IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 January 2024, accepted 1 March 2024, date of publication 7 March 2024, date of current version 25 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3374882

RESEARCH ARTICLE

Overlapping Community Detection Based on Weak Equiconcept

SUNQIAN SHI, MENGYU YAN, AND JINHAI LI
Faculty of Science, Kunming University of Science and Technology, Kunming 650500, China
Data Science Research Center, Kunming University of Science and Technology, Kunming 650600, China
Corresponding author: Mengyu Yan (yanmengyu2016@163.com)

ABSTRACT Community discovery refers to the process of searching for clusters in a network that are sparsely connected to other nodes and formed dense connections internally. In many real networks, some communities often overlap with each other, meaning that a node may belong to multiple communities simultaneously. Revealing these (overlapping) community structures is an important issue in complex network analysis, as it helps to better analyze the characteristics and organizational structure of the network. Community expansion methods are very important in the study of community detection problems. However, one of the key issues in developing effective community expansion methods is that the position of seed nodes greatly affects the performance of these algorithms, resulting in low robustness of these algorithms. Meanwhile, it is also difficult for them to provide high-quality results for community detection task. To solve the above problem, this paper proposes a seed selection method based on the weak equiconcepts in a network formal context, which integrates the attribute information of nodes during random walk to detect overlapping communities. Specifically, the weak equiconcepts are constructed by establishing a network formal context to obtain seed sets, and an improved PageRank clustering algorithm is used to expand these seed sets to better reveal the overlapping community structure in the network. Experiments show that seed selection is helpful to improving the performance of overlapping community detection algorithms.

INDEX TERMS Overlapping community discovery, network formal context, weak equiconcept, random walk.

I. INTRODUCTION

The purpose of community discovery is to explore the potential community structure in complex networks. The research results in this field have important theoretical significance and practical application value for understanding the topology of networks and the analysis of inter-community behavior patterns. The results of the community's findings have been widely used in various fields and tasks. For example, community discovery based on online social behavior was able to effectively determine the relationship between users and was used for the task of spammer detection [1], community discovery based on human brain network could help identify functional parts of the brain that play a role or have pathologies [2], image interpretation based on community discovery was able to generate better image semantic descriptions by introducing communities [3], and community detection could predict the absence of links in link prediction [4]. Most of the communities in the real network are not independent of each other, but overlap with each other, and such communities were called overlapping communities [5], [6], which means that the interaction between communities is more complex. In order to solve the problem of detecting overlapping communities, scholars have proposed various methods and algorithms, such as the classic clique percolation method [7], the community expansion algorithm GCE by [8], link density-based methods like the fast greedy modularity-based hierarchical community detection [9], probabilistic graphical models [10], spectral clustering methods [11], a label propagation community detection algorithm [12], and fuzzy methods for detection [13].

As far as we know, community expansion methods are important in the detection of overlapping communities. There are three main problems with the existing approaches:

The associate editor coordinating the review of this manuscript and approving it for publication was N. Ramesh Babu.

1) The seed node is manually specified, which affects the stability of the algorithm.
2) The selected seeds are not representative and cannot fully cover the core information of the community.
3) The fringe communities of the network are not fully detected.

In order to solve the above three problems, this paper proposes a new seed selection method based on Formal Concept Analysis (FCA) [14]. It is an effective computational intelligence method used for characterizing relationships between nodes in a graph. FCA is a powerful knowledge discovery theory that provides techniques for efficiently discovering fine-grained knowledge (called formal concepts) from binary relation and organizing them into lattice-based structures (called concept lattices). The formal context is an important input, which is composed of the triples of objects, attributes, and their relationships. In addition, two key operations were defined for extracting the common attributes/public objects for a given set of objects/attributes. In recent years, many scholars have successfully integrated FCA with specific tasks such as data mining [15], machine learning [16], [17], [18], knowledge discovery [19], [20], cloud computing [21], complex network [22], [23], [24], [25], and so on.

In the combination of FCA and complex network, adjacency matrices are used to describe network topology, while a formal context is used to describe the relationship between objects and attributes, and the formal context can be regarded as a modified adjacency matrix from the data point of view. Therefore, FCA and complex network are both based on the (modified) adjacency matrix to make data analysis and knowledge discovery. Furthermore, studying the generation and propagation of networks and concepts under a unified framework allows for the full utilization of their complementary advantages. Besides, the interpretability of complex network structures can be improved through FCA. Some scholars have paid much attention to their complementary researches. For instance, Hao et al. [23], [24] combined graph networks with FCA, conducting a thorough analysis of social networks and proposing a k-clique detection method applied to community discovery. Gao et al. [22] established an equivalence relationship between FCA and critical structures in graph networks, introducing a method for detecting key structures. Moreover, Yang et al. [25] extended static networks to dynamic ones, reinforcing the integration of FCA with networks.

This paper also focuses on the study of integrating FCA and complex network from the perspective of weak equiconcept to propose a novel seed selection method, which can not only represent the most critical structure of the network, but also automatically select the most representative seed nodes. This overcomes the problem of manually specifying seed nodes in the traditional methods, so as to improve the automation and universality of the division. We integrate the attribute information and topological structure of a network to expand communities by means of selecting seeds and detect the fringe communities of the network. Finally, the initial extended community is optimized, the information entropy of a network formal context is used to determine the parameters, and the final community division results are obtained by adjusting some communities and nodes. The main contributions of this paper are summarized as follows:

1) A more representative seed selection strategy is proposed by using the weak equiconcepts, which are derived from the network formal context.
2) The PageRank node clustering algorithm is improved by integrating the node attribute information and topological structure of the network, so as to ensure that the local community obtained by seed expansion has a large overall node influence, and the nodes in the local community have high attribute homogeneity and structural similarity.
3) Furthermore, the preliminary community division is further optimized by information entropy of a network formal context, and the algorithm's performance is validated on real networks. Experimental results show that the method proposed in this paper can identify more accurate community structures compared to the existing algorithms.

The outline of this paper is organized as follows: Section II provides the preliminary knowledge. We review some related work in local community detection in Section III. In Section IV, we put forward an overlapping community detection algorithm, named WEOCD (Weak Equiconcept for Overlapping Community Detection), a stability coefficient to guide the selection of seeds based on the weak equiconcept, and a overlapping degree between communities to estimate the quality of detected local communities for community optimization strategies. Section V conducts some comparative experiments to show the effectiveness of the proposed WEOCD algorithm. Finally, Section VI concludes this paper and presents future work.

II. PRELIMINARY

In FCA, the formal context matrix is used to describe the relationship between objects and attributes, while complex networks utilize adjacency matrices to depict topological structures of networks, which illustrate the relationships between nodes and edges in a network, with a value of "1" indicating the connection of them and "0" otherwise. Similarly, in a formal context, if edges in the network are considered as attributes, whether an object possesses an attribute can also be represented by a matrix, with values of "1" or "0". By combining the advantages of them, a network formal context is constructed by using a modified adjacency matrix.

Definition 1 [26]: A triplet (U, A, I) is termed a formal context, consisting of a non-empty finite set of objects $U = \{x_1, x_2, \cdots, x_n\}$, a non-empty finite set of attributes $A = \{a_1, a_2, \cdots, a_m\}$, and a binary relation I on the Cartesian

product $U \times A$. We denote

$$xI = \{a \in A \mid (x, a) \in I\}$$
$$Ia = \{x \in U \mid (x, a) \in I\} \quad (1)$$

where xI represents a collection of all attributes owned by an object x, Ia represents a collection of all objects with attribute a, and $(x, a) \in I$ indicates that object x has attribute a.

Definition 2 [26]: For a formal context (U, A, I), two operators are defined for $X \subseteq U, B \subseteq A$:

$$f(X) = \{a \in A \mid \forall x \in X, (x, a) \in I\} = \bigcap_{x \in X} xI$$
$$g(B) = \{x \in U \mid \forall a \in B, (x, a) \in I\} = \bigcap_{a \in B} Ia \quad (2)$$

where $f(X)$ represents the set of all attributes commonly possessed by all objects in X, and $g(B)$ denotes the set of all objects that possess all the attributes in B.

Definition 3 [26]: Given a formal context (U, A, I), if $f(X) = B$ and $g(B) = X$, then the ordered pair (X, B) is termed a formal concept, or simply a concept. Moreover, X is referred to as the extent of the concept (X, B), and B as the intent of the concept (X, B). Organizing all concepts according to the partial order relation $\prec$, we form a concept lattice $L(U, A, I)$.

Definition 4 [27]: A quadruple (U, M, A, I) is termed a network formal context, where $M = \{M_1, M_2, \cdots, M_k\}$, M_i is the network's i-th order adjacency matrix, $A = \{a_1, a_2, \cdots, a_m\}$ is a non-empty finite set of attributes, and $I = \{I_1, I_2, \cdots, I_k, I_{k+1}\}$, with $I_1, I_2, \cdots, I_k$ being binary relations on the Cartesian product $U \times U$, and I_{k+1} being a binary relation on the Cartesian product $U \times A$.

Definition 5 [22]: A network $G = \{V, E\}$ consists of n nodes, denoted as $x_1, x_2, \cdots, x_n$. Let $K' = (k_{ij})_{n \times n}$ be the modified adjacency matrix, which is defined as:

$$k_{ij} = \begin{cases} 1, & \text{if nodes } v_i \text{ and } v_j \text{ have a connection and } i \neq j, \\ 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, $MFC(G) = (U, U, I)$ is equivalent to the modified adjacency matrix of the graph $G = \{V, E\}$, denoted as $MFC(G) = K'$. Based on the properties of K', $MFC(G)$ also has the following characteristics:

1) $MFC(G)$ is symmetric;
2) all diagonal elements are equal to 1.

In a network $G = \{V, E\}$, the network formal context of G is denoted as (U, M, A, I), and a simplified version leads to a modified adjacency matrix $MFC(G) = (U, U, I)$.

Definition 6 [22]: For a network formal context (U, U, I), if a pair (X, B) satisfies $f(X) = B$, $g(B) = X$ and $X = B$, then (X, B) is referred to as an equipconcept. In this case, X is known as the extent, and B as the intent. If the number of objects in the equiconcept is $k = |f(X)|$, we call this equiconcept as k-equiconcept.

Definition 7 [28]: In a network $G = \{V, E\}$, the conductance of cluster $C(C \subset V)$ is defined as:

$$cond(C) = \frac{links(C, \overline{C})}{\min(links(C, V), links(\overline{C}, V))}$$

where $\overline{C} = V - C$ is the complement of C in V, and $links(C, \overline{C})$ is the number of edges between the clusters C and $\overline{C}$.

