

# Density-based rough set model for hesitant node clustering in overlapping community detection

Jun Wang<sup>1,\*</sup>, Jiaxu Peng<sup>1</sup>, and Ou Liu<sup>2</sup>

1. School of Economics and Management, Beihang University, Beijing 100191, China;

2. School of Accounting and Finance, The Hong Kong Polytechnic University, Hong Kong 999077, China

**Abstract:** Overlapping community detection in a network is a challenging issue which attracts lots of attention in recent years. A notion of hesitant node (HN) is proposed. An HN contacts with multiple communities while the communications are not strong or even accidental, thus the HN holds an implicit community structure. However, HNs are not rare in the real world network. It is important to identify them because they can be efficient hubs which form the overlapping portions of communities or simple attached nodes to some communities. Current approaches have difficulties in identifying and clustering HNs. A density-based rough set model (DBRSM) is proposed by combining the virtue of density-based algorithms and rough set models. It incorporates the macro perspective of the community structure of the whole network and the micro perspective of the local information held by HNs, which would facilitate the further “growth” of HNs in community. We offer a theoretical support for this model from the point of strength of the trust path. The experiments on the real-world and synthetic datasets show the practical significance of analyzing and clustering the HNs based on DBRSM. Besides, the clustering based on DBRSM promotes the modularity optimization.

**Keywords:** density-based rough set model (DBRSM), overlapping community detection, rough set, hesitant node (HN), trust path.

**DOI:** 10.1109/JSEE.2014.00125

## 1. Introduction

Insensibly but rapidly, huge data networks are forming in a wide variety of fields ranging from bio-engineering databases to the state information center or social network services, thus one of the primary data mining techniques is to divide a large data network into communities [1], which attracts researchers bound for clustering. In this paper, we regard the “cluster” the same as “community”. Both traditional clustering methods such as  $k$ -means clustering [2] and modern methods including the greedy technique in modularity maximum [3] have been applied to community

detection in network.

Furthermore, most actual networks are made of highly overlapping cohesive groups of nodes, instead of separated communities [4]. Thus, the study on overlapping community detection brings about fresh batches of approaches.

Baumes et al. proposed a method to find the overlapping community [5] which absorbs advantages of two efficient heuristics: the iterative scan (IS) and the rank removal (RaRe). A different method, the clique percolation method [4] is one of the most popular techniques which have been extended to the analysis of weighted, directed and bipartite graphs [6]. However, it has limitation when dealing graphs with just a few cliques. Huang et al. proposed a clustering algorithm called DenShrink [7]. By combining the advantages of density-based clustering [8–11] and modularity optimization methods [3,12], DenShrink efficiently reveals the embedded hierarchical and overlapping community structure in large-scale weighted undirected networks and identifies hubs and outliers. And unlike the traditional density-based clustering methods, it is parameter-free. However, we find that it has difficulty in dealing with a hesitant node (HN).

The HN is a new notion proposed in this paper, which is defined in Section 2.

An HN contacts with multiple communities, just as a hub. However, unlike the active and central hub [13,14], the connections are not necessarily strong and sometimes even weak and accidental. Thus some HNs between communities are not qualified hubs. Besides, HNs are not rare to find, such as new entrants, inactive members in organizations [15,16] and low frequently requested websites or commodities [17–19]. Actually, they might be like the long tail, which are just beginning to show their power [18]. Theoretical analysis and experimental results show that an HN might grow up to be a connector among multiple communities, and then becomes a hub; or it tends to be merged into a certain community, and then it is an attached

Manuscript received October 24, 2013.

\*Corresponding author.

This work was supported by the National Natural Science Foundation of China (71271018).

node. Misjudging their roles or simply leaving them alone provokes undesirable effects on the “growth” of HNs and leads to inaccurate analysis of the stability of the community structure. Fig. 1 shows the possible development trend of HNs. We use the Venn diagram to present the inclusion relation among the set of HNs, the attached node, the hub and the bridge [20]. The bridge is a special kind of hub.

