International Journal of Innovative Computing, Information and Control
Volume **18**, Number **3**, June **2022**

ICIC International ©2022 ISSN 1349-4198
pp. **883–900**

NETWORK REPRESENTATION LEARNING BASED ON RANDOM WALK OF CONNECTION NUMBER

Xiao Chen[1], Ying Wang[1,*], Hui Dong[2], Xiao Pan[2] and Jia Li[3]

[1]Research Center of Marine Sciences
Hebei Normal University of Science and Technology
No. 360, Hebei Avenue, Qinhuangdao 066004, P. R. China
chenxiao0604@hevttc.edu.cn; *Corresponding author: wy3748@hevttc.edu.cn

[2]School of Economics and Management
Shijiazhuang Tiedao University
No. 17, North Second Ring Road, Shijiazhuang 050043, P. R. China
{ 1181992484; 26458784 }@qq.com

[3]Information Technology Center
Baoding University of Technology
No. 1689, South Second Ring Road, Baoding 071000, P. R. China
49011415@qq.com

Received October 2021; revised February 2022

Abstract. *Existing methods based on random walks do not consider the high-order topological structure and the uncertainty relationship between nodes. Thus, the vector representations of nodes are inaccurate. To solve the above problems, we propose a novel algorithm TANE-RWCN (Topic-Attention Network Embedding based on Random Walk of Connection Number). Firstly, the topic-attention network is described as an identical-discrepancy-contrary system based on the high-order topology structure. Secondly, a random walk strategy suitable for the network is designed based on the social relationship between users and the user's preference for the topics; then, the transition probability model is constructed. Thirdly, based on the random walks and the skip-gram model, the vector representations of nodes are obtained. Finally, the task of overlapping community discovery based on Fuzzy C-Means and TANE-RWCN is verified by experiments. Compared with the classical algorithms, TANE-RWCN has higher modularity value.*
Keywords: Network representation learning, Topic-attention network, Connection number, Random walk, Transition probability

1. **Introduction.** At present, complex network analysis has been widely used in various fields [1], such as transportation networks, power networks, biological networks and social networks. With the growth of network scale and complexity, the methods based on the adjacency matrix present some problems, such as high dimensions, sparseness, and coupling. As a result, the development of various network analysis tasks encounters bottlenecks. An independent, low-dimensional and dense representation method is urgently needed. Fortunately, network representation learning [2] can solve the above problems. Moreover, with the development of deep learning technology, network representation learning has become one of the current research hotspots.

The DeepWalk model [3] based on the Word2Vec [4] was first proposed in 2014. Then, many advanced models [5-8] are proposed. According to the number of entity types in networks, the existing methods can be divided into homogeneous and heterogeneous network representation learning. According to the different learning methods, the existing

DOI: 10.24507/ijicic.18.03.883

methods can be divided into matrix decomposition and neural network model. According to the number of neural network layers, the existing methods can be divided into shallow learning and deep learning. According to the different learning strategies, the existing methods can also be divided into random walks and user-defined loss functions, and so on.

This paper mainly focuses on a special kind of heterogeneous network, the Topic-Attention Network (short for TAN) [9], as shown in Figure 1. We study the shallow network representation learning based on random walks and the skip-gram model. The existing methods based on random walks realize the sampling of the node's neighbors in networks [10]. In the existing researches on heterogeneous network representation learning based on random walks, most of them are based on meta-paths, such as Metapath2vec [5], HIN2Vec [11], and ESim [12]. However, the selection of meta-path is difficult, and requires the prior knowledge of domain experts or proposes (general or task-specific) strategies to combine a set of predefined meta-paths (e.g., meta-paths shorter than a predefined length). The meta-path also limits the flexibility of walking. Therefore, the methods of random walks based on non-meta-path have been proposed, such as JUST (JUmp and STay) [6]. At this time, the random walk is realized by the transition probability based on the tightness between nodes. No matter based on meta-path or not, the most of existing methods based on random walks only consider the low-order proximity and certainty of networks, and do not consider the high-order proximity and the uncertainty relationship. As a result, the node's representation cannot be described accurately.

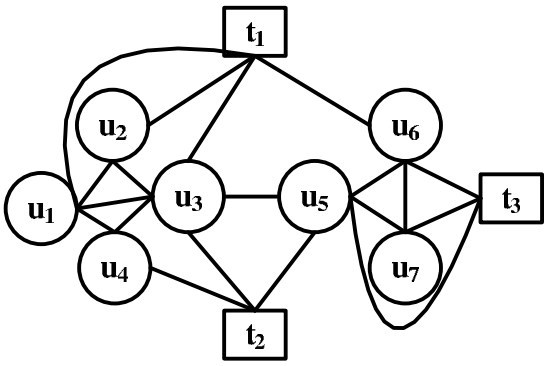


FIGURE 1. Topic-attention network

The characteristics of network structure are mainly divided into local and global. The low-order (1-order or 2-order) proximity between nodes only reflects the local structure of networks. A pair of nodes with edges (1-order) or common neighbors (2-order) indicates that the pair of nodes has a certain relationship. However, the tightness between each node pair is different. For example, a pair of nodes with common neighbors can or cannot become friends in the future depending on the aggregation ability of their common neighbors, which reflects a kind of uncertainty. The high-order ($\geq$ 3-order) proximity between nodes reflects the global structure of networks. The tightness of this pair of nodes is determined by all nodes on the path, which also reflects a kind of uncertainty. Therefore, to further improve the accuracy of vector representations, how to consider the high-order proximity and the uncertainty is the main research content of this paper.

Our paper first introduces the idea of multi-element connection number in set pair analysis theory [13], and describes the TAN as a certainty (identical or contrary)-uncertainty (discrepancy) system to solve the uncertainty relationship between nodes. Second, we design a random walk strategy according to the topological structure and high-order proximity. Then, we build a transition probability model based on the strategy. Finally,

the vector representations of nodes are obtained based on random walks and the skip-gram model in TAN, and applied to overlapping community discovery tasks.

The main contributions of this paper are summarized as follows.

1) We describe TAN as an Identical-Discrepancy-Contrary (short for IDC) system by the multi-element connection number. The multi-element relationship between nodes is constructed based on the L-step reachability, including the identical, discrepancy (partial identical, partial discrepancy and partial contrary) and contrary.

2) To better express the relationship between various entities of TAN, a suitable random walk strategy is proposed in the IDC system. The strategy combines the user's preference for topics and high-order proximity.

3) The transition probability model is constructed based on the random walk strategy; where considering the node's degree, path length, and the multi-element connection number between nodes are considered.

4) Through extensive experiments, we evaluate TANE-RWCN by conducting the task of overlapping community discovery based on the Fuzzy C-Means (short for FCM) algorithm, using two real datasets. The experimental results show that the TANE-RWCN outperforms the existing models.

The reminders of this paper are organized as follows. Section 2 introduces the related work. The definitions of the topic-attention network and connection number, and the problem statement of network representation learning based on random walks are described in Section 3. We present the framework of the TANE-RWCN algorithm in Section 4 and show the experimental results in Section 5. The last section concludes the paper and forecasts future work.

2. **Related Works.**

At present, the methods of representation learning on complex networks are mainly divided into three categories: random walk, decomposition and deep learning.

1) In the methods based on random walks, a node's neighbors are sampled by describing the transition probability between nodes. For example, Node2vec model [14] combines BFS and DFS walks with DeepWalk. The model improves the node classification task, since the node vectors contain local and global information. Metapath2vec model is proposed on heterogeneous networks based on the random walk of meta-path. HIN2Vec model learns the vector representations of nodes and meta-paths, from the complex relationship between various types of nodes. ESim model performs random walks based on meta-paths on heterogeneous networks, and learns the vector representations of nodes that appear in a meta-path instance by maximizing the probability of the meta-path instance. HINE model [15] calculates the proximity based on meta-paths, and learns node representations.

