International Journal of Innovative
Computing, Information and Control
Volume **16**, Number **4**, August **2020**

 ISSN 1349-4198
pp. 1371–1384

NETWORK EMBEDDING OF TOPIC-ATTENTION NETWORK BASED ON SET PAIR ANALYSIS

Jingfeng Guo[1], Hui Dong[1,*], Tingwei Zhang[1] and Xiao Chen[2]

[1]College of Information Science and Engineering
Yanshan University
No. 438, West Hebei Avenue, Qinhuangdao 066004, P. R. China
[*]Corresponding author: DongHui_DH0620@163.com

[2]College of Network Technology Center
Hebei Normal University of Science and Technology
No. 360, West Hebei Avenue, Qinhuangdao 066004, P. R. China

Received January 2020; revised May 2020

Abstract. *Concerning the problem that heterogeneous network embedding only considers social relations in structure and ignores semantics, combining the social relationship between users and the preference of users for topics, a network embedding algorithm based on topic-attention network was proposed. Firstly, according to the characteristics of the topic-attention network and combining with the idea of the identical-discrepancy-contrary (determination and uncertainty) of set pair analysis (SPA) theory, the transition probability model was given. Then a random walk algorithm based on two types of nodes was proposed by using the transition probability model, which obtained relatively high-quality random walk sequences. Finally, the embedding vector space representation of the topic-attention network was obtained by modeling based on two types of nodes in the sequences. After theoretical analysis, experimental results on the Douban dataset show that the modularity of the proposed algorithm is 0.5871 when the number of the overlapping communities is 11, which is nearly 6.5% higher than that of metapath2vec algorithm. The random walk algorithm combined with the transition probability model is more comprehensive in analyzing the connection relationship between nodes in the network, and can capture more detailed information in the network.*

Keywords: Topic-attention network, SPA, Transition probability, Random walk, Network embedding

1. **Introduction.** As the most intuitive representation of network, adjacency matrix has high time complexity of operation and space complexity of storage, which makes many existing algorithms unable to be applied to large-scale network. The analysis of large-scale network urgently needs a reasonable and efficient network representation. Therefore, network embedding has become research focus in recent years.

Network embedding [1] is different from the traditional network analysis method based on adjacency matrix. It combines the characteristics of network structure and semantics to learn the latent, low-dimensional and dense representations of the network vertices. Since network embedding vector is easy to be processed by machine learning algorithms, it has attracted wide attention from academia and industry, and is also effectively applied to tasks such as node classification, clustering and link prediction.

Social networks are divided into two categories based on whether the types of nodes and edges they contain are the same: homogeneous networks and heterogeneous networks. At

DOI: 10.24507/ijicic.16.04.1371

present, there are many and deeper network embedding methods for homogeneous network, while the methods of heterogeneous networks represent learning in terms of matrix decomposition [2], custom loss function [3,4] and neural network [5-7]. Among them, the method based on neural network [5] involves a combination of a random walk algorithm based on transfer probability model and skip-gram model to learn node vector representation in heterogeneous network. However, most of the definitions of transfer probability models in this method only consider basic connection relationships in the network (such as direct connection relationships between nodes), and only maintain first-order proximity, failing to capture the global network structure. Aiming at the above problems, this paper mainly conducts related research on presentation learning based on the sub-network of heterogeneous network (topic-attention network) [8-10]. The network consists of two types of nodes and edges (Figure 1), where u_i and t_k represent the user and topic nodes, respectively. In view of the characteristics of the structure and semantics between the nodes and edges in heterogeneous networks, based on the identical-discrepancy-contrary (determination and uncertainty) ideas in SPA theory, embedding for topic-attention network algorithm was proposed. The main contributions are as follows.

- A transfer probability model based on SPA theory is proposed. According to several cases of connection relationship between two types of nodes (users and topics) in the 4th order [11], based on the identical-discrepancy-contrary (determination and uncertainty) ideas in SPA theory, construct a transfer probability model between nodes in topic-attention network.
- A user-topic walk (U-T walk) algorithm based on users and topic nodes is proposed. Based on the transfer probability from user to user and between user and topic, this algorithm gives the random walk strategy, and realizes the flexible walks between two kinds of nodes to obtain the walk sequence.
- A presentation learning algorithm (TANE) based on topic-attention network is proposed. The algorithm combines the skip-gram model to train the walk sequences to get the U-T walk algorithm, and obtains the high-quality network embedding vector space representation.
- The proposed algorithm combines the fuzzy C-means (FCM) clustering algorithm for overlapping community discovery on two real-world datasets, and using the modularity as the evaluation index. Experimental results verify the effectiveness of the proposed algorithm, and the result of clustering is relatively high.

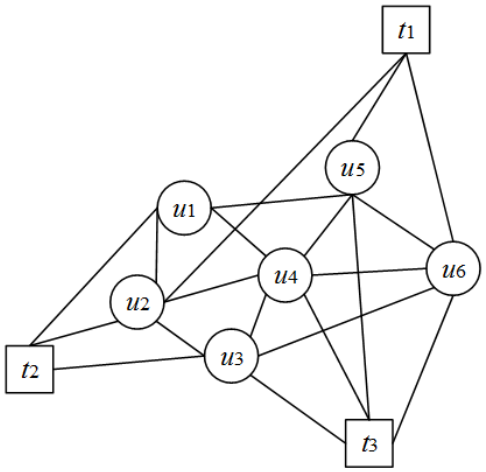


FIGURE 1. Topic-attention network

In summary, this paper mainly uses the random walk algorithm based on the set pair analysis transfer probability model and the skip gram model to study the representation learning of the topic-attention network.

In the rest of this paper, we first review the related work of representation learning in Section 2, and describe the characteristics and related definitions of topic-attention network in Section 3. We present the proposed framework of TANE algorithm in Section 4 and show experimental results in Section 5. Finally, we conclude the paper in Section 6.

2. **Related Work.** With the rapid growth of social network users, the traditional network representation methods hit the wall, and network representation methods combined with deep learning have gradually emerged.

