Omega(i)\cap R|}{|\Omega(i)\cup R|} \end{matrix}\right] i(t) - \\ \left[\begin{matrix} \beta_0 + e^{\frac{1}{at+1}+b} + \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} + \\ \theta \times \left\{\begin{matrix} \frac{\sum_{j\in\Gamma(i)} \times \sum_{z=1}^{k} \alpha_z \times \frac{c_{(j,z)} - c_{(j,z)_{min}}}{c_{(j,z)_{max}} - c_{(j,z)_{min}}}}{|\Gamma(i)|} \times l\,n(hI(t)) + \\ \left[\left(\frac{v_1}{\sum_{i=1}^{q} v_i}\right)\left(\frac{v_2}{\sum_{i=1}^{q} v_i}\right)\left(\frac{v_3}{\sum_{i=1}^{q} v_i}\right)\cdots\left(\frac{v_q}{\sum_{i=1}^{q} v_i}\right)\right]_{1\times q} \times \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{q,1} & \cdots & a_{q,n} \end{bmatrix}_{q\times n} \times \\ [1,1,\cdots,1]_{n\times 1}^{T} \times \frac{|\Omega(i)\cap R|}{|\Omega(i)\cup R|} \end{matrix}\right\} \end{matrix}\right] s(t)i(t) \\ \frac{di(t)}{dt} = \left[\begin{matrix} \beta_0 + e^{\frac{1}{at+1}+b} + \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} + \\ \theta \times \left\{\begin{matrix} \frac{\sum_{j\in\Gamma(i)} \times \sum_{z=1}^{k} \alpha_z \times \frac{c_{(j,z)} - c_{(j,z)_{min}}}{c_{(j,z)_{max}} - c_{(j,z)_{min}}}}{|\Gamma(i)|} \times l\,n(hI(t)) + \\ \left[\left(\frac{v_1}{\sum_{i=1}^{q} v_i}\right)\left(\frac{v_2}{\sum_{i=1}^{q} v_i}\right)\left(\frac{v_3}{\sum_{i=1}^{q} v_i}\right)\cdots\left(\frac{v_q}{\sum_{i=1}^{q} v_i}\right)\right]_{1\times q} \times \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{q,1} & \cdots & a_{q,n} \end{bmatrix}_{q\times n} \times \\ [1,1,\cdots,1]_{n\times 1}^{T} \times \frac{|\Omega(i)\cap R|}{|\Omega(i)\cup R|} \end{matrix}\right\} \end{matrix}\right] s(t)i(t) - \\ \left[\begin{matrix} \gamma_0 - \alpha\left(e^{\frac{1}{at+1}+b} + \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) - \\ \frac{\sum_{j\in\Gamma(i)} \times \sum_{z=1}^{k} \alpha_z \times \frac{c_{(j,z)} - c_{(j,z)_{min}}}{c_{(j,z)_{max}} - c_{(j,z)_{min}}}}{|\Gamma(i)|} \times l\,n(hI(t)) + \\ k \times \left[\left(\frac{v_1}{\sum_{i=1}^{q} v_i}\right)\left(\frac{v_2}{\sum_{i=1}^{q} v_i}\right)\left(\frac{v_3}{\sum_{i=1}^{q} v_i}\right)\cdots\left(\frac{v_q}{\sum_{i=1}^{q} v_i}\right)\right]_{1\times q} \times \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{q,1} & \cdots & a_{q,n} \end{bmatrix}_{q\times n} \times \\ [1,1,\cdots,1]_{n\times 1}^{T} \times \frac{|\Omega(i)\cap R|}{|\Omega(i)\cup R|} \end{matrix}\right] i(t) \end{cases} \tag{17}$$

where $s(t)$ and $i(t)$ are the proportions of nodes in the susceptible state and nodes in the infected state in the population, respectively. $s(t) + i(t) \equiv 1, (s(t) \geqslant 0, i(t) \leqslant 1)$. The initial state of the node is state S. The initial state of the node state is shown in Equation (18).

$$\begin{cases} S_0 = N - I_0 \\ I_0 = \left[\beta_0 + \delta(0) + \varphi(0)_{\mu,\sigma} + \theta(E(i,0) + F(i,0))\right] \times N \\ \end{cases} \tag{18}$$

where N is the total number of nodes in the network, S_0 is the number of susceptible infectors in the network, and I_0 is the number of infected individuals in the network. It is calculated as the number of nodes infected in the network at the moment $t = 0$ with propagation probability as $\beta(i, 0)$. The spread of the information can be represented by the spread threshold, which we define as Λ representation, as shown in Equation (19).

$$\Lambda = \frac{\beta(i,t)}{\gamma(i,t)} \tag{19}$$

where $\beta(i, t)$ represents the propagation rate of node i at time t and $\gamma(i, t)$ represents the recovery rate of node i at time t.

Definition of Similarity Index in the Scientist Cooperation Network

If we want to use the similarity between nodes to perform link prediction, we must set a premise that the greater the similarity between two nodes, the greater the probability of the existence of connected edges between nodes. This section constructs a new similarity index by looking at the properties of the scientist cooperation network. In the prediction model of the scientist cooperation network, we first calculate the prediction accuracy of this metric for the cooperation network using the similarity index based on node attributes. Second, we calculate the prediction accuracy of this index for the cooperation network using the similarity index based on information spread factors. Finally, the hybrid weighted index based on node attributes and spread factors is constructed by combining the attribute similarity index as the base indicator with the proposed similarity indicator based on information spread factors. The three similarity indexes are described in detail in this chapter.

Similarity index based on the node attributes

The degree of similarity of the preexisting attributes between scientists influences the search and matching of partners. Similar research fields and similar institutions, as well as the same job titles, represent the common pursuit of research areas among individuals. Therefore, individuals with similar attributes have a higher probability of connection and they are more likely to produce cooperation in the scientist cooperation network. In contrast, scientists with lower degrees of similarity have a relatively lower probability of cooperation. In this article, attribute similarity is defined as the number of common attribute labels shared between nodes. Based on this, we propose a formula for calculating the similarity index based on node attributes, as shown in Equation (20).

$$sim_{ij}^{(1)} = \frac{|\Omega(i) \cap \Omega(j)|}{|\Omega(i) \cup \Omega(j)|}, (0 \leq sim_{ij}^{(1)} \leq 1) \tag{20}$$

where $\Omega(i)$ represents the set of attributes of node i, $\Omega(j)$ represents the set of attributes of j, $sim_{ij}^{(1)}$ represents the probability of individual generating cooperative links based on the similarity of node attributes, and $sim_{ij}^{(1)} = sim_{ji}^{(1)}$.

Similarity index based on information spread factors

The SIS model which takes into account the psychological effects and personal attributes of scientists involved in the process of cooperation is used to simulate the process of information spread in the network. After the information spread in the network, the nodes will be infected with several states because the nodes are infected with different degrees of information. We put the nodes infected with different states into different sets and only consider the nodes in the same set that have a higher degree of similarity to each other. If there is no infection of a state, the similarity index is 0.