III. RELATED WORKS

Community expansion-based methods start from specific nodes and incrementally expand, relying on a local community metric. The community is optimized by continuously assessing changes in this metric. The selection of seeds and the methodology for measuring local community metrics are crucial for such algorithms. This section gives an overview of current researches on community expansion methods and discusses challenges in the field.

A. SEED SELECTION

The initial step in community expansion is seed selection, and it has been extensively studied by numerous scholars. Andersen and Lang [29] have demonstrated the feasibility of identifying a robust seed set within communities. Lancichinetti et al. [30] adopted a strategy of randomly selecting a node as the seed, introducing an element of uncertainty to the algorithmic outcome. Baumes et al. [31] proposed a global approach through link clustering, considering the degree of nodes from a global perspective but inadvertently neglecting peripheral communities, thereby affecting the accuracy of the results. Lee et al. [8] employed k-cliques as candidate seed sets. Whang et al. [32] introduced methods for appropriate seed set selection, such as the use of 'Graclus centers' and 'Spread hubs' algorithms, coupled with PageRank clustering for expansion. Gao et al. [33] suggested the employment of graph topological metrics for node ranking and neighborhood inflation for seed selection. Lastly, Gao et al. [22] utilized formal context methods for the selection of local key structures.

B. COMMUNITY EXPANSION

Starting from the detected seeds as initial communities, adjacent nodes are added to expand these into local communities. Common community expansion methods include quality function-based approaches [34], [35], [36], [37] and influence propagation methods [38], [39], [40], [41]. The community structure in networks is defined by quality functions, which assess the quality of community partitions [42]. In the Local Fitness Maximization (LFM) community detection method proposed by Lancichinetti et al. [30] and the Greedy Clique Expansion (GCE) method by Lee et al. [8], community expansion optimization was achieved by greedily maximizing a local fitness function. The central idea of influence propagation methods is to score each node using an influence assessment mechanism and propagate this scoring throughout

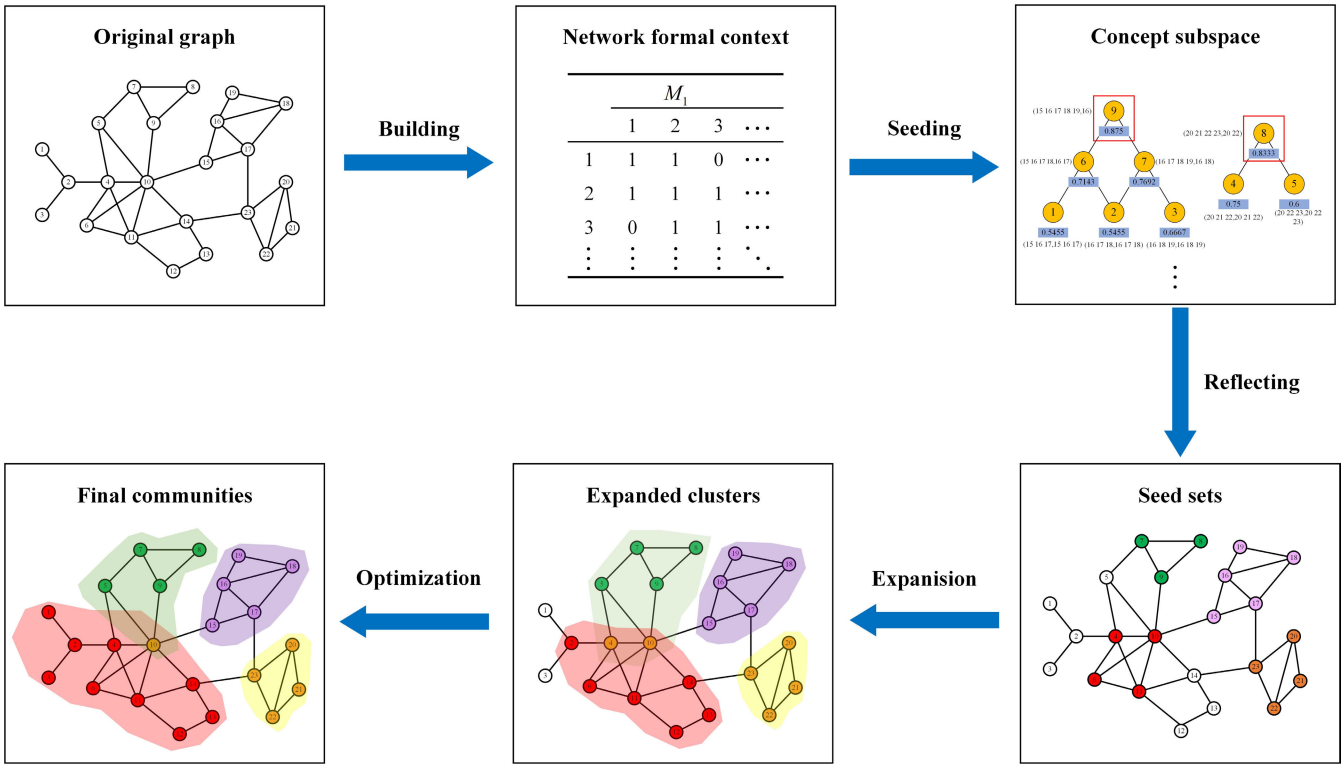


FIGURE 1. WEOCD method overview diagram.

the network. Raghavan et al. [43] introduced the Label Propagation Algorithm (LPA), which is based on an epidemic spread model. Building on LPA, Gregory [12] developed a method for detecting overlapping communities, termed the Community Overlap Propagation Algorithm (COPRA). Additionally, Andersen et al. [28] proposed a seed expansion method based on random walks.

In the stage of seed expansion, Andersen's forward search uses residual $\hat{r}(v_0, u)$, which represents the probability of being distributed at node u, and uses reserve $\hat{\pi}(s_0, u)$, where α represents the probability of permanently staying at node u. In each push operation, $d(u)$ is the degree of node u, it selects the one with the maximum $\hat{r}(v_0, u)/d(u)$, and transfers a part of α to $\hat{\pi}(s_0, u)$, as a reserve for u. Then, the algorithm transfers the other $1-\alpha$ parts to the neighbors of u. For each neighbor v_t of u, the residual is $(1-\alpha)\cdot(\hat{r}(v_0, u)/d(u))$. When the maximum residue/degree ratio drops below the error parameter ε, the process ends. Finally, the forward search uses the reserved $\hat{\pi}(v_0, u)$ as the estimate of $\pi(v_0, u)$. Next, using the sweep operation, each node v_0 around the seed obtains a PageRank score $\hat{\pi}(v_0, u)$, which measures the proximity between the node and the seed. Then, the nodes are sorted in a descending order by $\hat{\pi}(v_0, u)/d(v_0)$, and the set of the top p nodes with the best derivative in the sequence is the community where the seed is located.

While the community expansion method can yield high-quality local communities, several significant issues require substantial resolution. Firstly, some community expansion methods necessitate pre-execution parameter setting, posing challenges in obtaining the most suitable parameters and being time-consuming. Secondly, these methods primarily focus on expanding the seed into communities that closely resemble real-world communities. However, during the expansion process, certain nodes may end up on the periphery of a community or even get excluded, indicating an insufficient seed-centered expansion with a local community focus. Furthermore, in the forward search process previously mentioned, probabilities are uniformly distributed to neighboring nodes, overlooking the influence relationships between them. The propagation of information does not disperse uniformly to each neighbor, being influenced by the mutual attributes of these nodes.

IV. MOTIVATION AND THE PROPOSED ALGORITHM

A. MOTIVATION FOR WEOCD METHOD

As discussed in Section III, while overlapping community detection algorithms have achieved considerable success in community expansion methods, three limitations persist: presetting parameters, non-representative seed selection, and information loss during random walks. Addressing these aspects, this section proposes a novel overlapping community detection method within the FCA framework, termed the Weak Equiconcept for Overlapping Community Detection (WEOCD) method. Not only does the WEOCD method eliminate the requirement for pre-defined parameters, but also it effectively extracts more representative nodes as the seed set.

Algorithm 1 WE Menthod

Input: Graph $G = \{V, E\}$, $S \leftarrow \emptyset$, $H \leftarrow \emptyset$
Output: Seed set S

Construct a formal context $FC(G)$ according to modified adjacency matrix
Build a concept lattice $L(FC(G))$
for each concept $(X, B) \in L(FC(G))$ **do**
  if $X = B$ and $|X| = |B|$ **then**
    $H \leftarrow H \cup \{(X, B)\}$
  end if
end for
for each equiconcept $(X', B') \in H$ **do**
  if $X_1 \cap X_2 \neq \emptyset$ **then**
    $CS_1 \leftarrow \{(X_1, B_1), (X_2, B_2)\}$
  else
    $CS_1 \leftarrow (X_1, B_1)$, $CS_2 \leftarrow (X_2, B_2)$
  end if
end for
Calculate concept space $CS_1, CS_2, \cdots, CS_m$
for each concept space CS_i **do**
  constructing lattice structure using a partial order relationship $\prec$
end for
Calculate EC and EC_k in every $CS_1, CS_2, \cdots, CS_m$ according to Definition 10
for each concept space $CS_i \in \{CS_1, CS_2, \cdots, CS_m\}$ **do**
  for each weak equiconcept or equiconcept $(X_i, B_i) \in CS_i$ **do**
    calculate EC and EC_k
    select (X_i, B_i) of the max stability, $S \leftarrow X_i$
  end for
end for
return seed set S

Specifically, we first introduce a new seed selection approach called the WE method, which incorporates a novel metric to measure the tightness of objects within concepts. By selecting weak equiconcepts, WE can efficiently detect the locations of peripheral communities. Furthermore, we enhance the PageRank node clustering algorithm, incorporating the influence of node attributes during random walks to extend the seed set and form preliminary community partitions. Finally, we utilize information entropy within the network formal context to measure the overlapping between communities. The merging of highly overlapping communities reduces redundancy, enhancing the accuracy and interpretability of overlapping community detection.

Fig. 1 illustrates the process of the WEOCD method, which comprises four distinct stages: constructing the network formal context, seed selection, seed expansion, and community optimization. In the initial stage, the algorithm constructs an equivalent network formal context for a complex network with attributes. During the seeding phase, optimal seeds are identified by filtering through the weak equiconcept. The seed expansion stage is viewed as an enhancement in the clustering process of forward propagation, integrating attribute preferences between nodes to facilitate seed expansion. Finally, the algorithm discovers and refines overlapping communities by minimizing the conductance and information entropy within the network formal context.

B. SEEDS SELECTION STRATEGY

Definition 8: For the equiconcepts $H_1 = (X_1, B_1)$ and $H_2 = (X_2, B_2)$, if $X_1 \cap X_2 \neq \emptyset$, then H_1 and H_2 are referred to as neighboring concepts.