DeepWalk algorithm [12] introduced the technology of deep learning into homogeneous network embedding for the first time. This algorithm used the idea of Word2vec [13] model to take nodes as words and random walk sequences as sentences. It was trained to obtain the embedding vector space representation of the network, which proved its effectiveness in multi-label classification task. Node2vec algorithm [14] combined with breadth first search (BFS) and depth first search (DFS) searching algorithm, considering the local and global information improves the random walk in DeepWalk algorithm strategy; LINE algorithm [15] improves the first-order nearest neighbor sparse problem of DeepWalk by describing the second-order proximity of the neighbors of two nodes that are not directly connected. The results of the two algorithms in multi-label classification task were significantly improved. MPNE algorithm [16] combined with Motif structure for higher-order network embedding has achieved better results in multi-label classification task. SDNE algorithm [17] combined with the neural network model to capture the highly non-linear relationship in the network, and finally took the hidden layer weight as the embedded representation of the network. Outstanding experimental results are obtained in the link prediction task.

Heterogeneous network embedding research started gradually from PTE algorithm [3], which embodies the predictive information in the final embedding. Unlike convolutional neural network (CNN) or recurrent neural network (RNN) models, which directly nest a complex predictive model, the PME algorithm [4] used the metric learning method to simultaneously capture the first-order and second-order relations in the network, learning the vector representations of nodes and relations in the object and relational space respectively; both metapath2vec and metapath2vec++ algorithms [5] are based on the idea of DeepWalk, using meta-path to guide random walk, and using heterogeneous skip-gram model to train to get the embedding vector space representation of the network; HIN2Vec algorithm [6] combined with the neural network model learned the relationship (meta-path) representation while getting the embedding vector space representation of the network; DVNE algorithm [7] used the neural network model combined with Wasserstein distance to change the method of measuring the similarity between two distributions, not only satisfying the triangle inequality, but also maintaining the transitivity of proximity well in undirected graph. The experimental results of the above algorithms in multi-label classification, clustering and link prediction are obviously improved. At present, there are many restrictions on the random walk with bias in heterogeneous network embedding based on the meta-path [18] random walk strategy. Therefore, it is necessary to explore new theories and methods for heterogeneous network representation. At present, the heterogeneous representation learning based on random walk is mostly based on the walk strategy of symmetrical meta-path (e.g., in the paper co-author network, the author-the-author), and there are more restrictions on random walk strategies based on the symmetric

meta-path; the topic-based modeling for social network analysis has achieved good results [19]. Therefore, it is necessary to explore a new random walk mode based on theoretical knowledge to study topic-attention network representation learning.

3. **Topic-Attention Network.** This paper mainly focuses on the research of network representation learning based on the topic-attention network modeling. The specific introduction of the topic-attention network is as follows.

The topic-attention network [10] is defined as a binary group $G = (V, E)$, where $V = \{U, T\}$, $U = \{u_1, u_2, \ldots, u_n\}$ represents the user node set, and $T = \{t_1, t_2, \ldots, t_m\}$ represents the topic node set; $E = \{EUU, EUT\}$ represents the edge set, where $EUU = \{(u_i, u_j)|u_i, u_j \in U\}$ represents the relationship between the user and user, and $EUT = \{(u_i, t_k)|u_i \in U, t_k \in T\}$ represents the relationship between the user and topic.

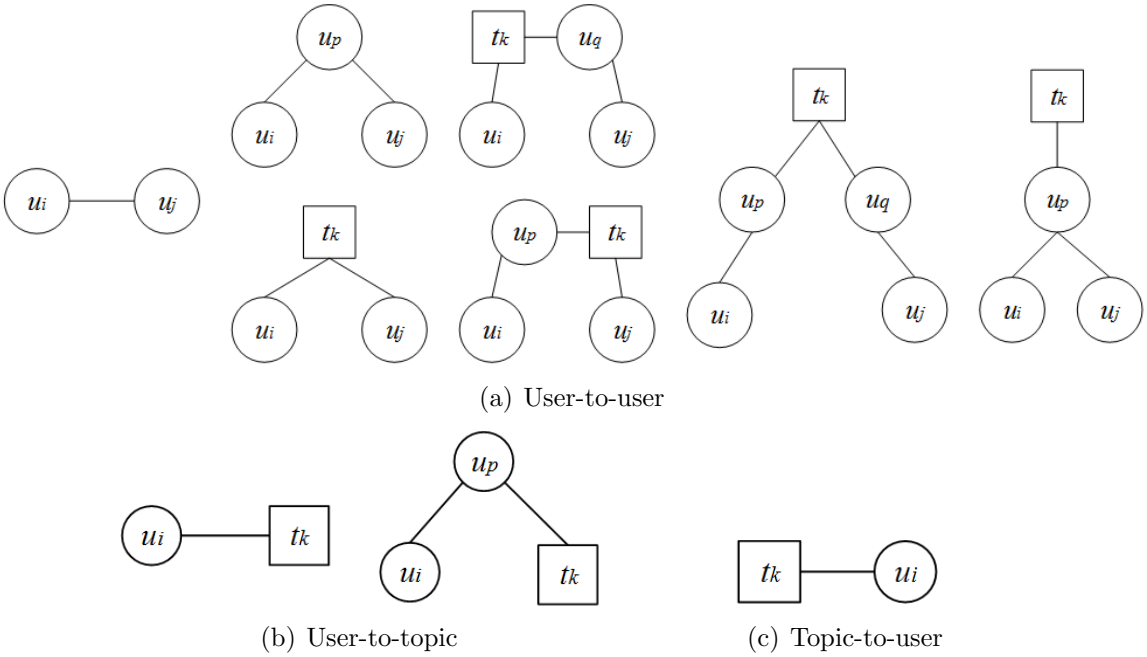
In the topic-attention network, $N(u_i)_1 = NU(u_i)_1 \cup NT(u_i)_1$ and $N(u_i)_2 = NU(u_i)_2 \cup NT(u_i)_2$ represent the first-order and the second-order neighbor sets of the node u_i, respectively (including user neighbor set $NU(u_i)_m$ and topic neighbor set $NT(u_i)_m$, $m = 1, 2$); $CN(u_i, u_j)_1 = CNU(u_i, u_j)_1 \cup CNT(u_i, u_j)_1$ and $CN(u_i, u_j)_2 = CNU(u_i, u_j)_2 \cup CNT(u_i, u_j)_2$ represent $N(u_i)_1 \cap N(u_j)_1$ and $N(u_i)_2 \cap N(u_j)_2$, respectively; $CN(u_i, u_j)_{1\cap 2} = CNU(u_i, u_j)_{1\cap 2} \cup CNT(u_i, u_j)_{1\cap 2}$ and $CN(u_i, u_j)_{2\cap 1} = CNU(u_i, u_j)_{2\cap 1} \cup CNT(u_i, u_j)_{2\cap 1}$ represent $N(u_i)_1 \cap N(u_j)_2$ and $N(u_i)_2 \cap N(u_j)_1$, respectively, and other relevant definitions are detailed in [10]. The network is a model combining social relationships and interests, and the edges between nodes have obvious semantic characteristics.