To quantify the degree of similarity of nodes in the same set, we define the value of the spread factors as ε, and $0 < \varepsilon < 1$. We propose the information spread factors of similarity index calculation formula, as shown in Equation (21), and $sim_{ij}^{(2)} = sim_{ji}^{(2)}$. The set I represents the set of nodes infected with the same state. Then the probability that any two nodes in the set produce cooperation that is ε.

$$sim_{ij}^{(2)} = \begin{cases} \varepsilon & i,j \in I \\ 0 & i,j \notin I \end{cases} \tag{21}$$

Hybrid weighted similarity index for information spread factors based on node attributes

On the basis of the similarity index of node attributes, we consider the information spread factor, and construct the hybrid weighted similarity index for information spread factors based on node attributes to predict the cooperation relationship. Compared with the traditional single-link prediction index in the prediction of scientists' cooperation, the mixed weighted similarity index takes not only the nodes' own attribute information, but also the evolution in the network from a global perspective into account. We calculate the probability of collaboration between scientists from different perspectives to provide more accurate predictions. The calculation is shown in Equation (22).

$$Sim_{i,j} = \lambda * sim_{ij}^{(1)} + (1-\lambda) * sim_{ij}^{(2)}, 0 \leqslant \lambda \leqslant 1 \tag{22}$$

$sim_{ij}^{(1)}$ is the similarity index based on node attributes, $sim_{ij}^{(2)}$ is the similarity index based on information spread factors, λ is used to adjust the proportion of attribute similarity index and information spread factors similarity index, and $Sim_{i,j} = Sim_{j,i}$.

Simulation Experimental Design and Analysis

Data set division method

First, the known connected edge E is divided into two parts: the training set $E^T$and the test set E^P. Only the information in the training set can be used when calculating the fractional values, obviously $E = E^T \cup E^P$ and $E^T \cap E^P = \phi$. We define U as the full set of edges, here the edges belonging to U but not to E^T are called nonexistent edges, and the edges belonging to U but not to E^T are called unknown edges. There are various ways to divide the data set, and the division of the data set is described in detail in Zhu et al.[34]

To test and compare the performance of the link prediction algorithm based on spread factors, a portion of the simulated network needs to be selected as the training set for information spread. Since information is spread according to its neighboring nodes, the information cannot be spread to the set of unconnected nodes if the delineated training set is not connected. Combining the properties of propagation, we need to ensure that the training set remains connected after sampling. The specific process of dividing the test set and training set is as follows.

1. In this experiment, 90% of the edges in the simulated network are used as the training set E^P, and 10% of the edges are used as the test set E^P. The number of edges in E^P is calculated from the number of edges in the simulated network, which we denote as number$_1$.
2. We select a randomly connected edge e in the network and delete the connected edge e.
3. We need to determine whether the network after deleting this connection is fully connected. If the answer is yes, we put this connecting edge into the test set. If the answer is no, we put it back again and return to step 1.
4. Repeat steps 2–3 and monitor the number of edges in E^P. When the number of edges is equal to number$_1$, the test set and training set are divided.

Evaluation index

There are various measures of the accuracy of link prediction algorithms, and the common ones are AUC,[35] accuracy,[36] and ranking score.[37] In this article, AUC is chosen as a measurement tool for the accuracy of link prediction algorithms, and the AUC[38] is calculated as follows.

$$A \cup C = \frac{n' + 0.5n''}{n} \tag{23}$$

AUC can be understood as the comparison of the probability of a high score calculated by the link prediction algorithm for a randomly selected edge in the test set versus a randomly selected edge among nonexistent edges. So, the degree to which AUC is greater than 0.5 measures the degree to which the algorithm is more accurate than the method of randomly selecting connected edges.

Experimental simulation and analysis

Construction of cooperation simulation network for scientist. After analyzing the crawled scientist cooperation network, we construct a virtual network model. The goal is to construct a simulated network $G = (V, E)$ that matches the characteristics of the scientist cooperation network, where $V = \{v_1, v_2, v_3 \cdots v_n\}$ represents the set of nodes of the network G containing N nodes, and N is the number of nodes of the network G. E is the set of edges that represent the edges already present in the network, $E = \{(v_i, v_j) | i, j = 1, 2 \cdots N$ and $v_i \neq v_j\}$.

By observing the real scientist cooperation network, we can find that the scientist cooperation network is characterized by small-world phenomenon and scale-free property, and the degree distribution of the network obeys power-law distribution. However, the underlying small-world and scale-free networks cannot reflect the distribution characteristics of node attributes. Therefore, we construct the simulation network. The process of building the simulation network is shown below.

1. A network structure containing N nodes without concatenated edges is constructed, and each node follows the six virtual attributes of the real scientist cooperation network: research area, current institution, current title, number of publications, number of citations of published articles, and number of years of research time. These attributes are used for later studies.

2. The similarity of attributes between all nodes in the network is calculated and the values of similarity are ranked. The formula for calculating the similarity y between nodes is shown below, where $\Omega(i)$ represents the set of attributes of node i and $\Omega(j)$ represents the set of attributes of j.

$$y = \frac{|\Omega(i) \cap \Omega(j)|}{|\Omega(i) \cup \Omega(j)|} \tag{24}$$

3. Connected edges are created based on observing the sparsity of connected edges in the real network. We consider the possibility of cooperation between scientists in similar research fields and scientists in different fields in the real network. So, we add random edges to the set of edges with high similarity and randomly select a certain proportion of edges to be connected. We simulate the cooperation phenomenon in the real scientist network by controlling the proportion of similarity-linked edges and random-linked edges.

We use the above steps to construct the simulated network with the number of nodes $N = 3000$ and $N = 6000$, and visualize the network with Gephi software. We can see more intuitively the similarity between the simulated network and the real network by the visualized graph. In Figure 1, the same color of the nodes represents the nodes with high similarity to the research field and the institutions they belong to. The neighbors of some nodes are also marked with different colors to simulate the cooperation of some

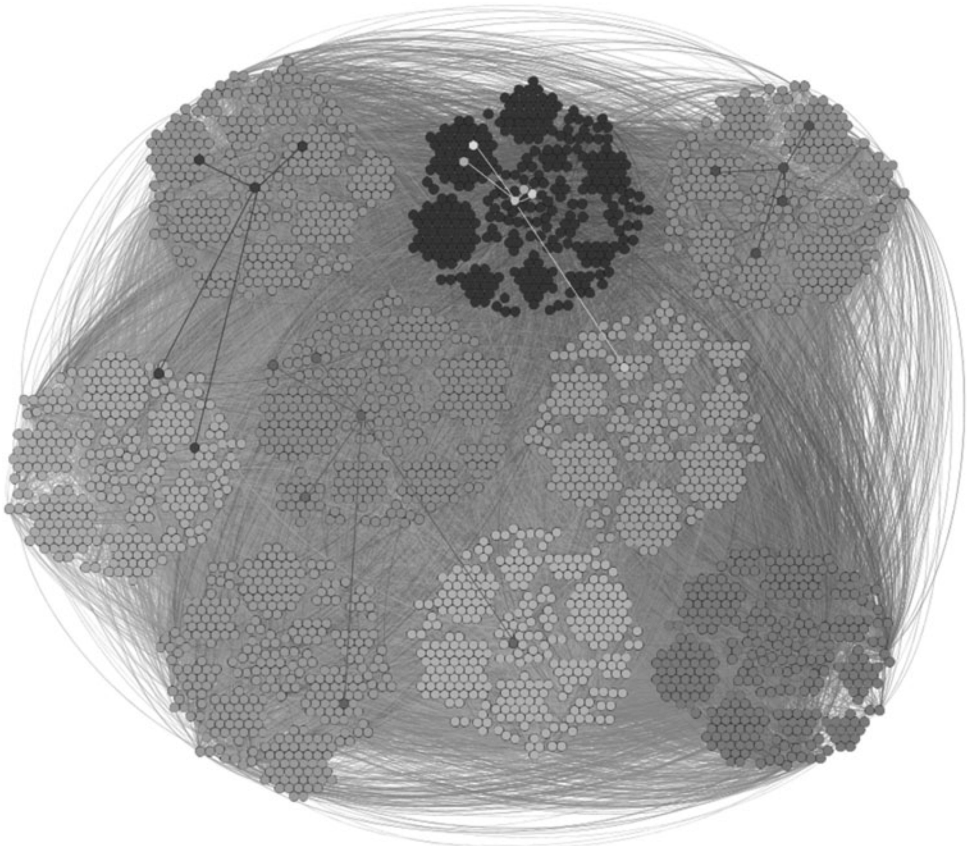

FIG. 1. Cooperation of nodes in the simulated network.

