IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 23, 2022, accepted February 6, 2022, date of publication February 11, 2022, date of current version March 1, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3151112

Heterogeneous Attention Concentration Link Prediction Algorithm for Attracting Customer Flow in Online Brand Community

SHUGANG LI, BOYI ZHU, HE ZHU, FANG LIU, YUQI ZHANG, RU WANG, AND HANYU LU
School of Management, Shanghai University, Baoshan, Shanghai 200444, China

Corresponding author: He Zhu (zhuhe_star@163.com)

This work was supported by the Chinese National Natural Science Foundation under Grant 71871135.

ABSTRACT Attracting users from a mature large online product community to a new small one by friend recommendation is vital for new product marketing in social network. However, the traditional link prediction algorithms for friend recommendation cannot get high accuracy because of the network sparsity and scale-free problems when attracting customer flow between large and small circles. In order to better adapt to the link prediction of node pairs between circles of different sizes, we propose a collaborative combined link prediction algorithm (CCLPA), which can deeply extract user attention concentration (AC) features in sparse networks. CCLPA possesses three distinctive merits. Firstly, different edges in the network are assigned different attention, and heterogeneous attention concentration indexes (HACIs) within and beyond triadic closure structure are defined accordingly. Second, a random forest (RF) model is designed to adaptively select the appropriate HACIs for a given circle structure, so as to avoid the impact of scale-free problem on link prediction accuracy between different circles. Third, according to the collaboration of the selected indexes and their sensitivity to the circle structure, appropriate sensitive collaborative heterogeneous attention concentration index (SCHACI) is built to avoid the negative impact of blind combination of indexes on predicted performance. Experimental results on Twitter confirm the effectiveness of our proposed method in attracting customer flow in online brand community.

INDEX TERMS Accurately attracting customer flow, attention concentration, collaborative indicator, friend recommendation, HACI, link prediction, new product marketing.

I. INTRODUCTION

Users in social networks gather together due to common interests, hobbies, occupations, positions or careers to form a community (or cluster) [1], that is, circle. Marketers can use these circles to establish corresponding online brand communities [2]. A large number of studies have shown that friendships have a great influence on the willingness of consumers to join the brand community [3], intention to share e-word-of-mouth information, brand attitude [4] and actual purchase behavior [5]. Aral and Walker [6] studied the adoption of a movie App by 1.3 million Facebook users, confirming the influence of friendships on consumers' decision-making. When Facebook users used the movie App to share information and opinions about movies, actors, directors and movie industries, automatically generated notifications were sent randomly to users' friends on Facebook to remind them that their friends were using the movie App. The results showed that the product adoption rate had increased by 13%. The study also found that the variables reflecting relationship strength, such as the number of common Facebook pages and the number of common Facebook groups joined, were positively correlated with the product adoption rate.

When marketing new products, the marketing community established in the social network is generally on a smaller scale. To achieve a good marketing effect, it is necessary to establish friendships between users in the new product community and influential users in the community with mature large products. With the help of the influence of friendships, some users in the large-scale community are attracted to the new community of small-scale, that is, attracting customer flow. Through the transfer of customer flow, we can promote users in the mature product circle to buy new products, so as to realize the promotion of new products.

The associate editor coordinating the review of this manuscript and approving it for publication was Dongxiao Yu.

When user nodes in different circles establish links, they will face the problem of network sparsity, which is characterized by the average degree of the network is far less than the number of nodes [7]. Besides, prediction links between users in circles of large and small scales face the problem of scale-free [8], that is, degrees of nodes in large circles are larger, while the degrees of nodes in small circles are smaller. The major method for friend recommendation in social networks is the scoring link prediction algorithm (SLPA) [9]. However, most of these algorithms are fixed, rigid and based on large networks, which consider neither the different attention users assigning to each edge nor the self-adaptive construction of suitable SLPA for the specific circle structure. As a result, these SLPAs cannot deeply extract the possible friendships in sparse networks, and worse yet, their performances leave much to be desired when the degrees of node pairs fluctuate greatly.

Therefore, to fill the gap of previous studies, we consider the heterogeneity of attention concentration (AC) allocated by nodes to each edge and propose a collaborative combined link prediction algorithm (CCLPA) from the perspective of heterogeneous AC of users in triadic closure structure. Firstly, in order to fully describe the possibility of friendships between node pairs in sparse networks, considering that different edges in the network are assigned different attention, heterogeneous attention concentration indexes (HACIs) within and beyond triadic closure structure are constructed. Then, for the sake of overcoming the influence of scale-free problem on prediction accuracy, a random forest (RF) model [10] is designed to select suitable HACIs for the specific network according to the shortest distance between nodes and the AC of nodes. Next, we build the suitable composite HACI which can deeply extract user AC features in circle structure from three aspects. To begin with, three kinds of suitable HACIs are selected for each network, namely the most, second and third suitable HACIs. Secondly, the logistic regression (LR) model is developed to identify the appropriate collaborative HACIs for each suitable HACI, and subsequently, three collaborative suitable composite HACIs are built, which can avoid the negative impact of the blind combination of indicators on prediction performance. Finally, three collaborative suitable composite HACIs are merged into sensitive collaborative heterogeneous attention concentration index (SCHACI) according to their sensitivity to the circle structure, so as to deeply extract user AC features in sparse networks and accurately predict the possible links between small and large circles. According to the predicted links, users in small and large circles can be recommended to become friends, and subsequently, accurately attracting customer flow is achieved.

The remainder of this study is organized as follows: Section 2 introduces the link prediction; Section 3 offers friend recommendation; Section 4 explains the CCLPA; Section 5 offers the experimental design and the analysis of the results; Section 6 gives conclusions of this study.

II. LINK PREDICTION

Link prediction is the fundamental problem in social network analysis [11]. The existing link prediction techniques can be roughly divided into four categories: similarity-based method, probabilistic and maximum likelihood-based method, dimensionality reduction-based method and algorithm-based method [12]–[14].

Among these methods, the similarity-based method is the most widely studied, which can be applied to large-scale networks. However, the traditional link prediction methods based on local similarity are mostly based on large networks, and their performance in sparse networks is not satisfactory. Some scholars have proposed algorithms for sparse networks. Shang *et al.* [15] constructed heterogeneity index (HEI), homogeneity index (HOI) and heterogeneity adaptation index (HAI), and proposed a link prediction algorithm to solve the network sparsity and scale-free problems faced in link prediction, but the algorithm could only get good performance in regular tree networks with high heterogeneity. Zhang *et al.* [16] proposed a link prediction framework, AdaSim, by introducing an Adaptive Similarity function using features obtained from network embedding based on random walks. Experimental results showed that AdaSim was robust to different sparsities of the networks. Nguyen and Mamitsuka [17] transformed the link prediction problem into a binary classification problem, and used the kernel function to represent the potential network characteristics, so that the method could be extended to large-scale networks. They also proved that this method could be well applied to sparse networks. However, these algorithms did not abundantly consider the heterogeneity of the network, so they could not fully mine the information contained in the network, and could not ensure accurate prediction in some special cases (such as link prediction between circles of different scales proposed in this study).

In recent years, a few scholars have developed link prediction algorithms based on heterogeneous networks by considering the weights of edges and combining them with a variety of network features. For example, Ozcan and Oguducu [18] proposed a new method called multivariable time series link prediction for evolving heterogeneous networks by combining the node connection information, local similarity indicators and global similarity indicators of time, link and multitype relationships. Bütün *et al.* [19] proposed a new link prediction method combining with directed, weighted and time information of links based on the neighborhood-based link prediction method. Kuo *et al.* [20] devised a novel unsupervised framework to predict the opinion holder in a heterogeneous social network without any labeled data. Liu *et al.* [21] used three zero models to describe the topological structure and link weights of the network, and generated a general link prediction method by combining them. Aghabozorgi *et al.* [22] measured the similarity of nodes based on the recent activities of nodes and the weights of edges, and proposed a supervised link prediction method that

took network features and node similarities as its feature sets. Lü and Zhou [23] used local similarity indicators to estimate the possibility of links in the weighted network, including common neighbors (CN), Adamic ADAR indicators (AA) and resource allocation indicators (RA). However, the indicators selected by them could not prove that the prediction performance of weighted links was better than unweighted links. Shang *et al.* [24] found that shifting attention from the direct link weight between nodes to the link weight between nodes and common neighbors could improve the performance of the algorithm. Shang *et al.* [25] proved that the weight value of network structure and the number of common neighbors played an important role in link prediction. Similarly, the link prediction algorithm proposed in this study not only considers the weight of the link, but also fully considers the heterogeneous AC assigned by nodes to different edges when constructing the HACIs, which helps to fully mine the information contained in the network.