4. **TANE: Topic-Attention Network Embedding.**

4.1. **Modeling based on set pair analysis theory.** The core idea of SPA theory [20] is to regard the relation between things as a system, which is referred to as the system of identical-discrepancy-contrary (determination and uncertainty). Among them, determination includes the identical and contrary, and usually includes both identity and opposition aspects of the thing, in which only the identical aspect is considered; uncertainty is discrepancy, usually referring to the macro and micro impact factors of things. The key of constructing the topic-attention network model is how to build the deterministic and uncertain relationships between nodes.

Topic-attention network consists of two types of nodes: user and topic, and three types of inter-node relationships: user-to-user, user-to-topic, topic-to-user. Based on the topology of the network, it is regarded as a deterministic relationship that is the relationship between two user nodes through the nodes directly connected to them, for example, users are more likely to become friends if they have common friends or interests; it is regarded as an uncertain relationship that is the relationship between two user nodes through the nodes indirectly connected with them, for example, it is more uncertain for users to become friends by sharing common interests with their friends, which may be converted into a deterministic relationship under certain conditions. As shown in Figure 2, the three connections within the second-order between nodes (u_i, u_j), (u_i, t_k), (t_k, u_i) as part of the determination while other cases are modeled as the uncertain part.

4.2. **Transition probability model.** Network analysis based on transfer probability model in traditional homogeneous social networks has taken account of the second-order of nodes with BFS and DFS, but it is not possible to capture the global network structure accurately. The diversity of realistic networks (e.g., networks include multiple types of nodes and edges) and the complexity of network relationships (e.g., friends are connected by common interests) are the focus of research on the topic-attention network for a comprehensive and accurate definition of the transfer probability model.



(a) User-to-user

(b) User-to-topic

(c) Topic-to-user

FIGURE 2. Modeling involving the connection between nodes

For the relationships between the two types of nodes in the topic-attention network in Figure 2, a transition probability model is constructed by integrating the network structure, topic preferences, the idea of determination and uncertainty. The specific definitions are as follows.

Definition 4.1. *(User-to-user transfer probability) Given the topic-attention network* $G = (V, E)$, $\forall u_i, u_j \in U$, $t_k \in T$, *considering* $N(u_i)_1$, $CN(u_i, u_j)_1$, $CNT(u_i, u_j)_{1\cap 2}$, $CNT(u_i, u_j)_{2\cap 1}$, *and* $CNT(u_i, u_j)_2$, *the transfer probability from user-to-user is denoted as follows:*

$$P(u_j|u_i) = \alpha P_{ij}(1) + \beta \sum_{m=1}^{2} P_{ij}(2) + \sum_{n=1}^{2} \gamma_n P_{ij}^n(3) + \delta P_{ij}(4) \tag{1}$$

where $P_{ij}(1)$ *indicates whether the two user nodes are directly connected, and we have*

$$P_{ij}(1) = \frac{S}{d_i} \tag{2}$$

where d_i *represents the degree of node* u_i*;* S *as a tag, and* $S = 1$ *indicates that there is a direct connection between two nodes.*

$P_{ij}(2)$ *indicates that two user nodes establish a connection through common first-order neighbors, and we have*

$$P_{ij}(2) = \frac{\chi_1 \left|CNT(u_i, u_j)_1\right| + \chi_2 \left|CNU(u_i, u_j)_1\right|}{d_i + d_j - \left|CNT(u_i, u_j)_1\right| - \left|CNU(u_i, u_j)_1\right|} \tag{3}$$

where $|CNT(u_i, u_j)_1|$ *represents the number of common topics at first-order neighbors to the user nodes* u_i *and* u_j*, and* $|CNU(u_i, u_j)_1|$ *represents the number of common users at first-order neighbors to the user nodes* u_i *and* u_j*. The same applies in the following equations.*

$P_{ij}^n(3)$ *indicates that two user nodes establish a connection through the common topics* $N(u_i)_1 \cap N(u_j)_2$ *and* $N(u_i)_2 \cap N(u_j)_1$*. It is equal to the sum of Equation (4) and Equation*

(5), and we have

$$P_{ij}^{1}(3) = P_{ij}(3)_{1\cap 2} = \sum_{t_k \in CNT(u_i,u_j)_{1\cap 2}} \frac{1}{d_i} \times \frac{|CNT(t_k,u_j)_1| + |CNU(t_k,u_j)_1|}{d_k + d_j - |CNT(t_k,u_j)_1| - |CNU(t_k,u_j)_1|} \tag{4}$$

$$P_{ij}^{2}(3) = P_{ij}(3)_{2\cap 1} = \sum_{t_k \in CNT(u_i,u_j)_{2\cap 1}} \frac{|CNT(u_i,t_k)_1| + |CNU(u_i,t_k)_1|}{d_i + d_k - |CNT(u_i,t_k)_1| - |CNU(u_i,t_k)_1|} \times \frac{1}{d_k} \tag{5}$$

where $|CNT(t_k,u_j)_1|$ represents the number of common topics at first-order neighbors to the topic node t_k and user node u_j. The connection relationship between topic-to-topic is not considered here, which is recorded as 0. The same applies in the following equations.

