INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 11(4):538-552, August 2016.

Detecting Topic-oriented Overlapping Community Using Hybrid a Hypergraph Model

G.L. Shen, X.P. Yang, J. Sun

Gui-lan Shen*; Xiao-ping Yang
Information School, Renmin University
Beijng, China
guilan.shen@buu.edu.cn; yang@ruc.edu.cn
*Corresponding author: guilan.shen@buu.edu.cn

Jie Sun
Business School, Beijing Union University
Beijing,China
jie.sun@buu.edu.cn

Abstract: A large number of emerging information networks brings new challenges to the overlapping community detection. The meaningful community should be topic-oriented. However, the topology-based methods only reflect the strength of connection, but ignore the consistency of the topics. This paper explores a topic-oriented overlapping community detection method for information work. The method utilizes a hybrid hypergraph model to combine the node content and structure information naturally. Two connections for hyperedge pair, including real connection and virtual connection are defined. A novel hyperedge pair similarity measure is proposed by combining linearly extended common neighbors metric for real connection and incremental fitness for virtual connection. Extensive experiments on two real-world datasets validate our proposed method outperforms other baseline algorithms.

Keywords: information network, overlapping community detection, topic-oriented, hybrid hypergraph model.

1 Introduction

Community is considered to be a fundamental property of complex network. Despite the variety of complex networks, community often accounts for the functionality of the system [1]. Research of recent years shows that the structure of community is not disjoint. Overlapping is an important property of many real-world networks, i.e., they are naturally characterized by multiple community memberships. For example, a person could join in several hobby groups in social networks; one academic paper could cover a number of fields, etc. It is therefore a very essential work to develop approach for efficient overlapping community detection, which will contribute to the links prediction, collaborative recommendation and influence propagation in many application fields.

Although numerous techniques have been developed for overlapping community detection in recent years, most of them only focus on the structure information for real network. It is well understood, however, that there exist a large quantity of real networks with node content or semantic information, which is referred to as information network, such as www, scientific citation network, and social network. The meaningful detected community of the information network should be topic-oriented, which has two characteristics: the nodes inside one community should have dense connections and consistent or similar topics. Communities identified via those topological methods often incorporate different topics since stronger connections represent the interactions that occur across several different topics, which would confuse the meanings of the topic-oriented community [2].

In this paper, we propose a topic-oriented overlapping community method for information network which combines node content information and link information. Firstly, information network is modeled as a hybrid hypergraph composed of hyperedge that features the collection of nodes with common attributes. For different information network, the node attribute could be represented by interest, tab, word or topic. Secondly, a hyperedge pair similarity calculation method is proposed, which combines the content information by calculating common neighbors of hyperedge pair and structure information by measuring the link relationships between the nodes involved into two hyperedges. Finally, an agglomerative hierarchical clustering algorithm is applied to partition the hybrid hypergraph model into different topic-oriented overlapping communities.

Compared with the existing methods, our method can identify communities from the perspective of both content and link structure for information network. From this result, we can easily find more meaningful communities, such as topics, research fields and so forth. Moreover, with the inherent characteristics of hybrid hypergraph model, overlapping of communities could be identified easily.

We proceed to report our work in the rest of the paper as follows. We discuss the related works in Section 2. In section 3, we propose our approach for identifying the meaningful overlapping communities based on hybrid hypergraph model for information network. In order to verify our approach, we conducted extensive experiments. The experimental design and results analysis are given in Section 4. Finally, a conclusion is drawn in Section 5.

2 Related works

Overlapping community detection using topology. Some methods have been proposed to detect overlapping communities in a network.

LFM proposed by Lancichinetti et al [3] is a kind of algorithms utilizing local expansion and optimization. This method relies on a local benefit function that characterizes the quality of a densely connected group of nodes. LFM expands a community from a random seed node to form a natural community until the fitness function

$$f(c) = \frac{k_{in}^c}{(k_{in}^c + k_{out}^c)^{\alpha}} \tag{1}$$

is locally maximal, where k_{in}^c and k_{out}^c are the total internal and external degree of the community c, and α is the resolution parameter controlling the size of the communities. After detecting one community, LFM randomly selects another node not assigned to any community to expand another new community. This method obviously can identify the overlapping community, since they allow a single node to be put into different community owing to different optimization process.

In real-world network, it's difficult to decide how many communities that a single node should be put in, however it's very clear whether the edge incident on the node is in the community or not. Now, researcher suggests using links to defining community [4], owing to an edge only is in one community, but the nodes connected by edge could be put into different communities. Some methods [5–7] using line graph and link partitioning to detect overlapping community have been proposed. Among them, Ahn [6] partitions links into clusters via hierarchical clustering of edge similarity. Given a pair of links e_{ik} and e_{jk} incident on a node k, the edge pair similarity can be computed via the Jarracd Index defined as,

$$sim(e_{ik}, e_{jk}) = \frac{|Nb_{+}(i) \cap Nb_{+}(j)|}{|Nb_{+}(i) \cup Nb_{+}(j)|} \tag{2}$$

Where $Nb_{+}(i)$ is the inclusive neighbors of a node i, which the set contains the node itself and its neighbors. With this similarity, single-linkage hierarchical clustering is then used to build a link dendrogram. Cutting this dendrogram at a special threshold yields link communities. Although the link partitioning method can detect the overlapping communities naturally, there is no guarantee that is provides high quality detection for information network because the method only relies on links of the network while ignores the node content totally.