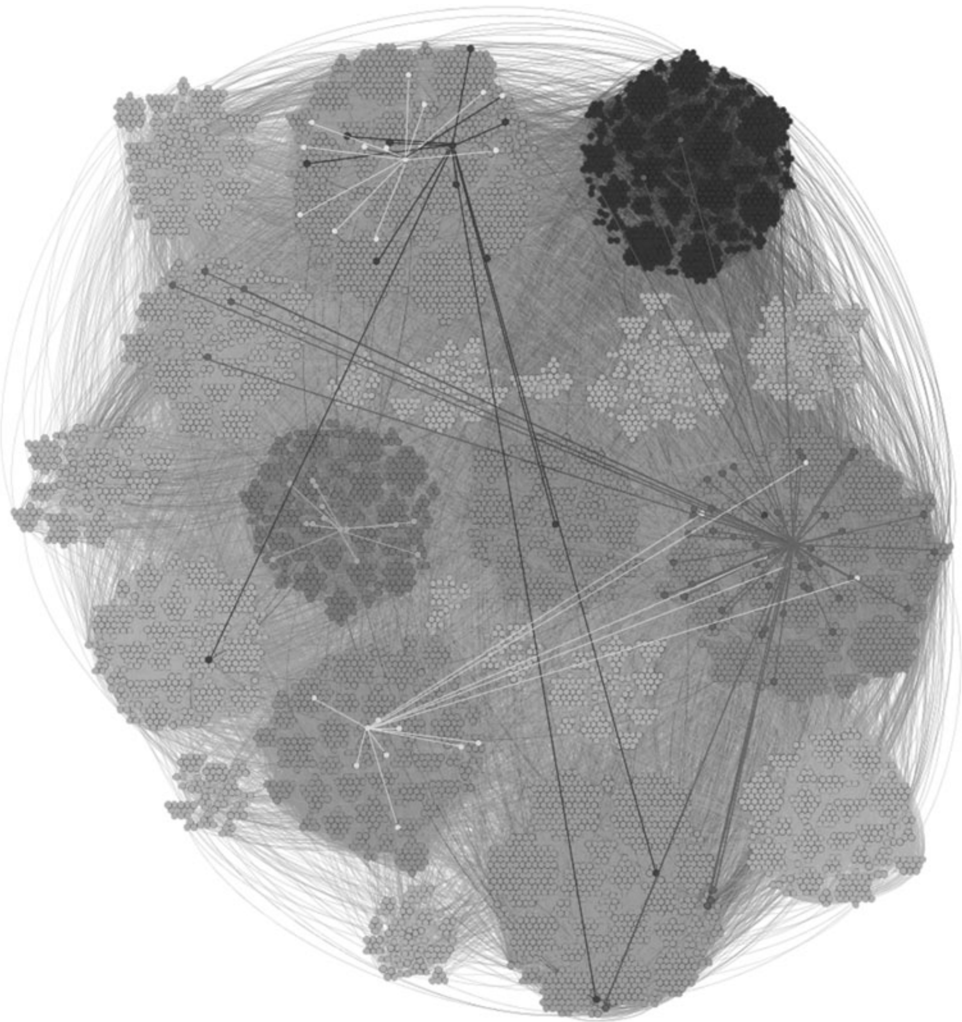

FIG. 2. Cooperation of some nodes in the real network.

scientists who have produced collaborative relationships in the real network. Nodes of the same color represent groups of scientists who have had collaborative relationships. Figure 2 shows the cooperation of the real network. It can be found that there are some scientists who collaborate in the same field and the similarity between them is relatively high.

There are some scientists who collaborate across fields and the similarity between them is relatively low. Therefore, it is found that the phenomenon of cooperation in the simulated network can be well represented in the network of scientists' cooperation by comparison. Figure 3 shows the degree distribution of the simulated network graph and the real scientist cooperation network with a power-law distribution.

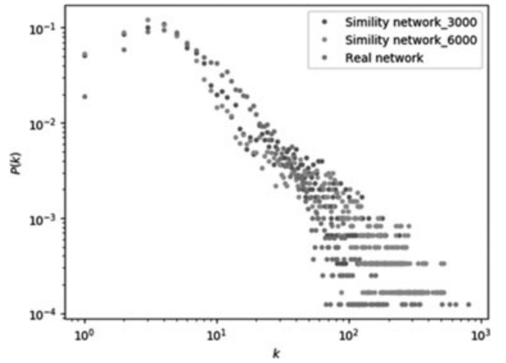

FIG. 3. Degree distribution of the real and simulated networks.

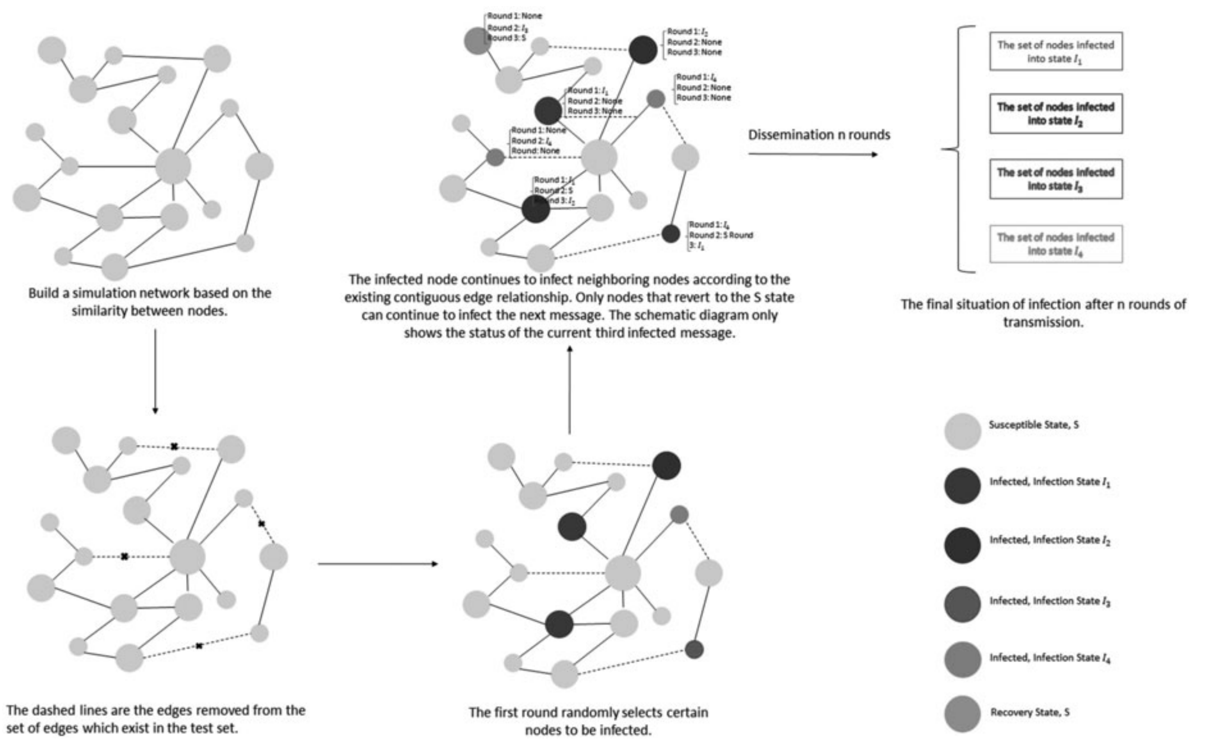
Round 1: None
Round 2: I_3
Round 3: S
Round 1: I_2
Round 2: None
Round 3: None
Round 1: I_4
Round 2: None
Round 3: None
Round 1: I_1
Round 2: None
Round 3: None
Round 1: None
Round 2: I_4
Round: None
Round 1: I_1
Round 2: S
Round 3: I_2
Round 1: I_4
Round 2: S Round
3: I_1
Dissemination n rounds
The set of nodes infected into state I_1
The set of nodes infected into state I_2
The set of nodes infected into state I_3
The set of nodes infected into state I_4
Build a simulation network based on the similarity between nodes.
The infected node continues to infect neighboring nodes according to the existing contiguous edge relationship. Only nodes that revert to the S state can continue to infect the next message. The schematic diagram only shows the status of the current third infected message.
The final situation of infection after n rounds of transmission.
The dashed lines are the edges removed from the set of edges which exist in the test set.
The first round randomly selects certain nodes to be infected.
Susceptible State, S
Infected, Infection State I_1
Infected, Infection State I_2
Infected, Infection State I_3
Infected, Infection State I_4
Recovery State, S