$P_{ij}(4)$ indicates that two user nodes establish a connection through the common second-order topic neighbors, and we have

$$P_{ij}(4) = \sum_{t_k \in CNT(u_i,u_j)_2} \frac{|CNT(u_i,t_k)_1| + |CNU(u_i,t_k)_1|}{d_i + d_k - |CNT(u_i,t_k)_1| - |CNU(u_i,t_k)_1|} \times \frac{1}{d_k} \times \sum_{u_p \in CNU(t_k,u_j)} \frac{1}{d_p} \tag{6}$$

where α, β, γ_1, γ_2, δ represent the weight coefficients of $NU(u_i)_1$, $CN(u_i,u_j)_1$, $CNT(u_i,u_j)_{1\cap 2}$, $CNT(u_i,u_j)_{2\cap 1}$ and $CNT(u_i,u_j)_2$, respectively, and $\alpha+\beta+\gamma_1+\gamma_2+\delta=1$. χ_1 and χ_2 represent the weight coefficients of common users and topics when the two user nodes establish connection through $CN(u_i,u_j)_1$ and $\chi_1+\chi_2=1$.

Definition 4.2. *(User-to-topic transfer probability) Given the topic-attention network $G=(V,E)$, $\forall u_i,u_p \in U$, $t_k \in T$, considering $NT(u_i)_1$, $CNU(u_i,t_k)_1$, the transfer probability from user-to-topic is denoted as follows:*

$$P(t_k|u_i) = \varepsilon P_{ik}(1) + \theta P_{ik}(2) \tag{7}$$

where $P_{ik}(1)$ indicates whether the user node and the topic node are directly connected, and the calculation method is the same as Equation (2); $P_{ik}(2)$ indicates the connection between the user node and the topic node through $CNU(u_i,t_k)_1$, and the calculation method is the same as Equation (3); ε and θ represent the weight coefficients of $NT(u_i)_1$ and $CNU(u_i,t_k)_1$, respectively, and $\varepsilon+\theta=1$.

Definition 4.3. *(Topic-to-user transfer probability) Given the topic-attention network $G=(V,E)$, $\forall u_i,u_j \in U$, $t_k \in T$, considering $NU(t_k)_1$, the transfer probability from topic-to-user is denoted as follows:*

$$P(u_i|t_k) = P_{ki}(1) \tag{8}$$

where $P_{ki}(1)$ indicates whether the topic node and the user node are directly connected, and the calculation method is the same as Equation (2).

4.3. **TANE algorithm framework.** The core idea of TANE algorithm is to obtain the transfer probability model from the above definition and apply it to the random walk algorithm to obtain the sequences of random walk. By training the sequences, the vector representation of two types of nodes can be obtained, depending on the needs of two types of node vector representation, i.e., vector representation of user nodes in the network for overlapping community discovery tasks, or reasonable recommendation of users in the community based on the distribution of topic nodes in the network.

4.3.1. *User-topic walk algorithm.* According to the above three transfer probability models, a random walk algorithm based on topic-attention network is proposed in Algorithm 1. Wherein, line 1) represents the initialization random walk path; line 2) represents the selection of the starting node is random, which can be the user or the topic nodes; and line 3) to line 5) represent respectively the selection of the current node ($v_{current}$), calculating the transfer probability to the next node (v_{next}) under the given $v_{current}$, and the selection method of the v_{next} within length L of the path walk.

Algorithm 1 User-topic walk algorithm (U-T walk)

Input: graph $G = (V, E)$, starting node u, walk length L, transfer probability P
Output: random walk sequences with user and topic nodes
Begin:

1) init path = $\emptyset$
2) path = $[u]$
3) while len(path) $< L$ do
4) $v_{current}$ = path$[-1]$
5) $P = P(v_{next} \mid v_{current})$
6) v_{next} = randoms.choices((P.keys, P.values), $k = 1$)
7) path.append(v_{next})
8) end while
9) return path

End

In the topic-attention network, the next node selection based on transfer probability models is divided into two strategies: 1) if $v_{current}$ is a user node, v_{next} is a user or topic node; 2) if $v_{current}$ is a topic node, v_{next} must be a user node. The selection of v_{next} by non-uniform sampling with weights (i.e., transfer probability) is more inclined to nodes with high transfer probability (i.e., a large number of common users and topics with $v_{current}$).

The u_1 and t_3 in Figure 1 are taken as an example to illustrate the above algorithm. The parameter setting of the transfer probability model is consistent with Karate dataset, and the transfer probability of other nodes is shown in Table 1. It can be known from Table 1 that if u_1 is taken as the $v_{current}$, the selection of the v_{next} is more likely to the nodes with a relatively high transfer probability of u_2, u_4, t_1, t_2. As can be seen from $P(u_5 \mid u_1) < P(u_6 \mid u_1)$ that the transfer probability between inter-node is high not because the direct connections exist between nodes, but also because indirect connections are rich. In summary, the selection of the v_{next} for the random walk needs to consider the indirect connection of the nodes.

TABLE 1. Transition probability with nodes u_1 and t_3 as $v_{current}$

P	u_2	u_3	u_4	u_5	u_6	t_1	t_2	t_3
u_1	0.19	0.08	0.18	0.05	0.06	0.12	0.23	0.10
t_3	0	0.25	0.25	0.25	0.25	0	0	0

4.3.2. *TANE.* Topic-attention network embedding (TANE) algorithm, which combines the transfer probability model and U-T walk algorithm in Algorithm 2. Wherein, line 1) represents the initialization of the random walk path set; line 2) to 8) respectively represent that on the premise that each node is n times as the start node, parameters such as the start node and the length of the walk path are set by U-T walk algorithm, and

Algorithm 2 Topic-attention network embedding (TANE)

Input: graph $G = (V, E)$, window size w, walks per vertex n, embedding size d, walk length L, transfer probability P

Output: matrix of vertex representation $\Phi \in R^{|V| \times d}$

Begin:

1) init $W = \emptyset$
2) for $i = 0$ to n do
3) rand. shuffle(V)
4) for each $v \in V$ do
5) path = UT-walk (G, v, L, P)
6) W.append(path)
7) end for
8) end for
9) Φ = Word2vec$(W, d, w, sg = 1, hs = 1)$

End

multiple relatively high-quality random walk sequences are obtained by using the transfer probability models; line 9) uses the Word2vec model to train the random walk sequences to get the embedding vector space Φ of the topic-attention network.