In network link prediction, we can get better algorithm performance by shifting attention from the direct links between nodes to the common neighbors of node pairs. For example, existing studies have proved that algorithms with more common neighbor effects could achieve better performance in Facebook network, Contact network and E-mail network. This showed that if two users had more common friends, they were more likely to establish friend relationships in the future [24]. Guimera & Sales-Pardo [26] reconstructed the network by observing the missing and false links of the network based on the impact of nodes on their common neighbors. Vallès-Català *et al.* [27] constructed prediction indicators through the common neighbors of node pairs, indicating that in order to pursue the best link prediction results, over fitting problems might occur. Lü *et al.* [28] proposed a local path index (LP index) based on the common neighbors of node pairs to estimate the possibility of links between nodes. A large number of simulation experiments on networks showed that LP index had higher efficiency and effectiveness than two widely used common neighbor indexes: CN and Katz index. In addition, there were many other indicators used to predict the links between node pairs based on their common neighbors in existing studies, such as average compute time (ACT), random walk with restart (RWR), matrix forest index (MFI), etc [14]. Based on this, we fully consider node pairs and their common neighbors when constructing indicators. Moreover, in order to further make full use of the large range of network information and mine the possible links in extremely sparse networks, we also consider the role of the neighbors of the common neighbors, that is, the indirect common neighbors, when constructing the HACIs.

Many scholars have proposed friend recommendation algorithms in social networks based on link prediction. For example, Cheng *et al.* [29] proposed an extensible friend recommendation framework, combined with seven information sources of personal characteristics, network structure characteristics and social characteristics. Chen *et al.* [30] combined social impact, used learning ranking technology to analyze user behavior, and proposed a learning-based recommendation method to recommend informative friends for users. Ma *et al.* [31] proposed local friend recommendation indexes and mixed friend recommendation indexes based on weak group structure. Yu *et al.* [32] applied algebraic connectivity to the existing Friends-of-Friends recommendation algorithms, and realized the relevance of recommendation and dissemination of content. Ghasemian *et al.* [33] proposed a stacking model, considered 203 link prediction algorithms, and applied them to 550 different real-world networks including social networks. They believed that the performance of the stacking model was significantly better than the single prediction model. Guimerà's study also confirmed the effectiveness of the stacking model [34]. However, a recent large-scale experiment by Muscoloni *et al.* showed that the performance of the above stacking model was not necessarily better than a carefully designed single algorithm [35].

Although the researchers have considered the heterogeneity of networks and believed that the weights of edges were different, they did not consider the differences of nodes' attention assigning to different edges. Therefore, it was not conducive to deeply extract user AC features in sparse networks. Based on the heterogeneous AC, we develop HACIs within and beyond triadic closure structure. Moreover, CCLPA is developed, which can construct suitable composite HACIs according to the specific network characteristics, and accurately predict the friendships between small and large circles.

III. NEW PRODUCT MARKETING BASED ON FRIEND RECOMMENDATION

A. NEW PRODUCT MARKETING IN THE BRAND COMMUNITY

$H(G, P)$ is adopted to represent the online brand community, where G denotes the user node set in the community, P represents the edges set, and the edge indicates the users' friendships. There are some product circles in the brand community, which are developed through their shared interests in products. Suppose there are two kinds of products Π_1 and Π_2 in brand community Ω. The circle of product Π_1 is a small circle established when marketing new products, and the circle of product Π_2 is a mature large circle. Define G_{Π_1} as the user set of product circle Π_1 and G_{Π_2} as the user set of product circle Π_2. In the activity of attracting customers from large to small circles, if product Π_1 is sold to users in G_{Π_2}, users in G_{Π_1} can be recommended to users in G_{Π_2} as friends [13].

Fig. 1 shows the schematic diagram of attracting new consumers from large to small circles of brand community Ω, in which Fig. 1(a) shows the initial network and Fig. 1(b) describes the network with predicted links. In Fig. 1(a), users belonging to G_{Π_1} are 1, 2, 3, 4, and users belonging to G_{Π_2} are 5, 6, 7, 8, 9, 10, 11, 12, 13. Assume that user 3 is a marketer, the purpose of friend recommendation is to make influential users in G_{Π_2} and their friends buy product Π_1.

Because node 8 has the largest degree, it is selected as the most influential node in G_{Π_2}. Then, based on the link prediction algorithm, the possibility of friendships between user 8 and 3 is predicted, and if possible, recommends user 3 to become the friend of user 8, as shown in Fig. 1(b). User 3 encourages user 8 and his friends to buy product Π_1 by the influence of friendships, and realizes accurately attracting customer flow from the mature product circle to the new one [13]. However, the challenge of attracting customer flow from large to small circles is how to overcome the network sparsity and scale-free problems. So, this study proposes CCLPA to solve these problems.

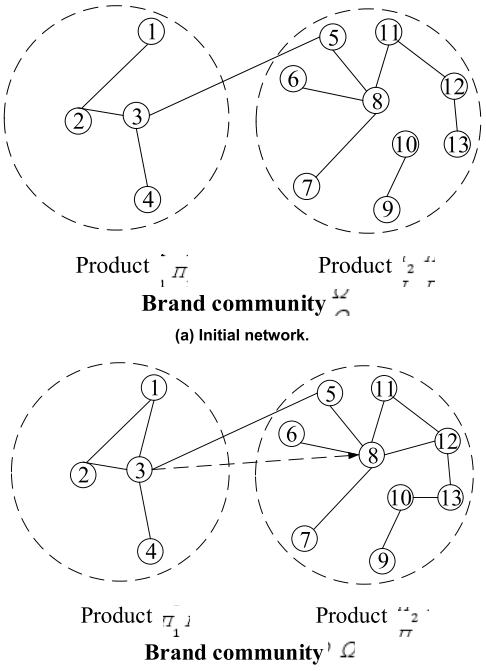

FIGURE 1. Friend recommendation diagram.

B. HACI FOR FRIEND RECOMMENDATION

Attracting users from a mature product community to a new one is an effective way of new product marketing. In order to overcome the problem of network sparsity when recommending friends among different scale circles, the hidden feature structure information in the sparse network is deeply extracted by constructing a variety of HACIs. Specifically, suppose that the degree of node i is k_i (node degree refers to the number of edges connected with the node), based on degree k_i, the attention assigned by user i to its each link in the social network is $1+\frac{\sigma}{k_i}$, where σ is a constant. Therefore, based on the above methods, different attention is assigned to each edge and HACIs are proposed.

Triadic closure structure was proposed in complex network researches by Newman [36], and it has been widely adopted by many scholars [14], [15], [37], [38]. It refers to social properties contained in the triple composed of three nodes X, Y and Z, as shown in Fig. 2, that is, if there is a connection between node pair (X, Y) and (X, Z), then it is easier for Y and Z to establish friendship. In order to fully describe the possibility of establishing friendships between node pairs in sparse networks, we construct HACIs within and beyond triadic closure structure. HACIs within triadic closure structure only contain the information of the node's direct common neighbors, and HACIs beyond triadic closure structure involve the information of the node's indirect common neighbors.

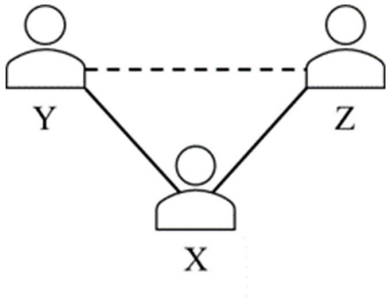

FIGURE 2. Triadic closure structure.

1) HACI WITHIN TRIADIC CLOSURE STRUCTURE

a: HACI BETWEEN NODE PAIRS AND THEIR DIRECT COMMON NEIGHBORS

PA, CN, Salton, Jaccard, Sorenson, HPI, HDI and LHN are redefined according to the heterogeneous AC, that is, considering that the attention assigned to various edges is different. And they are divided into two categories based on the network structure, one is HACIs between node pairs, the other is HACIs between node pairs and their direct common neighbors, as shown in Table 1.

In Table 1, the basic principle of the definition of HACIs is that the higher the AC of common neighbors is, the more possible the node pairs have friendships and the greater the scores of the node pairs are, vice versa. In Table 1, $\Gamma(\cdot)$ represents the neighbor sets of nodes, and $w(x,y)^{\sigma}$ indicates the AC assigned by node x to node y, namely $1+\frac{\sigma}{k_x}$. $s(x)=\sum_{y\in\Gamma(x)} w(y,x)^{\sigma}$ represents the AC of node x, namely the sum of the attention assigned by neighbors of node x.

b: HACI BETWEEN DIRECT COMMON NEIGHBORS AND NEIGHBORS OF NODE PAIRS

According to heterogeneous AC, we define the HACI between direct common neighbors and neighbors of node pairs, the formula is shown in Table 1.