FIG. 4. Schematic diagram of simulated information spread in the training set.

We conclude that the constructed network is consistent with the small-world and scale-free characteristics embodied in the real network.

Simulating information spread in the network. After the test set and training set are divided, the multifactor-coupled scientist-based collaborative propagation model is simulated in the training set, and the simulation process is shown in Figure 4. Figure 5 shows the information infection density plot after simulating the propagation in the training set network. Since there are certain articles with more collaborating authors, we select the top four subauthors for simulation and statistics after observing the real network. Thus, we divide the authors into first, second, third, and fourth authors according to their cooperation, corresponding to the I_1, I_2, I_3, and I_4 states of the nodes. The remaining scientists who are not involved in information spread (no publications related to this field), the state of these nodes is S. Scientists will publish different articles with different ranks of authorship. Thus, in the simulation experiment, each node will be infected with one or more states.

The authorship of articles published as first, second, third, and fourth author derived from the simulation propagation is put into the sets $C1$, $C2$, $C3$, and $C4$. We get the nodes infected with different states and put the nodes infected with the same state into the same set after the end of the simulation spread.

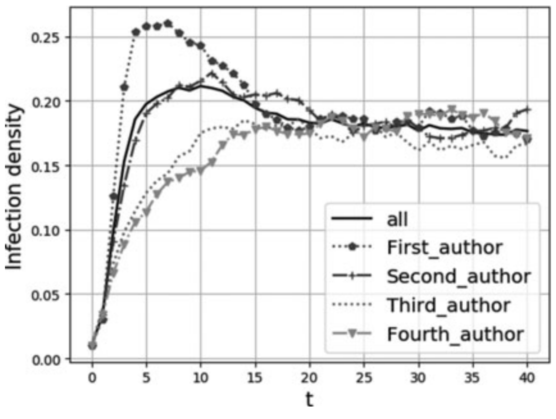

FIG. 5. Diagram of density change of information infection in the training set.

It can be seen from the information infection density diagram that all simulated infections have a tendency of first increasing and then decreasing, and then becoming stable after a period of time. This is because in the process of scientists' cooperation at the beginning, the high level of interest in the involvement of everyone in the publication of articles will reach a peak in the number of cooperations within a certain period of time due to the rise of a certain research area. After a period of time, scientists' cooperation becomes less influenced by the group effect, and thus, the share of individual scientist's influence becomes larger. Therefore, the situation of thesis cooperation gradually tends to be relatively stable.

Figure 6 shows the statistics of real network cooperations. Among the real scientist cooperation network, the research in related fields is still high due to the fact that most of the crawled scientists and published articles belong to communication- and computer-related fields. For the scientist cooperation network, the time span of scientists' cooperation to publish articles is long, and it may take a year or more from prepreparation to publication, and so, there is a certain delay in response to the information. Therefore, the cooperation trend that we observe for real networks is still on the rise. However, every research field will eventually be replaced by other new scientific technologies. So, we have reason to believe that after a few years, the number of publications related to this field will fall back to a relatively stable value in a fluctuating range.

Comparison of algorithm performance in the simulation network

1. Prediction of cooperation relationship based on similarity index of node attributes
   Before information spread, the fraction of connected edges between nodes in a simulated network with 3000 nodes is calculated using a similarity index based on node attributes. The distribution of attribute similarity values is shown in Figure 7.
   Using the AUC evaluation index, 672,400 independent comparison experiments are conducted and 10 sets of experiments are done to take the average. Finally, the AUC value is calculated by using the similarity index of node attributes, and the result is 0.7645.

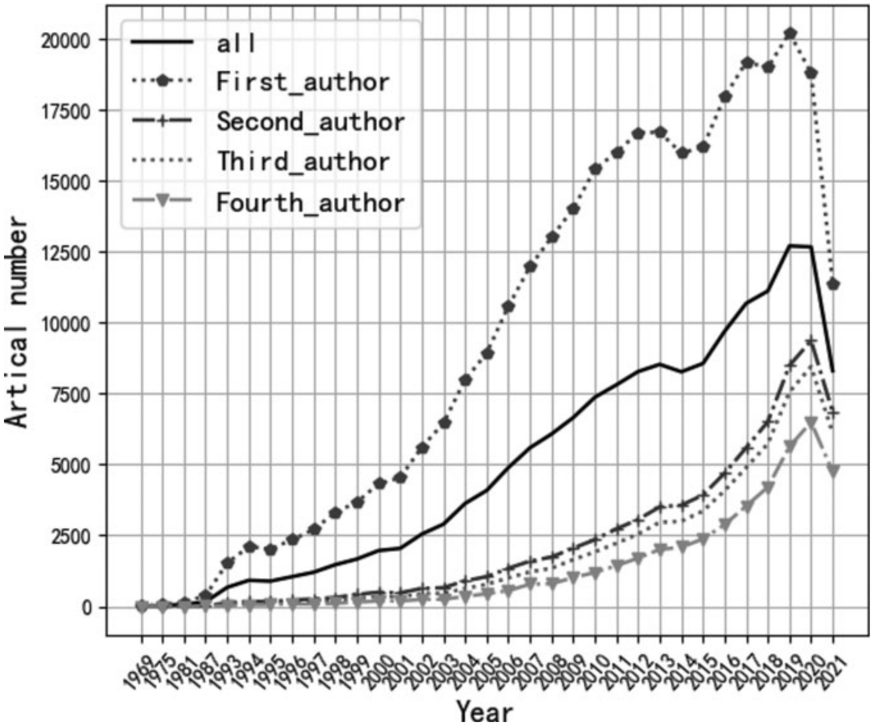

FIG. 6. Statistical chart of real network cooperation.