Definition 9: In the network formal context (U, M_1, B, I), for neighboring concepts H_1 and H_2, the following holds:

$$X_3 = X_1 \cup X_2 \tag{3}$$

$$B_3 = B_1 \cap B_2 \tag{4}$$

$H_3 = (X_3, B_3)$ is referred to as the weak equiconcept. Denote $\delta = \frac{|X_3|}{|f(B_3)|}$, where $0 < \delta < 1$, and (X_3, B_3) is a weak equiconcept at the degree of δ.

Similarly, (X_3, B_3) and (X_4, B_4) are weak equiconcepts. For concept (C, D), $C \subseteq X_3 \cup X_4$, $D \subseteq B_3 \cap B_4$, it follows that:

$$C^{\diamond} = \left\{ c_i \in U \mid 0 < \frac{|C|}{|f(D)|} < \delta \right\} \tag{5}$$

$$D^{\diamond} = \left\{ d_i \in B \mid 0 < \frac{|D|}{|g(C)|} < \delta \right\} \tag{6}$$

the concept (C, D) is also a weak equiconcept.

Within the same concept subspace, where k represents the number of objects in an equiconcept, a $(k+1)$-weak equiconcept is necessarily a parent concept of a k-equiconcept, and a k-equiconcept is a child concept of a $(k+1)$-weak equiconcept.

Definition 10: For an equiconcept $H_1 = (X_1, B_1)$, the stability coefficient of the equiconcept is defined as:

$$EC(H_1) = \frac{|X_1| \cdot (|X_1| - 1)}{\sum_{x_i \in X_1}^{|X_1|} \deg(x_i)} \tag{7}$$

which is to measure the internal cohesion of the cluster formed by the objects of the equiconcept and the extent of connections outside the cluster.

The stability coefficient of the weak equiconcept $H_3 = (X_3, B_3)$ is defined as:

$$EC_w(H_3) = \frac{2 \times links(H_3)}{\sum_{x_j \in X_3}^{|X_3|} \deg(x_j)} \tag{8}$$

where $links(H_3)$ is the number of edges among the objects within the concept H_3, and $\deg(x_i)$ is the degree of the object x_i.

Property 1: In the context of k-equiconcepts, a higher stability coefficient implies that the extent of the concept is more likely to belong to a singular community, leading to smaller communities with clearer boundaries. When these

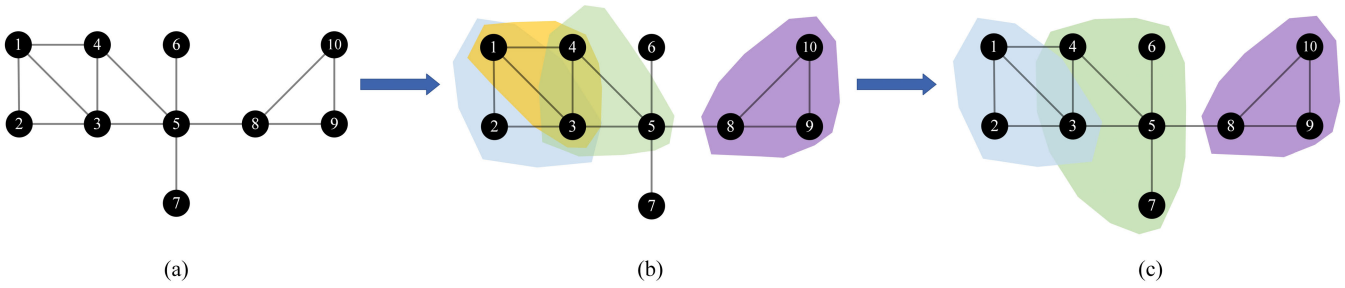


FIGURE 2. In the social network graph of 10 users, the equiconcept evolves into a weak equiconcept.

are used as seed sets, they impact the quality of the communities formed during expansion. Conversely, a lower stability coefficient results in more ambiguous community boundaries, and objects within these concepts are more likely to be part of multiple communities.

Proof: For a subgraph $G' = \{V', E'\} \subseteq G = \{V, E\}$, the subgraph modularity can be simplified as $Q_c = \frac{L_c}{L} - \left(\frac{d_c}{2L}\right)^2$, where L_c represents the edges within the community C, L is the number of edges in G, and d_c is the sum of the degrees of nodes within the community C.

Given two k-equiconcepts C_1 and C_2, serving as seed sets and also considered as two small communities. Thus, we have $L_{c_1} = L_{c_2}$. Assume the stability value is $C^k_{1stability} > C^k_{2stability}$. From the definition of the stability coefficient, it follows $d_{c_1} < d_{c_2}$. Modularity of subgraphs C_1 and C_2 are as follows:

$$Q_{c_1} = \frac{L_{c_1}}{L} - \left(\frac{d_{c_1}}{2L}\right)^2 \tag{9}$$

$$Q_{c_2} = \frac{L_{c_2}}{L} - \left(\frac{d_{c_2}}{2L}\right)^2 \tag{10}$$

Subtracting Equation (9) from Equation (10), we obtain:

$$Q_{c_1} - Q_{c_2} = \frac{1}{4L^2}[(d_{c_2})^2 - (d_{c_1})^2] > 0 \tag{11}$$

This implies $Q_{c_1} > Q_{c_2}$, meaning that an equiconcept with a higher stability coefficient has a higher modularity and is more characteristic of a community. □

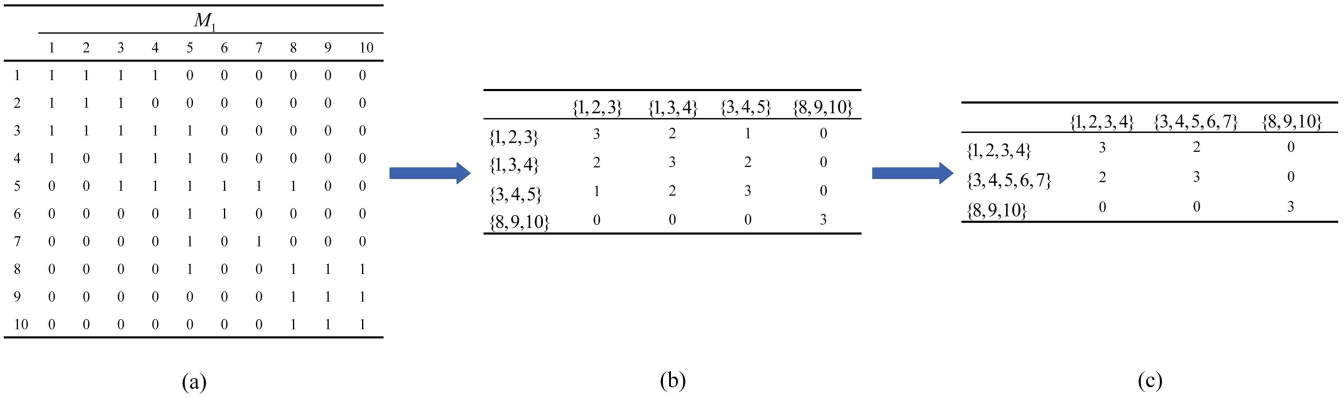
Property 2: The formation of a $(k+1)$-weak equiconcept, constructed from k-equiconcepts, is characterized by changes in the stability coefficient. An increase in the stability coefficient indicates that the newly added objects strengthen the connections among existing internal objects, thereby enhancing the clarity of community boundaries in the resulting communities. Conversely, a decrease in the stability coefficient suggests that the addition of new objects negatively impacts the stability of the internal community, leading to more ambiguous community structures.

Proof: The proof is similar to that of Property 1, so it is neglected here. □

As shown in Algorithm 1, equiconcepts are constructed by using the network formal context. Further, these concepts are stratified based on the number of objects they contain. Within each stratum, equiconcepts are sorted by their stability values from high to low, forming a sequence of concepts for each level. This process facilitates the selection of seed sets characterized by tight internal connections and sparse external links. Sequential generation of weak equiconcepts within each stratum effectively curbs the emergence of multiple highly overlapping seed sets, which can enhance the accuracy of subsequent community expansion. For seeds with exceptionally low stability coefficients, a threshold control is employed to mitigate assortative mixing, which often occurs when high-degree nodes connect to low-degree nodes. This algorithm, named WE (Weak Equiconcept) method, can select seeds directly, without prior knowledge of the actual number of communities. Moreover, as the selection of seed nodes involves no additional parameters, it yields stable community detection outcomes. This approach overcomes the shortcomings of existing community expansion-based methods, which often suffer from poor stability in overlapping community detection results.

Example 1: Fig. 2 illustrates a social network of 10 users and the development of initial and final communities, highlighting friendships among these users through connecting edges. Fig. 3 presents the process of the formal context matrix transformation corresponding to Fig. 2. Here, the symbol "1" in the matrix indicates that users 1 and 2 are first-order adjacent, in other words, they are directly connected, while "0" signifies that users 1 and 2 are not first-order adjacent, indicating no direct connection between them. Fig. 4 displays the corresponding Hasse diagram.

Fig. 3(a) shows its corresponding simplified network formal context, considering only 1-th association matrix M_1, while in Fig. 3(b), equiconcepts are generated by Definitions 2 and 5. We consider each equiconcept's objects as an initial seed set, objects 1, 3, and 4 are part of more than one seed set, with object 3 belonging to three seed sets. Based on the decomposition of the concept subspaces in Fig. 4, we observe changes in the stability coefficient. Equiconcepts 1 and 2 merge into weak equiconcept 7, with the stability value of concept 7 being higher than both 1 and 2, thus selecting concept 7. Similarly, concepts 1 and 3 form weak equiconcept 11, which is also a weak equiconcept of concepts 7 and 8. However, its stability coefficient is lower



M_1

	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	0	0	0	0	0	0
2	1	1	1	0	0	0	0	0	0	0
3	1	1	1	1	1	0	0	0	0	0
4	1	0	1	1	1	0	0	0	0	0
5	0	0	1	1	1	1	1	1	0	0
6	0	0	0	0	1	1	0	0	0	0
7	0	0	0	0	1	0	1	0	0	0
8	0	0	0	0	1	0	0	1	1	1
9	0	0	0	0	0	0	0	1	1	1
10	0	0	0	0	0	0	0	1	1	1

(a)

	{1,2,3}	{1,3,4}	{3,4,5}	{8,9,10}
{1,2,3}	3	2	1	0
{1,3,4}	2	3	2	0
{3,4,5}	1	2	3	0
{8,9,10}	0	0	0	3

(b)

	{1,2,3,4}	{3,4,5,6,7}	{8,9,10}
{1,2,3,4}	3	2	0
{3,4,5,6,7}	2	3	0
{8,9,10}	0	0	3

(c)

FIGURE 3. The process of seed sets selection based on the changes in the network formal context corresponding to Fig. 2.