2) HACI BEYOND TRIADIC CLOSURE STRUCTURE

Considering the sparsity of the local network caused by the connection between different scale circles, **HACIs between node pairs and their indirect common neighbors** are developed innovatively. Specifically, attention assigned to common neighbors by the friends of common neighbors will change the link of common neighbors, and then indirectly affect the attention that the common neighbors assign to the target node pairs.

a: TA1

TA1 represents attention that indirect common neighbors assign to node pairs. The fewer friends the indirect common neighbors have, the more attention they assign to their common neighbors, which indicating a higher score of AC, as shown in formula (1).

$$S_{xy}^{TA1} = \sum_{\Gamma(z), z \in \Gamma(x) \cap \Gamma(y)} \frac{w\left(\Gamma\left(z\right), z\right)^{\sigma}}{s(\Gamma\left(z\right))} \tag{1}$$

b: TA2

TA2 represents attention assigned to node pairs by direct and indirect common neighbors. The smaller the clustering coefficients are, the higher AC indirect common neighbors have. The fewer friends the common neighbors have, the more attention the common neighbors assign to node pairs, which indicating the higher AC, as shown in formula (2).

$$S_{xy}^{TA2} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w\left(z, x\right)^{\sigma} + w\left(z, y\right)^{\sigma}}{2 * s\left(z\right) * \left(\alpha * c\left(z\right) + \beta * \left(1 - c\left(z\right)\right)\right)} \tag{2}$$

where $c\left(z\right) = \frac{2 * e_z}{k_z * (e_z - 1)}$ is the clustering coefficient of node z, e_z represents the number of edges connected between the neighbors of node z. α and β are constants.

c: TA3

Obviously, the more friends the nodes have, the more attention is dispersed. In TA3, the attention dispersion of the nodes and the attention dispersion of the common neighbor nodes are combined, as shown in formula (3).

$$S_{xy}^{TA3} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \left(\frac{w\left(z, x\right)^{\sigma} + w\left(z, y\right)^{\sigma}}{2 * s\left(z\right) * \left(1 - c\left(z\right)\right)} + \frac{1}{s\left(x\right) * c\left(x\right)} + \frac{1}{s\left(y\right) * c\left(y\right)} \right) \tag{3}$$

d: RTA

In fact, the links of node pairs are usually affected by indirect and direct common neighbors and the AC of nodes in node pairs. Therefore, a combined HACI is obtained by integrating HACIs of direct and indirect common neighbors of users, so as to fully describe the characteristics of attention distribution between nodes, and overcome the shortcomings of low accuracy of prediction caused by the sparsity of the circle structure. In HACIs of node pairs and their direct common neighbors, RA* is considered in RTA because of its higher performance. In HACIs of node pairs and their indirect common neighbors, TA2 and TA3 are considered in RTA. Therefore, RTA is constructed, as shown in formula (4).

$$S_{xy}^{\mathrm{RTA}} = \delta * S_{xy}^{RA*} + \varepsilon * S_{xy}^{\mathrm{TA2}} + \theta * S_{xy}^{\mathrm{TA3}} \tag{4}$$

where δ, ε, θ are the weight of S_{xy}^{RA*}, S_{xy}^{TA2}, and S_{xy}^{TA3}, respectively.

IV. CCLPA

In this study, CCLPA is proposed from the perspective of heterogeneous AC of users in triadic closure structure. In order to describe the direct and potential friendships in sparse networks accurately, HACIs within and beyond triadic closure structure are constructed. Then, for the sake of overcoming the influence of scale-free problems, RF model is designed to select suitable HACIs according to the shortest distance between nodes and the AC of nodes. For the sake of deeply extracting user AC features in circle structure and improving the calculation efficiency of the algorithm, we build the suitable composite HACI from three aspects, namely select three kinds of suitable HACIs for each network by RF, construct three collaborative suitable composite HACIs using LR model, and build SCHACI by combining three collaborative suitable composite HACIs according to their sensitivity to the network structure, and subsequently, friendships between users in small and large circles are predicted. Fig. 3 shows the structure of CCLPA.

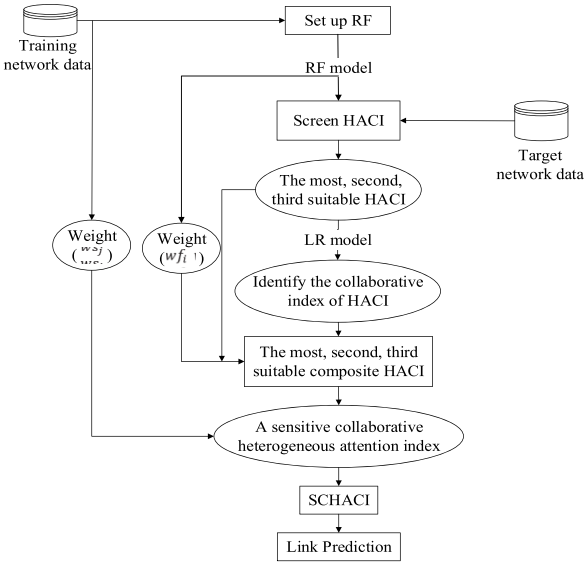

FIGURE 3. Structure of CCLPA.

A. ALGORITHM EVALUATION

The area under the curve (AUC) has unique sorting characteristics (only focusing on the sorting of indicators rather than the predicted value of the model), is insensitive to whether the positive and negative samples are balanced (the calculation method of AUC considers the classification ability of the learner for both positive and negative samples, and can still make a reasonable evaluation in the case of unbalanced samples), and can evaluate the link prediction method from an overall perspective. Therefore, AUC is widely used in the accuracy measurement of link prediction algorithm [15], [27], [39].

AUC value can be interpreted as the probability that randomly selected missing links get higher scores than nonexistent links. In this study, it can be defined in formula (5), where τ represents independent comparisons, τ' denotes the times of the linked node pairs having higher scores, and the larger AUC indicates the higher accuracy of the algorithm.

$$AUC = \frac{\tau' + 0.5(\tau - \tau')}{\tau} \tag{5}$$

TABLE 1. HACIs.

	HACI	Formula
HACI between node pairs and their direct common neighbors.	PA*	$S_{xy}^{PA*} = s(x) * s(y)$
	CN*	$S_{xy}^{CN*} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{w(z,x)^{\sigma} + w(z,y)^{\sigma}}{2}$
	Salton*	$S_{xy}^{Salton*} = \frac{\sum_{z\in\Gamma(x)\cap\Gamma(y)} w(z,x)^{\sigma} + w(z,y)^{\sigma}}{2 * \sqrt{s(x) * s(y)}}$
	Sorenson*	$S_{xy}^{Sorenson*} = \frac{\sum_{z\in\Gamma(x)\cap\Gamma(y)} w(z,x)^{\sigma} + w(z,y)^{\sigma}}{2 * (s(x) + s(y))}$
	HDI*	$S_{xy}^{HDI*} = \frac{\sum_{z\in\Gamma(x)\cap\Gamma(y)} w(z,x)^{\sigma} + w(z,y)^{\sigma}}{2 * max\{s(x), s(y)\}}$
	LHN*	$S_{xy}^{LHN*} = \frac{\sum_{z\in\Gamma(x)\cap\Gamma(y)} w(z,x)^{\sigma} + w(z,y)^{\sigma}}{2 * s(x) * s(y)}$
	RA*	$S_{xy}^{RA*} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{w(z,x)^{\sigma} + w(z,y)^{\sigma}}{2 * s(z)}$
HACI between direct common neighbors and neighbors of node pairs.	RAA*	$S_{xy}^{\mathrm{RAA}*} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{w(z,x)^{\sigma} + w(z,y)^{\sigma}}{2 * s(z)} * \max\{s(x), s(y)\}$
HACI between node pairs and their indirect common neighbors.	TA1	$S_{xy}^{TA1} = \sum_{\Gamma(z),z\in\Gamma(x)\cap\Gamma(y)} \frac{w(\Gamma(z),z)^{\sigma}}{s(\Gamma(z))}$
	TA2	$S_{xy}^{\mathrm{TA2}} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{w(z,x)^{\sigma} + w(z,y)^{\sigma}}{2 * s(z) * (\alpha * c(z) + \beta * (2 - c(z)))}$
	TA3	$S_{xy}^{\mathrm{TA3}} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \left(\frac{w(z,x)^{\sigma} + w(z,y)^{\sigma}}{2 * s(z) * (2 - c(z))} + \frac{1}{s(x) * c(x)} + \frac{1}{s(y) * c(y)}\right)$
	RTA	$S_{xy}^{\mathrm{RTA}} = \delta * S_{xy}^{RA*} + \varepsilon * S_{xy}^{\mathrm{TA2}} + \theta * S_{xy}^{\mathrm{TA3}}$

Fig 4 shows the example of the AUC calculation process. In Fig. 4(a), in order to test the accuracy of the algorithm, we need to select some existing links as probed links. For example, we select (1, 3) and (4, 5) as probed links, as shown by the dotted lines in Fig. 4 (b). The algorithm can only be trained by using the information contained in the solid lines in Fig. 4 (b). Assume that the algorithm assigns scores S_{12}, S_{13}, S_{14}, S_{34} and S_{45} to all unobserved links. Then, we need to compare the scores of probed links and non-existent links [14], and accordingly, calculate AUC using formula (5).

In this study, AUC is applied to three scenes: 1) After screening HACIs in RF model, the most, second and third suitable HACIs are determined according to the AUC value; 2) When identifying collaborative HACIs, if the combination of HACIs and other candidate HACIs has a better AUC value than the original HACIs, then the candidate indexes are considered as the collaborative indexes of the original HACIs; 3) When HACI and its collaborative HACI are combined, the combined weights are determined according to the AUC value.

B. SELECTING HACI BASED ON RF

The sparsity of network circles is diverse, which may lead to over-fitting of the model when identifying HACIs, that is, the model is too accurate to adapt to specific circles but cannot adapt to other circles reliably [27]. Because RF has high prediction accuracy and good tolerance for outliers and noise, and is not prone to over-fitting [40], it is selected to adaptively screen HACIs for specific circles.

Network Characteristic Indexes

In this study, in order to accurately describe the direct and potential friendships in the network, the following two indicators are considered from the perspective of the radiation range of the network itself and the indicators related to the heterogeneous AC.