There are many other methods to detect overlapping community. For example, the ones based on subgraphs, such as CPM [8], CPMw [9] etc.al, treat community structure as the composition of adjacent subgraphs, as one node can belong to several subgraphs. However, these methods are usually considered to solve the pattern matching of complex networks rather than finding communities. In addition, the methods extended Girvan and Newman's divisive clustering algorithm [10], such as CONGA [11], CONCO [12], allow a node to split into multiple copies.

Despite the use of different techniques, the above methods can always detect overlapping dense connections in network. However, they only focus on the topology information but ignore the content information that contributes to improve the quality of the community [13].

Topic-oriented community detection using topology and content. Based on the assumption that the content information can improve the quality of the detected community, various approaches have been combined the links and contents for community detection. Some approaches have combined content information with structure information for community discovery. One of them is generative probabilistic modeling which considers both contents and links as being dependent on one or more latent variables, and then estimates the conditional distributions to find community assignments. PLSA-PHITS [4], Community-User-Topic model [15] and PMC [16] are three representatives in this category. Other fusing the content and structure methods, such as SA-clustering [17] via augmenting the underlying network to take into account the content information, heuristic algorithm CKC [18] to solve the connected k-Center problem, subspace clustering algorithm [19] on graphs with feature vectors.

Different from those methods using topology, these methods account for the content information of the nodes, so the division results for the network is more cohesive in the topics. However, considering the content of nodes, the complexity of the algorithm is greatly increased which will lead to some new challenges, such as how to deal with the high dimensional sparse for node attributes. Furthermore, those methods are not designed for overlapping communities.

3 Methodology

In this section, we present our method for fusing structure and content via hybrid hypergraph to detect the overlapping communities for information networks. Firstly, we present the definitions, and then introduce how to build the hybrid hypergraph model for information network. Thirdly, we give the method that how to measure hyperedge pairs similarity. Finally, we briefly introduce algorithmic details of HLP (Hyper Link Partition), a novel method extended from link graph partition algorithm.

3.1 Definitions

Link partitioning method is a kind of topology methodology based on classic graph theory. It's very simple and distinct to model information networks as simple graphs, in which nodes indicate entity object, and links indicate the binary relationships between node pairs. However, real information networks are characterized by node content attribute, hence simple graph is not suitable for representing the content information. As the generalization of simple graph, hypergraph can represent the multiple relationships for nodes in finite set, and describe the relationship between general discrete structures, which overcome the defect of the knowledge represented by the simple graph. Hypergraph characterized by one hyperedge incident on any number of nodes, is a graph in generalization. The definition of hypergraph is provided as follows.

Definition 1. A hypergraph [20], $H = (V, E)$ is defined as a set of vertices $V = \{v_1, v_2, \cdots, v_n\}$, and a set of hyperedges $E = \{e_1, e_2, \cdots, e_m\}$, where:

(1) $e_1 \neq \phi (i = 1, 2, \cdots, m)$

(2) $\bigcup\limits_{i=1}^{m} e_i = V$

According to definition, a hyperedge essentially is the set of vertices which are independent. That is, hypergraph cannot represent the original topology of vertices. Thus, we give the definition of hybrid hypergraph.

Definition 2. A Hybrid Hypergraph, $HH = (V, E, \varepsilon, \psi)$ is defined as a set of vertices $V = \{v_1, v_2, \cdots, v_n\}$, a set of hyperedges $E = \{e_1, e_2, \cdots, e_m\}$ and a set of edges ε where:

(1) $e_i \neq \emptyset \; (i = 1, 2, \cdots, m)$

(2) $\bigcup\limits_{i=1}^{m} e_i = V$

(3) $\psi(\varepsilon_i) = (v_i, v_j) \; (i, j = 1, 2, \cdots, n)$

3.2 Modeling information network as hybrid hypergraph

We need to extend a structural graph with tuples describing node attributes. This can be formally expressed as a quad $AG = \{V, \varepsilon, F_V, \psi\}$, where each node v is associated with a feature vector $f(v)$. F_V is the set of features for all nodes, where, $f(v) \subseteq F_V, v \in V$. Feature selection is an important issue in system anomaly detection applications [21]. With different node attributes in different information network, the feature vector $f(v)$ can be varied as a topic, a keyword, a place, an author, an activity. The number of features of F_V is m, formally, $m = \left| \bigcup\limits_{v \in V}^{f(v)} \right|$.

So, the question of how to build the information network as the hybrid hypergraph model is simplified as how to map a quad attribute graph AG into a quad HH. Here, we take each feature f_i as a basic unit to build hyperedge e, when $f_i \in f(v)$, the node $v \in e$

We use incidence matrix and adjacent matrix to represent the data structures with related to hybrid hypergraph.