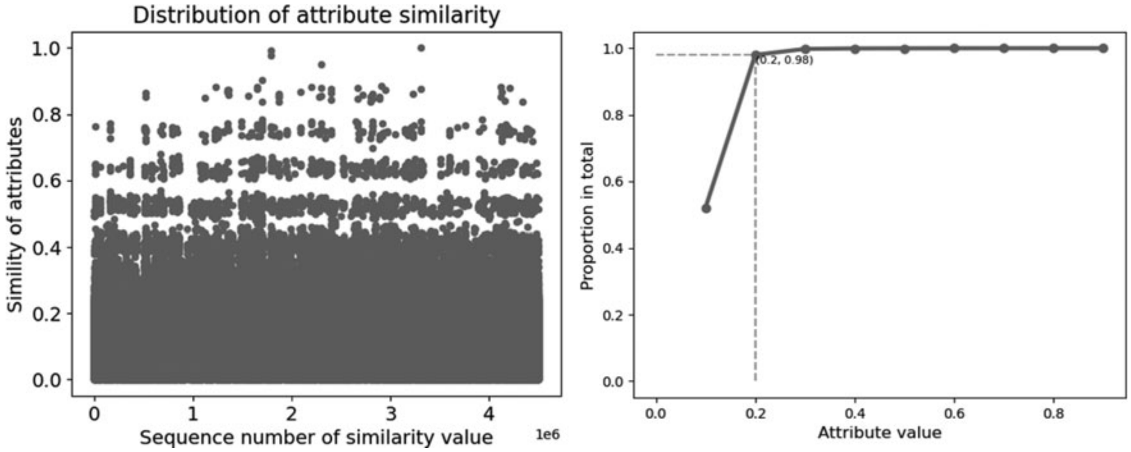

FIG. 7. Distribution of attribute similarity values in the simulation network.

2. Prediction of cooperation relationship based on similarity index of information spread factors

   After the simulation of network spread, the infection situation set of nodes is collected. We use the similarity index based on information spread factors to calculate the edge score between nodes in the simulation network with 3000 nodes. The average AUC is 0.6545, and the prediction effect of this algorithm needs to be improved obviously.
3. Prediction of cooperation relationship based on a mixed weighted similarity index of node attributes and information spread factors

   After the infection situation set of nodes is collected, how to add spread factors to the link prediction algorithm is the focus of our thinking. In this article, only this case of similarity between nodes in the set of infections with the same state is considered. Combined with the observed distribution of attribute similarity values, we consider the value of the weight occupied by the spread factors (ε) to be on the same order of magnitude as the value of the attribute. By observing Figure 7, it is found that 99% of the attribute values are less than 0.9, so the values of ε are set to [0.1, 0.9], respectively, for the experiments, and the percentage of spread factors in the algorithm is controlled by adjusting the parameter λ. The obtained experimental results are shown in Figure 8.

   We conducted 672,400 independent comparison experiments and monitored the variance of the accuracy during the experiments, thus ensuring that the experimental calculation error is within a reliable range. By adjusting the parameter λ, the trend change of AUC is drawn. When the value of λ is 0, it represents the case where only the node infection status (spread factors) is utilized, thus transforming the algorithm into one that only considers the spread factors, and the prediction accuracy is obviously very low. In addition, when the value of λ is 1, it represents the use of node attributes only to calculate the similarity between nodes, at which time the AUC value is 0.7645.

   It is found by Figure 8 that we can find the relatively optimal value to improve the prediction accuracy of the link by adjusting the values of ε and parameter λ. At this time, the AUC value is 0.8245, which corresponds to λ of 0.3 and $\varepsilon = 0.3$. Compared with the similarity calculated by using attributes only, the AUC value of 0.7645 is much improved. The simulation results show that adding spread factors into the link prediction algorithm in this way can improve the accuracy of the link prediction algorithm, so as to achieve more accurate recommendations for scientists to find partners. The above results show that the proposed link prediction algorithm considering spread factors is superior to the algorithm only considering node attributes under specific weights, but the importance of node attributes cannot be ignored.

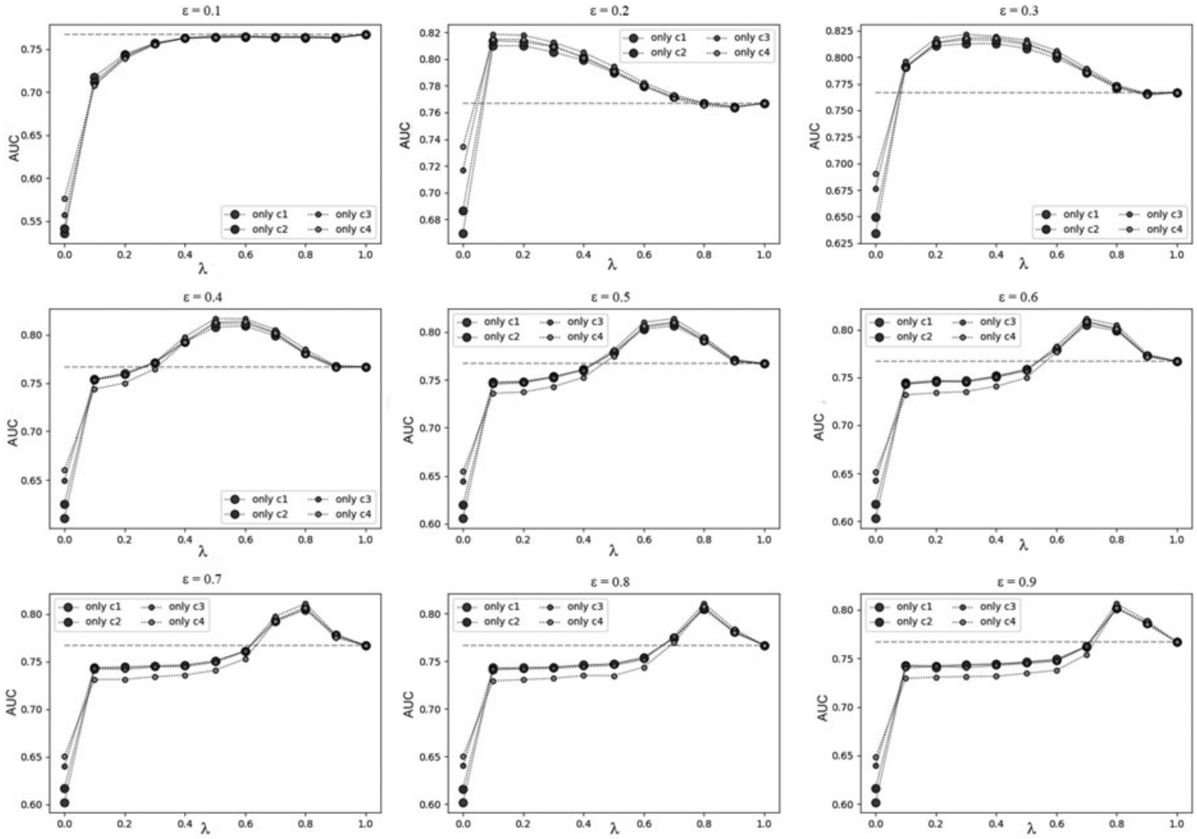

FIG. 8. The direction of AUC variation in the simulation network.

Experiments and Analysis of the Real Network

To verify that adding information spread factors to the link prediction algorithm can improve the correctness of the accuracy of the link prediction algorithm in the real scientist cooperation network. In this section, the effectiveness of the proposed hybrid weighted link prediction algorithm combined with information spread is verified by statistical data of real network and link prediction algorithm in a simulation experiment.