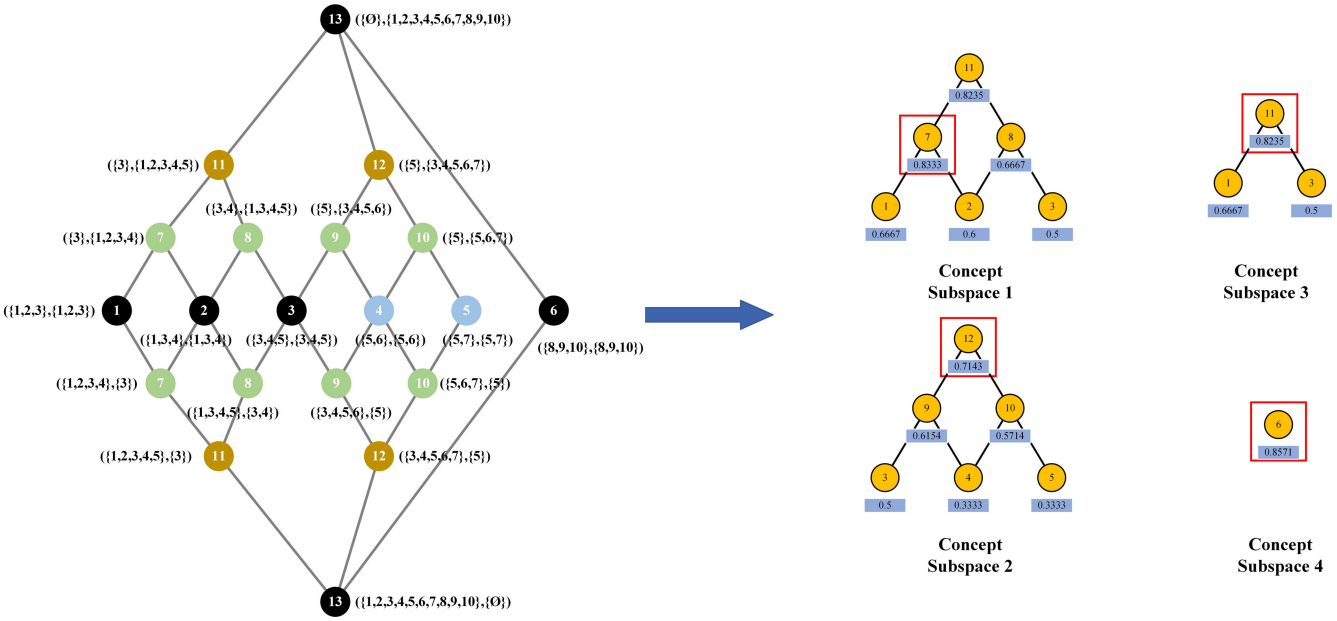


FIGURE 4. The Hasse diagram constructed from equiconcepts enables the decomposition of larger concept spaces into smaller, distinct equiconcept subspaces. This decomposition is guided by the stability coefficients of the equiconcepts. Given the symmetrical nature of the Hasse diagram, the decomposition can be illustrated using only the upper half of the concept subspaces. In this process, the concept with the highest stability coefficient in each concept subspace is selected. The objects within this chosen concept are then considered as a seed set.

TABLE 1. Comparative analysis of extended modularity in seed sets selected using different methods about Fig. 2(a) network.

Methods	Ref.	Seed Sets	*EQ*
NEB on degree centrality	[33]	{3,4,5,6,7,8} , {1,2,3} , {9,10}	13.54
NEB on betweenness centrality	[33]	{1,2} , {3,4,5,6,7,8} , {9} ,{10}	6.21
NEB on closeness centrality	[33]	{1,2} , {3,4,5,6,7,8} , {9,10}	12.72
NEB on eigenvector centrality	[33]	{1,3,4,5} , {2} , {6,7} , {8,9,10}	19.82
k-clique	[8]	{1,2,3} , {1,3,4} , {3,4,5} , {8,9,10}	17.68
Seed set strategy	[44]	{5} , {3} , {8} , {1} , {2}	-
Key structure identification algorithm	[22]	{8,9,10} , {5,6} , {5,7} , {1,2,3} , {1,3,4} , {3,4,5}	22.67
WE method	-	{1,2,3,4} , {3,4,5,6,7} , {8,9,10}	**34.99**

than that of concept 7, so concept 11 is not chosen, but concept 7 is. Weak equiconcept 12, formed by concepts 9 and 10, is selected, along with concept 6. In Fig. 2(c), there are three small communities in blue, green, and purple, composed of objects from concepts 7, 12, and 6, forming three seed sets. Object 3, as overlapping node in two seed sets, demonstrates their multiple social relationships, belonging to two social circles.

Algorithm 2 Community Expansion

Input: Network formal context $NFC = \{U, M_1, A, I\}$, seed set S, teleport probability α, and error ϵ
Output: Preliminary community division $\overline{U}$

$\hat{\pi}(s,t) \leftarrow 0, s \in S$
$\hat{r}(s,s) \leftarrow 1, \hat{r}(s,t) \leftarrow 0$
for each $s \in S$ **do**
  calculate $sim(s, neighbor(s))$ and $d_s(t)$ according to Definition 11
end for
for any $\hat{r}(s,t) \geq \varepsilon d_s(t)$ **do**
  $\mu \leftarrow \hat{r}(s,t) - \frac{\varepsilon}{2} d_s(t), \hat{\pi}(s,t) \leftarrow \hat{\pi}(s,t) + (1-\alpha) \cdot \mu, \hat{r}(s,t) \leftarrow \frac{\varepsilon}{2} d_s(t)$
end for
for object $g \in V, (g,t) \in E$ **do**
  $\hat{r}(s,g) = \hat{r}(s,g) + \frac{sim(g,t)}{d_s(t)} \cdot \alpha \cdot \mu$
end for
return $\hat{\pi}(s,t)$ as the estimator for $\pi(s,t), s \in V$
Using sweep operation: sort objects by decreasing $\frac{\hat{\pi}(s,t)}{d_s(t)}$, select the first p elements from the sequence, calculate the conductance of them, and let community be the set of objects that reaches the minimum value
return community division $\overline{U}$

This approach is contrasted with several common seed selection methods: directly selecting k-cliques, using various centrality measures as an importance metric for nodes, sorting them, and then forming a seed set from the highest node and its neighbors. Nodes selected in this process are not used in subsequent neighborhood expansions for other seed sets, continuing until no nodes are left to select. This neighborhood expansion method is concisely referred to as NEB in Table 1. Four types of centrality are considered for comparison: degree, betweenness, closeness, and eigenvector centralities. The seed selection strategy proposed in [33] is also compared, using the cohesion measure EQ from Section IV to assess the cohesiveness of the seed sets, as shown in Table 1, where "-" represents a negative EQ value.

When comparing the first seven seed selection algorithms, where each seed set is considered an initial community, the extended modularity EQ is used to measure these small communities. The seed set chosen by WE method exhibits the highest EQ. Possessing high modularity in the initial step of community expansion influences the quality of the communities expanded from these seeds in later stages. This indicates the excellent quality of seeds selected under the equiconcept, highlighting the effectiveness of the WE method in enhancing the overall community detection process.

C. COMMUNITY EXPANSION BY SELECTING SEEDS

Definition 11: The Jaccard coefficient is used to measure the homogeneity of attributes of a node. The Jaccard similarity coefficient treats the attributes of two nodes as two sets. For nodes $s, g, t \in V$, $sim(s,g) = \frac{|A_s \cap A_g|}{|A_s \cup A_g|}$. When

Algorithm 3 Community Merging

Input: Preliminary community division $\overline{U}$
Output: New community division U

$U \leftarrow \emptyset$
Calculate overlap for $\overline{U}$ according to Definition 12
Calculate $\overline{CO}$ for $\overline{U}$ according to Definition 13
for each $\overline{U_i}, \overline{U_j} \in \overline{U}$ **do**
  if $\mathrm{CO}(U_i, U_j) > \overline{CO}$ **then**
    $\overline{U_{ij}} = \overline{U_i} \cup \overline{U_j}, U \leftarrow U \cup \overline{U_{ij}}$
  else if $U = \emptyset$ **then**
    $U \leftarrow \{\overline{U_i}, \overline{U_j}\}$
  else
    $U \leftarrow \{U, \overline{U_i}, \overline{U_j}\}$
  end if
  if $\mathrm{CO}(\overline{U_i}, \overline{U_{j_0}}) \geq \cdots \geq \mathrm{CO}(\overline{U_i}, \overline{U_{j_k}}) \geq \overline{CO}$ **then**
    $\overline{U_{ij_0}} = \overline{U_i} \cup \overline{U_{j_0}}, U \leftarrow U \cup \overline{U_{ij_0}}$
  end if
end for
return U

considering attribute homogeneity in node propagation, the preference of $sim(s,g)$ for nodes is taken into account. Therefore, we have:

$$d_s(t) = \sum_{(g,t) \in E} sim(s,g) \tag{12}$$

$$\hat{r}(s,g) = \hat{r}(s,g) + \frac{sim(g,t)}{d_s(t)} \cdot \alpha \cdot \mu \tag{13}$$

where A_s is the attribute set of node s, $\hat{r}(s,g)$ is the residual quantity of nodes s and g, which is a neighbor of seed nodes s, μ updated to $\hat{r}(s,t) - \frac{\varepsilon}{2} d_s(t)$, and $\hat{r}(s,t)$ is the residual quantity between the source node and the target node.

Through the algorithm discussed in the previous section, a set of seeds is obtained. This section focuses on expanding these seeds into preliminary communities. During the process of information propagation in nodes, a portion of the information is stored with a certain probability for the node itself, while another portion is likely to be transmitted to neighboring nodes. Since nodes have attributes, those with more shared attributes have a higher probability of information transfer, indicating a preference between nodes. Considering this important aspect, the push operation is modified to better align with real-world scenarios. This approach incorporates the attribute-based preferences of nodes into the community expansion process, enhancing the relevance and accuracy of the detected communities, as shown in Fig. 5.

As shown in Algorithm 2, the forward push process represents the message transmission between source and target nodes. In this process, for any given node $t \in V$, it temporarily holds a forward storage $\hat{\pi}(s,t)$ and a forward residual $\hat{r}(s,t)$. These values are continually updated through forward push operations. Unlike the average distribution of residuals to each neighbor of t, for nodes with attributes, those with more shared attributes exhibit greater homogeneity.