The $N_v \times M_E$ Incidence matrix for a hybrid hypergraph HH, say I, is defined as that

$$I_{ve} = \begin{cases} 1, & if \; v \in e \\ 0, & otherwise \end{cases} \tag{3}$$

The $M_E \times N_V$ transposed matrix for I, say K, is defined as that

$$K_{ev} = \begin{cases} 1, & if \; v \in e \\ 0, & otherwise \end{cases} \tag{4}$$

The $N_V \times N_V$ adjacent matrix for a hybrid hypergraph HH, say A, is defined as that

$$A_{ij} = \begin{cases} 1, & if \; (i, j) \in \varepsilon \\ 0, & otherwise \end{cases} \tag{5}$$

The $M_E \times M_E$ similarity matrix for a hybrid hypergraph HH, say Sim, is defined as that

$$Sim_{ij} = simlarity(e_i, e_j) \;\; e_i, e_j \in E \tag{6}$$

3.3 Similarity for hyperedge pairs

In this section, we present the method of how to calculate the similarity of hyperedge pairs in HH. This technique was originally introduced by link partitioning algorithm and local expansion algorithm for the purposes of identifying the overlapping communities in network. However, Link partitioning algorithm and local expansion algorithm both only focus on topological information, but ignore the content information.

We argue the HH model for information network can fuse the content information and structure information naturally. According to the definition of HH, a hyperedge can be regarded as a sub-community characterized by a special feature, because the hyperedge is a set of nodes with the same feature. We know clearly that the sub-community represented by the hyperedge is different from the final detected community. The reason is that one node is usually associated with more than one feature, but a hyperedge only reflect one of the features. Nevertheless, we can determine whether the hyperedge pairs have the same topic or not by exploring the link relationship between them. If the node only has one feature, moreover, such as topic, our work can be simplified as the work presented by zhao [22].

The relationship between two hyperedges is the link between two node sets which is more complicated, rather than just that between the two nodes in the simple graph. The link type for two hyperedges comes into two kinds: one kind is shared common nodes; the other is that the nodes in one hyperedge are connected with those in the other hyperedge. In order to better illustrate this, we offer two formal definitions as follows.

Definition 3. Given two hyperedge e_i and e_j, is real connection, where $e_i \cap e_j \neq \emptyset$.

For instance, two hyperedges e_1, e_2 are shared with two common nodes v_2, v_3 (Fig1.a), that is $e_1 \cap e_2 = \{v_2, v_3\} \neq \emptyset$, therefore the link type of e_1, e_2 is real connection.

Definition 4. Given two hyperedge e_i and e_j, is virtual connection, where

(1) $e_i \cap e_j \neq \emptyset$

(2) $\forall \varepsilon \in \{< v_m, v_n > |\exists v_m \in e_i, \exists v_n \in e_j\} \neq \emptyset$

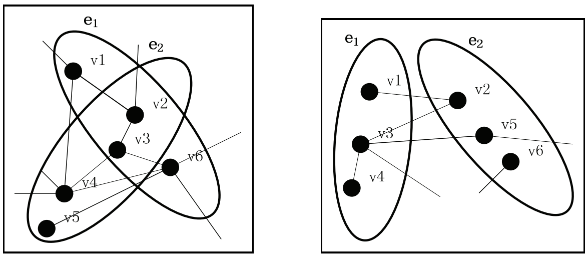
Or

(3) $e_i \cap e_j \neq \emptyset$

(4) $\forall \varepsilon \in \{< v_m, v_n > |\exists v_m \in e_i - e_i \cap e_j, \exists v_n \in e_j - e_i \cap e_j\} \neq \emptyset$

For instance, as shown in Fig1.b $e_1 \cap e_2 = \emptyset$, and, there are links $< v_1, v_2 >$, $< v_2, v_3 >$ between nodes for e_1 and e_2, this is one case of virtual connection. Another case is also shown in Fig1.a, $e_1 \cap e_2 = \{v_2, v_3\} = \emptyset$, considering edge $< v_4, v_6 >$, which match the conditions (4) $v_4 \in e_1 - e_2 \cap e_2$ and $v_6 \in e_2 - e_1 \cap e_2$. Therefore, hyperedge e_1 and e_2 has both real connection and virtual connection.

Obviously, when measuring the hyperedge similarity, both virtual connection and real connection are meaningful. Virtual connection reflects the structure information of the original network essentially, while real connection reflects the number of nodes with common attributes. Intuitively, hyperedge similarity is dependent on the tightness of hyperedge pairs. As mentioned above, hyperedge is the group of nodes; therefore, measuring the tightness of hyperedge pair can be converted into how to measure the tightness of two groups of nodes. If there are a lot of links between two sets of nodes, the two sets are strongly tight. For instance, in citation network, if article a1 with keyword k1 cited article a2 with keyword k2, then k1 and k2 have certain correlation or similarity. Similarly, the more articles involving the keyword k1 are cited by the articles involving k2 keyword, that is, k1 set has strong tightness with k2 set, which turns out that k1 and k2 have higher correlation.



(a) Real connection (b) Virtual connection

Figure 1: Two kinds of link type for hyperedge

To better quantify the tightness between the two groups of nodes, we extend the local fitness function in LFM [3]. Local fitness function for a given local community S, is formally given as

$$f_s = \frac{k_{in}^s}{k_{in}^s + k_{out}^s} \tag{7}$$

Where k_{in}^s and k_{out}^s are the total internal and external degree of the community.