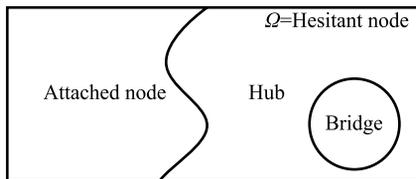


Fig. 1 Possible inner trend of HNs

In this paper, a density-based rough set model (DBRSM) is established for HN clustering after essential comparison between DBSCAN [9] and DenShrink and flexible application of rough set models [21]. On one hand, DBRSM inherits the virtue of DenShrink algorithms which are able to obtain the community structure of the whole graph; on the other hand, it combines the local information held by HNs with rough set models when refining the clustering process of DenShrink. The whole process can be divided into two steps. First, qualified hubs including bridges are detected. Then we use the neighborhood information of HNs, namely the membership degree, rather than merely the linkage of their own to cluster them into communities.

In this paper, we theoretically prove that DBRSM facilitates the “growth” of attached nodes in perspective of the strength of the trust path [22] which represents the strength of information dissemination from communities to HNs. Besides, our experiments on the real-world and synthetic datasets show that DBRSM can also promote modularity maximization.

The remainder of the paper is structured as follows. Section 2 is the theoretical foundation including related definitions and algorithms. Section 3 introduces the details of the proposed DBRSM. Section 4 presents the experiments, which shows the efficiency of our method in HNs clustering and modularity optimization. Finally, we give a conclusion and provide future research directions in Section 5.

## 2. Theoretical foundation

### 2.1 Definition of hesitant node

**Definition 1** (Hesitant node) Let  $G = \langle V, E, \omega \rangle$  denote a weighted undirected network.  $V$  is the set of nodes in network.  $E$  is the set of edges connecting any two nodes that communicate with each other.  $\omega(e)$  is the weight of

any edge  $e \in E$ . A node  $h \in V$  is called a hesitant node in  $G$  if it satisfies the following two properties:

- (i) The node  $h$  contacts with multiple communities;
- (ii) The similarity between node  $h$  and any of its adjacent community is at a low level that node  $h$  could not be clustered into multiple communities with increasing modularity.

The definition of the HN summarizes the feature of nodes which show the implicit community structure at some level. It is even similar to the main character discussed in the long tail theory [18] or the power laws [23]. But here, our research background is complex network. Note that the “hubs” detected by the DenShrink algorithm have the property listed in Definition 1, thus they are HNs as well and then gotten to be the object of this study.

### 2.2 Essential comparison of DBSCAN and DenShrink

It can be observed that not all HNs are qualified hubs. The basic reason is found through the essential comparison of DBSCAN and DenShrink.

DBSCAN is one of the most famous density-based clustering approaches which have been widely used in data mining for their ability of finding clusters of arbitrary shapes even with the nodes arbitrarily distributed. However, its effectiveness is limited, since the values of necessary input parameters have significant impact on the clustering outcome while they are usually difficult to determine [1]. Besides, it is unable to detect overlapping communities.

DenShrink, a functional extension to traditional density-based clustering approaches, overcomes those two weaknesses with the parameter free clustering process. But it fails to deal with HNs properly.

According to the process of DBSCAN, any two nodes within a cluster have symmetric relation i.e., density-connected. And the density-connected relation is based on an asymmetric relation, i.e., density-reachable. The asymmetric relation is flexible since the strength of it can be adjusted by changing the input parameter, i.e., Eps, maximum radius of the neighborhood.