2) Methods based on the decomposition of matrix or network, which use the idea of divide and conquer. For example, MERP model [16] considers both the accuracy and efficiency of social recommender systems. The basic idea is to generate user social features by SNE and user latent features with PMF, and then these two features are combined to predict ratings. HERec model [17] transforms a heterogeneous network into multi-homogeneous networks based on the meta-path extraction. Then, it fuses the vector representations of multi-homogeneous networks through fusion functions. EOE model [18] transforms the academic network into a word co-occurrence network and an author co-occurrence network. The model learns vector representations of the node pairs within and between subnets.

3) Methods based on deep learning model. For example, SDNE model [19] constructs a multi-layer nonlinear function. The model is composed of the first-order proximity

supervised component and the second-order proximity unsupervised component. It is robust to sparse networks. DIME model [20] obtains the information of heterogeneous networks under different meta-paths, and jointly encodes the information by an autoencoder. Different networks are fused through a transition matrix. SHINE model [21] transforms a heterogeneous network into an emotional relation network, a social relation network and a user attribute network. It uses multi-depth auto-encoders to extract the user's non-linear representation from the three subnetworks, and uses aggregation functions to fuse them.

At present, studies based on homogeneous network are more extensive and in-depth, while few studies are based on heterogeneous networks. In the research of heterogeneous network representation learning, most of them are guided by meta-paths. However, different meta-paths have different semantics. Different meta-paths will get different results. Methods based on meta-paths can only preserve the network structure related to the meta-paths. Therefore, it is necessary to study representation learning based on non-meta-path according to the characteristics of different heterogeneous networks.

3. Preliminaries and Problem Statement.

3.1. Topic-attention networks.

Definition 3.1. *The topic-attention network [7] is defined as a two-tuple* $TAN = (V, E)$. $V = \{U, T\}$ *is the node set, where* $U = \{u_1, u_2, \ldots, u_n\}$ *is the user node set and* $T = \{t_1, t_2, \ldots, t_m\}$ *is the topic node set.* $E = \{EU, EUT\}$ *is the edge set, where* $EU = \{(u_i, u_j)|u_i, u_j \in U\}$ *is the relationship between users, and* $EUT = \{(u_i, t_j)|u_i \in U, t_j \in T\}$ *is the relationship between the users and the topics.*

In TAN, if $e(u_i, u_j) \in EU$, then $|e(u_i, u_j)| = 1$; otherwise $|e(u_i, u_j)| = 0$. The relationship of other node pairs can be obtained in the same way. $N(u_i)_L^{TAN} = NU(u_i)_L^{TAN} \cup NT(u_i)_L^{TAN}$ is the L-order neighbor set of u_i, where $NU(u_i)_L^{TAN}$ and $NT(u_i)_L^{TAN}$ are the L-order user neighbor set and topic neighbor set of u_i respectively. $CN(u_i, u_j)_{L\cap M}^{TAN} = N(u_i)_L^{TAN} \cup N(u_j)_M^{TAN} = CNU(u_i, u_j)_{L\cap M}^{TAN} \cup CNT(u_i, u_j)_{L\cap M}^{TAN}$ is the common $L \cap M$-order neighbor set of u_i and u_j, where $CNU(u_i, u_j)_{L\cap M}^{TAN}$ and $CNT(u_i, u_j)_{L\cap M}^{TAN}$ are the common $L \cap M$-order user neighbor set and topic neighbor set of u_i and u_j, respectively. When $L = M$, it is abbreviated as the common L-order neighbor set of u_i and u_j. The descriptions of the topic's neighbor set, common neighbor set, and common neighbor set of the user and topic are similar. Detailed definitions and explanations can be found in [7].

3.2. Multi-element connection number.

The connection number was proposed by Zhao in the IDC system of set pair analysis theory [13]. The purpose is constructing set pair $H = (A, B)$ to study the relationship between entities from the perspective of the identical, discrepancy and contrary. That is $\mu(A, B) = I + Di + Cj$. I, D and C represent the identical degree, discrepancy degree and contrary degree between entities respectively. Therefore, it is also called the three-element connection number.

When the discrepancy changes to the identical or contrary, some elements in the discrepancy may be biased to become the identical, others may be biased to be the contrary, and the rest may continue to be discrepancy. The three-element connection number can be further extended to five-element. That is, the discrepancy is further divided into three parts: partial identical, partial discrepancy and partial contrary. The definition of the five-element connection number is as follows.

Definition 3.2. *Given a set pair* $H = (A, B)$*, under a specific background, the attributes of two sets* A *and* B *are analyzed, and* $N\,(= I + D + C)$ *attributes are obtained. Among them,* I *is the number of identical attributes,* C *is the number of contrary attributes, and*

$D\,(= DI + DD + DC)$ is the number of discrepancy attributes. DI, DD and DC represent the number of partial identical, partial discrepancy and partial contrary attributes, respectively. Then, the five-element connection number μ is shown in Formula (1).

$$\mu(A, B) = I + D(Ii_I + Di_D + Ci_C) + Cj = I + DIi_I + DDi_D + DCi_C + Cj \tag{1}$$

In Equation (1), $i_I \in (0, 1]$ is the probability that the discrepancy tends to be identical, $i_C \in [-1, 0)$ is the probability that the discrepancy tends to be the contrary, and $i_D \in [-1, 1]$ is the difference marker. The larger the value of $|i|$ is, the greater the conversion probability is.

In the five-element connection number, if one or more of them is ignored, it can also be called the compressed or reduced connection number. Therefore, we can flexibly define or characterize the components of the connection number according to the actual problems.

3.3. **Problem formulation.** Given a $TAN = (V, E)$, our goal is to learn the mapping function $f : V \longrightarrow R^{|V|\times d}$ in the network based on a random walk strategy and skip-gram model. f is a mapping function. V is the node set. $R^{|V|\times d}$ is the vector representations of nodes, and its size is $|V| \times d$, where $d \ll |V|$.

The purpose of random walk is to generate a walk path $\{v_1, v_2, \dots v_i, v_j, \dots\}$ based on a specific sampling strategy. For each current node $v_i \in V$, $v_j \in N(v_i)_L$ is the next node in the path. $N(v_i)_L \subset V$ as the candidate set of the next nodes, is defined through a neighborhood sampling strategy. Note that the probability of v_i to v_j is usually different from that of v_j to v_i in the process of random walking.

The core tasks of random walks are to design a random walk strategy and build a transition probability model. Most traditional random walk strategies are based on the 1-order (edge) or 2-order (common neighbor) relationship, without considering the high-order (≥ 3) relationship between nodes. However, according to the six-degree separation theory, any two nodes can be reached within 6-steps. Therefore, the higher-order path can be considered. For example, in TAN, users can establish connections sharing common topics or neighbors (2-order). Users can establish connections with new friends through friends' recommendations from friends (3-order). Users can establish connections by following the common topic with friends' friends (4-order). Users can also establish connections with new topic through friends' recommendations (2-order), etc. In this paper, considering the influence and computational complexity of transition probability, we choose the 1-4 order topological structure in the network as the research objects.

In TAN, we consider two entities (users and topics), three relationship (users to users, users to topics, and topics to users), the influence of topics and the higher-order proximity constructing a new random walk strategy, as follows.

1) When the current node is the user, the next node can be selected according to the user's preference. The candidate set of next nodes can be the users with 1-4 order reachability, or the topics with 1-2 order reachability. Since the user has subjective consciousness, the next node is realized by non-uniform sampling.

2) When the current node is the topic, the candidate set is only the users. Since the topic does not have subjective awareness, the next node is realized by uniform sampling.