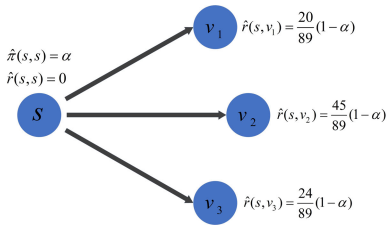


FIGURE 5. The probability distribution is influenced by attribute homogeneity, in which the attribute similarity between the source node s and nodes v_1, v_2, and v_3 is assumed to be $sim(s, v_1) \mathcal{D} \frac{1}{3}$, $sim(s, v_2) \mathcal{D} \frac{3}{4}$ and $sim(s, v_3) \mathcal{D} \frac{2}{5}$, during the residual propagation process of the source node s.

Algorithm 4 Node Adjustment

Input: Node set U_{nodes}
Output: Final community division H

for each $v \in U_{\text{nodes}}$ **do**
  Sort nodes by degree: degree(v_1) $>$ degree(v_2) $>$ $\cdots$ $>$ degree(v_s)
end for
for each U_{i_j} $(j = 1, \cdots, k)$ connected with v **do**
  Calculate $IV(\{v\} \cup U_{i_1}), IV(\{v\} \cup U_{i_2}), \cdots, IV(\{v\} \cup U_{i_k})$ according to Definition 14
end for
$\delta \leftarrow IV(\{v\} \cup U_{i_j}) - IV(U_{i_j})$
if $\delta \geq 0$ **then**
  Select U_{i_j} with min(δ), then update $U_{i_j} \leftarrow U_{i_j} \cup \{v\}$
else
  Select U_{i_j} with max(δ), then update $U_{i_j} \leftarrow U_{i_j} \cup \{v\}$
end if
return H

Considering this, the forward residual $\hat{r}(s, t)$ is temporarily stored in the neighbors of other t nodes. The current residual held by t is distributed to the neighbors of t based on different "attractions" between nodes, allocating probabilities according to attribute similarity.

D. COMMUNITY MERGING

This section proposes a target function value to assess whether two communities can be merged into a larger one. This value focuses on the degree of structural and attribute overlap between communities. The greater the overlap, the more reasonable it is to merge the two communities into one. By merging communities with high overlap, the number of communities can be reduced while retaining relevance and similarity, thus yielding more meaningful community partition results. This approach helps to reduce the fragmentation of communities, enhancing the accuracy and coherence of community division, and making community detection more practical and applicable.

Definition 12: The overlapping degree between communities U_i and U_j is defined as:

$$CO(U_i, U_j) = \theta \cdot \left(1 - \frac{|U_i \cap U_j|}{|U_i \cup U_j|}\right) + (1 - \theta) \cdot IE(U_i \cup U_j) \tag{14}$$

where U_i and U_j are the initially partitioned i-th and j-th communities, respectively, and $IE(A_i)$ is the information entropy of the formal context (U_i, A_i, I_i), which is defined as:

$$IE(A_i) = \frac{1}{|U_i|} \sum_{x \in U_i} \left(1 - \frac{|g_{A_i} f_{A_i}(x)|}{|U_i|}\right)$$

Definition 13: Based on the degree of overlapping between communities, determine whether two communities can be merged into one. The average value of community overlapping $CO(U_i, U_j)$ as follows:

$$\overline{CO} = \frac{1}{q} \sum_{U_i, U_j \subseteq U} CO(U_i, U_j) \tag{15}$$

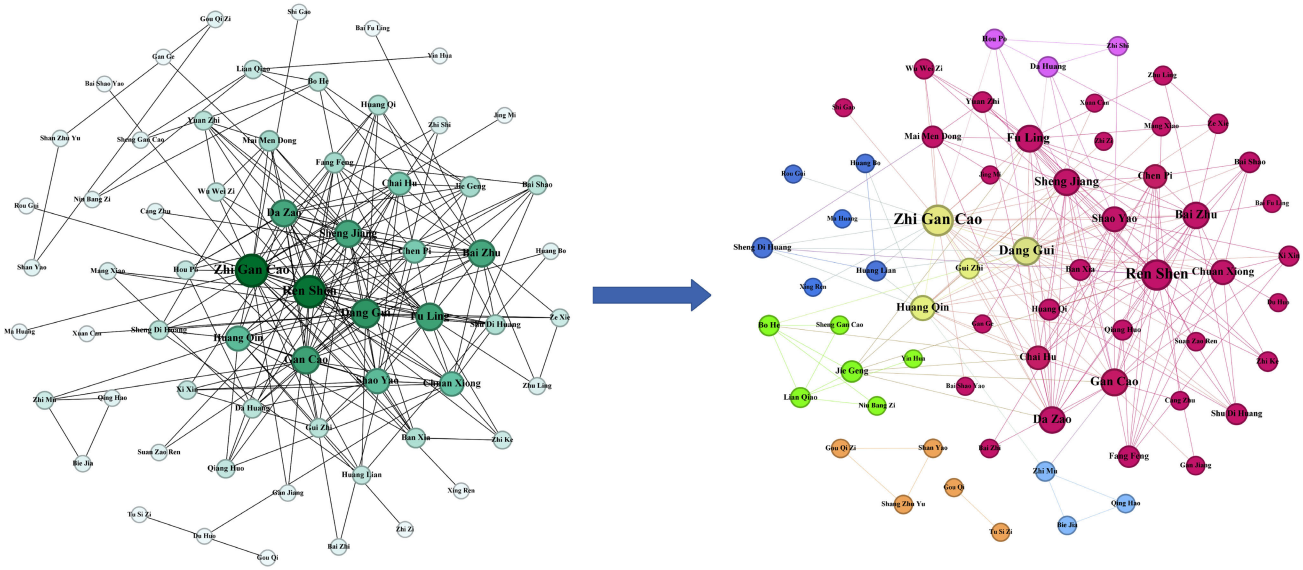
where q is the number of communities in the initial partition.

As shown in Algorithm 3, calculate the overlapping matrix based on Definition 11, similar to the network formal context in Fig. 3(b). Only consider the average of the elements in the upper triangle of the matrix. Set a threshold for the average value to filter out elements with low overlapping. Each element corresponds to the strength of overlap between communities. Keep the elements that indicate the need for merging operations between communities.

Theorem 1: Given $\theta = 0.5$, if $\mu < \overline{CO}$ holds, the community partition tends to decrease after merging $CO(U_i, U_j)$.

Proof: Assuming that $G = (V, E)$ is initially divided into q communities $\{U_1, U_2, \cdots, U_q\}$, each with a unique formal context represented by (U_1, A_1, I_1), (U_2, A_2, I_2), $\cdots$, (U_q, A_q, I_q), the corresponding information entropy for $IE(A_1), IE(A_2), \cdots, IE(A_q)$, and community U_i after merging with U_j is represented by $CO(U_i, U_j)$, where q represents the number of communities before merging.

$$\begin{aligned}
&\frac{1}{q}\sum_{r=1}^{q} IE(A_r) - \frac{1}{q-1}\left(\sum_{r=1, r\neq i,j}^{q} IE(A_r) + IE(A_i \cup A_j)\right) \\
&> \frac{1}{q}\sum_{r=1}^{q} IE(A_r) \\
&\quad - \frac{1}{q-1}\left(\sum_{r=1, r\neq i,j}^{q} IE(A_r) + IE(A_i) + IE(A_j)\right) \\
&= \left(\frac{1}{q} - \frac{1}{q-1}\right)\sum_{r=1, r\neq i,j}^{q} IE(A_r) \\
&\quad - \left(\frac{2}{q} - \frac{1}{q-1}\right)(IE(A_i) + IE(A_j))
\end{aligned}$$



(a) (b)

FIGURE 6. The TCMs network graph, and the results of the TCMs community division based on WEOCD method.

$$= \frac{q-2}{q\cdot(q-1)}\left(IE(A_i)+IE(A_j)-\frac{\sum_{r=1,r\neq i,j}^{q} IE(A_r)}{q-2}\right)$$

Given $IE(A_i) > \frac{1}{q}\cdot\sum_{r=1}^{q} IE(A_r)$ and $IE(A_j) > \frac{1}{q}\cdot\sum_{r=1}^{q} IE(A_r)$, we have

$$\begin{aligned} IE(A_i)+IE(A_j) &\geq \frac{1}{q}\cdot\sum_{r=1}^{q} IE(A_r) \\ &\geq \frac{1}{q-2}\cdot\sum_{r=1,r\neq i,j}^{q} IE(A_r) \end{aligned}$$

Thus,

$$\frac{q-2}{q\cdot(q-1)}\left(IE(A_i)+IE(A_j)-\frac{\sum_{r=1,r\neq i,j}^{q} IE(A_r)}{q-2}\right) > 0$$

In other words,

$$\frac{1}{q}\sum_{r=1}^{q} IE(A_r) - \frac{1}{q-1}\left(\sum_{r=1,r\neq i,j}^{q} IE(A_r) + IE(A_i \cup A_j)\right) > 0$$

is true. □

E. ADJUSTMENT OF NODES

After the community merging, there may still exist isolated nodes and nodes that haven't been assigned to any community. It is necessary to determine whether these nodes can form a new community or be assigned to existing ones.

Definition 14: For an unassigned node U_{nodes}, when choosing to connect to the nearest community U_i, the following criteria can be used to measure the impact of the node joining the community U_i:

$$IV = \frac{cond(\{u\}\cup U_i)}{2} + \frac{IE(A'_i)}{2} \tag{16}$$

where $IE(A'_i)$ is the information entropy of the formal context of new cluster $\{u\}\cup U_i$.

We give Algorithm 4 for undivided node adjustment. An isolated node is not connected to other nodes. At this time, the node can exist as an independent community. For nodes that are not isolated nodes and are not divided into any communities, we select connections in turn. The nearest community or communities, as judged by Definition 12, choose to join the largest community.

Example 2: The paper conducted preprocessing on a dataset of 144 prescription formulas from Traditional Chinese Medicine scientist Zhang Zhongjing's ''Treatise on Cold Damage and Miscellaneous Diseases''. In this preprocessing, Traditional Chinese Medicine (TCMs) were treated as nodes. If two TCMs appeared together in the same prescription formula at least 3 times, a connection was established between those two medicines. Additionally, the attributes of the TCMs were considered, including the four natures (''qi'') and five flavors (''wei'') of TCMs, along with the attributes ''ping'' and ''dan''. In total, there were 11 attributes for the medicinal nodes. The final dataset includes 63 medicinal nodes and 202 edges representing co-occurrences in prescription formulas, as shown in Fig. 6(a).

For the drug classification task in Example 2 using the WEOCD algorithm, the partition results are depicted in

TABLE 2. The division of medicinal herbs.