Fitness function f_s can measure the internal and external tightness for a local community. In topology detecting community methods, it has shown a good performance, and was expanded or applied by more researchers [23,24]. Inspired by this thought, we propose the fitness function for a hyperedge e to be given as

$$f_e = \frac{k_{in}^e}{k_{in}^e + k_{out}^e} \tag{8}$$

Where k_{in}^e and k_{out}^e are the total internal and external degree of the hyperedge. The total of k_{in}^e and k_{out}^e is the total degree of all nodes in this hyperedge. Similar to f_s, f_e can measure the internal and external tightness for hyperedge. Whether to merge two hyperedge into a larger hyperedge depends on how the changed numbers of virtual connection influence the fitness of the combined hyperedge. We define incremental fitness for combined hyperedge as the similarity for virtual connection of two hyperedges. Given two hyperedge e_i and e_j

$$Sim_{virtual}(e_i, e_j) = \Delta f_{ij} = f_{e_i} \cup f_{e_i} = \frac{k_{in}^{e_ie_j}}{k_{in}^{e_ie_j} + k_{out}^{e_ie_j}} \tag{9}$$

Where $k_{in}^{e_ie_j}$ is the numbers of virtual connection between e_i and e_j. The total of $k_{in}^{e_ie_j}$ and $k_{out}^{e_ie_j}$ is the total degree of all nodes in merged hyperedge. For instance, in Fig2, considering two hyperedge e_1 and e_4, $k_{in}^{e_1e_4} = 4$, $k_{in}^{e_1e_4} + k_{out}^{e_1e_4} = 32$, so, $\Delta f_{14} = f_{e_1} \cup f_{e4} = \frac{1}{8}$. Another example in this figure is, $k_{in}^{e3e4} = 3$ and$k_{in}^{e3e4} + k_{out}^{e3e4} = 41$, so, $\Delta f_{34} = f_{e_3} \cup f_{e4} = \frac{3}{41}$.

Whether to merge two hyperedges into a larger hyperedge depends on the value of incremental fitness function.

Virtual connection reflects the tightness of hyperedge pair via the structure information, while, real connection reflects the semantic similarity of hyperedge via the content information. In our method, we do not directly calculate the similarity of the features implied by hyperedges, such as that of keyword or topic, instead, we evaluate the similarity via CN metric. CN(Common Neighbors) [25] is also called structural equivalence, namely, the nodes are similar if they share a lot of common neighbors. CN is one of the most widely used metrics when measuring the similarity in local community detection methods. Therefore, we extend the CN metric to make it suitable for measuring the similarity between two hyperedges, or, two groups of nodes. To define clearly the neighbors set for hyperedges, we give the following definitions.

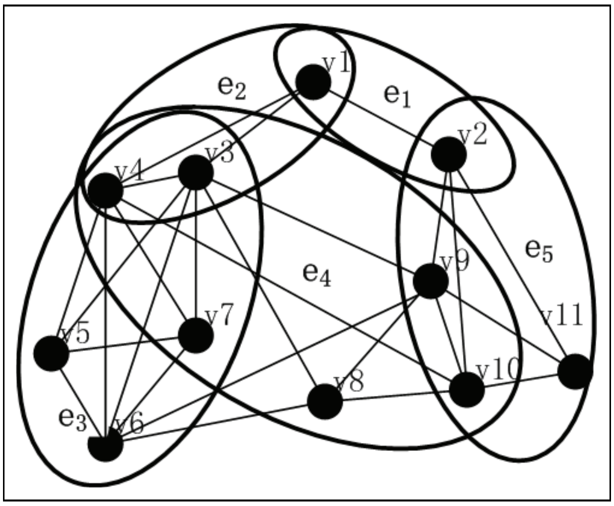


Figure 2: A sample of hybrid hypergraph model

Definition 5 (Inductive nodes set)**.** Given two hyperedge e_i, e_j in HH, where e_i, e_j has real connection, the Inductive nodes set for e_i related to e_j, say $p_j(e_i)$ is formally as $p_j(e_i) = \{v|v \in e_i - e_i \cap e_j\}$.

In order to illustrate the neighbors for hyperedges, we specially designate those neighbors related to Inductive nodes set.

Definition 6 (Extended Common Neighbors of Inductive nodes set)**.** Given two hyperedge e_i, e_j and inductive nodes set $p_j(e_i)$ in HH, the extended neighbors of $p_j(e_i)$ including $p_j(e_i)$, say $n_+(p_j(e_i))$,is formally as $n_+(p_j(e_i)) = \{x|d(v,x) \leq 1, v \in p_j(e_i)\}$

Where $d(x,y)$ is the distance of two nodes, formally as follows:

$$d(x,y) \leq 1 \ if \ x \in e_i, y \in e_j \ and \ e_i \cap e_j \neq \emptyset \tag{10}$$

In Fig2, $e_1 \cap e_2 = \{v_1\}$, $p_2(e_1) = \{v_2\}$, $p_1(e_2) = \{v_3, v_4\}$, so we can calculate the extended common neighbors of inductive nodes sets $n_+(p_1(e_2)) = \{v_1, v_3, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}\}$, $n_+(p_2(e_1)) = \{v_1, v_2, v_9, v_{10}, v_{11}\}$.