4. **Representation Learning Based on Random Walks of Connection Number in TAN.** To realize the vector representations based on random walks and multi-element connection number in TAN. The focus of this paper is mainly divided into two aspects. 1) How to describe the TAN as an IDC system to solve the uncertainty relationship between nodes. 2) How to establish a transition probability model based on the IDC system, the multi-element connection number, the network topology and the random walk strategy

to better express the tightest relationship between various entities. The details are as follows.

4.1. **Identical-discrepancy-contrary system based on topic-attention network.** Based on the idea of a five-element connection number, and drawing on previous research on IDC system, a TAN is described as an IDC system. The main task is to find the relationship and attributes of identical, partial identical, partial discrepancy, partial contrary and contrary between nodes.

In TAN, for any pair of nodes, the IDC relationship is determined according to the L-step ($L = 1, 2, \dots$) accessibility in a network, as the following.

1) If a node is reachable from another node in 1-step, they have the identical relationship. The nodes pair are their identical attributes.

2) If a node is reachable from another node in 2, 3 or 4-steps, they have the discrepancy relationship. The intermediate nodes in this path are their discrepancy attributes.

3) If a node cannot be reachable from another node within 4-steps, they have the contrary relationship. The intermediate nodes in this path are their contrary attributes.

For users to users, the IDC relationship is shown in Figure 2. We can see from Figure 2, the discrepancy attributes are further refined into partial identical, partial discrepancy and partial contrary.

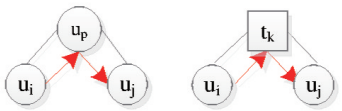
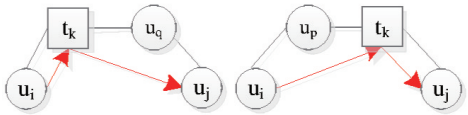
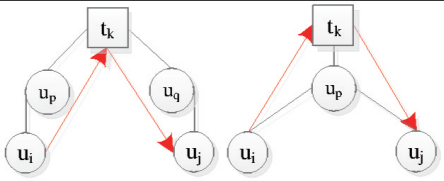
Relationship		Topological Structures	
Identical		1-order	
Discrepancy	Identical	2-order	
	Discrepancy	3-order	
		4-order	
	Contrary	3,4-order	Other network structures of 3,4-order
Contrary		>=5-order	Network structures of 5-order or higher

FIGURE 2. Topic-attention network

1) The common 1-order users and topics are regarded as partially identical attributes.

2) The common $1 \cap 2$, $2 \cap 1$ and 2-order topics are regarded as partially discrepancy attributes.

3) The remaining user nodes in 3 or 4-order network structures are regarded as partially contrary attributes.

Thus, the five-element connection number between users is shown in Formula (2).

$$\begin{aligned} \mu(u_i, u_j) &= I + DIi_I + DDi_D + DCi_C + Cj \\ &= |e(u_i, u_j)| + (|CNT(u_i, u_j)_1| + CNU(u_i, u_j)_1) \times i_I \end{aligned}$$

$$
\begin{aligned}
&+(|CNT(u_i,u_j)_{1\cap 2}|+|CNT(u_i,u_j)_{2\cap 1}|+|CNT(u_i,u_j)_2|)\times i_D\\
&+(|N(u_i)_1|+|N(u_i)_2|+|N(u_j)_1|+|N(u_j)_2|-|CN(u_i,u_j)_1|\\
&-|CNT(u_i,u_j)_{1\cap 2}|-|CNT(u_i,u_j)_{2\cap 1}|+|CNT(u_i,u_j)_2|)\times i_C\\
&+\left(\sum_{L\geq 3 \text{ or } M\geq 3}|CN(u_i,u_j)_{L\cap M}|\right)\times j
\end{aligned}
\tag{2}
$$

According to the characteristics of the TAN, the multi-element relationship between users and topics is simplified. For the relationship of users to topics, we consider whether they have edges, and whether they have common 1-order user neighbors. For the relationship of topics to users, we only consider whether there is an edge. Their connection number can be constructed in the same way, which is omitted here. Finally, TAN has been described as an IDC system. The following main task is to build the transition probability model based on the IDC system and random walk strategy.

4.2. **Transition probability model.** To improve the accuracy of the vector representations, this paper uses the multi-element connection number to describe the tightness between various entities in TAN, and gives a transition probability model based on set pair theory.

Given a TAN, $\forall u_i, u_j \in U$, $\forall t_k \in T$, according to the three categories of relationship between nodes: (u_i, u_j), (u_i, t_k), and (t_k, u_i), the transition probability is constructed as follows.

4.2.1. *Transition probability from users-to-users.* In a TAN-IDC system, the core task is to construct the transition probability model, where we transform the five-element connection number between nodes into the corresponding transition probability. For users to users, we transform Equation (2) to its transition probability, that is $P(u_j|u_i) = P_I(u_j|u_i)+P_D(u_j|u_i)i+P_C(u_j|u_i)j = P_I(u_j|u_i)+P_{DI}(u_j|u_i)i_I+P_{DD}(u_j|u_i)i_D+P_{DC}(u_j|u_i)i_C +P_C(u_j|u_i)j$.

In this paper, we expect to obtain the maximum transfer probability between nodes based on TAN-IDC. That is, it is expected to consider the contribution of the identical attribute, and all other attributes convert to the identical attributes. However, the contrary and partially contrary attributes cannot be converted to be identical. Therefore, we ignore C and DC, and only consider the influence of I, DI and DD. Moreover, we want to convert DI and DD to I. Therefore, let $i_I = i_D = 1$, and $i_C = j = 0$. When describing the transition probability, we need to consider not only the influence of the number of identical, discrepancy and contrary attributes, but also the influence of the node degree, and path length. Therefore, based on the five-element connection number between users and the 1-4 order proximity in the network, as shown in Figure 2, the transition probability of users to users is constructed, denoted as $P(u_j|u_i)$, as shown in Formula (3).

$$
\begin{aligned}
P(u_j|u_i) &= w_I P_I(u_j|u_i) + w_D P_D(u_j|u_i)\\
&= w_I P_I(u_j|u_i) + w_D(\delta_I P_{DI}(u_j|u_i) + \delta_D P_{DD}(u_j|u_i))
\end{aligned}
\tag{3}
$$

In Equation (3), $P_I(u_j|u_i)$, $P_{DI}(u_j|u_i)$ and $P_{DD}(u_j|u_i)$ represent the transition probability of u_i arriving at u_j through identical, partially identical and partially discrepancy attributes, respectively. w_I and w_D represent the weights of the identical and discrepancy respectively, and $w_I + w_D = 1$. δ_I and δ_D represent the weight of the partially identical and partially discrepancy respectively, and $\delta_I + \delta_D = 1$.

$P_I(u_j|u_i)$ is the transition probability under the 1-order connected edge, as shown in Formula (4).

$$P_I(u_j|u_i) = P_{ij}(1) = \frac{|e(u_i, u_j)|}{d(u_i)} \tag{4}$$

In Equation (4), if $(u_i, u_j) \in EU$, $P_I(u_j|u_i)$ is the inverse of the degree of u_i; otherwise, it is 0.

$P_D(u_j|u_i) = \delta_I P_{DI}(u_j|u_i) + \delta_D P_{DD}(u_j|u_i)$ is the transition probability when 2-4 order structures.