Community	Indications
Red	Mainly regulating qi and blood, promoting blood circulation, resolving stasis, clearing heat, and nourishing blood. Corresponding symptoms: upper respiratory tract infections, blood deficiency, qi deficiency.
Blue	Adjunctive treatment for damp-heat conditions and supplementary therapy for blood-heat conditions.
Puple	Treatment for constipation, abdominal pain, and bloating caused by heat accumulation and qi stagnation. Mainly used for the syndrome of heat-induced constipation.
Green	Treatment for cold, fever, cough, and phlegm-heat caused by wind-heat invasion.
Orange	Mainly nourishes kidney yin and replenishes kidney yang. Corresponding symptoms: liver and kidney yin deficiency.
Sky blue	Mainly used to balance yin and yang, clear heat, detoxify, and nourish yin and moisten dryness.
Yellow	Overlapping nodes between the red and blue communities.

Fig. 6(b), with the primary indications and effects of TCMs in each community detailed in Table 2.

F. COMPLEXITY ANALYSIS

This section discusses the time complexity of constructing formal concept lattices. In the proposed formal context, the number of objects is denoted as $|V|$, and the number of attributes is denoted as $|V|$. N represents the total number of concepts. U_C represents all the classified nodes in the network. The time complexity analysis is as follows.

1) During the seed selection phase, the construction of the network formal context has a time complexity of $O(|V|^3 + |V|^2 N)$.

2) The time complexity of the seed expansion phase can be represented as $O\left(\sum_{i=1}^{q} links(C_i, U_C)\right)$.

3) The process of community merging is divided into two parts. In the first part, calculations are performed on the community's topological structure, considering the largest community $C_{\max}$ in terms of the number of nodes. In the second part, information entropy is also influenced by $C_{\max}$. Overall, the time complexity can be expressed as $O(|C_{\max}|)$.

4) The final step of community optimization involves adjusting nodes, considering the possibility of nodes moving to other communities. Therefore, the time complexity is proportional to the number of community partitions C_i, and it can be expressed as $O(|C_i|)$.

In summary, for the four steps mentioned above, the overall time complexity is given by $O(|V|^3 + |V|^2 N)$.

V. EXPERIMENTS

Here, experiments on 7 real-world datasets are conducted to validate the effectiveness of the proposed algorithm.

A. DATESETS AND EVALUATIONS METRICS

1) DATESETS

The datasets used in this study hold significant relevance and importance in research.

Participant dataset is a participant-project network formed during the 2013 Santa Fe Complex System Summer School, consisting of 61 nodes and 224 edges. Nodes represent participants, while edges represent collaborative relationships among participants in a project. Attributes are the academic background of the participant. Each participant's background consists of four different categories of subjects: math & physics, life sciences & ecology, social sciences & economics, and computing & programming. The Medcine dataset, as illustrated in Example 2, represents pharmaceutical prescription data. It serves as a valuable resource for exploring patterns and relationships in prescription records. WebKB dataset comprises four subnetworks collected from four different universities: Cornell, Texas, Washington, and Wisconsin. Each subnetwork includes multiple communities, web pages, binary word attributes (with 1703 dimensions), and edges. It provides valuable information about web page categorization. Cora dataset contains 2708 scientific publications categorized into 7 classes. With 5429 edges, each publication is described using binary word attributes (with 1433 dimensions), indicating the presence or absence of specific words in the document. Citeseer dataset consists of 3312 scientific publications categorized into 6 classes. Each publication is associated with a 3703-dimensional binary word attribute vector, denoting the presence or absence of words from a dictionary. The Facebook dataset is constructed from Facebook "circles", containing 4039 nodes (users) and 88234 edges (friendship connections). It offers insights into social network analysis. The Deezer Europe dataset represents a network of deezer users from european countries. Edges in this network denote follower relationships between users based on their shared liking of music artists. Node features are extracted based on users' preferred artists. All information of each dataset is shown in Table 3.

2) EVALUATION METRICS

- Average P, R and F_1.
  In this section, we use Macro-Precision, Macro-Recall, and Macro-F1 to evaluate the local community detection result. The Precision and Recall for class i ($i = 1, \cdots, q$) can be represented as follows:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \tag{17}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{18}$$

TABLE 3. Basic information of selected datasets.

Dataset	Nodes	Edges	Attributes	Average degree $\bar{k}$	Cluster coefficient C
Participant	61	224	4	3.6721	0.6729
Medcine	63	202	11	3.2063	0.5329
Webkb	807	1608	1703	1.8335	0.2159
Cora	2708	5429	1433	2.0048	0.2408
Citeseer	3312	4732	3703	1.4287	0.1447
Facebook	4039	88234	175	21.8455	0.6055
Deezer Europe	28281	92752	3000	3.2797	0.1412

Calculate the average Precision and Recall across all classes:

$$\text{Precision}_{\text{macro}} = \frac{\sum_{i=1}^{q} \text{Precision}_i}{q} \tag{19}$$

$$\text{Recall}_{\text{macro}} = \frac{\sum_{i=1}^{q} \text{Recall}_i}{q} \tag{20}$$

$F1$ scores:

$$F1_{\text{macro}} = 2 \cdot \frac{\text{Precision}_{\text{macro}} \cdot \text{Recall}_{\text{macro}}}{\text{Precision}_{\text{macro}} + \text{Recall}_{\text{macro}}} \tag{21}$$

- Expanded modularity.
  Due to the fact that the modularity function is only applicable to non-overlapping community detection, this paper adopts the extended modularity to assess the quality of overlapping community structure partition. A larger value of the extended modularity indicates a better community partitioning result. The extended modularity function is defined as follows:

$$EQ = \frac{1}{2m} \cdot \sum_{i=1}^{q} \sum_{u \in C_i, v \in C_i} \frac{1}{Q_u Q_v} [A_{uv} - \frac{k_u k_v}{2m}] \tag{22}$$

  where Q_u represents the community to which node u belongs, A_{uv} is the adjacency matrix, k_u is the degree of node u, and m represents the number of edges in the network.

B. EXPERIMENTAL RESULTS AND ANALYSIS

To validate the effectiveness of the proposed algorithm, several comparative algorithms were selected.

BIGCLAM: a method that considers only the network structure and specifies the number of communities as the true number of communities.

COPRA: a method based on label propagation. It is an overlapping community detection algorithm that uses membership scores to help nodes determine their belonging to multiple communities. It terminates when the remaining label sets in the network are the same after two consecutive iterations or when it reaches the maximum iteration limit. It exhibits instability due to the randomness in label selection.

LFM: a method as a representative of community expansion methods. LFM defines a fitness function for a subgraph of the network. It exhibits uncertainty in results due to the randomness in seed selection.

NISE: a greedy method for seed node selection and community generation. The parameter specifying the number of communities is set to be the true number of communities. It uses the sphub method for seed selection and PPR for community expansion.

EWKM: a method that considers both node attributes and subspaces. It requires the user to provide the number of communities in advance, and this parameter is set to be the true number of communities.

SAC: an algorithm that fuses attributes and topology, and has high complexity.

These comparative algorithms were chosen to evaluate the proposed algorithm's performance in terms of community detection, especially in handling overlapping communities. Comparing the results of these algorithms provides insights into the strengths and weaknesses of the proposed approach.

TABLE 4. Time complexities of comparative algorithms.

Algorithm	Complexity
BIGCLAM	$O(qn^2)$
COPRA	$O(vm\log(\frac{vm}{n}))$
LFM	$O(l{N_{cu}}^2 + qN_{nu})$
NISE	$O\left(\sum_{i=1}^{q} \text{links}(C_i, U_C)\right)$
EWKM	$O(hmnq)$
SAC	$O(n^3)$
WEOCD	$O(\|V\|^3 + \|V\|^2 N)$

Table 4 lists the time complexities of the mentioned methods, where q represents the number of community classifications, n is the number of nodes, m is the number of edges, v is the threshold used in the COPRA algorithm, l is the initial number of seed nodes, N_{cu} is the average number of nodes generated within a community, N_{nu} represents the average number of nodes in the neighborhood of a community or node, $\sum_{i=1}^{q} \text{links}(C_i, U_C)$ is the sum of node degrees within community C_i, and h is the number of iterations required for the clustering process to converge.

In this section, we use the EQ value to evaluate the density and strength of the communities formed by expanding the

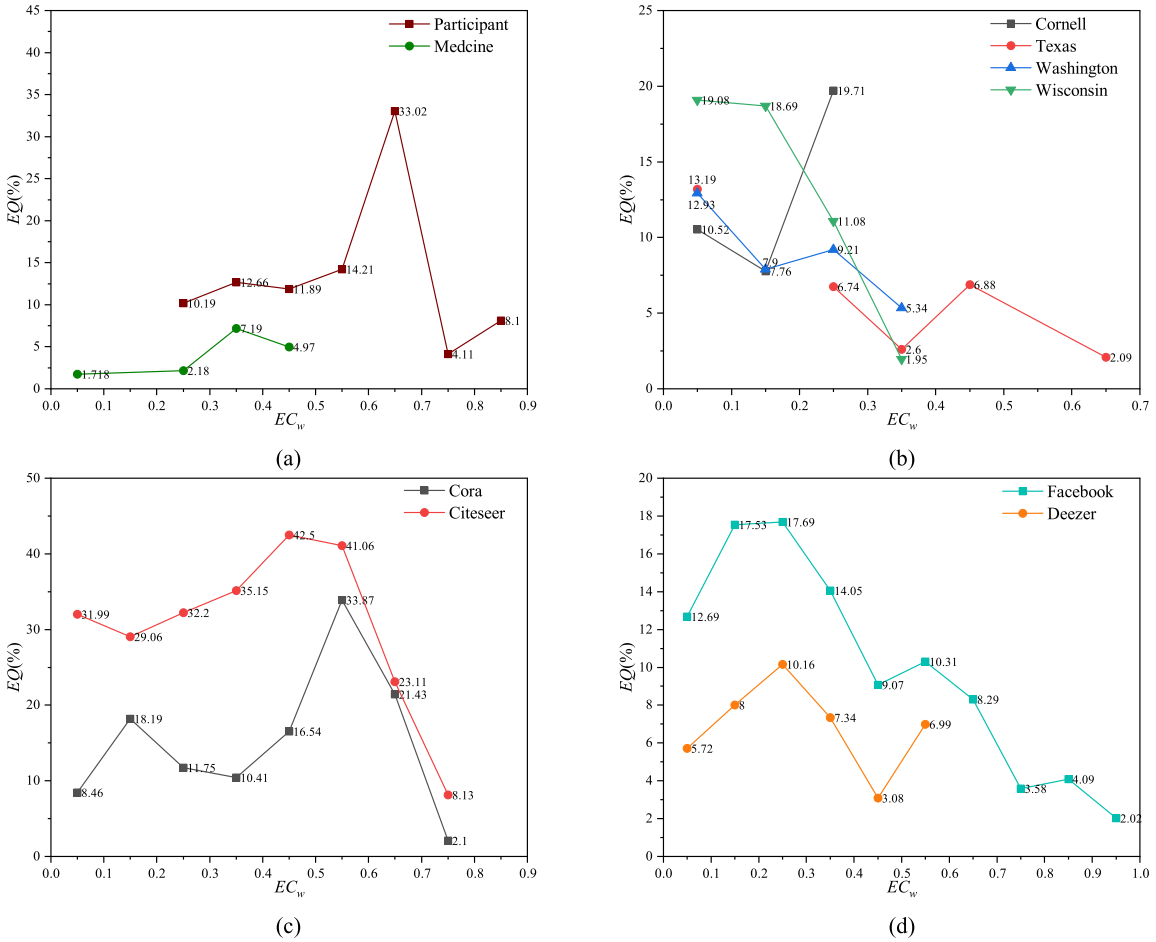


FIGURE 7. The relationship of the community modularity and stability coefficient of each order.

seed sets under each stability value. It can be observed that the trends of EQ values for all four datasets show a wave-like pattern. According to Properties 1 and 2, for networks with larger average degree or clustering coefficient, the EQ values for each segment tend to be smaller compared to networks with lower average degree. Within each segment of a dataset, if the segment with the highest stability coefficient shows a decreasing trend, it indicates that the nodes in this network are less likely to exhibit characteristics of small, marginal communities. If the EQ values of communities expanded by stability coefficient in the range of 0-0.1 are relatively higher than the average level, it suggests that in this network, nodes tend to aggregate with communities that have higher degrees due to the cohesive nature of the network.