The DenShrink algorithm is the iteration of two phrases. The first phrase is to find all the micro-communities (MCs). The MC is an isolated node or a sub-graph that consists of one or more connected dense pairs. Secondly, the MC whose merge increases the modularity is merged and regarded as a super-node in the following iteration. The process of iteration stops when there is no merge of the MC that increases the modularity. Finally, the MC which contains more than one node is regarded as the community, while the MC consists of an isolated node as the hub. Compared with DBSCAN, nodes within an MC also

hold symmetric relation with each other. However, unlike the flexible relation defined in the DBSCAN, the relation of nodes in the MC extends from the relation of nodes in the dense pair [7], i.e., a rigorous equivalence relation. The dense pair is a pair of nodes with the largest similarity from each other. Two nodes form a dense pair only when they have high similarity with each other which is not less than their surrounding links. That is to say, without flexible input parameters, DenShrink is so rigorous that it misses the clustering of HNs which always keep low similarity with other nodes. What's more, the DenShrink regards the HNs as hubs. It is not accurate. Because some of the HNs connecting with multi-community remain independent in the last iteration of DenShrink for keeping low similarity with all their adjacent nodes, but not for their role of real com-

munity junction.

Fig. 2 shows an example of community detection results given by the DenShrink algorithm. The tagged value on the edge is the structural similarity [24] between two adjacent nodes. Three communities are detected by the DenShrink algorithm because the mergence of super-nodes in  $MC_1$  which is consisted of  $Hub_2$ ,  $Community_1$  and  $Community_2$  reduces the modularity. Nodes 1, 2, 11, 12 are treated as hubs. However, if considering the measurement of centrality degree [25,26] and predicted trust of the information distribution path [14,22,27] through the node, we will see nodes 1 and 11 are less qualified hubs compared with nodes 2 and 12. So, DenShrink has difficulties in distinguishing qualified hubs from attached nodes.

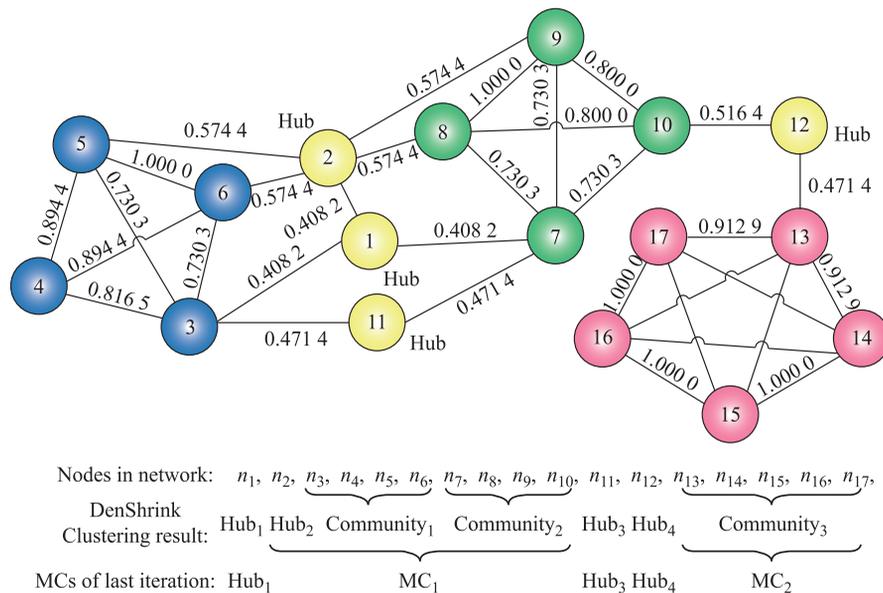


Fig. 2 Example of communities detected by DenShrink

### 2.3 Related definitions based on rough set theory

The classical rough set theory, first proposed by Pawlak [21], attracted great attention for its fundamental role in data classification and rule extraction problems. Later, some extension versions of the rough set theory based on various kinds of binary indiscernibility relations are proposed. To use the rough set model, the precondition is the definition of binary indiscernibility relations. In this paper, a density based tolerance relation is proposed in the following Definition 2, which is based on the notion of tolerance relation [28] in the rough set theory and adjusted to the application background of network analysis.