$P_{DI}(u_j|u_i)$ is the transition probability of $CNU(u_i, u_j)_1$ and $CNT(u_i, u_j)_1$ under the 2-order structures, as shown in Formula (5).

$$P_{DI}(u_j|u_i) = \delta_{I1} P_{ij}^{CNT}(2) + \delta_{I2} P_{ij}^{CNU}(2) = \frac{\delta_{I1}|CNT(u_i, u_j)_1| + \delta_{I2}|CNU(u_i, u_j)_1|}{d(u_i) + d(u_j) - |CN(u_i, u_j)_1|} \tag{5}$$

$P_{DD}(u_j|u_i)$ is the transition probability of $CNT(u_i, u_j)_{1\cap 2}$ and $CNT(u_i, u_j)_{2\cap 1}$, under the 3-order structures and $CNT(u_i, u_j)_2$ under the 4-order structures, as shown in Formula (6).

$$\begin{aligned} P_{DD}(u_j|u_i) &= \delta_{D1} P_{ij}(3)_{1\cap 2} + \delta_{D2} P_{ij}(3)_{2\cap 1} + \delta_{D3} P_{ij}(4)_{2\cap 2} \\ &= \delta_{D1} \sum_{t_k \in CNT(u_i,u_j)_{1\cap 2}} \frac{1}{d(u_i)} \times \frac{|CNT(t_k, u_j)_1|}{M1} \\ &\quad + \delta_{D2} \sum_{t_k \in CNT(u_i,u_j)_{2\cap 1}} \frac{|CNT(u_i, t_k)_1|}{M2} \times \frac{1}{d(t_k)} \\ &\quad + \delta_{D3} \sum_{t_k \in CNT(u_i,u_j)_2} \frac{|CNT(u_i, t_k)_1|}{M1} \times \frac{1}{d(t_k)} \times \sum_{u_p \in CNU(t_k,u_j)_1} \frac{1}{d(u_p)} \end{aligned} \tag{6}$$

4.2.2. *Transition probability from users-to-topics.* For users to topics, the complete expression of the transition probability is $P(t_k|u_i) = P_I(t_k|u_i) + P_{DI}(t_k|u_i)i_I + P_{DD}(t_k|u_i)i_D + P_{DC}(t_k|u_i)i_C + P_C(t_k|u_i)j$. Similarly, we expect to obtain the maximum transfer probability value; thus, let $i_I = i_D = 1$, and $i_C = j = 0$. Based on the network structure, users usually walk to topics that they are directly following or recommended by friends. For example, $\{u_i, t_k\}$ (1-order proximity, that is identical) or $\{u_i, u_q, t_k\}$ (2-order proximity, that is partial identical) in Figure 2, the partial discrepancy attribute is an empty set. Therefore, based on the 1-2 order proximity of users to topics in the network, we ignore the C, DC and DD, and only consider the influence of I and DI; the transition probability of users to topics is constructed, denoted as $P(t_k|u_i)$, as shown in Formula (7).

$$\begin{aligned} P(t_k|u_i) &= \theta_I P_I(t_k|u_i) + \theta_D P_{DI}(t_k|u_i) = \theta_I P_{ik}(1) + \theta_D P_{ik}(2) \\ &= \theta_I \times \frac{|e(u_i, t_k)|}{d(u_i)} + \theta_D \times \frac{|CNU(u_i, t_k)_1|}{d(u_i) + d(t_k) - |CN(u_i, t_k)_1|} \end{aligned} \tag{7}$$

In Equation (7), $P_I(t_k|u_i)$ and $P_{ID}(t_k|u_i)$ represent the transition probability of u_i arriving at t_k through the identical and partially identical attributes, respectively. θ_I and θ_D represent the weights of the identical and discrepancy respectively, and $\theta_I + \theta_D = 1$.

$P_I(t_k|u_i)$ is the transition probability under the 1-order connected edge. If $(u_i, t_k) \in EUT$, $P_I(t_k|u_i)$ is the inverse of the degree of u_i; otherwise, it is 0.

$P_{DI}(t_k|u_i)$ is the transition probability of $CNU(u_i, t_k)_1$ under the 2-order structures.

4.2.3. *Transition probability from topics-to-users.* When constructing the transition probability of topics to users, we only consider the influence of I. Since the topic has no subjective awareness, and it can only walk to the users directly connected, based on the 1-order relationship between topics and users in the network, the transition probability of topics to users is constructed, denoted as $P(u_i|t_k)$, as shown in Formula (8).

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

4.3. **Example.** To understand the above transition probability model, taking Figure 1 as an example, the calculation methods of (u_3, u_1), (u_3, t_2) and (t_2, u_3) are given.

For users to users, when taking (u_3, u_1) as the research object, first obtain the 1-order and 2-order neighbor sets of each node. That is, $NT(u_3)_1 = \{t_1, t_2\}$, $NU(u_3)_1 = \{u_1, u_2, u_4, u_5\}$, $NT(u_1)_1 = \{t_1\}$, and $NU(u_1)_1 = \{u_2, u_3, u_4\}$. Secondly, the number of identical, discrepancy and contrary attributes between nodes is calculated, as shown in Table 1. $(u_3, u_1) \in EU$, then the number of identical attributes is $|e(u_3, u_1)| = 1$. $CNT(u_3, u_1)_1 = \{t_1\}$, then the number of partial identical topics is $|CNT(u_3, u_1)_1| = 1$. $CNU(u_3, u_1)_1 = \{u_2, u_4\}$, then the number of partial identical users is $|CNU(u_3, u_1)_1| = 2$. $CNT(u_3, u_1)_{1\cap 2} = \{t_2\}$, $CNT(u_3, u_1)_{2\cap 1} = \emptyset$ and $CNT(u_3, u_1)_2 = \emptyset$, then the number of partial discrepancy topics is $|CNT(u_3, u_1)_{1\cap 2}| + |CNT(u_3, u_1)_{2\cap 1}| + |CNT(u_3, u_1)_2| = 1$. Therefore, the transition probability from u_3 to u_1 is

$$P(u_3|u_1) = w_I \times \frac{1}{6} + w_D \times \left[\delta_I \times \left(\delta_{I1} \times \frac{1}{6+4-3} + \delta_{I2} \times \frac{2}{6+4-3}\right) + \delta_D \times \left(\delta_{D1} \times \frac{1}{6} \times \frac{1}{3+4-1} + 0\right)\right].$$

Similarly, we can obtain the identical, discrepancy and contrary attributes of (u_3, t_2) and (t_2, u_3), as shown in Table 1, and their transition probabilities are $P(t_2|u_3) = \theta_I \times 1/6 + \theta_{ID} \times 2/(6+3-2)$ and $P(u_3|t_2) = 1/3$, respectively.

Table 1. Parameters setting of TANE-RWCN

	Node pairs	(u_3, u_1)	(u_3, t_2)	(t_2, u_3)
Identical attributes	Edge	1	1	1
Partial identical attributes	Common 1-order topics	1	–	–
	Common 1-order users	2	2	–
Partial discrepancy attributes	Common $1 \cap 2$-order topics	1	–	–
	Common $1 \cap 2$-order topics	0	–	–
	Common 2-order topics	0	–	–

4.4. **Algorithm description.** The detailed algorithm TANE-RWCN (Topic Attention Network Embedding based on Random Walk of Connection Number) is shown in Algorithm 1.

Algorithm 1: TANE-RWCN
Input: TAN, n, L, d, ω, w_I, w_D, δ_I, δ_D, δ_{I1}, δ_{I2}, δ_{D1}, δ_{D2}, δ_{D3}, θ_I, θ_D;
Output: $\Phi \in R^{|V|\times d}$.
1) init Walks = $\emptyset$
2) for $i = 0$ to n do
3) rand.shuffle(V)
4) for each v_i in V do
5) $Path(v_i)_L$ = RW-CN($TAN, v_i, L, w_I, w_D, \delta_I, \delta_D, \delta_{I1}, \delta_{I2}, \delta_{D1}, \delta_{D2}, \delta_{D3}, \theta_I, \theta_D$)
6) $Walks = Walks + Path(v_i)_L$
7) $\Phi = Word2vec(Walks, d, \omega, sg = 1, hs = 1)$

Algorithm 1 is mainly divided into three steps. Step 1 initializes the random walk path set Walks to be empty, as shown in line 1). Step 2 obtains the final path set Walks by

random walk, as shown in lines 2)-6). In step 3, we put the Walks into the skip-gram model for training, so that the vector representation of each node in TAN is obtained, as shown in line 7).