Considering the average degree and clustering coefficient for the seven chosen datasets, we obtain that the average degree is $\overline{k}_{Facebook} > \overline{k}_{Participant} > \overline{k}_{Deezer} > \overline{k}_{Medcine} > \overline{k}_{Cora} > \overline{k}_{Webkb} > \overline{k}_{Citeseer}$ and the clustering coefficient is $C_{Participant} > C_{Facebook} > C_{Medcine} > C_{Cora} > C_{Webkb} > C_{Citeseer} > C_{Deezer}$. In Fig. 7(a), Participant dataset has the highest clustering coefficient, and EC_w has seven stages. EQ ranging from 0.6 to 0.7 is considered very high, with lower values at both ends. Nodes tend to cluster with nodes with high surrounding degrees, and weak equiconcepts with EC_w as the intermediate stage are prioritized as seeds. Medcine dataset has four stages where weak equiconcepts can be detected, and there is not much difference in EQ between each stage. As a small dataset, when selecting seeds, we need to consider the weak equiconcepts of the stability coefficients for all stages. The structure in Fig. 7(b) is relatively complex, where Webkb contains four subgraphs: Cornell, Texas, Washington and Wisconsin, encompassing an average of four phases of isopotential concepts. The differentiation in community modularity arising from the expansion of weak equiconcepts is notably pronounced. This complexity is attributable to the intricate inter-node relationships within the four subgraph datasets, characterized by extensive interactions and a diversity of relationship types. In seed selection, the strategy involves initially choosing from the weak equiconcepts with larger stability coefficients, proceeding in a sequential manner. In Fig. 7(c), the citation networks Citeseer and Cora demonstrate a truncation in their stability coefficients, signifying the scarcity of communities that are disconnected from external influences.

TABLE 5. Comparison of selected algorithm in terms of P, F_1 and R.

Algorithm	P						
	Participant	Medcine	Webkb	Cora	Citeseer	Facebook	Deezer
BIGCLAM	0.5	0.5593	0.4013	0.1715	0.2081	0.2775	0.5015
COPRA	0.3772	0.4784	0.3161	0.7305	0.7111	0.3085	0.5102
LFM	0.4371	0.5351	0.3683	0.4262	0.4241	0.3668	0.5005
NISE	0.4588	0.6437	0.3944	0.6956	0.6177	0.5263	0.5349
EWKM	0.4868	0.3208	0.4149	0.1735	0.1854	0.3954	0.5
SAC	0.42	0.6298	0.2885	0.5035	0.242	0.3245	-
WEOCD	**0.7173**	**0.9954**	**0.4669**	**0.7325**	**0.7501**	**0.7408**	**0.5963**
Algorithm	F_1						
	Participant	Medcine	Webkb	Cora	Citeseer	Facebook	Deezer
BIGCLAM	0.5151	0.4564	0.4301	0.131	0.1727	0.3045	0.3674
COPRA	0.3291	0.5058	0.3096	0.719	0.7245	0.3783	0.4819
LFM	0.4387	0.5681	0.3683	0.4138	0.4447	0.4066	0.3532
NISE	0.4651	0.6894	0.4035	0.7091	0.6851	0.4562	0.5349
EWKM	0.4305	0.3787	0.4088	0.1268	0.097	0.4331	0.5
SAC	0.4074	0.8778	0.2845	0.5196	0.226	0.1531	-
WEOCD	**0.7360**	**0.9738**	**0.5019**	**0.7526**	**0.7701**	**0.7973**	**0.5963**
Algorithm	R						
	Participant	Medcine	Webkb	Cora	Citeseer	Facebook	Deezer
BIGCLAM	0.7185	0.6561	0.6994	0.589	0.3875	0.8434	0.5436
COPRA	0.3291	0.8333	0.6907	0.7286	0.7593	**0.9321**	0.5156
LFM	0.6844	0.8506	0.5911	0.6905	0.6391	0.8452	0.5622
NISE	0.5429	0.8472	0.4441	0.7564	**0.8864**	0.8421	0.5349
EWKM	0.4325	0.8298	0.7193	**0.8601**	0.6935	0.7769	0.5
SAC	0.7357	0.6414	0.6591	0.5461	0.3303	0.3786	-
WEOCD	**0.7663**	**0.9583**	**0.7215**	0.7824	0.8225	0.9309	**0.599**

This phenomenon underscores the academic sphere's reliance on communication and mutual learning for producing high-quality scholarly papers. Operating in isolation is neither advisable nor conducive to positive development. In Fig. 7(d), Facebook exhibits the highest average degree and the second highest clustering coefficient. However, as the modularity expands, the stability value of each segment consistently decreases. In social networks, aside from isolated nodes, there are also peripheral nodes within marginal communities. Such social connections often prove to be more stable. Conversely, nodes with high stability coefficients tend to be more versatile, participating in multiple communities and social circles. Deezer, as a music-based social platform, has the lowest clustering coefficient. This observation, alongside the performance of the stability coefficient, suggests that social interactions on music platforms are relatively independent. The common interest in music might be the primary reason for users to follow each other. This analysis provides insights for the current study. In networks with high average degree and clustering coefficient, selecting seed sets based on low stability coefficients is not advisable. For communities emanating from high stability coefficient seed sets, it is viable to consider them as distinct groups. However, in networks with low clustering coefficients, it is meaningful to incorporate the concept of low stability coefficients into the part of the seed set.

In Table 5, the values of P, F_1, and R are compared across 7 datasets among 6 algorithms, where "-" indicates excessively long computational time (> 7 days). It can be seen from the table that for indicators P and $F1$, our algorithm is the best compared to all the selected algorithms on all the datasets; for indicator R, our algorithm is better than all the compared algorithms on four datasets, and it is ranked second on the rest datasets. Therefore, our algorithm is satisfactory in achieving overlapping community detection task.

In addition, for our purpose, we continue to make a detailed analysis of the relationship between the performance of our algorithm and how to select seeds of networks. Notably, our algorithm demonstrates excellent performance on the Paticipant, Medicine, Webkb, and Deezer datasets, attributed to the consideration of both edge communities and the most representative seeds during the seed selection process. Detailed discussion is as follows:

(1) Through the selection of seeds for Medcine dateset by WE method, we are able to screen out the most representative drug set as the seed set for expansion, as well as the marginal drug community. Also as a small data set, Paticipant has the highest clustering coefficient. Therefore, for the weak equiconcepts with small stability coefficients, we did not choose them as seeds, as it inevitably includes numerous nodes with high degrees. These nodes with high degrees frequently appear in various orders of weak equiconcepts, and we selected the weak equiconcept satisfying the maximum EC_w. There have been some nodes with low degrees but they existed in weak equiconcepts with high EC_w, and these nodes were detected as fringe communities.

(2) Webkb dataset lacks high homogeneity, and high-degree nodes may not share the same category as their neighbors. For instance, if a teacher has many students, teacher's neighbors are not in the same category, resulting in a relatively small community. In seed selection, we prioritized seeds formed by the highest stability coefficient, avoiding their expansion in subsequent steps. Note that we rarely considered seeds with low stability coefficients due to the dataset's heterogeneity. Overall, performance is not uniformly high across all methods, but our approach in this paper proves to be the most effective.

(3) As a social dataset, Deezer is relatively independent among users and is not as dense as Facebook. This data took a long time for SAC method to run, but no results were obtained because it was terminated forcibly, so it was not evaluated successfully in the experiments. In comparison with the other six methods in terms of running time, the accuracy of finding communities is more advantageous. This dataset has fewer stability factor orders than Facebook, and all of them were selected as the seed set. The WEOCD method also performs better on networks that do not have obvious community characteristics.

Note that, for indicator R, our algorithm is ranked second on datasets Cora, Citeseer, and Facebook. This is because the Cora and Citeseer datasets have fewer fringe communities, and few people publish papers independently without collaborating with other scholars. For the Facebook dataset, it has a large number of isolated nodes that have not been classified, which to some extent reduces the performance of our algorithm on the indicator R.

VI. CONCLUSION

In this paper, an effective method for selecting seed sets of a network has been proposed within the framework of FCA. Firstly, a representation of a network based on FCA has been implemented to generate weak equiconcepts that characterize the network structure. Subsequently, concept subspaces have been filtered, and concepts that are more community-oriented and representative have been selected from these subspaces. Furthermore, an improved personalized PageRank clustering algorithm has been used to expand the selected concepts. Finally, a community optimization scheme has been obtained. Experiments conducted on real social networks have demonstrated the efficiency and effectiveness of the proposed method.

Further researches include: (1) a faster method for weak equiconcept generation and concept stability coefficient calculation should be explored; (2) in this paper, we have focused solely on the single-layer relationship within the topology of complex networks. Note that the single-layer methods may not be able to fully reveal more complex multi-layer network structures and interactive behaviors between clusters. Therefore, it is still necessary to investigate multi-layer relationships in the network formal context, integrating attributes and topological structures more effectively into formal contexts and network analysis.

REFERENCES

[1] S. K. Gupta and D. P. Singh, "Seed community identification framework for community detection over social media," *Arabian J. Sci. Eng.*, vol. 48, no. 2, pp. 1829–1843, Feb. 2023, doi: 10.1007/s13369-022-07020-z.