1) INDICATORS RELEVANT TO THE SHORTEST DISTANCE BETWEEN NODES

a: NETWORK DIAMETER

The diameter of the network represents the maximum length of the path with the minimum resistance, which reflects the accessibility between nodes. In this study, the AC on the edge

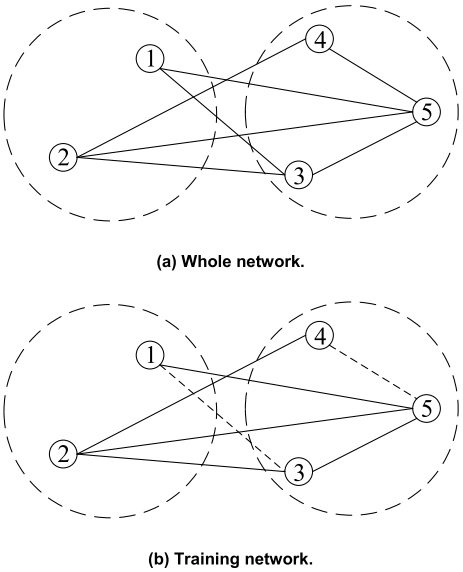

FIGURE 4. An illustration about the calculation of AUC value.

represents the close ties between nodes. The higher the AC is, the higher the accessibility between nodes is. The calculation of the network diameter is shown in formula (6).

$$d = \max_{i \neq j} L_{ij} \tag{6}$$

where $L_{ij} = \max_{i \neq j} l_{ij}$, $l_{ij} = \frac{1}{w_{ij}}$ represents the social distance of links between node i and j, w_{ij} represents the attention assigned by node i to the link between node i and j, if there is no direct path between node i and j, w_{ij} represents the sum of attention assigned to indirect paths and L_{ij} represents the length of the path with the minimum social distance between node i and j.

b: NETWORK EFFICIENCY

In the network, the smaller the social distance between nodes is, the smoother the communication between nodes will be, and the higher the network efficiency is. The calculation of network efficiency is shown in formula (7).

$$f = \frac{1}{Q \times (Q-1)} \sum_{i \neq j} \frac{1}{L_{ij}} \tag{7}$$

where Q represents the total number of nodes in the network.

2) INDICATORS RELEVANT TO THE AC OF NODES

a: AVERAGE NODE INTENSITY

The average node intensity represents the average AC allocated to each node in the network, as shown in formula (8).

$$\bar{s} = \frac{1}{Q} \sum_{i=1}^{Q} s_i \tag{8}$$

where $s_i = \sum_{j \in \Gamma(i)} w_{ji}$ indicates the AC of node i, w_{ji} represents the attention assigned by node j to the link between node i and j.

b: DEGREE HETEROGENEITY

Degree heterogeneity indicates the heterogeneity of AC of nodes in the network, as shown in formula (9).

$$h = \frac{1}{Q} \sum_{i=1}^{Q} (s_i - \bar{s})^2 \tag{9}$$

c: ASSORTATIVITY COEFFICIENT

The assortativity coefficient represents the matching characteristics of nodes' AC. If the assortativity coefficient is more than 0, then the network is called assortative, and nodes with similar AC tend to link with each other. If the assortativity coefficient is less than 0, then the network is heterozygous, and nodes with large AC differences tend to link with each other, as shown in formula (10).

$$r = \frac{M^{-1} \times \sum s_i s_j - \left[M^{-1} \times \sum \frac{1}{2}(s_i + s_j)\right]^2}{M^{-1} \times \sum \frac{1}{2}(s_i^2 + s_j^2) - \left[M^{-1} \times \sum \frac{1}{2}(s_i + s_j)\right]^2} \tag{10}$$

where M is the total number of connected edges in the network, s_i and s_j are the AC of node i and j, respectively.

RF Model

In RF, the independent variables are the network characteristic indicators, and the dependent variables are the HACIs with the largest, second and third AUC values in the training set. The structure of RF for classification is shown in Fig. 5. The RF classifier can be described as $h_i(x, \theta_i), i = 1, 2, \ldots, N$, where x represents the network characteristic index vector, N is the number of decision trees. θ_i represents the parameter vector of the i-th decision tree, which is the independent and identically distributed random vector determined by learning from the corresponding training set.

Because the CART decision tree has high classification accuracy and strong adaptability [41], we construct the decision tree based on the idea of CART algorithm, and the example of the CART decision tree is shown in Fig. 6. For the decision tree, it is assumed that the network training set is L and contains n samples. In each sample, HACIs with the largest, second and third AUC values are used as the label of the network. Considering the existence of K types of HACIs, we can obtain a partition of L as $\{E_1, E_2, \ldots, E_K\}$. The prior probability is $P_i = \frac{|E_i|}{|L|}$, and the Gini index used to classify L is $Gini(L) = 1 - \sum_{i=1}^{K} P_i^2$.

Use feature A (such as network efficiency) to divide the network in L, the sequence $\{A_1, A_2, \ldots, A_J\}$ can be obtained by sorting the value of feature A in ascending order. Define any $i\text{-}th (1 \leq i \leq J-1)$ segmentation point as $a_i = (A_i + A_{(i+1)})/2$, and divide L into two subsets $\{L_1, L_2\}$, where the value of feature A of the network in L_1 is $V(A, L_1) \in [A_1, a_i]$, similarly $V(A, L_2) \in (a_i, A_J]$. Corresponding to this division, the Gini index of attribute A is defined, as shown in formula (11).

$$Gini_{split}(A, a_i) = \frac{|L_1|}{|L|} Gini(L_1) + \frac{|L_2|}{|L|} Gini(L_2) \tag{11}$$

According to formula (11), calculate the Gini index of each partition point in the sequence $\{A_1, A_2, \ldots, A_J\}$. Select the

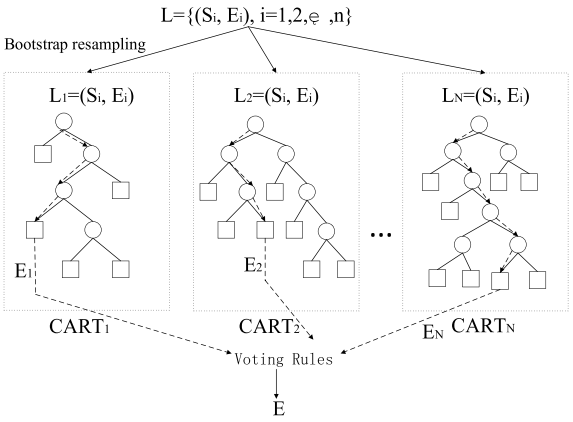

FIGURE 5. RF classification.

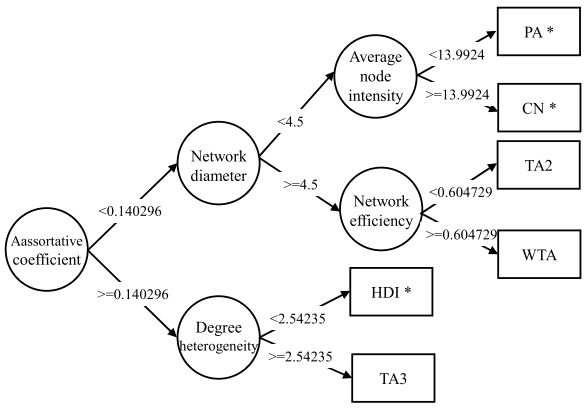

FIGURE 6. CART decision tree.

partition point with the smallest Gini index as the best branch threshold of attribute A, that is, $Threshold\,(A) = \min_{1\le i\le J-1}\{Gini_{split}\,(A, a_i)\}$.

To sum up, the steps of generating random forest are summarized as follows.

Step 1: Use bootstrap resampling technology to randomly retrieve n sample subsets from the same number of original training sample set, $L_j = \{(S_i, E_i)\,, i = 1, 2, \ldots n\}$, $j = 1, 2, \ldots, N$, where S_i and E_i represent the network characteristic set and the label of the i-th sample (i.e. HACI), respectively.

Step 2: For the j-th sample subset, calculate the network characteristic indexes related to the shortest path, degree and AC of nodes. Randomly select $I = int(\sqrt{k})$ indexes from k network characteristic indexes as the candidate segmentation feature subsets, calculate the partition Gini index of each candidate characteristic index, select the characteristic indexes with the smallest partition Gini index as the root node or higher-level node, and then use its best branch threshold to branch.

Step 3: Use the same method in step 2 to recursively establish tree branches for data subsets corresponding to branches with different characteristics, until all sample data of each branch belong to the same type of HACI.

Step 4: Repeat step 2 and step 3 in parallel to generate all N decision trees.

Step 5: Extract decision rules. For each decision tree generated in step 4, decision rules can be mined directly, that is, the most, second and third suitable HACIs for the circle can be identified according to the characteristic indexes of the community network.

Step 6: Classify the new sample according to the number of votes of the RF, and select the category with the first, second and third largest number of votes as the most, second and third suitable HACIs of the community, respectively. The voting rules are shown in formula (12).

$$E = arg\,max_{1\le j\le k}\sum_{i=1}^{N} Z(h_i\,(x, \theta_i) = E_j) \qquad (12)$$

where $h_i\,(x, \theta_i)$ is the classified result of the i-th decision tree, $Z(\cdot)$ is the indicator function, $Z(\cdot) \in \{0, 1\}$. When the i-th decision tree selects the HACIs of class E_j, $Z\,(\cdot) = 1$, otherwise $Z\,(\cdot) = 0$.