5. Experiments.

5.1. Experimental setup.

5.1.1. *Datasets and baseline algorithms.* Two benchmark data sets and five classical algorithms we consider in our experiments are given as follows.

1) Zachary's Karate Club dataset is a classic dataset in the field of social network analysis, consisting of 34 nodes and 78 edges.

2) The Douban dataset is a network that contains two parts of information about the user's attention and movie comment attributes, captured by the crawler program. Process the dataset, and filter out information that users rated the movies below 3 points. In the final, the network has 2289 nodes (2253 users and 36 topics) and 70489 edges (34580 user-user edges and 35909 user-topic edges).

The five classical algorithms are DeepWalk, Node2vec, MPNE, PTE and metapath2vec. In order to make the experimental results more convincing, the parameters in the comparison algorithm are consistent with the original paper, and the dimension settings are 64, 128, 64, 100 and 100 respectively. The dimension of the TANE algorithm proposed in this paper is 64.

5.1.2. *The evaluation index.* In the experiment, node clustering is selected as the network analysis task, and modularity (Q) is selected as the evaluation index in Equation (9). The higher Q is, the better the quality of overlapping communities is divided.

$$Q = \frac{1}{2m} \sum_{i=1}^{|C|} \sum_{u \in C_i, v \in C_i} \frac{1}{O_u O_v} \times \left(A_{uv} - \frac{k_u k_v}{2m} \right) \tag{9}$$

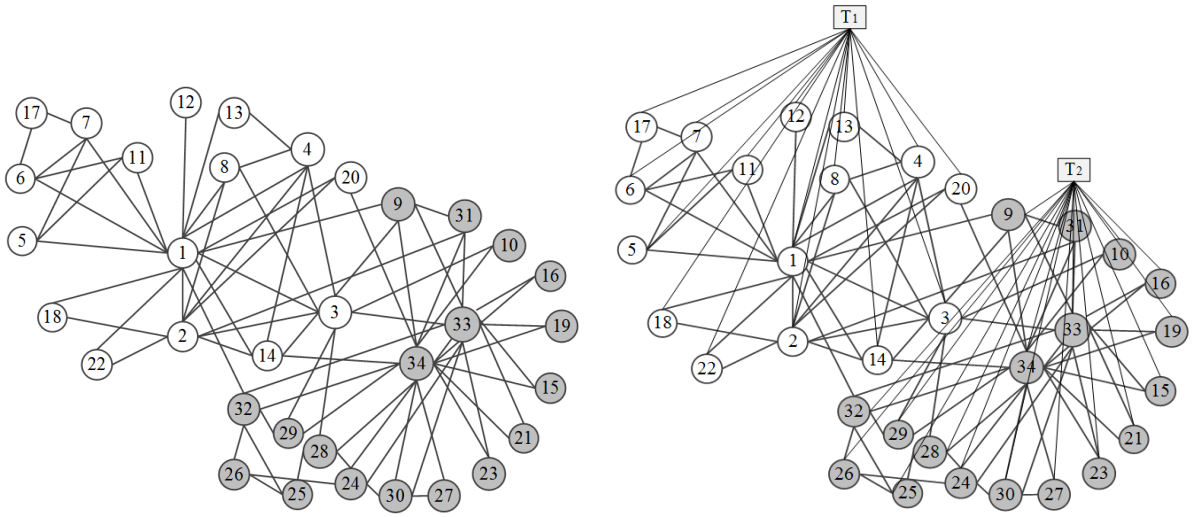
It should be noted that the experiment is to prove the importance of topic nodes to the stability of social relationships, and the topic only plays the role of a bridge. Therefore, the topic nodes in the network should be removed when clustering analysis and calculation of Q.

5.2. **Experimental results and analysis.** The experiment is mainly divided into two parts:

1) Verifying the correctness and rationality of the TANE algorithm;

2) Verifying the influence of the topic on the transfer probability and vector representation.

5.2.1. *Experiments based on Karate network.* Karate network is a topic-free real-world network (Figure 3(a)). By adding topics to Karate network (Figure 3(b)), two embedding vector space representations of this network are obtained by DeepWalk algorithm and TANE algorithm, and analysis of Euclidean distance of vector representation between nodes. Considering the relationship between nodes in the 4th order, Karate network is relatively dense, and the 4th order neighbors only involve one node. Therefore, Figures 4(a)-4(c) respectively show the variation trend of Euclidean distance between 34 nodes and the first-order, second-order and third-order neighbor vectors of each node under the two algorithms. It can be seen from Figure 4 that the variation trend of Euclidean distance is basically the same in the three conditions. In summary, the TANE algorithm conforms to the core idea of network embedding.