Based on Jaccard index, we propose the method that how to calculate the similarity for hyperedge pair e_i, e_j with real connection. Formally as:

$$Sim_{real}(e_i, e_j) = \frac{|n_+(p_j(e_i)) \cap n_+(p_i(e_j))|}{|n_+(p_j(e_i)) \cup n_+(p_i(e_j))|} \tag{11}$$

In fact, this definition is consistent with the similarity for the link pair in Ahn[14_6] research. For instance, in Fig2, $s(e_1, e_2) = \frac{3}{11}$. We combine the Jarracd index and fitness function to compute the similarity for hyperedge pair e_i, e_j, formally as:

$$Sim_{ij} = Sim(e_i, e_j) = \lambda Sim_{real}(e_i, e_j) + (1 - \lambda) Sim_{virtual}(e_i, e_j) \tag{12}$$

We give the detail of how to compute the similarity for Hyperedge pair in Algorithm 1.

3.4 Stop criterion

In our method, we adopt divisive hierarchical clustering to cluster the hyperedges. The results of hierarchical clustering are presented in a dendrogram. One important job in this method is

```
Algorithm 1: Computing the similarity for hyperedge pair
Input: $e_1$, $e_2$, $K$
Output: simValue
$j = 0$, $f = 0$
check $e_1$ and $e_2$ is whether real connection or virtual connection
If ($e_1 \cap e_2 \neq \emptyset$)
   //real connection
   for each $v \in V$
   if ($v \in e_1$ & &$v \in e_2$)
      $vComm \leftarrow v$
      //get extended neighbors of Inductive nodes set
   $NIV_1 \leftarrow veNeighbors(K, e_1, vComm)$
   $NIV_2 \leftarrow veNeighbors(K, e_2, vComm)$
If $NIV_1 \neq \emptyset$ && $NIV_2 \neq \emptyset$
   interNum=interSection($NIV_1, NIV_2$)
   unionNum=unionSection($NIV_1, NIV_2$)
   Jaccard = interNum / unionNum
   //fitness
   fin=blink($e_1, e_2$)
   ftotal=totalDegree($e_1$)+ totalDegree($e_2$)
   f=fin/ftotal
else
   fin=blink($e_1$, $e_2$)
   ftotal=totalDegree($e_1$)+ totalDegree($e_2$)
   f=fin/ftotal
   simValue = $\lambda$*Jaccard + (1 - $\lambda$)*f
```

Figure 3: Algorithm Similarity

deciding the stop criterion for clustering. We define partition density D, as a function of the dendrogram cut threshold. The maximum of D indicates the discovered hyperedge communities are well structured.

For a HH, $\{HP_1, HP_2, \cdots, HP_K\}$ is a partition of the hyperedges into K clusters. Cluster HP_K has m_K hyperedges, and $n_K = \left| \bigcup\limits_{e_i \in HP_K} \{v \in e_i\} \right|$ nodes. Then we define this as normalized form:

$$D_K = \frac{m_{HP_K} - 1}{m_K - 1} \tag{13}$$

Where, $m_{HP_k} = \sum\limits_{i=1}^{n_K} m_{HP_K}^i \Big/ n$ is the mean value for the nodes located in a number of hyperedges. $m_{HP_K}^i$ is the number of hyperedges in which node i is located. The larger D_k indicates the probability, at which nodes are clustering into one cluster, the tighter internal connection in the cluster is. The partition density D, is the average of D_k.

$$D = \sum_{i=1}^{K} D_i \Big/ K \tag{14}$$

The goal of hierarchical clustering is to find K clusters when partition density D is maximal. When D = 1, all clusters are merged into one community.

3.5 HLP algorithm

Based on the HH model for information network, we apply the previously-defined similarity algorithm to pairs of all hyperedges, and produce a similarity matrix S. In the process of clustering, the clustered hyperedge contain more than one initial hyperedge. For instance, $e_{i'}$ has n initial hyperedges, $e_{j'}$ has m initial hyperedges. The Jarracd similarity of $e_{i'}$, $e_{j'}$ is computed as:

$$Sim_{real}(e_{i'}, e'_{j}) = \sum_{t}^{n} \sum_{c}^{m} Sim_{real}(e_i^c, e_j^t) \Bigg/ n * m \tag{15}$$

We give a simple description of HLP in Algorithm 2. We obtain the clusters of merged hyperedges via this algorithm. The nodes located in each clustered hyperedges constitute the communities related to some special topics.

```
Algorithm2: Clustering for hyperedges
Input: $K$
Output: $K'$
$S \leftarrow 0$, $m = 0$, $n = 0$, $K' \leftarrow K$
$m$ = LengthofRow($K$)
$n$ = LengthofCol($K$)
//Initialize the similarity matrix for $K$
for(int $i = 0$; $i < m$; $i{+}{+}$)
for(int $j = 0$; $j < n$; $j{+}{+}$)
    $S(i, j)$ = Similarity$(i, j)$
    $S' \leftarrow S$
While $m > 1$
//find the coordinates for the maximal similarity in $S'$
    max$_i$ = maxI($S'$)
    $max_j$ = maxJ($S'$)
    $K' \leftarrow$ clustering($K'$(, max$_i$), K(, max$_j$))
If Density($K'$) = maxDensity($K'$)
    return $K'$
```

Figure 4: Algorithm HLP

4 Experiments

In this section, we present experiments on real datasets to evaluate the performance of our method. We first applied our method to two datasets to choose the optimal Value of λ. Then we compared the performance of our method with two baseline methods. Before going to details, we first describe the datasets and the method to extract the node feature, and introduce the performance metric to be used in our experiments.