It should be noted that in the process of training RF model, the average AUC value wf_i of each HACI in the training set is taken as its weight, which is used as the weight of the combined model in section 4.4.

C. CHOOSING COLLABORATIVE HACI USING LR ALGORITHM

Since a single HACI cannot describe all network characteristics, three kinds of suitable HACIs identified by RF are combined with other HACIs to construct composite HACIs to describe the possibility of establishing friendships between nodes. Because the blind and random combination of each HACI will decrease algorithm accuracy, it is needed to screen out the collaborative HACIs. The principle of selecting collaborative HACIs is that if the AUC value of the combination of the suitable HACI and the candidate HACI is better than the suitable HACI, then the candidate HACI is considered to be its collaborative indicator. Because LR model is a classical binary algorithm, and it is very easy to achieve large-scale real-time parallel processing [42], we use LR model to identify collaborative HACIs.

Factors for Identifying Collaborative HACI

For a network sample, the total number of nodes in the network is O, and there are up to $O * (O - 1)/2$ node pairs in the network. HACIs C and D (such as PA* and TA2) are used to calculate the scores between all node pairs in the network, respectively, two scoring sequences can be obtained as $(C_1, C_2, \ldots, C_{O*(O-1)/2})$, $(D_1, D_2, \ldots, D_{O*(O-1)/2})$. In order to fully describe the collaborative relationship between HACIs, the similarity distance between any two HACIs (i.e. C and D) is described from the following three dimensions.

(1) The difference between two HACIs, which can be described by the Hamming distance and Jaccard distance, as shown in formula (13) and (14), respectively.

Hamming distance of HACIs C and D is shown in formula (13).

$$d_{Hamming} = \frac{1}{O * (O - 1)/2}\sum_{i=1}^{O*(O-1)/2}(C_i \oplus D_i) \qquad (13)$$

where if $|C_i - D_i| > \varphi$, then C_i is equal to D_i, otherwise they are different, φ is a constant.

The Jaccard distance of HACIs C and D are in formula (14), as shown at the bottom of the page.

(2) The similarity distance between HACIs, which can be represented by the square of Euclidean distance ($d_{squaredeuclidean}$) and Minkowski distance ($d_{minkowski}$), as shown in formula (15) and (16), respectively.

$$d_{squaredeuclidean} = \sum_{i=1}^{O*(O-1)/2} (C_i - D_i)^2 \quad (15)$$

$$d_{minkowski} = \left(\sum_{i=1}^{O*(O-1)/2} |C_i - D_i|^p \right)^{\frac{1}{p}} \quad (16)$$

where p is a constant.

(3) The degree of deviation between two HACIs, which can be described by the mean absolute difference (d_{MAD}) and mean square error (d_{MSE}), as shown in formula (17) and (18), respectively.

$$d_{MAD} = \frac{1}{O*(O-1)/2} \sum_{i=1}^{O*(O-1)/2} |C_i - D_i| \quad (17)$$

$$d_{MSE} = \frac{1}{O*(O-1)/2} \sum_{i=1}^{O*(O-1)/2} (C_i - D_i)^2 \quad (18)$$

LR Model

The training sample set is *[Hi, Ni], $i = 1, 2, \ldots, l$*, where the independent variable *Hi* is the above six similarity distances between the two HACIs in sample i, the dependent variable *Ni* indicates whether the two HACIs are collaborative. If the two HACIs are collaborative, *Ni = 1*, otherwise *Ni = 0*. The conditional probability $P(N_i = 1|H_i)$ in LR represents the probability that the two HACIs are collaborative, as shown in formula (19).

$$P(N_i = 1|H_i) = \frac{1}{1 + e^{-g(Hi)}} \quad (19)$$

where $g(Hi) = \beta_0 + \beta_1 H_{i1} + \beta_2 H_{i2} + \ldots + \beta_6 H_{i6}$, $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_6)$ are the regression coefficients of the independent variable, which can be obtained by the maximum likelihood estimation method, and when the probability is greater than 0.5, it indicates that there is a link.

D. SCHACI

In order to improve the performance of the combined HACIs, the average AUC value for the HACI calculated in section 4.2 is used as its weight to strengthen the role of high-performance HACI in link prediction when combining the three categories of suitable HACIs E_1, E_2 *and* E_3 identified by RF, with their collaborative HACIs (i.e. F_{ij}) screened by LR to construct the combination index P_i, as shown in formula (20).

$$P_i = \left[wf_i, wf_{i1}, wf_{i2}, \ldots, wf_{ij}\right] * \left[E_i, F_{i1}, F_{i2}, \ldots, F_{ij}\right]^T, \quad (i = 1, 2, 3, j = 1, 2, \ldots, l) \quad (20)$$

where wf_i indicates the weight of E_i, that is, its average AUC value. Similarly, wf_{ij} represents the weight of F_{ij}.

To further improve the performance of the combination model, it is necessary to assign different weights to the three types of combination indexes and combine them into a new composite index by linear combination. When determining the weight, in order to avoid the bias and one-sidedness caused by subjective experiences, and objectively reflect the different influence degrees of each composite index, the sensitivity coefficient method is adopted in this study to determine the weight of each combination index (i.e. ws_j) based on the training set, as shown in formula (21).

$$ws_j = \frac{F_j}{\sum_{i=1}^{3} F_i} + \Psi_j \quad (j = 1, 2, 3) \quad (21)$$

where $F_i = \frac{\lambda_i}{\bar{u}_i} (i = 1, 2, 3)$ represents the sensitivity coefficients of the three combined indexes, $\bar{u}_i = \frac{1}{B} \sum_{b=1}^{B} u_{bi} (i = 1, 2, 3)$ depicts the mean values of the composite indexes in formula (20), u_{bi} means composite index i in sample b, and B is the number of samples in the training set, and $\lambda_i = \sqrt[2]{\frac{1}{B-1} \sum_{b=1}^{B} (u_{bi} - \bar{u}_{bi})^2} (i = 1, 2, 3)$ denotes the standard deviations of the three composite indexes calculated from training samples, Ψ_j is a parameter.

The larger the sensitivity coefficient of composite index is, the greater the role it plays in the composite index. Based on this, SCHACI is proposed, as shown in formula (22).

$$S = [ws_1, ws_2, ws_3] * [P_1, P_2, P_3]^T \quad (22)$$

V. EXPERIMENTS AND RESULTS ANALYSIS

A. EXPERIMENTAL DESIGN

The purpose of this study is to recommend users in small product circles to users in large mature product circles, if there are possible friendships, then become fans of users in small circles, and realize the sales of new products to users in mature product circles. This is very similar to the directed relationship between users in the Twitter. Therefore, we use the Twitter directed data set obtained from ''Stanford Network Analysis Project'' website (http://snap.stanford.edu/data/ego-Twitter.html) to verify CCLPA. It should be noted that the HACIs constructed in this study are based on undirected network, which can be used not only for one-way link prediction between nodes, but also for two-way link prediction. Therefore, the proposed HACIs are fully applicable to Twitter.

Through 971 data sets in Twitter, we verified the validity of CCLPA. Each ego-network in the data set indicated an online brand community, in which the central node denoted the brand enterprise, and users were clustered into various product circles. Selected 300 networks with product circles from Twitter. In each experiment, 240 networks were randomly selected from 300 networks as training sets and the

$$d_{Jaccard} = 1 - \frac{\left|\left(C_1, C_2, \ldots, C_{O*(O-1)/2}\right) \cap \left(D_1, D_2, \ldots, D_{O*(O-1)/2}\right)\right|}{\left|\left(C_1, C_2, \ldots, C_{O*(O-1)/2}\right) \cup \left(D_1, D_2, \ldots, D_{O*(O-1)/2}\right)\right|} \quad (14)$$

TABLE 2. Statistical information of network samples in Twitter.

		Twitter	
Statistical indicators	Minimum	Mean	Maximum
Number of nodes	17	40.7600	114
Number of edges	102	266.7367	585
Mean joint strength	3.0750	13.3630	28.6111
The efficiency of the network	0.2703	0.6606	0.9433
Average number of node interfaces	2.7200	37.0155	225.2456
Average clustering coefficient	0.1223	0.6576	0.9171
Assortativity coefficient	-0.5794	-0.0972	-0.5686
The degree of heterogeneity	0.3183	3.2397	8.8716
Unicom group number	1	1.1400	4
Maximum unicom group size	17	40.2367	114

TABLE 3. Abbreviation of algorithm with parameters.