Among them, the step 2 can be divided into three small steps, as follows. (1) The number of cycles as the starting node is specified, as shown in line 2). (2) All nodes in the network are randomly sorted, and the node sequence V is obtained, as shown in line 3). For each v_i in V, the Random Walk based on the Connection Number (RW-CN) algorithm is called in turn, as shown in Algorithm 2; the $Path(v_i)_L$ with v_i as the starting node and walk length L is obtained; to obtain the final walking path set Walks, as shown in lines 4)-6).

Algorithm 2: RW-CN
Input: TAN, v_i, L, w_I, w_D, δ_I, δ_D, δ_{I1}, δ_{I2}, δ_{D1}, δ_{D2}, δ_{D3}, θ_I, θ_D;
Output: $Path(v_i)_L$.
1) init $Path(v_i)_L = \emptyset$
2) $Path(v_i)_L = v_i$
3) while $Len(Path(v_i)_L) < L$ do
4) $v_{current} = Path(v_i)_L[-1]$, $P = \emptyset$
5) if $Type(v_{current}) = U$ then
6) $NU = NU(v_{current})_1 \cup NU(v_{current})_2 \cup NU(v_{current})_3 \cup NU(v_{current})_4$
7) for v_{next} in NU do
8) $P(v_{next}|v_{current}) = P(u_j|u_i)$
9) $P = P \cup P(v_{next}|v_{current})$
10) $NT = NT(v_{current})_1 \cup NT(v_{current})_2$
11) for v_{next} in NT do
12) $P(v_{next}|v_{current}) = P(t_k|u_i)$
13) $P = P \cup P(v_{next}|v_{current})$
14) else
15) $NU = NU(v_{current})_1$
16) for v_{next} in NU do
17) $P(v_{next}|v_{current}) = P(u_i|t_k)$
18) $P = P \cup P(v_{next}|v_{current})$
19) $v_{next} = randoms.choices((P.keys, P.values), k = 1)$
20) $Path(v_i)_L = Path(v_i)_L + v_{next}$
21) return $Path(v_i)_L$

Algorithm 2 is mainly divided into four steps. Step 1 initializes an empty random walk path $Path(v_i)_L$, as shown in line 1). In step 2, v_i is added to $Path(v_i)_L$, as shown in line 2). Step 3 takes v_i as the starting node for L-steps walks, as shown in lines 3)-20). Step 4 returns $Path(v_i)_L$, as shown in line 21).

Among them, the step 3 can be divided into four small steps, as follows. (1) Obtain the current node and initialize the candidate set of the next node to be empty, as shown in line 4). (2) The type of the current node is determined, as shown in line 5). If it is a user node, obtain its 1-4 order user neighbor set, and calculate the transition probability according to Equation (3), as shown in lines 6)-9), continue to obtain its 1-2 order topics neighbor set, and calculate the transition probability according to Equation (7), as shown in lines 10)-13). Otherwise, if it is a subject neighbor set, obtain its 1-order user neighbor set, and calculate the transition probability according to Equation (8), as shown in lines 15)-18). (3) Based on the sampling strategy, the next node v_{next} is selected from the candidate neighbor set, as shown in lines 19). (4) Add v_{next} to the random walk sequence $Path(v_i)_L$, as shown in line 20).

5. Experiments.

5.1. Experimental setup.

5.1.1. *Datasets.* We use two real datasets in experiments. 1) Zachary's Karate club network, referred to as the Karate network, which is shown in Figure 3(a). Karate network consists of 34 nodes and 78 edges, and it has two real community structures of C_1 and C_2. 2) User movie reviews from Douban online in 2012, referred to as Douban Network. Douban network consists of 2289 nodes (2253 users and 36 movies) and 70489 edges (34580 users-to-users and 35909 users-to-topics).

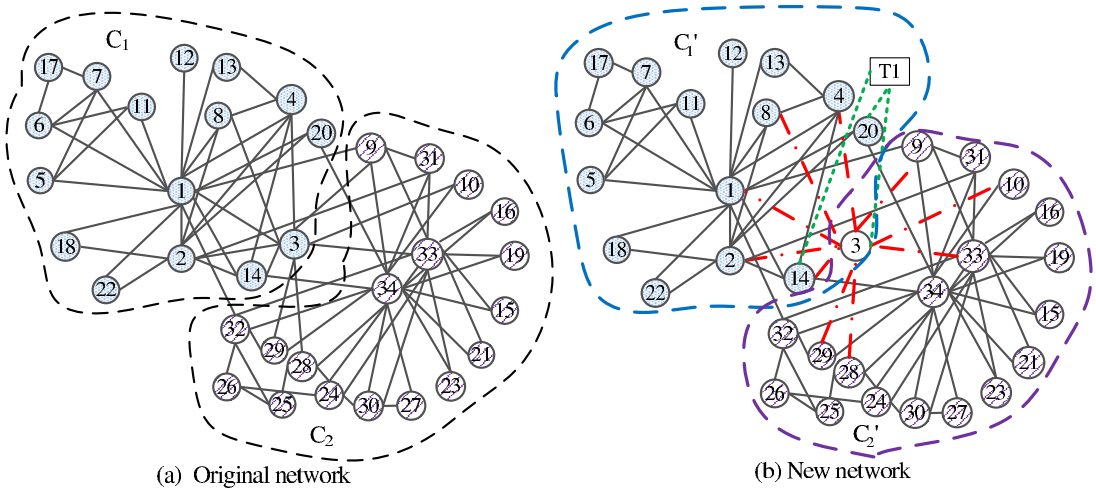


FIGURE 3. Karate network

5.1.2. *Comparison algorithms.* In experiments, 8 classic algorithms are compared with our work, which are COPAR [22], SLPA [23], DeepWalk [3], Node2Vec [4], Metapath2Vec [5], JUST [6] and TANE [2]. Among them, the first two are overlapping community algorithms based on label propagation, which can obtain overlapping communities directly. The rest are network representation learning algorithms, which learn vector representations of nodes, and then use Fuzzy C-Means (short for FCM) to mine overlapping communities.

To make the experimental results more convincing, the parameters in the comparison algorithms are consistent with those in the original papers. In Node2Vec, the bias parameters are $p = 1$ and $q = 1$. In Metapath2Vec, the range of the meta-path length is 1-4. In JUST, the stay probability is set to $\alpha \in [0.2, 0.5]$, and the number of node types recorded recently visited is set to $m = 1$. In TANE, the parameters are $\alpha = 0.35$, $\beta = 0.25$, $\gamma_1 = \gamma_2 = 0.15$, $\delta = 0.1$, $\chi_1 = 1/3$, $\chi_2 = 2/3$, $\varepsilon = 0.7$ and $\theta = 0.3$ on the Karate network; the parameters are $\alpha = 0.4$, $\beta = 0.3$, $\gamma_1 = \gamma_2 = \delta = 0.1$, $\chi_1 = 1/3$, $\chi_2 = 2/3$, $\varepsilon = 0.7$ and $\theta = 0.3$ on the Douban network.

In the above network representation learning algorithms, each node as the start node times $r = 10$, walk length $L = 40$. The dimensions d of DeepWalk, Node2Vec, Metapath2Vec, JUST, TANE and TANE-RWCN are 64, 128, 100, 128, 64, and 64 respectively. In TANE-RWCN, the rest parameters are shown in Table 2.