4.1 Datasets

Two real datasets used in our experiments are described in the following:

Cora Dataset: The Cora dataset [26] consists of the abstracts and references of about 34000 computer science research papers. Three subfields of Machine Learning (ML), Programming (PL) and Database (DB) are used and those articles without references to other articles in the set are removed. The detailed information about each subfield is shown in Table1.

Table 1: Cora dataset

Area	***#of subfields***	***#of total papers***	***#of used papers***	***#of links***
ML	7	4218	2708	5249
PL	9	4496	3292	7772
DB	9	1396	1060	2522

WebKB Dataset: The WebKB dataset [27] consists of about 6000 web pages from computer science departments of four schools. The web pages are classified into seven categories, including, staff(SA), course(CS),department(DP), faculty(FA), student(SD),project(PJ),other(OT). We select five categories excluding DP and OT, in that the two categories contain only a few pages. The detailed information about these web pages is shown in Table 2.

Table 2: WebKB dataset

School	***#of CS***	***#of FA***	***#of SD***	***#of PJ***	***#of SA***	***#of Total***	***#of links***
CN	44	34	18	128	21	245	304
TX	36	46	20	148	2	252	328
WA	77	30	18	123	10	258	446
WI	85	38	25	156	12	316	530

4.2 Preprocessing

For Cora, we treated each subfield as an independent dataset. After stemming and removing stopwords we were left with vocabulary of stemmed unique words for each subfield respectively. All words with document frequency less than 15 were removed. Taking DB dataset as an example, a vocabulary of size 5911 unique words was reduced to size 1086 words by removing the low document frequency words. The word cloud for preprocessed DB dataset is shown in Fig.5

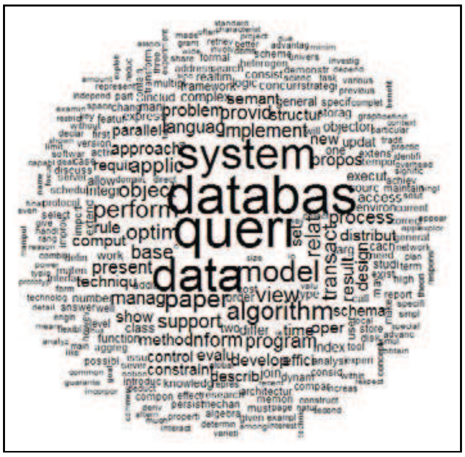


Figure 5: Word cloud for DB dataset

The same preprocessing is handled by WebKB. We extracted words whose document frequency is more than 10.

4.3 Evaluation metrics

In our paper, we focus on the topic similarity of the detected overlapping communities, therefore when the ground-truth community is known, we utilize two measures of purity and NMI to evaluate the quality of overlapping communities detected by different methods. Purity measures the internal topic similarity within the community, and NMI is the most widely used measure to account for overlapping communities.

Given the ground-truth community structure, $G = \{G_1, G_2, \cdots, G_S\}$ where G_S contains the set of nodes that are in the sth community. The community structure given by the algorithms is represented by $C = \{C_1, C_2, \cdots, C_S\}$, where C_k contains the set of nodes that are in the kth community.

The purity of C_i is defined as:

$$Purity(C_i) = \frac{1}{|C_i|} \max_j \{C_i \cap G_j\} \tag{16}$$

Usually, the detected community C_i includes nodes that belong to other G_j in the ground-truth. For C_i, we compute the intersection set with each standard community G_j, then take the maximum as the final purity for it.

The purity of C is defined as:

$$Purity(C) = \frac{1}{K} \sum_{i=1}^{K} Purity(C_i) \tag{17}$$

The higher the purity, the better the communities are partitioned from the perspective of topics.

The mutual information between G and C is defined as

$$MI(G, C) = \sum_{x \in G, y \in C} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{18}$$

The NMI(normalized mutual information) is defined by

$$NMI(G, C) = \frac{MI(G, C)}{\max(H(G), H(C))} \tag{19}$$

where $H(G)$ and $H(C)$ are the entropies of the partitions G and C. The higher the NMI, the closer the partition is to the ground truth.

4.4 Optimal value of λ

As we discussed in section 3, the parameter λ balances the Jaccard index and fitness function value when compute the similarity of hyperedge pairs. We perform experiments to study how the λ value affects the purity of detected communities. We set the step 0.1, the result is shown in Figure 4.