Algorithm	Parameters	Algorithm	Parameters
TA2a	$\alpha = 2$, β=7	RTAf	$\alpha = 7$, β =2, δ =7, ε =2, θ=2
TA2b	$\alpha = 2$, β=2	RTAg	$\alpha = 2$, β =7, δ =2, ε =7, θ=2
TA2c	$\alpha = 7$, β=2	RTAh	$\alpha = 2$, β =2, δ =2, ε =7, θ=2
RTAa	$\alpha = 2$, β =2, δ =2, ε =2, θ=2	RTAi	$\alpha = 7$, β =2, δ =2, ε =7, θ=2
RTAb	$\alpha = 2$, β =2, δ =2, ε =2, θ=2	RTAj	$\alpha = 2$, β =7, δ =2, ε =2, θ=7
RTAc	$\alpha = 7$, β =2, δ =2, ε =2, θ=2	RTAk	$\alpha = 2$, β =2, δ =2, ε =2, θ=7
RTAd	$\alpha = 2$, β =7, δ =7, ε =2, θ=2	RTAl	$\alpha = 7$, β =2, δ =2, ε =2, θ=7
RTAe	$\alpha = 2$, β=2, δ =7, ε =2, θ=2		

other 60 networks as test sets. In order to test the efficiency of CCLPA on attracting customer flow, the links between the nodes in the large and small circles were predicted in each network. Table 2 shows the mean, minimum and maximum of the statistical indicators of the selected network samples.

For a brief description, there are 25 algorithms shown in Table 1 and Table 3. Table 1 shows HACIs without parameters and Table 3 shows HACIs with specific parameter values. These parameters are obtained from a large number of experimental results, and the indexes using these parameters have high performance and robustness. In the formula of CCLPA, $\Psi_1 = -0.2, \Psi_2 = -0.2, \Psi_3 = 0.2, \varphi = 0.001$, while keeping the other parameters at default value.

To further reveal the predictive performance of CCLPA, we took the existing link prediction methods based on local similarity as the references. In addition, other combined algorithms based on LR were proposed as the benchmark methods for CCLPA, where 8 HACIs within triadic closure structure in Table 1 were used to predict the links between node pairs. All algorithms were applied in MATLAB with default settings.

Table 4 demonstrates the average AUC of all algorithms proposed in this study in 100 random experiments. Fig. 7 displays the performance comparison of different HACIs. Fig. 8 shows the performance comparison between single RF and non-combined HACIs. Table 5 and Fig. 9 show the average AUC of CCLPA and all reference methods. Among them, single RF denotes the optimal HACI selected by RF, CCLPAa represents the collaborative indicator, which combines the most, second and third suitable composite collaborative indicators with weights, i.e. SCHACI, and CCLPAb means directly combining these composite collaborative indicators without weights. CCLPAc indicates the weighted composite index, which is combined with the most suitable index and the collaborative index with high performance, CCLPAd means combining the indicators in CCLPAc without weights. CCLPAe represents the weighted composite index, which integrates the most suitable index and all collaborative indexes. Table 6 and Fig. 10 show the performance comparison between the newly defined HACIs and the original SLPA indexes.

B. ALGORITHMS PERFORMANCE COMPARISON

Table 5 and Fig. 9 display that CCLPA achieves the highest performance among various algorithms, that is, when accurately attracting customer flow in the online community, the CCLPA can accurately recommend friends between small and large circles of online community. Besides, the accuracy of CCLPAa is 0.918919 and the accuracy of LR is 0.712816.

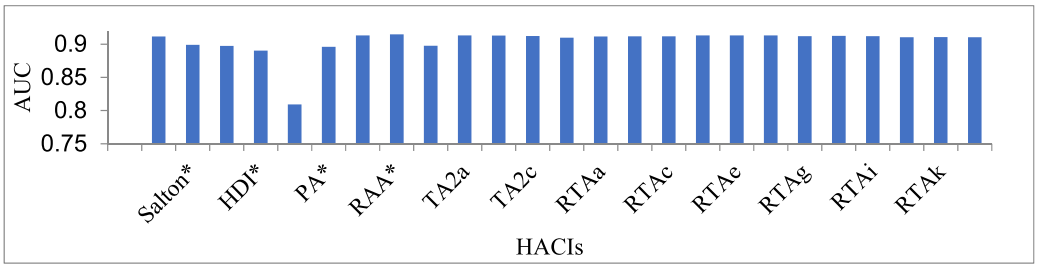

FIGURE 7. Performance comparison of all HACIs.

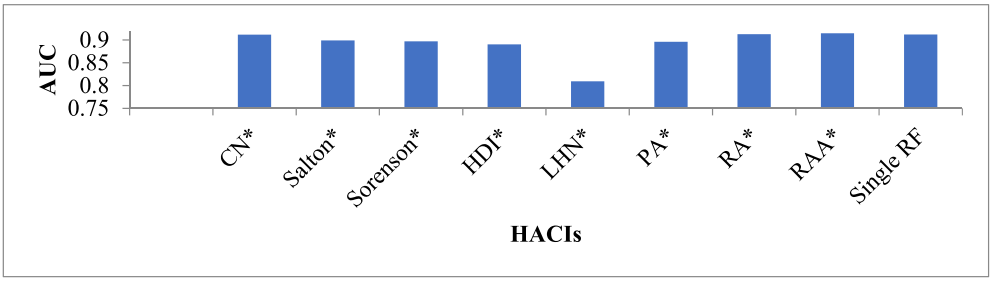

FIGURE 8. Performance comparison between single RF and non-combined HACIs.

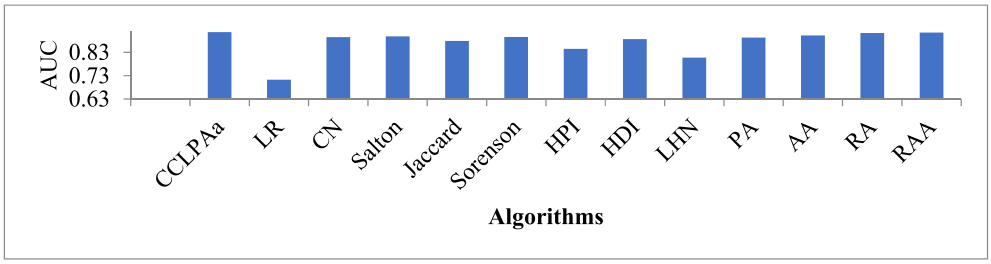

FIGURE 9. Performance comparison between CCLPAa and reference algorithms.

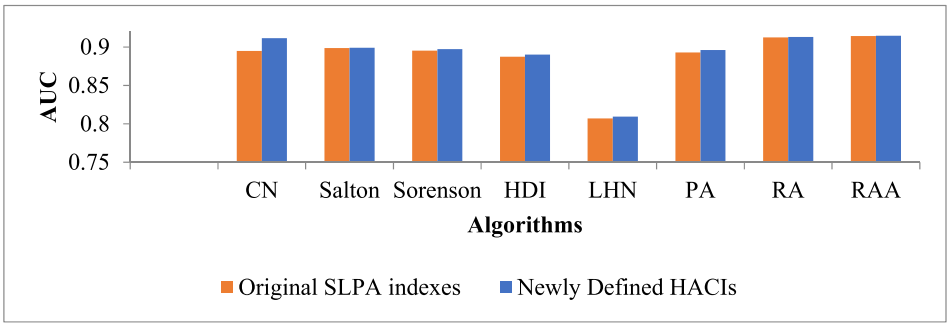

FIGURE 10. Performance comparison between the newly defined HACIs and the original SLPA indexes.

It can be seen that the CCLPA framework proposed in this study is far better than the LR framework.

Table 6 and Fig. 10 verify that the newly defined HACIs in this study perform better than those in the original SLPA, which displays that from the perspective of heterogeneity AC, the hidden feature structure information in the sparse network can be extracted in-depth, and effectively overcome the scale-free problem of recommending users in large circles to small circles.

At the same time, Fig. 7 shows that the accuracy of 19 kinds of algorithms from RA* to RTAl is significantly higher than that of the other 7 kinds of HACIs. In addition to RA*, these optimal HACIs are all newly proposed, which demonstrates that the proposed HACIs within and beyond triadic closure structure based on the principle of heterogeneity AC on the edges can effectively overcome the network sparsity problem in predicting the user friendships between various circles.

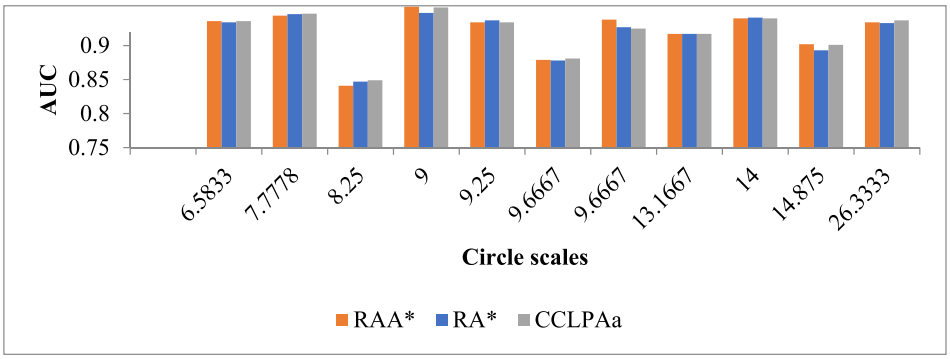

FIGURE 11. Algorithms performance in various circle scales.

TABLE 4. Performance of all algorithms ($\sigma = 1.0$).