As the data set of the existing cooperation network of scientists in the internet rarely contains the attribute information of scientists, it is impossible to use the attributes of scientists to analyze. Therefore, we use the internet as a huge carrier and the web crawler technology to obtain data sets with node attributes. This website mainly records the communication engineering scientists and computer related scientists from 1990 to the present part of the cooperation.

The contents of the crawl include the author's coauthors, the articles produced by the collaboration, the publication time of the article, the order of the authors of the cooperative articles, the author's personal information, including nationality, current affiliation, and research area. Some scientists only have the ID (a unique identification of a scientist) value of the node, but there is no record of the attribute of the scientist. So we sift through the data and finally analyze the attributes of 7747 scientists and 104,736 collaborations between these scientists.

Figure 9 is a schematic diagram of the number of articles. The ordinate of the figure is the number of articles published, and the abscissa is the time when the articles are published. We can see when a certain event occurs by looking at the graph of the number of articles published. The number of articles published in the related field is sharply increasing, and then in the following years, the number of articles published increases more and more slowly and tends to fall back. By looking at the data in Figure 9, we can see that the number of publications fluctuates from year to year.

After the concept of 3G network was introduced, a large number of scientists flocked to the field of wireless

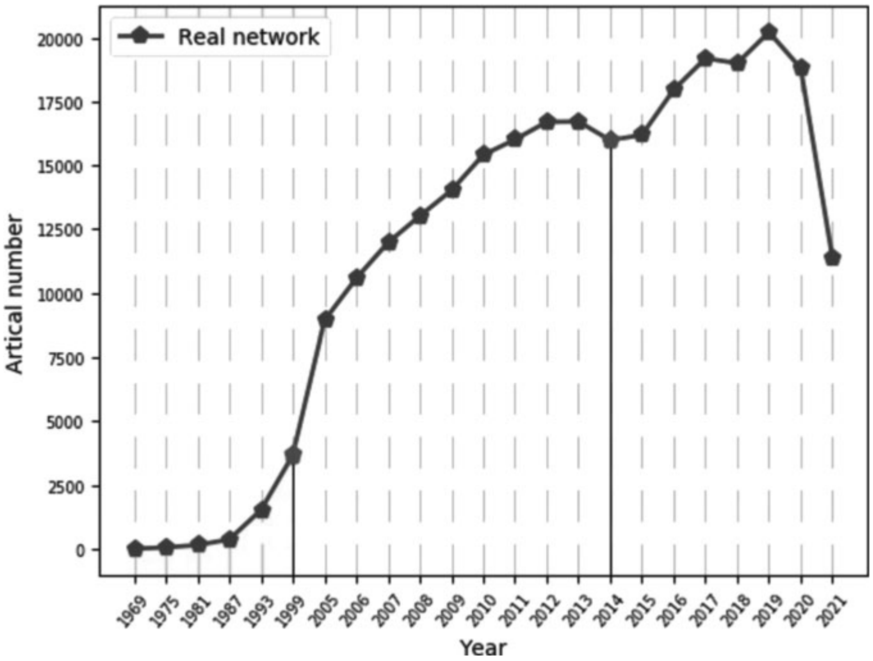

FIG. 9. Trend in the number of articles.

communication, and the number of publications increased substantially in the following years. On May 13, 2014, Samsung Electronics announced that it had pioneered the development of the first mobile transmission network based on 5G core technology. Therefore, we find that when the 5G wireless communication technology started to be proposed, the number of articles has increased substantially compared with the previous ones. This means that there are many scientists who are actively responding to the call to start researching about wireless communication technologies and collaborating to publish articles.

There is a tendency for the number of articles to decrease after the year 2000. Two reasons are analyzed as follows. One is that the number of articles published with wireless communication as the main content has decreased due to the impact of the global epidemic in 2020. The second point is that the wireless communication industry is steadily evolving and the discovery of new areas of relevance is becoming more difficult. This confirms that in the network of scientists' cooperation, the cooperation is influenced by the information spread factors.

Figure 10 provides a count of the number of authors who publish in different positions. By looking at the percentage of the number of each scientist who publish articles with the first seven authors, it is found that the majority of cases are involved in publishing articles as the first four scientists. Therefore, in the scientist cooperation network, only the nodes published by the previous four authors are selected as the nodes we consider to participate in information spread.

Figure 11 is the distribution diagram of the attribute similarity values of the edges in the real scientist network calculated according to the similarity index based on the node attributes. It can be observed that the connected edges with attribute similarity values below 0.09 are around 91%. Therefore, the value of ε is set in the range of [0.01, 0.09], and the variation of AUC is obtained by adjusting the percentage of spread factors as λ, as shown in Figure 12.

When the value of λ is 0, it represents the situation where only the similarity index of the information spread factors is used. When the value of λ is 1, it means that only the similarity index of node attributes is used to calculate the similarity between nodes, and the calculated AUC value is 0.8850. The situation of cooperation is added to the calculation of similarity, and the relative optimal value of AUC is calculated to be 0.8985, corresponding to the λ value of 0.7 and ε of 0.05. Adding spread factors to the real network and the accuracy of link prediction

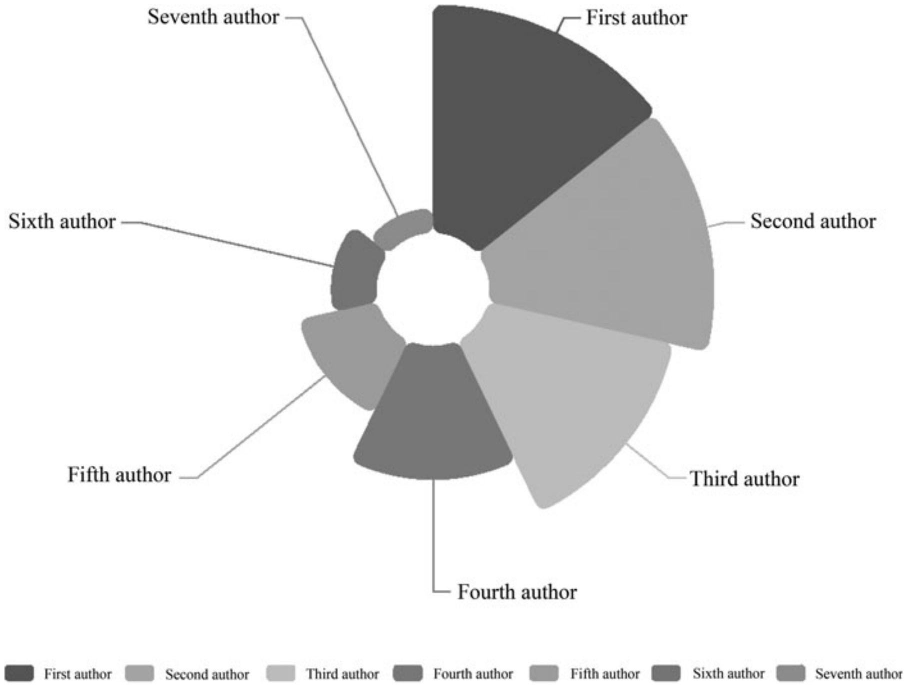

FIG. 10. Percentage of the number of scientists who published in the previous seven positions.

algorithms are improved compared with the link prediction algorithm without spread.

Although the improvement is not as large as that of the simulation network, it is because in the scientist cooperation network, the composition of the network and the formation of cooperation are not only affected by the attributes of the nodes, but also affected by many other factors, which also reflects the complexity of the network. However, through experimental analysis, it can be found that in the process of cooperation between scientists, the objects of cooperation between scientists tend to be close to their own research fields, and authors who have published articles with a higher ranking among previous scientific research results.