[2] T. Schuurman and E. Bruner, "Modularity and community detection in human brain morphology," *Anatomical Rec.*, vol. 307, no. 2, pp. 345–355, Aug. 2023, doi: 10.1002/ar.25308.

[3] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," in *Proc. 31th Int. Conf. Mach. Learn.*, Nov. 2014, vol. 32, no. 2, pp. 595–603.

[4] V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," *ACM Comput. Surv.*, vol. 49, no. 4, pp. 1–33, Dec. 2016, doi: 10.1145/3012704.

[5] X. Teng, J. Liu, and M. Li, "Overlapping community detection in directed and undirected attributed networks using a multiobjective evolutionary algorithm," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 138–150, Jan. 2021, doi: 10.1109/TCYB.2019.2931983.

[6] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Comput. Surv.*, vol. 45, no. 4, pp. 1–35, Aug. 2013, doi: 10.1145/2501654.2501657.

[7] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, Jun. 2005, doi: 10.1038/nature03607.

[8] C. Lee, F. Reid, A. McDaid, and N. Hurley, "Detecting highly overlapping community structure by greedy clique expansion," 2010, *arXiv:1002.1827*.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008, doi: 10.1088/1742-5468/2008/10/p10008.

[10] D. Jin, Z. Yu, P. Jiao, S. Pan, D. He, J. Wu, P. S. Yu, and W. Zhang, "A survey of community detection approaches: From statistical modeling to deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 2, pp. 1149–1170, Feb. 2023, doi: 10.1109/TKDE.2021.3104155.

[11] M. Mitrović and B. Tadić, "Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, no. 2, Aug. 2009, Art. no. 026123, doi: 10.1103/physreve.80.026123.

[12] S. Gregory, "Finding overlapping communities in networks by label propagation," *New J. Phys.*, vol. 12, no. 10, Oct. 2010, Art. no. 103018, doi: 10.1088/1367-2630/12/10/103018.

[13] S. Zhang, R.-S. Wang, and X.-S. Zhang, "Identification of overlapping community structure in complex networks using fuzzy -means clustering," *Phys. A, Stat. Mech. Appl.*, vol. 374, no. 1, pp. 483–490, Jan. 2007, doi: 10.1016/j.physa.2006.07.023.

[14] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*. Berlin, Germany: Springer, 1999.

[15] M. Kaytoue, S. O. Kuznetsov, A. Napoli, and S. Duplessis, "Mining gene expression data with pattern structures in formal concept analysis," *Inf. Sci.*, vol. 181, no. 10, pp. 1989–2001, May 2011, doi: 10.1016/j.ins.2010.07.007.

[16] Y. Shi, Y. Mi, J. Li, and W. Liu, "Concurrent concept-cognitive learning model for classification," *Inf. Sci.*, vol. 496, pp. 65–81, Sep. 2019, doi: 10.1016/j.ins.2019.05.009.

[17] Y. Shi, Y. Mi, J. Li, and W. Liu, "Concept-cognitive learning model for incremental concept learning," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 2, pp. 809–821, Feb. 2021, doi: 10.1109/TSMC.2018.2882090.

[18] D. Guo, W. Xu, Y. Qian, and W. Ding, "M-FCCL: Memory-based concept-cognitive learning for dynamic fuzzy data classification and knowledge fusion," *Inf. Fusion*, vol. 100, Dec. 2023, Art. no. 101962, doi: 10.1016/j.inffus.2023.101962.

[19] M. Yan and J. Li, "Knowledge discovery and updating under the evolution of network formal contexts based on three-way decision," *Inf. Sci.*, vol. 601, pp. 18–38, Jul. 2022, doi: 10.1016/j.ins.2022.04.010.

[20] Y. Wan and L. Zou, "An efficient algorithm for decreasing the granularity levels of attributes in formal concept analysis," *IEEE Access*, vol. 7, pp. 11029–11040, 2019, doi: 10.1109/ACCESS.2019.2892016.

[21] F. Hao, G. Pang, Z. Pei, K. Qin, Y. Zhang, and X. Wang, "Virtual machines scheduling in mobile edge computing: A formal concept analysis approach," *IEEE Trans. Sustain. Comput.*, vol. 5, no. 3, pp. 319–328, Jul. 2020, doi: 10.1109/TSUSC.2019.2894136.

[22] J. Gao, F. Hao, Z. Pei, and G. Min, "Learning concept interestingness for identifying key structures from social networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 4, pp. 3220–3232, Oct. 2021, doi: 10.1109/TNSE.2021.3107529.
[23] F. Hao, G. Min, Z. Pei, D.-S. Park, and L. T. Yang, "K-clique community detection in social networks based on formal concept analysis," *IEEE Syst. J.*, vol. 11, no. 1, pp. 250–259, Mar. 2017, doi: 10.1109/JSYST.2015.2433294.
[24] F. Hao, Z. Pei, and L. T. Yang, "Diversified top-k maximal clique detection in social Internet of Things," *Future Gener. Comput. Syst.*, vol. 107, pp. 408–417, Jun. 2020, doi: 10.1016/j.future.2020.02.023.
[25] Y. Yang, F. Hao, B. Pang, G. Min, and Y. Wu, "Dynamic maximal cliques detection and evolution management in social Internet of Things: A formal concept analysis approach," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 3, pp. 1020–1032, May 2022, doi: 10.1109/TNSE.2021.3067939.
[26] R. Wille, "Restructuring lattice theory: An approach based on hierarchies of concepts," in *Ordered Sets*, I. Rival, Ed. Dordrecht, The Netherlands: Springer, Sep. 1982, pp. 445–470.
[27] M. Fan, S. Luo, and J. Li, "Network rule extraction under the network formal context based on three-way decision," *Appl. Intell.*, vol. 53, no. 5, pp. 5126–5145, Jun. 2022, doi: 10.1007/s10489-022-03672-4.
[28] R. Andersen, F. Chung, and K. Lang, "Local graph partitioning using PageRank vectors," in *Proc. 47th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2006, pp. 475–486, doi: 10.1109/FOCS.2006.44.
[29] R. Andersen and K. J. Lang, "Communities from seed sets," in *Proc. 15th Int. Conf. World Wide Web*, May 2006, pp. 223–232, doi: 10.1145/1135777.1135814.
[30] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New J. Phys.*, vol. 11, no. 3, Mar. 2009, Art. no. 033015, doi: 10.1088/1367-2630/11/3/033015.
[31] J. Baumes, M. Goldberg, and M. Magdon-Ismail, "Efficient identification of overlapping communities," *Intell. Secur. Inform.*, vol. 3495, pp. 27–36, May 2005.
[32] J. J. Whang, D. F. Gleich, and I. S. Dhillon, "Overlapping community detection using neighborhood-inflated seed expansion," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 5, pp. 1272–1284, May 2016, doi: 10.1109/TKDE.2016.2518687.
[33] Y. Gao, H. Zhang, and Y. Zhang, "Overlapping community detection based on conductance optimization in large-scale networks," *Phys. A, Stat. Mech. Appl.*, vol. 522, pp. 69–79, May 2019, doi: 10.1016/j.physa.2019.01.142.
[34] R. Kanawati, "Empirical evaluation of applying ensemble methods to ego-centred community identification in complex networks," *Neurocomputing*, vol. 150, pp. 417–427, Feb. 2015, doi: 10.1016/j.neucom.2014.09.042.
[35] R. Zhang, L. Li, C. Bao, L. Zhou, and B. Kong, "The community detection algorithm based on the node clustering coefficient and the edge clustering coefficient," in *Proc. 11th World Congr. Intell. Control Autom.*, Jun. 2014, pp. 3240–3245, doi: 10.1109/WCICA.2014.7053250.
[36] P. Liakos, A. Ntoulas, and A. Delis, "Scalable link community detection: A local dispersion-aware approach," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 716–725, doi: 10.1109/BIGDATA.2016.7840664.
[37] J. Zhu, B. Chen, and Y. Zeng, "Community detection based on modularity and k-plexes," *Inf. Sci.*, vol. 513, pp. 127–142, Mar. 2020, doi: 10.1016/j.ins.2019.10.076.
[38] K. Kloster and D. F. Gleich, "Heat kernel based community detection," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 1386–1395, doi: 10.1145/2623330.2623706.
[39] K. He, Y. Sun, D. Bindel, J. Hopcroft, and Y. Li, "Detecting overlapping communities from local spectral subspaces," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 769–774, doi: 10.1109/ICDM.2015.89.
[40] Y. Yao, W. Wu, M. Lei, and X. Zhang, "Community detection based on variable vertex influence," in *Proc. IEEE 1st Int. Conf. Data Sci. Cyberspace (DSC)*, Jun. 2016, pp. 418–423, doi: 10.1109/DSC.2016.99.
[41] X. You, Y. Ma, and Z. Liu, "A three-stage algorithm on community detection in social networks," *Knowl.-Based Syst.*, vol. 187, Jan. 2020, Art. no. 104822, doi: 10.1016/j.knosys.2019.06.030.
[42] J. X. Yang and X. D. Zhang, "Finding overlapping communities using seed set," *Phys. A, Stat. Mech. Appl.*, vol. 467, pp. 96–106, Feb. 2017, doi: 10.1016/j.physa.2016.10.006.
[43] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, no. 3, Sep. 2007, Art. no. 036106, doi: 10.1103/physreve.76.036106.
[44] Y. Li, K. He, K. Kloster, D. Bindel, and J. Hopcroft, "Local spectral clustering for overlapping community detection," *ACM Trans. Knowl. Discovery Data*, vol. 12, no. 2, pp. 1–27, Jan. 2018, doi: 10.1145/3106370.

SUNQIAN SHI received the B.Sc. degree in mathematics from Tianjin University of Commerce, Tianjin, China, in 2019. She is currently pursuing the M.Sc. degree with the Kunming University of Science and Technology, Kunming, China. Her research interests include community detection, concept-cognitive learning, and formal concept analysis.

MENGYU YAN received the Ph.D. degree in science from Kunming University of Science and Technology, Kunming, China, in 2023. She is currently a Lecturer with Kunming University of Science and Technology. Her research interests include complex networks, concept-cognitive learning, and graph node classification.

JINHAI LI received the M.Sc. degree in science from Guangxi University, Guangxi, China, in 2009, and the Ph.D. degree in science from Xi'an Jiaotong University, Xi'an, China, in 2012. He is currently a Professor with Kunming University of Science and Technology, Kunming, China. Up to now, he has published more than 40 articles in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, and *Pattern Recognition*. His research interests include big data, cognitive computing, granular computing, and concept learning. He is an Area Editor of *International Journal of Approximate Reasoning* and an Associate Editor of *International Journal of Machine Learning and Cybernetics*.