TABLE 2. Parameters setting of TANE-RWCN

	w_I	w_D	δ_I	δ_D	δ_{I1}	δ_{I2}	δ_{D1}	δ_{D2}	δ_{D3}	θ_I	θ_D
Karate network	0.7	0.3	0.7	0.3	0.7	0.3	0.4	0.4	0.2	0.7	0.3
Douban network	0.5	0.5	0.5	0.5	0.5	0.5	0.4	0.4	0.2	0.7	0.3

5.1.3. *Evaluating indicator.* When overlapping communities are divided [23], the extended modularity EQ [24] is the evaluation indicator, as shown in Formula (9). In Equation (9), O_i is the number of communities to which v_i belongs, and other parameters are similar to those in modularity Q.

$$EQ = \frac{1}{2m} \sum_{c} \sum_{i,j \in C_c} \frac{1}{O_i O_j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \tag{9}$$

EQ integrates the node overlap to Newman's Q. By dividing the contribution of node overlap in the community, the contribution rate of non-overlapping nodes is increased. When each node can belong to only one community, EQ will degenerate to Q. The higher the EQ value, the better the quality of community division.

5.2. **Experimental results and analysis.** We divided the experiments into three parts: 1) In the Karate network, the influence of topics on vector representation is verified. 2) In the Douban network, the accuracy and effectiveness of the TANE-RWCN algorithm are verified. 3) Sensitivity analysis of parameters.

5.2.1. *Experiments on Karate network.* We verify the influence of topics on vector representation and community division in the Karate network by adding topics from two perspectives: specified locations and random locations. The ways of adding topics are shown in Table 3. Among them, lines 1, 2 and 4 add one or two topics at the specified location; lines 3 and 5 add one or two topics randomly. The specific analyses are as follows.

First, at the boundary area of the two communities, topic T_1 is added to u_3, u_{14} and u_{20} in community C_1. Secondly, the overlapping community discovery based on representation learning is carried out under the new network. Finally, two new communities C_1' and C_2' are obtained. At this time, u_3 is an overlapping node belonging to the two communities, as shown in Figure 3(b).

In Figure 3(b), u_3 has 10 1-order user neighbors. Among them, 5 neighbors belong to C_1', which is $\{u_1, u_2, u_4, u_8, u_{14}\}$, and 5 neighbors belong to C_2', which is $\{u_9, u_{10}, u_{28}, u_{29}, u_{33}\}$. We can see that u_3 is close to both communities. At the same time, the TANE-RWCN mainly uses the transition probability model for random walks to learn the vector representation of nodes. Therefore, the key to which community the node belongs is the transition probability value.

This section analyzes the transition probability associated with u_3 from two aspects: whether this node is the starting node or not, as shown in Table 4. Because there are many node pairs in the 1-4 order range, only the top 10 are shown in the table. Among them, P_Y and P_N represent the transition probability values of whether considering the topic or not respectively. For user nodes, non-uniform sampling is used for random walks. That is, the larger the transition probability value is, the greater the possibility of being v_{next} is. We can see from Table 4, when u_3 is the starting node, the probability of it walking to u_9, u_{10}, u_{28}, u_{29} and u_{33} is higher. When the u_3 is the next node, the probability of walking

Table 3. Adding topics to Karate network

No.	Topics	C_1	C_2
1	1	$T_1 - u_3, u_{14}, u_{20}$	None
2	1	None	$T_1 - u_9, u_{28}, u_{29}, u_{32}$
3	1	$T_1 - u_3, u_9, u_{10}, u_{14}, u_{20}, u_{32}$	
4	2	$T_1 - u_3, u_{14}, u_{20}$	$T_2 - u_9, u_{28}, u_{29}, u_{32}$
5	2	$T_1 - u_1, u_3, u_9, u_{15}, u_{19}, u_{25}$, $T_2 - u_3, u_7, u_{11}, u_{23}, u_{25}, u_{32}$	

TABLE 4. Probability value of the top 10 nodes

u_3 as the current node				u_3 as the next node									
u_3				u_9		u_{10}		u_{28}		u_{29}		u_{33}	
u_j	P_N	u_j	P_Y	u_j	P	u_j	P	u_j	P	u_j	P	u_j	P
35	0.087	14	0.089	31	0.158	**3**	**0.353**	24	0.183	32	0.241	34	0.093
4	0.083	35	0.086	33	0.154	34	0.350	34	0.178	34	0.237	9	0.072
2	0.080	4	0.083	**3**	**0.151**	35	0.075	**3**	**0.177**	**3**	**0.236**	30	0.067
8	0.080	2	0.081	34	0.147	29	0.042	25	0.175	35	0.060	31	0.067
14	0.079	8	0.080	1	0.143	28	0.032	35	0.050	10	0.042	24	0.067
1	0.079	1	0.079	35	0.043	9	0.025	10	0.032	28	0.025	15	0.063
9	**0.075**	**9**	**0.075**	14	0.025	14	0.024	29	0.025	9	0.021	16	0.063
33	**0.070**	**33**	**0.070**	10	0.025	15	0.021	26	0.025	14	0.020	19	0.063
10	**0.068**	**28**	**0.069**	15	0.025	16	0.021	30	0.021	33	0.016	21	0.063
28	**0.068**	**29**	**0.069**	16	0.025	19	0.021	9	0.018	15	0.016	**3**	**0.062**

TABLE 5. EQ value of algorithms in Karate network

	1	2	3	4	5
COPRA	0.1286	0.1273	0.1247	0.1235	0.1093
SLPA	0.2904	0.2836	0.2810	0.3339	0.2939
DeepWalk	0.3573	0.3235	0.3196	0.3383	0.3031
Node2vec	0.3603	0.3021	0.1819	0.2961	0.2756
Metapath2vec	0.3627	0.3497	0.3090	0.3391	0.3044
JUST	0.3566	0.3603	**0.3304**	0.3403	0.2960
TANE	0.3608	0.3626	0.3218	0.3189	0.3045
TANE-RWCN	**0.3688**	***0.3722***	0.3269	**0.3410**	**0.3062**

from nodes u_9, u_{10}, u_{28}, u_{29} and u_{33} is also higher. In conclusion, u_3 as an overlapping node is reasonable.

Meanwhile, it can be seen from Table 4 that adding topic nodes can increase the transition probability between nodes. The tightness between nodes increases. Since fewer topics are added in Figure 3(b), the increase is not obvious. Other situations in Table 3 can be analyzed in the same way, which is omitted here.

For the 8 algorithms, overlapping communities are divided into 5 cases, and the EQ values are shown in Table 5. Compared with the other 7 algorithms, the TANE-RWCN algorithm has larger EQ value. Compared with the 2 traditional algorithms, TANE-RWCN increases by 24.49% at most. Compared with COPRA, the increase range is 19.69%-24.49%. Compared with SLPA, the increase range is 0.71%-8.86%. Compared with the 5 representation learning algorithms, TANE-RWCN increases by 14.50% at most, and the increase range is 1.0%-7.0% in most cases. Therefore, TANE-RWCN can capture the information in the network more comprehensively and effectively.

5.2.2. *Experiments on Douban network.* In the Douban network, the elbow rule is used to determine the clustering number K. The clustering number is taken as $K \in [5, 13]$. In experiments of over-lapping community mining, we randomly determine K initial clustering centers, experiments are repeated 10 times, and the average value of EQ is used as the experimental result, which is shown in Table 6. Among them, the number of overlapping communities of SLPA is not set artificially, so the number of clusters needs not be considered here.