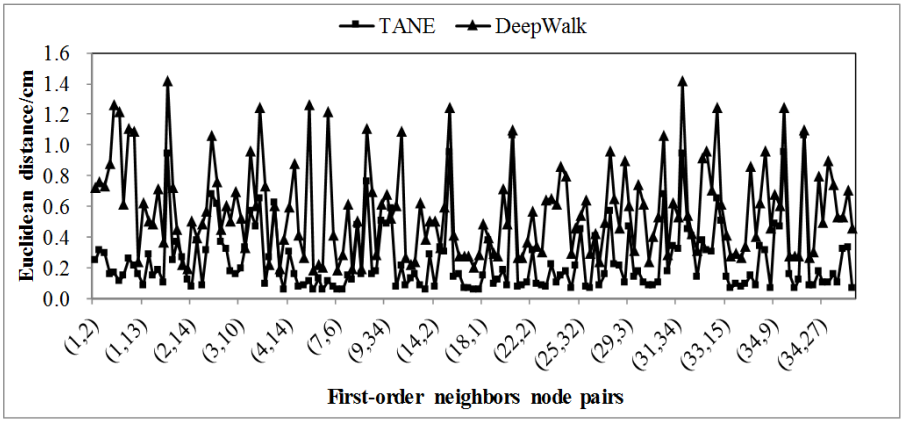
(a) Original network (b) Adding T_1 and T_2 topics

FIGURE 3. Karate network

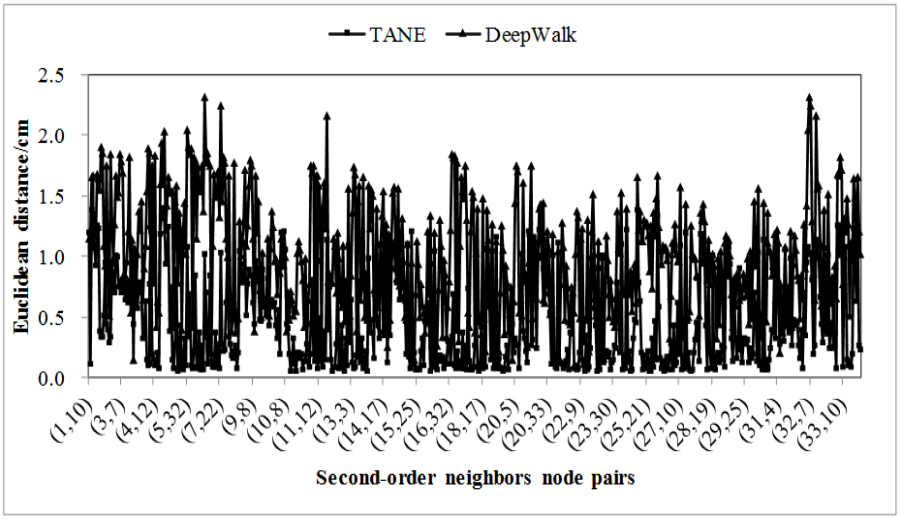
5.2.2. *Analysis based on clustering algorithm.* After the above analysis, the FCM algorithm is used to complete overlapping community detection task on Douban network. After comparison, the Elbow Rule was selected to determine the number of clusters K, and the core index is the sum of the squared errors (SSE) in Equation (10); the core idea is that as the number of clusters K increases, SSE will gradually decrease and tend to be flat. The value of K corresponding to the inflection point is the true number of clusters of the data.

$$SSE = \sum_{i=1}^{K} \sum_{p \in C_i} |p - m_i|^2 \tag{10}$$

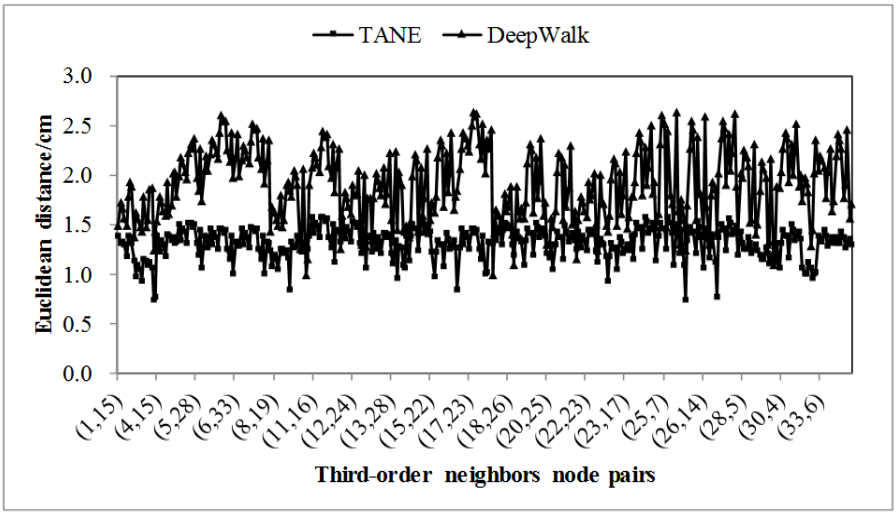
Firstly, the K value of Karate network is determined by the Elbow Rule in Figure 5(a). The inflection point is 4, which is consistent with the maximum Q value when modularity algorithm is used to cluster into four categories in the DeepWalk paper. Therefore, the method has certain rationality.



(a) First-order neighbors

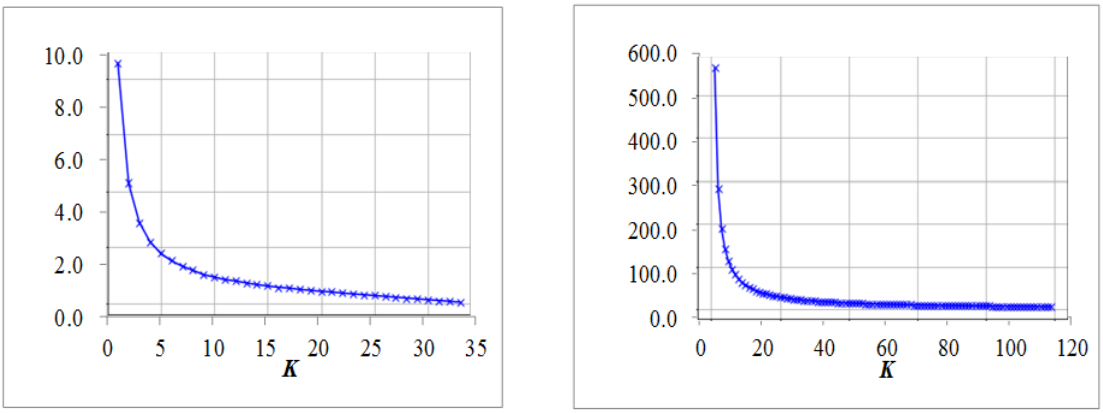


(b) Second-order neighbors



(c) Third-order neighbors

FIGURE 4. Comparison of Euclidean distance between neighbors of the order in two algorithms