Algorithm	Average AUC	Algorithm	Average AUC
CN*	0.911637	RTAd	0.913180
Salton*	0.899254	RTAe	0.913220
Sorenson*	0.897392	RTAf	0.913174
HDI*	0.890239	RTAg	0.911959
LHN*	0.809052	RTAh	0.912389
PA*	0.896363	RTAi	0.912025
RA*	0.912902	RTAj	0.910553
RAA*	0.914690	RTAk	0.910626
TA1	0.897774	RTAl	0.910552
TA2a	0.913126	Single RF	0.912492
TA2b	0.912895	CCLPAa	0.918919
TA2c	0.912194	CCLPAb	0.916718
TA3	0.909874	CCLPAc	0.916720
RTAa	0.911684	CCLPAd	0.916658
RTAb	0.911838	CCLPAe	0.914570
RTAc	0.911705		

Table 4 displays that the performance of CCLPAa is better than unweighted CCLPAb, which shows that the weight setting based on AUC proposed in this study is effective. The performances of CCLPAa and CCLPAb are better than single RF and other combination algorithms, which verifies that the proposed mechanism to select collaborative HACI based on LR and the mechanism of integrating the most, second and third suitable composite HACI into a new SCHACI are helpful to solve the impact of scale-free on the accuracy of the algorithm. The performance of CCLPAc is better than CCLPAd, which confirms that the method to set the weight of each collaborative combination index is effective. The performance of CCLPAa is better than CCLPAc, which indicates that it is reasonable to recombine the most, second and third suitable collaborative combination indexes with different weights into a new composite index. CCLPAa creates a performance advantage over the unweighted CCLPAb, which demonstrates that the sensitivity coefficient method for setting weights proposed in this study is valid. The performances of CCLPAa, CCLPAb, CCLPAc and CCLPAd are better than the single RF and other combined algorithms, which illustrates that the mechanism to select HACI based on LR is efficient.

TABLE 5. Performance comparison between CCLPA and reference methods.

Algorithm	Average AUC	Algorithm	Average AUC
CCLPAa	0.918919	HDI	0.887315
LR	0.712816	LHN	0.807096
CN	0.895020	PA	0.892906
Salton	0.898752	AA	0.902040
Jaccard	0.878748	RA	0.912553
Sorenson	0.895441	RAA	0.914313
HPI	0.845366		

TABLE 6. Performance comparison between the newly defined HACIs and the original SLPA indexes.

Algorithm	Average AUC	Algorithm	Average AUC
CN*	0.911637	CN	0.895020
Salton*	0.899254	Salton	0.898752
Sorenson*	0.897392	Sorenson	0.895441
HDI*	0.890239	HDI	0.887315
LHN*	0.809052	LHN	0.807096
PA*	0.896363	PA	0.892906
RA*	0.912902	RA	0.912553
RAA*	0.914690	RAA	0.914313

TABLE 7. Algorithms performance in different network scales.

Network ID	Node count	Circle count	Node count per circle	RA*	RAA*	CCLPAa
124296976	32	7	4.571429	0.933588	0.936475	0.934317
51694885	31	7	4.428571	0.938933	0.936928	0.942804
31631020	48	4	12	0.868109	0.873948	0.872158
62759402	25	2	12.500000	0.957444	0.963469	0.963512
96483973	28	2	14	0.930119	0.935777	0.937756
190696559	74	5	14.800000	0.901863	0.905412	0.905387
198327282	119	8	14.875000	0.893640	0.903467	0.901914
176872879	33	3	11	0.865545	0.874763	0.876925
314316607	50	3	16.666667	0.930587	0.936949	0.937909
1608991	79	12	6.5833333	0.933205	0.935273	0.935770
133663120	39	2	19.500000	0.887529	0.893510	0.899600
152388029	55	3	18.333333	0.946216	0.946894	0.942873
93906304	49	6	8.1666667	0.944146	0.944117	0.946885
Average AUC				**0.917763**	**0.922075**	**0.922908**

In addition, it can be concluded from Fig. 8 that the accuracy of RF is better than other non-combined HACIs, which confirms that selecting the optimal HACIs based on network characteristics is valuable.

Finally, we collected paired samples of any two algorithms' AUC in 100 experiments and used F-test to determine whether the variance of the two samples was statistically equal. The results demonstrate that all p-values are less than 5% of the significance level, which confirms the significant differences between the algorithms. These results verify that CCLPA can effectively overcome the network sparsity and scale-free problems existing in the link prediction between different circles, accurately forecast the user friendships between small circles and large circles in the brand community, and finally effectively attract customer flow.

C. ANALYZING CCLPA PERFORMANCE IN DIFFERENT NETWORK SCALES

Although the overall performance of CCLPA has been confirmed in the previous section, it is still necessary to analyze the performance of CCLPA on recommending friends in various scales of circles and different node densities. Accordingly, networks with product circles from 2 to 9 in Twitter were selected. In addition, HACIs with higher accuracy, namely RA* and RAA*, were chosen from Table 4 for comparison with the performance of CCLPA. Table 7 and Fig. 11 show the performance of CCLPA, RA* and RAA* in different networks.

It can be observed in Table 7 that, compared with the other two algorithms, the average AUC value of CCLPA is the largest, which confirms that CCLPA has superior robustness and can produce excellent performance in networks with more or fewer product cycles. Fig. 11 demonstrates that CCLPAa has high precision whether in the large circle or the small circle. However, when the circle is small, the prediction performance of RAA* is the worst. When the circle is the middle or maximum, the performance of RA* is the worst. Conversely, CCLPA can make good friend recommendations, no matter the circle is large or small.

VI. CONCLUSION AND FUTURE WORKS

In the early stage of sales, the marketing community established for new products is generally small, it is needed to attract customers from the large mature product communities to the new one. In order to attract users accurately, it is essential to predict the friendships between the small circles formed for new products and the large circles formed for mature products. The existing researches on link prediction are usually based on the SLPA, which ignore that the AC on edges is different, and cannot overcome the impact of network sparsity and scale-free problems on link prediction accuracy between large and small circles. CCLPA is proposed to deeply extract the hidden feature structure information in the sparse network and overcome the fluctuation of algorithm accuracy caused by the scale-free network by self-adaptive construction of SCHACI. Compared with the existing researches, the

distinctive aspects of this study are mainly reflected in the following.

Firstly, the existing algorithms do not consider the heterogeneity of attention on the edge. In this study, from the perspective of heterogeneous AC on each edge, the HACIs are proposed to comprehensively extract the features of attention heterogeneity in the sparse network.

Secondly, the existing link prediction algorithms are always fixed, that is, the single algorithm is applied for various networks, while the method in this study can select appropriate HACIs according to different network characteristics.

Thirdly, the existing composite link algorithms combine all the SLPAs in a blind way, but our algorithm screens out the collaborative HACIs suitable for the given circle structure for each HACI, and integrates them into a composite one.

Finally, in the existing link prediction algorithms, the prediction results are given based on the intuitionistic network topologies. CCLPA can deeply extract implicit information in sparse networks by constructing two-level composite HACIs, namely collaborative HACIs and SCHACI, as a result, the possible friendships between small and large circles' users can be predicted accurately.

Under the reliable friendships prediction results of CCLPA, the influential users in the mature product circle can be recommended to the users in the new product circle. Under the influence of friends, users in mature product circles are driven to purchase new products and realize users transfer for new products. And subsequently, the value of CCLPA in online marketing will be fully exploited. The experimental results of online brand communities in the Twitter confirm that the CCLPA proposed in this study has excellent performance and superior robustness, which provides a strong theoretical support for marketers to achieve accurately attracting customer flow in online brand communities.

This research is only suitable for static network link prediction, and there is a need to further explore the link prediction in dynamic networks with links generating and breaking. In the future, the prediction framework proposed in this study will be applied to dynamic networks, and efforts will be made to improve the CCLPA so that the algorithm can predict the new and breaking links.

ACKNOWLEDGMENT

The authors would like to thank Shanghai University.

REFERENCES

[1] D. H. Lee and P. Brusilovsky, "How to measure information similarity in online social networks: A case study of citeulike," *Inf. Sci.*, vols. 418–419, pp. 46–60, Dec. 2017.

[2] H. Weijo, J. Bean, and J. Rintamäki, "Brand community coping," *J. Bus. Res.*, vol. 94, pp. 128–136, Jan. 2019.

[3] M. Palazon, M. Sicilia, and M. Lopez, "The influence of 'Facebook friends' on the intention to join brand pages," *J. Product Brand Manage.*, vol. 24, no. 6, pp. 580–595, 2015.

[4] M. Renton and H. Simmonds, "Like is a verb: Exploring tie strength and casual brand use effects on brand attitudes and consumer online goal achievement," *J. Product Brand Manage.*, vol. 26, no. 4, pp. 365–374, Jul. 2017.

[5] P. A. Voyer and C. Ranaweera, "The impact of word of mouth on service purchase decisions: Examining risk and the interaction of tie strength and involvement," *J. Service Theory Pract.*, vol. 25, no. 5, pp. 636–656, 2015.

[6] S. Aral and D. Walker, "Tie strength, embeddedness, and social influence: A large-scale networked experiment," *Manage. Sci.*, vol. 60, no. 6, pp. 1352–1370, Jun. 2014.

[7] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach, "Convex and network flow optimization for structured sparsity," *J. Mach. Learn. Res.*, vol. 13, no. 9, pp. 2681–2720, Sep. 2011.