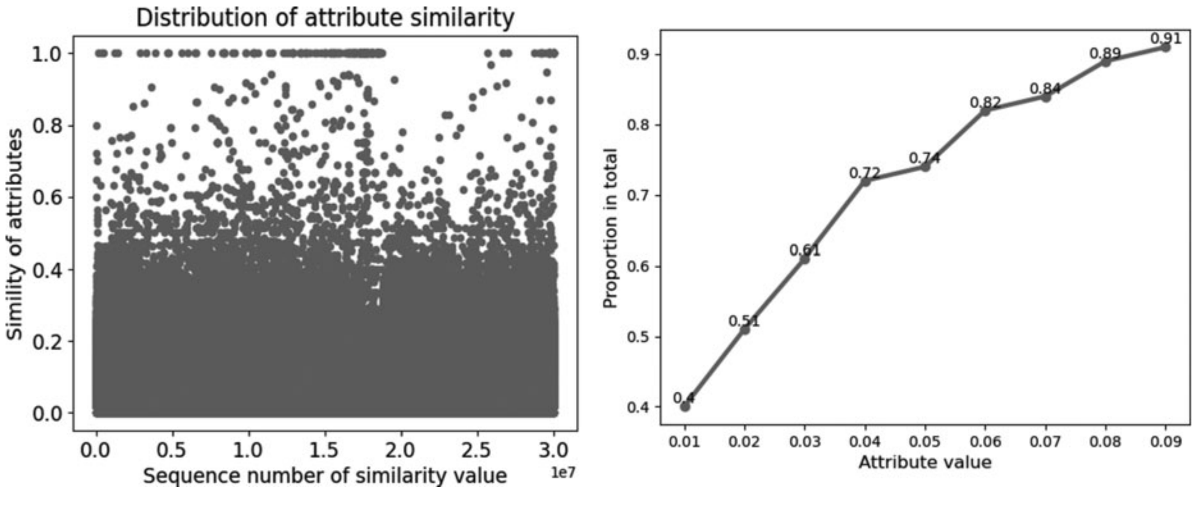

FIG. 11. Distribution of the attribute similarity values in the real network.

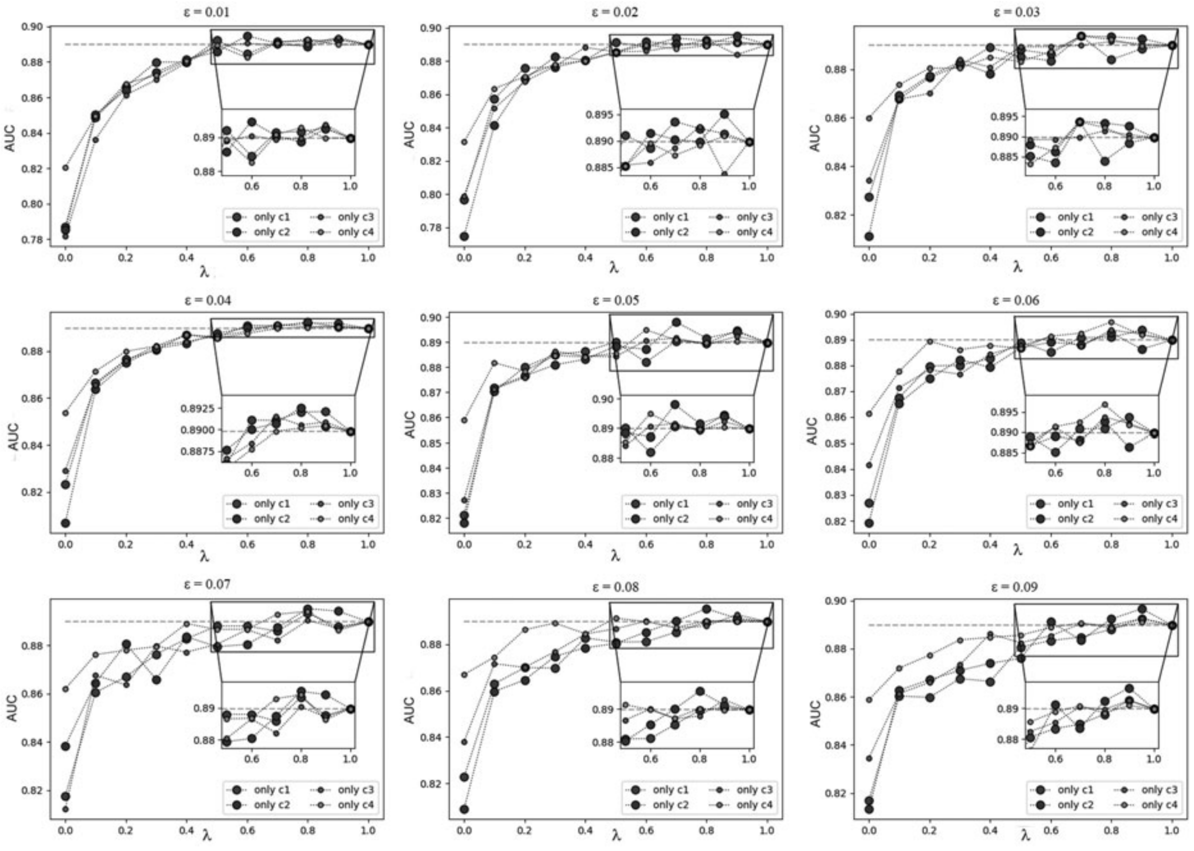

FIG. 12. Trend diagram of AUC changes in the real network.

Conclusion

This article studies the spread of information on the scientist cooperation network and the characteristics of the information in the process of spread. We add the influence of the information on the nodes in the network after the information spread into the link prediction algorithm, and design the hybrid weighted similarity index for information spread factors based on node attributes. It is proved that adding propagation factors into the link prediction algorithm will improve the accuracy of the link prediction algorithm, which has a certain theoretical significance and practical value for scientists to seek the recommendation of partners and predict cooperation.

In this article, a simulation network with small-world and scale-free characteristics is first constructed, and the improved SIS model is used to simulate the phenomenon of scientists' cooperation on the simulation network. The simulated cooperation is then further statistically and analytically analyzed, and the statistical results are added to the link prediction algorithm. We find that the accuracy of link prediction is improved by adding the information propagation on the network to the algorithm of link prediction. In the real network, the proposed idea is verified by analyzing the node attributes of the scientist cooperation network and the situation of the cooperation.

In the process of future research, the accuracy and universality of link prediction algorithms should be verified in networks with other characteristics. How to add the influencing factors of information spread into the link prediction algorithm and improve the accuracy of the link prediction algorithm is also a work that needs to be carefully studied in the future.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

This research is supported by the National Natural Science Foundation of China (No. 61863025), the Program for International S&T Cooperation Projects of

Gansu province (No. 144WCGA166), the Program for Longyuan Young Innovation Talents, and the Doctoral Foundation of LUT.