**Definition 2** (Density based tolerance relation) Let  $G = \langle V, E, \omega \rangle$  denote a weighted undirected network. A binary density based tolerance relation  $T$  is defined as any

relation between two nodes that form a dense pair. Let us define the density based tolerance relation more precisely:

$$T = \{(u, v) | u \leftrightarrow_{\varepsilon} v, u \in V, v \in V\}. \quad (1)$$

Any  $(u, v) \in T$  can also be denoted as  $uTv$ . Wherein,  $u \leftrightarrow_{\varepsilon} v$  represents that  $u$  and  $v$  form a dense pair with the similarity  $\varepsilon$ . It can be observed that tolerance relation has reflexivity and symmetry but not transitivity.

**Definition 3** (Density based tolerance class) Let  $G = \langle V, E, \omega \rangle$  denote a weighted undirected network. For any  $u \in V$ , let  $T(u)$  denote the set  $\{v \in V | uTv\}$ .  $T(u)$  is the density based tolerance class of  $u$ . Particularly, for any  $u \in V$ , there is  $u \in T(u)$ .

In the background of social network, the set  $T(u)$  consists of members in the network which have dense relation

with  $u$ .

**Definition 4** (Upper approximation) Let  $G = \langle V, E, \omega \rangle$  denote a weighted undirected network. For any subset  $V' \subseteq V$ , let  $\overline{V'}$  denote the set  $\cup\{T(u)|u \in V'\}$ .  $\overline{V'}$  is the upper approximation of  $V'$ . For any  $V' \subseteq V$ , there is  $V' \subseteq \overline{V'}$ .

The definitions in this section form a theoretical foundation for our proposed model.

### 3. Proposed DBRSM

#### 3.1 Quantization basis—membership degree

DBRSM is able to cluster the attached nodes among the set of HNs. The quantization basis for the clustering of an attached node is the membership degree of the attached node in its adjacent communities which is computed as the following.

Let  $a_i$  denote an attached node in network  $G$ .  $C_j$  is one of  $a_i$ 's adjacent communities. Let  $M(i, j)$  denote the membership degree of  $a_i$  in  $C_j$ .  $M(i, j)$  is defined as

$$M(i, j) = \sum_{v_k \in C_j \wedge \sigma(i, k) \neq 0} \sigma(i, k) \times |T(v_k)| \quad (2)$$

where  $v_k$  is one of the adjacent nodes of  $a_i$  in community  $C_j$ ;  $\sigma(i, k)$  is the similarity between  $a_i$  and  $v_k$ ;  $T(v_k)$  is the density based tolerance class of node  $v_k$ ;  $|T(v_k)|$  is the cardinality of the set  $T(v_k)$  which is used to amplify the contribution of  $v_k$  in disseminating information to  $a_i$ . The membership degree is defined with the standpoint that the more dense neighbours the  $v_k$  has, the more reliable and attractive it is for the attached node  $a_i$ .

Then the membership degree is the weighted sum of  $|T(v_k)|$  with the weight  $\sigma(i, k)$ .  $T(v_k)$  is the tolerance class of node  $v_k$  which belongs to the neighborhood of  $a_i$  and the adjacent community  $C_j$ . And  $M(i, j)$  largely embodies the influence of  $a_i$ 's adjacent nodes in  $C_j$  on  $a_i$ .

#### 3.2 Theoretical support for DBRSM—the strength of the trust path

Membership degree  $M(i, j)$  represents the influence of  $C_j$  on  $a_i$  in significant measure. We prove this point in perspective of the strength of the information dissemination path from communities to attached nodes.

Community  $C_j$  attracts attached node  $a_i$  by spreading information to it. We quantify the efficiency of information dissemination by the level of trust [14,22,27] between a source user in  $C_j$  and the target user  $a_i$ . A weighted mean aggregation method has been proposed to compute the strength of the trust path [29–31]. In this section, we use this method to compute the level of trust. Besides, the shorter and stronger the trust paths are, the more important

they are for predicting the level of trust [31]. So we only consider paths starting from nodes within the following set  $S$ :

$$S = \overline{\tau(a_i)} \quad (3)$$

where the set  $\tau(a_i)$  is the structure neighborhood of node  $a_i$  containing  $a_i$  itself and its adjacent nodes:  $\tau(a_i) = \{v \in V|(a_i, v) \in E\} \cup \{a_i\}$ .  $S = \overline{\tau(a_i)}$  is the upper approximation of  $\tau(a_i)$ . In the network,  $S$  is composed of all the nodes which keep dense relation with certain one of  $a_i$ 's neighborhoods. With the premise above, if we mark the source node with  $u_s$ , there is  $u_s \in S$ .