(a) Karate (b) Douban

FIGURE 5. Determining the number of clusters by Elbow Rule

TABLE 2. Q value on the Douban dataset

Modularity	Algorithm	K								
		5	6	7	8	9	10	11	12	13
Q	DeepWalk	0.4144	0.4212	0.3908	0.3889	0.4012	0.4109	0.4315	0.4466	0.4551
	Node2vec	0.4298	0.4305	0.4314	0.4205	0.4213	0.4331	0.4653	0.4413	0.4528
	MPNE	0.4559	0.4422	0.4791	0.3950	0.4706	0.4967	0.5028	0.5062	0.4834
	PTE	0.4891	0.5028	0.5171	0.4985	0.5233	0.5201	0.5135	0.5046	0.5186
	metapath2vec	0.5014	0.5171	0.5104	0.5228	0.5211	0.5343	0.5237	0.5429	0.5372
	TANE	*0.5039*	*0.5433*	*0.5401*	*0.5516*	*0.5642*	*0.5866*	*0.5871*	*0.5724*	*0.5743*

Then, the Elbow Rule is applied to determining the K value of Douban network in Figure 5(b). For the accuracy of the experiment, K takes the value within the inflection point range, i.e., $K \in [5, 13]$.

Finally, the FCM algorithm is called based on the obtained embedding vector space representation and K value to get the degree value of each cluster that the node belongs to, and the threshold value λ is determined by using the difference between the maximum value of degree value and other values in Equation (11). In the experiment, $\lambda \in [1.5, 2.5]$. The selection of initial clustering centers is random, and the results change slightly. Ten experiments are conducted for each K value, and the average value is taken as the final result. As shown in Table 2, within the value range of K, the modularity value obtained by TANE algorithm is higher than comparison algorithm, that is, the community structure strength obtained by clustering is stronger and the quality of community division is better.

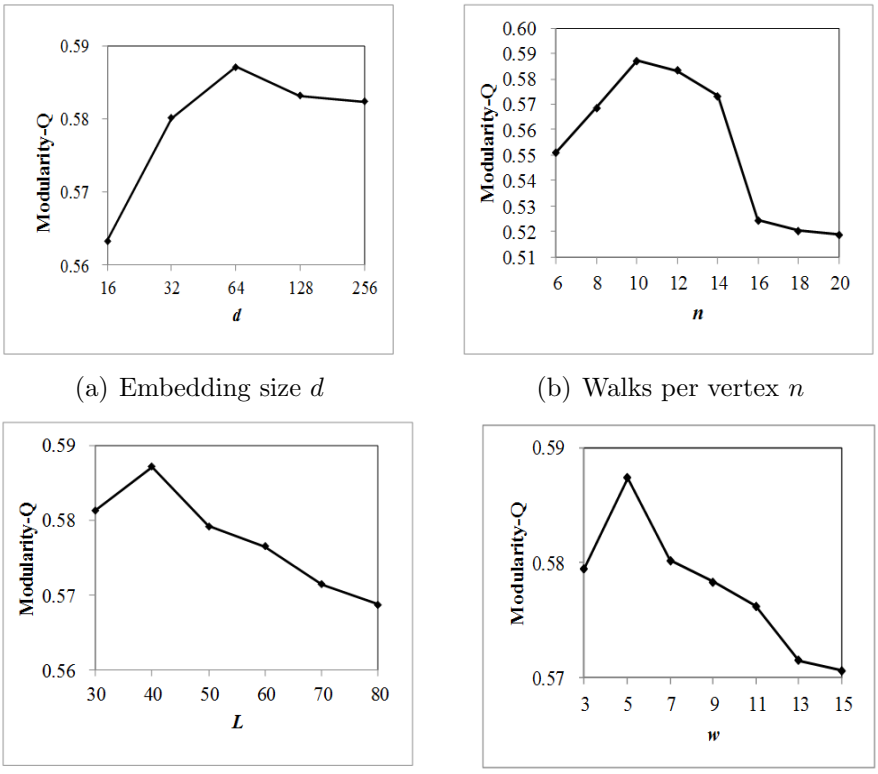
$$V = |\max(x_{ij}) - Others(x_{ij})| \leq \lambda \tag{11}$$

In summary, the network embedding based on TANE algorithm takes account of the information of the network structure and semantic, which has practical application value, such as rationalization and personalized recommendation in the recommendation system.

5.3. **Parameter sensitivity.** There are several parameters involved in the TANE algorithm, which are analyzed by the dataset of Douban. Firstly, use the Douban network to set five cases to analyze the parameters in the transfer probability in Table 3. The Q values obtained by the set of the 4th and 5th are relatively small, because the topic nodes are sparse in the network, to a large extent, the connection established by topic nodes seriously affects the extraction of information in the network. The Q values obtained by

TABLE 3. Transition probability model parameter analysis

n	Parameters									
	α	β	γ_1	γ_2	δ	χ_1	χ_2	ε	θ	Q_n
1	0.75	0.25	0	0	0	0	1	0.7	0.3	0.53
2	0.4	0.3	0.1	0.1	0.1	1/3	2/3	0.7	0.3	0.59
3	0.25	0.25	1/6	1/6	1/6	1/2	1/2	0.7	0.3	0.47
4	0.1	0.3	0.2	0.2	0.2	2/3	1/3	0.7	0.3	0.32
5	0	0.25	0.25	0.25	0.25	1	0	0.7	0.3	0.14



(a) Embedding size d (b) Walks per vertex n

(c) Walk length L (d) Window size w

FIGURE 6. Word2vec model parameter sensitivity analysis

the set of the 1st, 2nd and 3rd are relatively large, and it can be seen from $Q_2 > Q_1$ and $Q_1 > Q_3$ that under the premise of weighing the weight of each part, modeling based on the relationship between the two types of nodes is conducive to network embedding. Finally, select the second set in Table 3 for parameters based on the transfer probability model of the Douban dataset, and the parameters in the transfer probability model based on the Karate dataset are $\alpha = 0.35$, $\beta = 0.25$, $\gamma_1 = \gamma_2 = 0.15$, and the remaining parameters are the same as the Douban dataset, with the two networks different and the parameter settings slightly different.