The result shows that λ value is decided by the structure of network and the number of hyperedges. Jaccard index and fitness function value both affect the purity of the detected communities. It is proved that the structure and content information both have influence on the topical community detection. However, we observe that the characteristics of information networks detemine the value of parameter λ . Different network will lead to different λ . This is

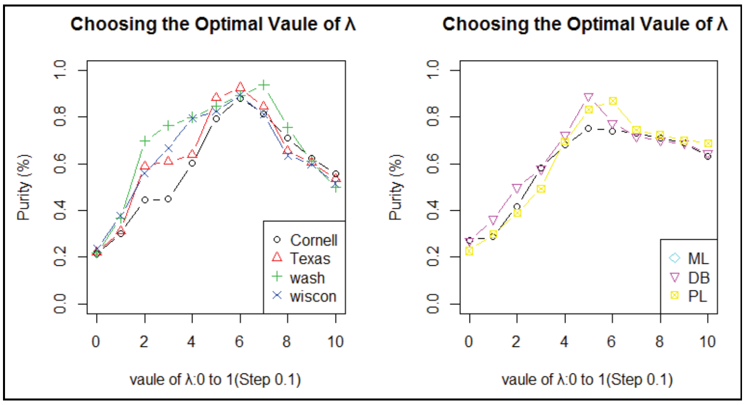


Figure 6: The result of choosing parameter λ

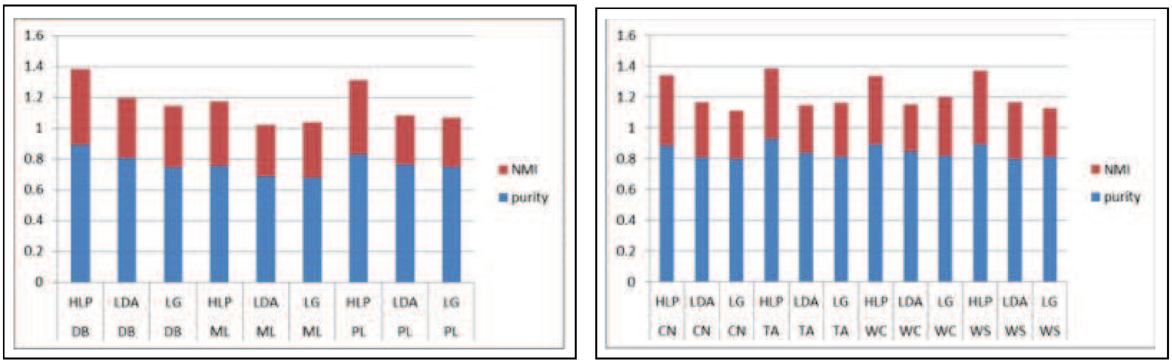
because there are noise problems to various degree, including both links and node content in the information network. For WebKB dataset,the best performace is achieved when "$\lambda = 0.6$", and for Cora dataset, the optimal value "$\lambda = 0.5$" which are used as default settings for the following experiments.

4.5 Results

To evaluate the effectiveness of HLP, we compare our method with two baseline methods: one is topology-based method, Line graph partition [6], the other is LDA to cluster the nodes by using content information only.

We use purity and NMI quantifying the performance each algorithm.

The details are shown in Figure 5, which illustrates HLP achieve the best performance in real information networks. From the results, we also can observe some interesting things. In some datasets, such as DB,PL,CN and WS, LDA algorithm can achieve better performance than Line Graph algorithm, In some other datasets, nevertheless, such as ML,TA and WC, the results are opposite. This confirms our assumption again that both node information and link information affect the quality of detected overlapping communities. Therefore, we are sure that the combination of node information and link information can improve the quality of overlapping community detection.



(a) Cora dataset

(b) WebKB dataset

Figure 7: The evaluations of community algorithms over two real information networks.

Conclusion

In this paper, we propose a topic-oriented overlapping community detection approach based on hierarchical clustering for hybrid hypergraph model, which can combine the content and structure information of information network naturally. Considering the complex of the hybrid hypergraph model, we classify the connections of hyperedges into real connection related to content information, and virtual connection related to structure information. We present incremental fitness to evaluate the tightness for hyperedge pairs in virtual connection. Meanwhile, we extend CN metric on hyperedge pairs to conduct the semantic similarity calculation in real connection. In order to balance the influence of two connections, we combine linearly the two measures for similarity of hyperedge pairs. The density function is employed to determine the appropriate number of communities. To evaluate the performance, we conducted experiments on two real datasets. Compared with the benchmark, Line graph partition algorithm focusing on topological detection, LDA focusing on clustering node contents, our approach gained a better performance in information network. Furthermore, the overlapping communities detected by our approach were more meaningful since they are topic-oriented.

Our approach has many potential applications. It can be applied to many kinds of information networks, where nodes contain content. With our method detecting the communities, we are able to improve the efficiency of collaborative scientific research, discover experts for each topic, and analyze topic-oriented influence propagation.

Future work includes qualifying the weight of each node in the hyperedge to improve the purity of detected communities. We also intend to take the time factor into account, so that we can detect the evolution communities.

Acknowledgment

This paper is supported by Natural Science Foundation of China (No.71572015), Scientific Research Project of Beijing Union University (No. Zk10201506), Beijing Higher Education Young Elite Teacher Project (No.YETP1503).