Thus we get (4) to compute the strength of the trust path from  $C_j$  to  $a_i$ , which is denoted by  $P(C_j, a_i)$ . And it is a modification of the weighted mean aggregation method to adapt the premise  $u_s \in S$ .

$$P(C_j, a_i) = \sum_{u_s \in \overline{\tau(a_i)}} P(u_s, a_i) \quad (4)$$

where  $P(u_s, a_i)$  is the strength of trust of the path from  $u_s$  to  $a_i$ . The following is going to introduce the computation of  $P(u_s, a_i)$ .

The source user  $u_s$  in set  $S$  is only one or two steps away from the target user  $a_i$ . We analyze the calculation of  $P(u_s, a_i)$  under the two conditions:

(i) If  $u_s \in \tau(a_i)$ , namely the source user  $u_s$  is one step away from the target user  $a_i$ , we have

$$P(u_s, a_i) = \sigma(u_s, a_i). \quad (5)$$

(ii) If  $u_s \in S - \tau(a_i)$ , namely, the source user  $u_s$  is two steps away from the target user  $a_i$ , we have

$$P(u_s, a_i) = \frac{\sum_{u_k \in \tau(u_s)} \sigma(u_s, u_k) \sigma(u_k, a_i)}{\sum_{u_k \in \tau(u_s)} \sigma(u_s, u_k)}. \quad (6)$$

The only contributions to the combination  $\sum_{u_k \in \tau(u_s)} \sigma(u_s, u_k) \sigma(u_k, a_i)$  come from nodes  $u_k \in \tau(a_i)$ . Besides, the similarities between  $u_s$  and each  $u_k \in \tau(u_s)$  are the same, since any  $u_s \in \tau(a_i)$  and any  $u_k \in \tau(u_s) \cap \tau(a_i)$  form a dense pair. Then we simplify (6) as follows:

$$P(u_s, a_i) = \sum_{u_k \in \tau(u_s) \cap \tau(a_i)} \sigma(u_k, a_i). \quad (7)$$

Now we get the formula for  $P(u_s, a_i)$ , the estimation of the strength of the trust path from  $u_s$  to  $a_i$ . Plug the expression of  $P(u_s, a_i)$  in (7) and (5) into (4) and group these terms together, we get the final expression of  $P(C_j, a_i)$  as

$$P(C_j, a_i) = \sum_{v_k \in C_j \wedge \sigma(i, k) \neq 0} \sigma(i, k) |T(v_k)|. \quad (8)$$

Thus  $P(C_j, a_i) = M(i, j)$  when we estimate the efficiency of information distribution of  $C_j$  to  $a_i$  based on the trust path from node  $u_s$ , belonging to set  $S$  to  $a_i$ . That is, when we adopt more desirable paths to evaluate the level of trust, as discussed above, we can use the membership

degree  $M(i, j)$  to represent the influence of  $C_j$  on  $a_i$ . In conclusion, the higher the membership degree of  $a_i$  in  $C_j$  is, the more reasonable it is to cluster the attached node  $a_i$  into the community  $C_j$ . Fig. 3 represents the calculation of  $P(C_j, a_i)$  in different linkage structures.