Secondly, several parameters are analyzed in the Word2vec model in Figure 6. Some parameters in the Word2vec model have little influence on the experimental results, and the TANE algorithm performs better when the vector dimension d is 64 and the context window ω is 5. The walks per vertex n and path length L have obvious effects on the experimental results, especially the parameter L. When $L > 40$, the Q value is in a downward trend and the algorithm performance decreases.

In summary, the parameters of the transfer probability model in the TANE algorithm are selected based on the sparseness of the two types of nodes in the network to weigh the probability of each part of the modeling. The parameters of the Word2vec model are considered in combination with the complexity and algorithm performance.

6. **Conclusions.** Based on the idea of determination and uncertainty in SPA theory, the transfer probability models are given, and the models are applied to the random walk algorithm to get relatively high-quality random walk sequences. The walk sequences are trained and the topic-attention network embedding is obtained. The experimental results show that compared with the four classical algorithms, when the number of divided communities is $5 \sim 13$, the Q value obtained by the TANE algorithm has a different degree of improvement, but the sparsity of the network makes the Q value change less obviously. When $K = 11$ on the Douban network, the Q value of the TANE algorithm is improved by nearly 6.5% compared with metapath2vec algorithm. In conclusion, from the structural and semantic aspects of the analysis of the network, we can more comprehensively capture the information in the network. Considering comprehensive rationality and experimental results, the proposed TANE algorithm has achieved better performance.

Next, based on the idea of TANE algorithm, increasing the scale of the data and considering the dynamics and complexity of the heterogeneous networks for network embedding relevant research is the focus of the next step.

Acknowledgment. This work is partially supported by National Natural Science Foundation of China (61472340), the National Youth Science Foundation of Hebei (F2017209070) the Doctoral Research Start-up Fund (Natural Science) of Hebei Normal University of Science and Technology (2019YB011) the Natural Science Foundation of Hebei (F2019203157) and the Key Project of Science and Technology Research Project in Hebei (ZD2019004). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] C. C. Tu, C. Yang, Z. Y. Liu et al., Network representation learning: An overview, *SCIENTIA SINICA Information*, vol.47, no.8, pp.980-996, 2017.

[2] M. Sun, Y. Zhao and Z. Liu, Representation learning for measuring entity relatedness with rich information, *Proc. of the 24th International Joint Conference on Artificial Intelligence*, Buenos Aires, pp.1412-1418, 2015.

[3] J. Tang, M. Qu and Q. Z. Mei, PTE: Predictive text embedding through large-scale heterogeneous text networks, *Proc. of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, pp.1165-1174, 2015.

[4] H. X. Chen, H. Z. Yin, W. Q. Wang et al., PME: Projected metric embedding on heterogeneous networks for link prediction, *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discover and Data Mining*, New York, pp.1177-1186, 2018.

[5] Y. Dong, N. V. Chawla and A. Swami, Metapath2vec: Scalable representation learning for heterogeneous networks, *Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, pp.135-144, 2017.

[6] T. Y. Fu, W. C. Lee and Z. Lei, HIN2Vec: Explore meta-paths in heterogeneous information networks for representation learning, *Proc. of the 17th ACM CIKM International Conference on Information and Knowledge Management*, New York, pp.1797-1806, 2017.

[7] D. Y. Zhu, P. Cui, D. X. Wang et al., Deep variational network embedding in Wasserstein space, *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discover and Data Mining*, New York, pp.2827-2836, 2018.

[8] H. Deng, J. Han, B. Zhao et al., Probabilistic topic models with biased propagation on heterogeneous information network, *Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discover and Data Mining*, New York, pp.1271-1279, 2011.

[9] A. Q. Li, A. Ahmed, S. Ravi et al., Reducing the sampling complexity of topic models, *Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, pp.891-900, 2014.

[10] X. Chen, J. F. Guo and C. Z. Fan, Research on the method of community discovery based on the topic-attention network of connection degree, *Journal of Computer Engineering and Applications*, vol.53, no.17, pp.85-93, 2017.

[11] Y. Z. Sun, J. W. Han, X. F. Yan et al., Pathsim: Meta path-based top-k similarity search in heterogeneous information networks, *Proc. of the 37th International Conference on Very Large Data Base*, Westin Seattle, pp.992-1003, 2011.

[12] B. Perozzi, R. AlRfou and S. Skiena, DeepWalk: Online learning of social representations, *Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, pp.701-710, 2014.

[13] X. Rong, Word2vec parameter learning explained, *arXiv Preprint, arXiv:1411.2738*, 2016.

[14] A. Grover and J. Leskovec, Node2vec: Scalable feature learning for networks, *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, pp.855-864, 2016.

[15] J. Tang, M. Qu, M. Wang et al., LINE: Large-scale information network embedding, *Proc. of the 24th International Conference on World Wide Web*, New York, pp.1067-1077, 2015.

[16] L. Xu, L. Huang and C. D. Wang, Motif-preserving network representation learning, *Journal of Computer Science and Technology*, vol.13, no.8, pp.1261-1271, 2019.

[17] D. Wang, P. Cui and W. Zhu, Structural deep network embedding, *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, pp.1225-1234, 2016.

[18] Y. Z. Sun, B. Norick, J. W. Han et al., Integrating meta-path selection with user-guided object clustering in heterogeneous information networks, *Proc. of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, pp.1348-1356, 2012.

[19] Y. Kim and N.-W. Cho, Research trends in social network analysis using topic modeling and network analysis, *ICIC Express Letters*, vol.12, no.1, pp.71-78, 2018.

[20] K. Q. Zhao, *Set Pair Analysis and Preliminary Application*, Zhejiang Science and Technology Press, Hangzhou, 2000.