TABLE 6. EQ value of algorithms in Douban network

	5	6	7	8	9	10	11	12	13
COPRA	0.0813	0.0967	0.0996	0.1003	0.1079	0.1195	0.1186	0.1139	0.1152
SLPA					0.1099				
DeepWalk	0.0785	0.0880	0.0813	0.0917	0.1041	0.1124	0.1222	0.1254	0.1164
Node2vec	0.0803	0.0797	0.1011	0.1158	0.1258	0.1165	0.1225	0.1135	0.1114
Metapath2vec	0.1034	0.1006	0.1093	0.1168	0.1156	0.1220	0.1231	0.1298	0.1256
JUST	0.1356	0.1384	0.1393	**0.1447**	0.1474	0.1439	0.1499	0.1458	**0.1531**
TANE	0.1266	0.1374	0.1388	0.1173	0.1277	0.1418	0.1436	0.1377	0.1344
TANE-RWCN	**0.1363**	**0.1393**	**0.1401**	**0.1447**	**0.1546**	**0.1509**	***0.1591***	**0.1565**	0.1478

We can see from Table 6 that compared with the classic algorithms, TANE-RWCN obtains larger EQ value. When the number of communities is 11, EQ is the highest, which is 0.1591. The EQ of TANE-RWCN increased by 5.96% at most, and the increase range is 0.5%-5.0% in most cases.

In summary, due to the integration of high-order topological structure, uncertainty relationship between nodes and user node preference for the topics, the result of overlapping communities divided based on TANE-RWCN algorithm is more reasonable.

5.2.3. *Parameter sensitivity analysis.* This paper mainly analyzes the sensitivity of parameters in the transition probability model and the skip-gram model.

1) Parameter analysis in the transition probability model.

In the transition probability model, we use 11 parameters. The transition probability of users to users involves 9 parameters, which are w_I, w_D, δ_I, δ_D, δ_{I1}, δ_{I2}, δ_{D1}, δ_{D2} and δ_{D3}. Among them, w_I, $w_D\delta_I\delta_{I1}$, $w_D\delta_I\delta_{I2}$, $w_D\delta_D\delta_{D1}$, $w_D\delta_D\delta_{D2}$ and $w_D\delta_D\delta_{D3}$ determine the weight of the identical, partial identical topic, partial identical neighbor, partial discrepancy common $1 \cap 2$-order topic, partial discrepancy common $2 \cap 1$-order topic and partial discrepancy common 2-order topic attributes respectively. The transition probability of users to topics involves 2 parameters, which are θ_I and θ_D.

In the Karate network (the first case in Table 3) and Douban network, 7 cases of parameter setting are used, and EQ is shown in Table 7.

From Table 7, 1) $EQ1$ has no value. Because only the partial discrepancy topics are considered, the transition probability is 0 in most cases. It is not reasonable to only consider partial discrepancy topics. 2) $EQ2 < EQ3 < EQ4 < EQ5$. With the increase of the identical and partial identical weights, and the decrease of the partial discrepancy weights, the value of EQ increases continuously. 3) $EQ5 > EQ6 > EQ7$. When the weight of discrepancy is 0, only the identical is considered, and only the 1-order relationship is considered in column "7"; the 1 and 2-order relationship is considered in column "6". It can be seen that considering only the identical and partial identical effects are the second.

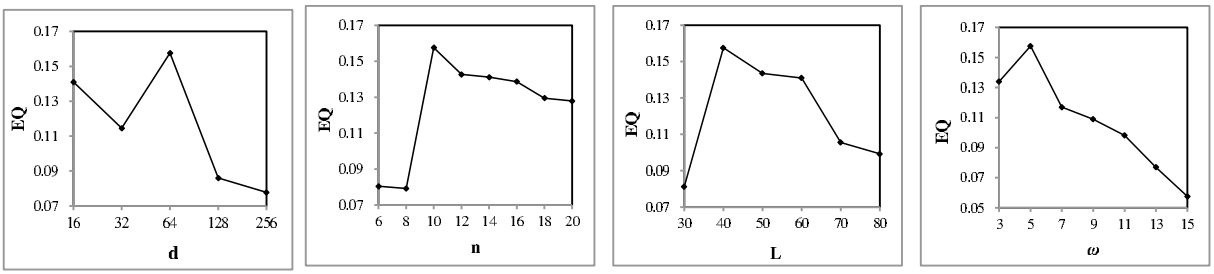
In summary, considering the identical, partial identical and partial discrepancy attributes, the lower-order (1-order and 2-order) relationship weights are larger. The higher-order (3-order and 4-order) relationship weights are smaller, and the transition probability is more reasonable and has the largest EQ value. In other words, modeling based on higher-order topology and uncertainty is more conducive to vector representation. In the experiments, the parameter values of column "5" in the two tables are selected as the weight of the transition probability.

TABLE 7. Parameter sensitivity analysis of the transition probability model

	Karate network							Douban network						
No.	1	2	3	4	5	6	7	1	2	3	4	5	6	7
w_I	0	0	0.3	0.5	**0.7**	0.8	1	0	0	0.2	0.3	**0.5**	0.8	1
w_D	1	1	0.7	0.5	**0.3**	0.2	0	1	1	0.8	0.7	**0.5**	0.2	0
δ_I	0	0.1	0.3	0.5	**0.7**	1	0	0	0.1	0	0.3	**0.5**	1	0
δ_D	1	0.9	0.7	0.5	**0.3**	0	0	1	0.9	1	0.7	**0.5**	0	0
δ_{I1}	0.7	0.7	0.7	0.5	**0.7**	0.7	0.7	0.7	0.7	0.7	0.7	**0.5**	0.7	0.7
δ_{I2}	0.3	0.3	0.3	0.5	**0.3**	0.3	0.3	0.3	0.3	0.3	0.3	**0.5**	0.3	0.3
δ_{D1}	0.4	0.4	0.4	0.4	**0.4**	0.4	0.4	0.4	0.4	0.4	0.4	**0.4**	0.4	0.4
δ_{D2}	0.4	0.4	0.4	0.4	**0.4**	0.4	0.4	0.4	0.4	0.4	0.4	**0.4**	0.4	0.4
δ_{D3}	0.2	0.2	0.2	0.2	**0.2**	0.2	0.2	0.2	0.2	0.2	0.2	**0.2**	0.2	0.2
θ_I	0.7	0.7	0.7	0.7	**0.7**	0.7	0.7	0.7	0.7	0.7	0.7	**0.7**	0.7	0.7
θ_D	0.3	0.3	0.3	0.3	**0.3**	0.3	0.3	0.3	0.3	0.3	0.3	**0.3**	0.3	0.3
EQ	–	0.334	0.347	0.363	**0.377**	0.365	0.355	–	0.079	0.118	0.136	**0.159**	0.108	0.089

2) Parameter analysis in the skip-gram model

In the skip-gram model, 4 parameters are involved, namely: the vector dimension d, the number of walks of starting node n, the random walk length L and the context window ω. In the Douban network, the experimental results are shown in Figure 4.



(a) d (b) n (c) L (d) ω

FIGURE 4. Parameter sensitivity analysis in skip-gram model

We can see from Figure 4 that d and n have less influence, while L and ω have great influence. Especially when $L > 40$ or $\omega > 5$, the EQ value is a downwards trend, and the algorithm performance is degraded. When $d = 64$, $n = 10$, $L = 40$ and $\omega = 5$, TANE-RWCN has the best overall effect, so the above parameter values are used in experiments.

In summary, in the transition probability model, the parameter values of each part are set based on the sparseness of networks. In the skip-gram model, the parameters are set based on the complexity of the network and algorithm performance.