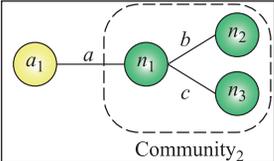
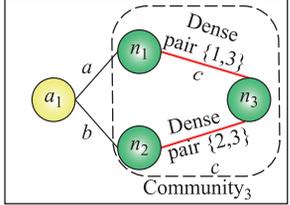
Linkage structure	$P(u_s, a_1)$	$P(C_j, a_1)$	$M(C_j, a_1)$
$u_s \in \tau(a_1)$ 	$P(n_1, a_1) = \sigma(n_1, a_1) = a$	$a$	$a$
$u_s \in S - \tau(a_1)$ 	$P(n_1, a_1) = \sigma(n_1, a_1) = a$ $P(n_2, a_1) = \frac{\sigma(n_1, a_1) \times \sigma(n_1, a_2)}{\sigma(n_1, a_2)} = a$ $P(n_3, a_1) = \frac{\sigma(n_1, a_1) \times \sigma(n_1, a_3)}{\sigma(n_1, a_3)} = a$	$3a$	$3a$
	$P(n_1, a_1) = \sigma(n_1, a_1) = a$ $P(n_2, a_1) = \sigma(n_2, a_1) = b$ $P(n_3, a_1) =$ $\frac{\sigma(n_1, a_1) \times \sigma(n_1, n_3) + \sigma(n_2, a_1) \times \sigma(n_2, n_3)}{\sigma(n_1, n_3) + \sigma(n_2, a_3)} = a + b$	$2a + 2b$	$2a + 2b$

Fig. 3 Calculation of  $P(C_j, a_i)$  in different linkage structures

### 3.3 Clustering of HNs

In the DBRSM model, the HNs are divided into three concrete types: the bridge, the hub and the attached nodes based on the clustering result of DenShrink. The potential hubs and bridges are scouted firstly; then the rest of the HNs, namely the attached nodes, are clustered based on the membership degree into communities which are detected by DenShrink. The processing procedure for HNs is listed as below.

As Section 2 mentioned, nodes which are left unclustered (they are the isolated points in the clustering result) and linking to multiple communities in the clustering result are the HNs. For each HN  $h_i$ , we set the following steps to make the concrete analysis.

**Step 1** Scout hubs. If NH  $h_i$  belongs to more than two tolerance classes (note that  $h_i \in T(h_i)$ , i.e.,  $h_i$  is included in a certain MC and is linking with multiple super-nodes with dense relation), then it is regarded as the hub since it keeps dense relation with multiple communities. In overlapping community detection, it belongs to the overlapping portion and is classified into its adjacent communities which have dense relation with it.

**Step 2** If HN  $h_i$  belongs to only one or two tolerance classes (note that  $h_i \in T(h_i)$ , i.e.,  $h_i$  and no more than one super-node form a dense pair), we cluster  $h_i$  under the

following two constraint conditions:

**Step 2.1** Scout bridges. If  $h_i$  is a bridge linking multiple communities which are connected only with the existence of the junction  $h_i$ , then  $h_i$  is shared by the corresponding adjacent communities. Therefore,  $h_i$  is a bridge (a special kind of hub) although its communications with the neighbor communities are not dense.

**Step 2.2** Cluster attached nodes. If  $h_i$  is not a bridge, then  $h_i$  is not a qualified hub. Instead, it is an attached node. That is because although it connects with multiple communities, the strength is not strong enough and it is not the necessary node of the communication of communities. Then we cluster  $h_i$  into the community in which  $h_i$  gets the largest membership degree.

The following pseudo code represents the process of DBRSM.

**Step 1** Scout hubs

```

for each  $C \in M\_C \wedge |C| > 1$ 
  for each  $v_1 \in C \wedge |v_i| = 1 \wedge degree(v_i) > 1$ 
    for each  $v_j \in C \wedge |v_j| > 1 \wedge \sigma(v_i, v_j) > 0$ 
       $v_j \leftarrow v_j \cup v_i;$ 
       $H \leftarrow H \cup v_i;$ 
    end for
  end for
end for
    
```

**Step 2**

```

for each  $C_i \in CR$ 
  if  $|C_i| = 1$  then
    Step 2.1: scout bridges;
    Step 2.2: classify attached nodes;
  end if
end for
return  $CR, H$ ;