Bibliography

[1] Cobanoglu B, Zengin A, Ekiz H, et al (2014); Implementation of DEVS Based Distributed Network Simulator for Large-Scale Networks [J]. *International Journal of Simulation Modelling* (IJSIMM), 13(2): 147-158.

[2] Ding Y. (2011), Community detection: topological vs. topical, *Journal of Informetrics*, DOI: 10.1016/j.joi.2011.02.006, 5(4): 498-514.

[3] Lancichinetti, Andrea, Santo Fortunato, and János Kertész (2009); Detecting the overlapping and hierarchical community structure in complex networks, *New Journal of Physics*, 11(3): 033015.

[4] Xie J., Kelley S., Szymanski B. K. (2013), Overlapping community detection in networks: The state-of-the-art and comparative study, *ACM Computing Surveys* (CSUR), 45(4): 43-79.

[5] Evans T. S., Lambiotte R. (2009), Line graphs, link partitions, and overlapping communities, *Physical Review E*, DOI:http://dx.doi.org/10.1103/PhysRevE.80.016105, 8(1): 92-105.

[6] Ahn Y. Y., Bagrow J. P., Lehmann S. (2010), Link communities reveal multiscale complexity in networks, *Nature*, doi:10.1038/nature09182, 466: 761-764.

[7] He, C., Ma, H., Kang, S., Cui, R. (2014), An Overlapping Community Detection Algorithm Based on Link Clustering in Complex Networks, *In Military Communications Conference (MILCOM) IEEE*, 865-870.

[8] Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005), Uncovering the overlapping community structure of complex networks in nature and society, *Nature*, 435: 814-818.

[9] Farkas, I., Ábel, D., Palla, G., & Vicsek, T. (2007), Weighted network modules, *New Journal of Physics*, 9: 80-198.

[10] Girvan M., Newman M. E. J. (2002), Community structure in social and biological networks, *Proc. of the National Academy of Sciences*, 99: 7821-7826.

[11] Gregory S. (2007), An algorithm to find overlapping community structure in networks, Knowledge discovery in databases: PKDD 2007, *Springer Berlin Heidelberg*, 91-102.

[12] Gregory S. (2008), A fast algorithm to find overlapping communities in networks, Machine learning and knowledge discovery in databases, *Springer Berlin Heidelberg*, 408-423.

[13] T. Yang, R. Jin, Y. Chi, S. Zhu (2009), Combining link and content for community detection: a discriminative approach, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris*, 927-936.

[14] Hric D., Darst R. K., Fortunato S. (2014), Community detection in networks: structural clusters versus ground truth, *arXiv preprint* arXiv:1406.0146.

[15] Hofmann, David Cohn Thomas (2001), The missing link-a probabilistic model of document content and hypertext connectivity, *Proceedings of the 2000 Conference on Advances in Neural Information Processing Systems*, Vancouver, 430-436.

[16] D. Zhou, E. Manavoglu, J. Li, C. Giles, and H. Zha (2006), Probabilistic models for discovering e-communities, *In Proceedings of the 15th international conference on World Wide Web*, Banff, 173-182.

[17] Yang, B., Di, J., Liu, J., Liu, D. (2013), Hierarchical community detection with applications to real-world network analysis, *Data & Knowledge Engineering*, 83: 20-38.

[18] M. Ester, R. Ge, B. Gao, Z. Hu, and B. Ben-Moshe (2006), Joint cluster analysis of attribute data and relationship data: the connected k-center problem, *Proceedings of the 2006 SIAM International Conference on Data Mining*, Maryland, USA, 25-46.

[19] Günnemann S, B. Boden, and T. Seidl (2011), Db-csc: a density-based approach forsubspace clustering in graphs with feature vectors, Machine Learning and Knowledge Discovery in Databases, *Springer Berlin Heidelberg*, 565-580.

[20] Berge, Claude (1989), *Hypergraphs: Combinatorics of Finite Sets*, North Holland.

[21] Zhou X. et al.(2014); Information-value-based feature selection algorithm for anomaly detection over data streams [J]. *Tehnički vjesnik*, 21: 223-232.

[22] Zhao, Z., Feng, S., Wang, Q., Huang, J. Z., Williams, G. J., Fan, J. (2012), Topic oriented community detection through social objects and link analysis in social networks, *Knowledge-Based Systems*, 26: 164-173.

[23] Darst R. K., Nussinov Z., Fortunato S. (2014), Improving the performance of algorithms to find communities in networks, *Physical Review E*, 89(3): 42-58.

[24] McAuley J., Leskovec J. (2014), Discovering social circles in ego networks, *ACM Transactions on Knowledge Discovery from Data* (TKDD), 8(1): 10-16.

[25] Lorrain F., White H. C. (1971) , Structural equivalence of individuals in social networks, *The Journal of mathematical sociology,*1: 49-80.

[26] Rayid Ghani (2014), CMU World Wide Knowledge Base (WebKB) project, Jan, 2001.[Online]. Available: http://www.cs.cmu.edu/∼webkb. [Accessed: April 9, 2014]