6. **Conclusions.** We describe TAN as an IDC system in this paper. The transition probability model is constructed based on the connection number and high-order topological structure. Then, a high-quality random walk sequence is obtained. Finally, the vector representations of the nodes are obtained by training the skip-gram model. The experimental results show that the topic increases the closeness between nodes and reduces the distance between nodes. Compared with other algorithms, TANE-RWCN obtains larger EQ value in both networks. In the Karate network, the EQ value of TANE-RWCN increases by up to 24.49%, and in the Douban network, it is up to 5.96%. Comprehensive consideration of

the high-order topology and uncertainty relationship can better retain the structural information, and improve the accuracy of vector representations. Next, the related research on network representation learning in large-scale and dynamic heterogeneous networks is the focus of the next work.

Acknowledgment. This work is supported by the S & T Program of Hebei (No. 203103-01D), the National Natural Science Foundation of China (No. 62172352, No. 61472340 and No. 61871465), the Fundamental Research Funds for the Hebei Province Universities (No. 2020JK005), the Doctoral Research Initiation Foundation of Hebei Normal University of Science and Technology of China (No. 2019YB011), the Hebei Provincial Natural Science Foundation of China (No. F2017209070 and No. F2019203157), and the S & T Program of Educational Commission of Hebei (No. ZD2019004). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] J. Su, F. Zhang and X. Yu, A novel NPD team formation method based on social network analysis approach, *International Journal of Innovative Computing, Information and Control*, vol.17, no.4, pp.1275-1285, 2021.

[2] J. Guo, H. Dong, T. Zhang and X. Chen, Network embedding of topic-attention network based on set pair analysis, *International Journal of Innovative Computing, Information and Control*, vol.16, no.4, pp.1371-1384, 2020.

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

[4] T. Mikolov, K. Chen, G. Corrado et al., Efficient estimation of word representations in vector space, *Computer Science*, arXiv: 1301.3781v3, pp.1-12, 2013.

[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*, Halifax, Canada, pp.135-144, 2017.

[6] R. Hussein, D. Yang and P. Cudré-Mauroux, Are meta-paths necessary?: Revisiting heterogeneous graph embeddings, *Proc. of the 27th ACM International Conference on Information and Knowledge Management*, Torino, Italy, pp.437-446, 2018.

[7] Y. L. Jin, G. J. Song and C. Shi, GraphLSP: Graph neural network with local structural patterns, *Proc. of the AAAI Conference on Artificial Intelligence*, vol.34, no.4, pp.4361-4368, 2020.

[8] S. H. Fan, X. Wang, C. Shi et al., One2Multi graph autoencoder for multi-view graph clustering, *Proc. of the WWW'20: The Web Conference 2020*, Taipei, Taiwan, pp.1-7, 2020.

[9] X. Chen, J. F. Guo and C. Z. Fan, Community discovering based on connection degree in topic-attention network, *Journal of Computer Engineering and Applications*, vol.53, no.17, pp.85-93, 2017.

[10] C. Shi and Y. Z. Sun, Research progress of heterogeneous network representation learning, *Communication of the CCF*, vol.14, no.3, pp.35-40, 2018.

[11] T. Fu, W. C. Lee and Z. Lei, HIN2Vec: Explore meta-paths in heterogeneous information networks for representation learning, *Proc. of the 2017 Conference on Information and Knowledge Management*, Pan Pacific Singapore, pp.1797-1806, 2017.

[12] J. B. Shang, M. Qu, J. L. Liu et al., Meta-path guided embedding for similarity search in large-scale heterogeneous information networks, *arXiv e-prints*, arXiv: 1610.09769, pp.2169-2174, 2016.

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

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

[15] Z. Huang and N. Mamoulis, Heterogeneous information network embedding for meta-path based proximity, *arXiv e-prints*, arXiv: 1701.05291, pp.1-10, 2017.

[16] M. H. Zhang, B. B. Hu, C. Shi et al., Matrix factorization meets social network embedding for rating prediction, *Proc. of the 2018 Conf. on APWeb-WAIM: Web and Big Data*, Macau, China, pp.121-129, 2018.
[17] C. Shi, B. Hu, W. X. Zhao et al., Heterogeneous information network embedding for recommendation, *IEEE Transactions on Knowledge & Data Engineering*, vol.31, no.2, pp.357-370, 2019.
[18] L. Xu, X. Wei, J. Cao et al., Embedding of embedding (EOE): Joint embedding for coupled heterogeneous networks, *Proc. of the 10th ACM International Conference on Web Search and Data Mining*, Phoenix, AZ, USA, pp.741-749, 2017.
[19] 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*, San Francisco, CA, pp.1225-1234, 2016.
[20] J. Zhang, C. Xia, C. Zhang et al., BL-MNE: Emerging heterogeneous social network embedding through broad learning with aligned autoencoder, *Proc. of the 2017 IEEE International Conference on Data Mining*, New Orleans, LA, USA, pp.605-614, 2017.
[21] H. W. Wang, F. Z. Zhang et al., SHINE: Signed heterogeneous information network embedding for sentiment link prediction, *Proc. of the 11th ACM International Conference on Web Search and Data Mining*, Los Angeles, CA, USA, pp.592-600, 2018.
[22] S. Gregory, Finding overlapping communities in networks by label propagation, *New Journal of Physics*, vol.12, no.10, 103018, 2000.
[23] J. R. Xie, B. K. Szymanski, X. M. Liu et al., SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process, *Proc. of the 11th IEEE International Conference on Data Mining Workshops*, Vancouver, BC, Canada, pp.344-349, 2011.
[24] Y. Lei, Y. Zhou and J. Shi, Overlapping communities detection of social network based on hybrid c-means clustering algorithm, *Sustainable Cities & Society*, vol.47, no.4, 101436, 2019.

Author Biography

Xiao Chen received the B.S. degree in information management and information systems from Hebei University of Science and Technology, Shijiazhuang, Hebei, China, 2006; the M.S. degree in computer application technology from Yanshan University, Qinhuangdao, China, 2010; and the Ph.D. degree in computer science and technology from Yanshan University, Qinhuangdao, China, 2017.

Dr. Chen is assistant research fellow at Hebei Normal University of Science and Technology, China. Her research interests include graph mining, social network analysis, big data, deep learning, etc. She has published over 30 papers in journals and conferences.

Ying Wang received the B.S. degree in marketing from China University of Petroleum, Beijing, China, 2005; the M.S. degree in management science and engineering from Yanshan University, Qinhuangdao, China, 2014.

Ms. Wang is lecturer at Hebei Normal University of Science and Technology, China. Her research interests include industrial cluster, innovation network, regional economics, etc.

Hui Dong received the B.S. degree in information management and information systems from Hebei University of Science and Technology, Shijiazhuang, Hebei, China, 2006; and the M.S. degree in computer science and technology from Yanshan University, Qinhuangdao, China, 2020.

Ms. Dong is currently pursuing the Ph.D. degree in computer science and technology from Shijiazhuang Tiedao University, China. Her research interests include social network analysis, network representation learning, deep learning, etc.

Xiao Pan received the B.S. degree in computer application technology from Yanshan University, Qinhuangdao, China, 2002; the M.S. degree in computer application technology from Yanshan University, Qinhuangdao, China, 2005. She was a visiting scholar in the Department of Computer Science, University of Illinois at Chicago, USA. She received her Ph.D. in computer science from Renmin University of China in 2010.

Dr. Pan is an associate professor at Shijiazhuang Tiedao University, China. Her research interests include data management on moving objects, location based social networks and privacy-aware computing.

Jia Li received the B.E. degree in computer science and technology from Hebei University of Science and Technology, Shijiazhuang, Hebei, China, 2006.

Ms. Li is medium electronic engineer at Baoding University of Technology, China. Her research interests include network security, computer application technology, teaching management, etc.