[8] A. Pachon, L. Sacerdote, and S. Yang, "Scale-free behavior of networks with the copresence of preferential and uniform attachment rules," *Phys. D: Nonlinear Phenomena*, vol. 371, pp. 1–12, May 2018.

[9] J. Xie, B. K. Szymanski, and X. Liu, "SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in *Proc. IEEE 11th Int. Conf. Data Mining Workshops*, Dec. 2011, pp. 344–349.

[10] W. Chen, Z. Sun, and J. Han, "Landslide susceptibility modeling using integrated ensemble weights of evidence with logistic regression and random forest models," *Appl. Sci.*, vol. 9, no. 1, p. 171, Jan. 2019.

[11] Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao, and H. Cao, "Link prediction and recommendation across heterogeneous social networks," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 181–190.

[12] H. M. Gomes, J. P. Barddal, F. Enembreck, and A. Bifet, "A survey on ensemble learning for data stream classification," *ACM Comput. Surv.*, vol. 50, no. 2, pp. 1–36, Mar. 2018.

[13] S. Li, X. Song, H. Lu, L. Zeng, M. Shi, and F. Liu, "Friend recommendation for cross marketing in online brand community based on intelligent attention allocation link prediction algorithm," *Expert Syst. Appl.*, vol. 139, Jan. 2020, Art. no. 112839.

[14] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A, Statist. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.

[15] K.-K. Shang, T.-C. Li, M. Small, D. Burton, and Y. Wang, "Link prediction for tree-like networks," *Chaos: Interdiscipl. J. Nonlinear Sci.*, vol. 29, no. 6, Jun. 2019, Art. no. 061103.

[16] C. Zhang, K.-K. Shang, and J. Qiao, "Adaptive similarity function with structural features of network embedding for missing link prediction," *Complexity*, vol. 2021, pp. 1–15, Nov. 2021.

[17] C. H. Nguyen and H. Mamitsuka, "Latent feature kernels for link prediction on sparse graphs," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1793–1804, Nov. 2012.

[18] A. Ozcan and S. G. Oguducu, "Multivariate time series link prediction for evolving heterogeneous network," *Int. J. Inf. Technol. Decis. Making*, vol. 18, no. 1, pp. 241–286, 2019.

[19] E. Bütün, M. Kaya, and R. Alhajj, "Extension of neighbor-based link prediction methods for directed, weighted and temporal social networks," *Inf. Sci.*, vols. 463–464, pp. 152–165, Oct. 2018.

[20] T.-T. Kuo, R. Yan, Y.-Y. Huang, P.-H. Kung, and S.-D. Lin, "Unsupervised link prediction using aggregative statistics on heterogeneous social networks," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2013, pp. 775–783.

[21] B. Liu, S. Xu, T. Li, J. Xiao, and X.-K. Xu, "Quantifying the effects of topology and weight for link prediction in weighted complex networks," *Entropy*, vol. 20, no. 5, p. 363, May 2018.

[22] F. Aghabozorgi and M. Reza Khayyambashi, "A new study of using temporality and weights to improve similarity measures for link prediction of social networks," *J. Intell. Fuzzy Syst.*, vol. 34, no. 4, pp. 2667–2678, Apr. 2018.

[23] L. Lü and T. Zhou, "Link prediction in weighted networks: The role of weak ties," *EPL (Europhysics Letters)*, vol. 89, no. 1, p. 18001, Jan. 2010.

[24] K.-K. Shang, M. Small, D. Yin, T.-C. Li, and W. Yan, "The key to the weak-ties phenomenon," *EPL (Europhys. Lett.)*, vol. 127, no. 4, p. 48002, Sep. 2019.

[25] K.-K. Shang, M. Small, X.-K. Xu, and W.-S. Yan, "The role of direct links for link prediction in evolving networks," *EPL (Europhys. Lett.)*, vol. 117, no. 2, p. 28002, Jan. 2017.

[26] R. Guimerà and M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 52, pp. 22073–22078, Dec. 2009.

[27] T. Vallès-Català, T. P. Peixoto, M. Sales-Pardo, and R. Guimerà, "Consistencies and inconsistencies between model selection and link prediction in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 97, no. 6, Jun. 2018, Art. no. 062316.

[28] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, no. 4, Oct. 2009, Art. no. 046122.
[29] S. Cheng, B. Zhang, G. Zou, M. Huang, and Z. Zhang, "Friend recommendation in social networks based on multi-source information fusion," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 5, pp. 1003–1024, May 2019.
[30] C. C. Chen, S.-Y. Shih, and M. Lee, "Who should you follow? Combining learning to rank with social influence for informative friend recommendation," *Decis. Support Syst.*, vol. 90, pp. 33–45, Oct. 2016.
[31] C. Ma, T. Zhou, and H.-F. Zhang, "Playing the role of weak clique property in link prediction: A friend recommendation model," *Sci. Rep.*, vol. 6, no. 1, p. 30098, Jul. 2016.
[32] Z. Yu, C. Wang, J. Bu, X. Wang, Y. Wu, and C. Chen, "Friend recommendation with content spread enhancement in social networks," *Inf. Sci.*, vol. 309, pp. 102–118, Jul. 2015.
[33] A. Ghasemian, H. Hosseinmardi, A. Galstyan, E. M. Airoldi, and A. Clauset, "Stacking models for nearly optimal link prediction in complex networks," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 38, pp. 23393–23400, Sep. 2020.
[34] R. Guimerà, "One model to rule them all in network science?" *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 41, pp. 25195–25197, Oct. 2020.
[35] A. Muscoloni, U. Michieli, and C. V. Cannistraci, "Adaptive network automata modelling of complex networks," *Preprints*, to be published.
[36] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 2, Jul. 2001, Art. no. 025102.
[37] S. Georg, T. Ulrich, and T. M. Kemple, "The social boundary," *Theory, Culture Soc.*, vol. 24, nos. 7–8, p. 53, Dec. 2007.
[38] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, May 2007.
[39] J. Li, J. Peng, S. Liu, X. Ji, X. Li, and X. Hu, "Link prediction in directed networks utilizing the role of reciprocal links," *IEEE Access*, vol. 8, pp. 28668–28680, 2020.
[40] C. Lei, J. Deng, K. Cao, L. Ma, Y. Xiao, and L. Ren, "A random forest approach for predicting coal spontaneous combustion," *Fuel*, vol. 223, pp. 63–73, Jul. 2018.
[41] F. J. H. Brims, T. M. Meniawy, I. Duffus, D. de Fonseka, A. Segal, J. Creaney, N. Maskell, R. A. Lake, N. de Klerk, and A. K. Nowak, "A novel clinical prediction model for prognosis in malignant pleural mesothelioma using decision tree analysis," *J. Thoracic Oncol.*, vol. 11, no. 4, pp. 573–582, Apr. 2016.
[42] M. Saliha, B. Ali, and S. Rachid, "Towards large-scale face-based race classification on spark framework," *Multimedia Tools Appl.*, vol. 78, no. 18, pp. 26729–26746, Sep. 2019.

SHUGANG LI received the Ph.D. degree from Shanghai Jiaotong University, Shanghai, China, in 2004.

From 2008 to 2014, he was an Associate Professor at Shanghai Jiaotong University. He was a Visiting Scholar with the University of Florida, Gainesville, FL, USA. He is currently a Professor with the School of Management, Shanghai University, Shanghai, China. He has been the Chairman of the Department of Information Management, since 2019. He presided over and participated in more than ten projects of NSFC, general projects, and Hong Kong, China, cooperation projects and local and enterprise horizontal projects. He has published two monographs and more than 60 papers, including more than 30 papers in IEEE TRANSACTIONS, *Information Sciences*, *Knowledge-Based Systems*, *Expert Systems with Applications*, *Computers & Operations Research*, *Computers & Industrial Engineering*, and other internationally renowned journals. He serves as on the Editorial Board of *Journal of Mathematics and Computer Science*.

BOYI ZHU is currently pursuing the doctor's degree with the School of Management, Shanghai University, Shanghai, China.

Her current research interests include complex network link prediction, business artificial intelligence, social e-commerce and product management, and internet commerce.

HE ZHU is currently pursuing the master's degree with the School of Management, Shanghai University, Shanghai, China.

Her current research interests include business artificial intelligence, social e-commerce and product management, and complex system modeling and evolutionary control.

FANG LIU is currently pursuing the doctor's degree with the School of Management, Shanghai University, Shanghai, China.

Her current research interests include consumer online reviews mining, business artificial intelligence, social e-commerce and product management, and complex network link prediction.

YUQI ZHANG is currently pursuing the doctor's degree with the School of Management, Shanghai University, Shanghai, China.

Her current research interests include social e-commerce consumer behavior, complex system modeling, and evolutionary control.

RU WANG is currently pursuing the doctor's degree with the School of Management, Shanghai University, Shanghai, China.

Her current research interests include social e-commerce development, user generated content, consumer online reviews mining, and business artificial intelligence.

HANYU LU is currently pursuing the doctor's degree with the School of Management, Shanghai University, Shanghai, China.

Her current research interests include consumer online reviews mining, social e-commerce development, complex network link prediction, and marketing management.