References

1. Karlovčec M, Mladenić D. Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. Scientometrics 2015;102(1):433–454; doi: 10.1007/s11192-014-1355-y.
2. Tang M, Mao X, Guessoum Z, et al. Rumor diffusion in an interests-based dynamic social network. ScientificWorldJournal 2013;2013:824505; doi: 10.1155/2013/824505.
3. Wang X, Zhou W, Li R, et al. Improving robustness of interdependent networks by a new coupling strategy. Physica A 2018;492:1075–1080; doi: 10.1016/j.physa.2017.11.037.
4. Wang X, Cao J, Li R, et al. A preferential attachment strategy for connectivity link addition strategy in improving the robustness of interdependent networks. Physica A 2017;483:412–422; doi: 10.1016/j.physa.2017.04.128.
5. Ma J, Zhu H. Rumor diffusion in heterogeneous networks by considering the individuals' subjective judgment and diverse characteristics. Physica A 2018;499:276–287; doi: 10.1016/j.physa.2018.02.037.
6. Wang X, Qin X. Asymmetric intimacy and algorithm for detecting communities in bipartite networks. Physica A 2016;462:569–578; doi: 10.1016/j.physa.2016.06.096.
7. Cui Y, Wang X. Detecting one-mode communities in bipartite networks by bipartite clustering triangular. Physica A 2016;457:307–315; doi: 10.1016/j.physa.2016.03.002.
8. Zhu H, Ma J. How the contact differences and individuals' similarity affect the rumor propagation process in complex heterogeneous networks. Int J Mod Phys C 2018;29(08):1850065; doi: 10.1142/S0129183118500651.
9. Ren G, Wang X. Epidemic spreading in time-varying community networks. Chaos 2014;24(2):023116; doi: 10.1063/1.4876436.
10. Yao H, Gao X. SE2IR invest market rumor spreading model considering hesitating mechanism. J Syst Sci Inf 2018;7(1):54–69; doi: 10.21078/JSSI-2019-054-16.
11. Wang X, Zhao T, Qin X. Model of epidemic control based on quarantine and message delivery. Physica A 2016;458:168–178; doi: 10.1016/j.physa.2016.04.0091.
12. Hosni AIE, Li K, Ahmad S. Minimizing rumor influence in multiplex online social networks based on human individual and social behaviors. Inf Sci 2020;512:1458–1480; doi: 10.1016/j.ins.2019.10.063.
13. Liu Z, Wang X, Wang M. Inhomogeneity of epidemic spreading. Chaos 2010;20(2):023128; doi: 10.1063/1.3445630.
14. Gebhart T, Funk RJ. The emergence of higher-order structure in scientific and technological knowledge networks. arXiv 2020; doi: 10.48550/arXiv.2009.13620.
15. Hui Z, Cai X, Greneche J, et al. Structure and collaboration relationship analysis in a scientific collaboration network. Chin Sci Bull 2011;56(34):3702–3706; doi: 10.1007/s11434-011-4756-9.
16. Zhou W, Wang X. Inhomogeneity of epidemic spreading with entropy-based infected clusters. Chaos 2013;23(4):043105; doi: 10.1063/1.4824316.
17. Wang X, Zhao T. Model for multi-messages spreading over complex networks considering the relationship between messages. Commun Nonlinear Sci Numer Simul 2017;48:63–69; doi: 10.1016/j.cnsns.2016.12.019.
18. Ren G, Wang X. Robustness of cooperation in memory-based prisoner's dilemma game on a square lattice. Physica A 2014;408:40–46; doi: 10.1016/j.physa.2014.04.022.
19. Lü L, Zhou T. Link prediction in complex networks: A survey. Physica A 2011;390(6):1150–1170; doi: 10.1016/j.physa.2010.11.027.
20. Altarelli F, Braunstein A, Dall'Asta L, et al. Bayesian inference of epidemics on networks via belief propagation. Phys Rev Lett 2014;112(11):118701; doi: 10.1103/PhysRevLett.112.118701.
21. Braunstein A, Ingrosso A, Muntoni AP. Network reconstruction from infection cascades. J R Soc Interface 2019;16(151):20180844; doi: 10.1098/rsif.2018.0844.
22. Choi E, Du N, Chen R, et al. Constructing disease network and temporal progression model via context-sensitive hawkes process. In: 2015 IEEE International Conference on Data Mining. IEEE: Atlantic City, NJ, USA, 2015; pp. 721–726; doi: 10.1109/ICDM.2015.144.
23. Daneshmand H, Gomez-Rodriguez M, Song L, et al. Estimating diffusion network structures: Recovery conditions, sample complexity soft-thresholding algorithm. International conference on machine learning 2014. PMLR 2014;32(2):793–801; doi: 10.48550/arXiv.1405.2936.
24. Duong Q, Wellman MP, Singh S. Modeling information diffusion in networks with unobserved links. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. IEEE: Boston, MA, USA, 2011; pp. 362–369; doi: 10.1109/PASSAT/SocialCom.2011.50.
25. Gomez-Rodriguez M, Leskovec J, Krause A. Inferring networks of diffusion and influence. ACM Trans Knowl Discovery Data 2012;5(4):1–37; doi: 10.1145/2086737.2086741.
26. Kamei T, Ono K, Kumano M, et al. Predicting missing links in social networks with hierarchical dirichlet processes. In: The 2012 International Joint Conference on Neural Networks (IJCNN). IEEE: Brisbane, Australia, 2012; pp. 1–8; doi: 10.1109/IJCNN.2012.6252619.
27. Netrapalli P, Sanghavi S. Learning the graph of epidemic cascades. ACM SIGMETRICS Perform Eval Rev 2012;40(1):211–222; doi: 10.1145/2318857.2254783.
28. Sefer E, Kingsford C. Convex risk minimization to infer networks from probabilistic diffusion data at multiple scales. In: 2015 IEEE 31st International Conference on Data Engineering. IEEE: Seoul, Korea (South), 2015; pp. 663–674; doi: 10.1109/ICDE.2015.7113323.
29. Tran L, Farajtabar M, Song L, et al. Netcodec: Community detection from individual activities. In: Proceedings of the 2015 SIAM International Conference on Data Mining (SDM). 2015; pp. 91–99; doi: 10.1137/1.9781611974010.11.
30. Shi H, Duan Z, Chen G. An SIS model with infective medium on complex networks. Physica A 2008;387(8–9):2133–2144; doi: 10.1016/j.physa.2007.11.048.
31. Nian F, Liu R, Cong A. An incentive mechanism model based on the correlation between neighbor behavior and distance. Int J Mod Phys C 2020;31(11):2050161; doi: 10.1142/S0129183120501612.
32. Nian F, Diao H. A human flesh search model based on multiple effects. IEEE Trans Network Sci Eng 2019;7(3):1394–1405; doi: 10.1109/TNSE.2019.2931943.
33. Xu J, Tang W, Zhang Y, et al. A dynamic dissemination model for recurring online public opinion. Nonlinear Dyn 2020;99(2):1269–1293; doi: 10.1007/s11071-019-05353-3.
34. Zhu Y, Lü L, Zhang Q, et al. Uncovering missing links with cold ends. Physica A 2012;391(22):5769–5778; doi: 10.1016/j.physa.2012.06.003.
35. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29–36; doi: 10.1148/radiology.143.1.7063747.
36. Herlocker JL, Konstan JA, Terveen LG, et al. Evaluating collaborative filtering recommender systems. ACM Trans Inf Syst 2004;22(1):5–53; doi: 10.1145/963770.963772.
37. Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation. Phys Rev E 2007;76(4):046115; doi: 10.1103/PhysRevE.76.046115.
38. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2005;27(8):861–874; doi: 10.1016/j.patrec.2005.10.010.

Cite this article as: Nian F, Qian Y, Yao Y (2023) Identification of potential cooperation relationships among scientists. *Big Data* 11:2, 87–104, DOI: 10.1089/big.2021.0398.