```

**Step 2.1 Scout bridges**

```

for each  $C_j \in CR \wedge \sigma(C_i, C_j) > 0$ 
  for each  $C_k \in CR \wedge \sigma(C_k, C_i) > 0$ 
    if  $\sigma(C_k, C_j) = 0$ 
       $C_j \leftarrow C_i \cup C_j$ ;
       $C_k \leftarrow C_i \cup C_k$ ;
       $H \leftarrow H \cup C_i$ ;
    end if
  end for
end for

```

**Step 2.2 Classify attached nodes**

```

if  $C_i$  is not a bridge;
  for each  $C_j \in CR \wedge \sigma(C_i, C_j) > 0$ 
    compute  $M(i, j)$ ;
  end for
   $k == \text{get}_j(\max M(i, j))$ ;
   $C_k \leftarrow C_i \cup C_k$ ;
end if

```

where  $H$  is the set of hubs. In Step 1,  $M\_C$  denotes the set of MCs in the last iteration of DenShrink.  $C$  is a micro-community belonging to MC. Each  $v \in C$  is a super-node which is a set consisting of one or more nodes. In Step 2, the  $CR$  is the set of communities detected by DenShrink.  $C_i \in CR$  is a community which can also be regarded as a super-node.  $\sigma(C_i, C_j)$  is the similarity between community  $C_i$  and  $C_j$ . The computation can be found in [7]. In Step 2.2,  $\text{get}_j(\max M(i, j))$  represents the value of  $j$  that maximizes the  $M(i, j)$ .

**3.4 Analysis of computational complexity**

In Step 1, we scout hubs directly from the independent nodes in micro-community detected by DenShrink, and then the running time on scouting hubs is linear with the amount of HNs. In Step 2.1 of scouting bridges, the computational complexity for each HN is  $O(D^2(h))$  where  $D(h)$  is the degree of the HN, no more than  $D(G)$ , the degree of network  $G$ , since we should judge whether any two of an HN's adjacent communities are still connected without the HN. In Step 2.2, we could calculate the membership degree for each HN with the computational complexity  $O(D(h))$  by employing the intermediate result of

DenShrink—the dense pair. In conclusion, the overall time complexity for HN clustering is  $O(h \cdot D^2(G))$  if there are  $h$  HNs in network  $G$ .

**4. Experiments**

In this section, we apply the DBRSM to the real world datasets and computer-generated data. The experiment result is compared with DenShrink, which is able to detect overlapping community and identify two kinds of special nodes: hubs and outliers.

**4.1 Evaluation on real world networks****4.1.1 Books about US politics**

Valdis Krebs compiled 105 books on US politics which were sold online by Amazon [32]. Each of them is assigned as “liberal”, “neutral” or “conservative” according to the book review. Thus they could be organized into three categories. While from the microcosmic point of view, these books possess a much richer community structure.

As shown in Fig. 4, the different political attitudes “liberal”, “neutral” and “conservative” are represented by hexagon, circular and triangle respectively. Fifteen subclasses detected by DBRSM are represented by different colors. They are subdivision of the three categories of books. Besides, three kinds of special nodes are detected. They are 3 bridges, 14 attached nodes and 2 outliers marked with square, diamond and upside-down triangle respectively.

However, DenShrink treats attached nodes and bridges as hubs, thinking both of them play significant roles in the contact among multiple communities and ignoring the clustering of attached nodes. This would lead to the loss of some valuable information in the real world.

Let us take a look at attached node 28 for example. It represents the book “All the Shah's Men” which stands on the neutral side. Nevertheless, there are still readers who brought this book and another liberal or conservative book at the same time. In our model, we figure out that it is more close to books on the liberal side. From this discovery, we detect an opportunity from the neutral to the liberal. In the case of other attached nodes, we also find the trends developing from the neutral to the conservative or from the liberal to the neutral. This kind of discovery might do a favor to study the political trend presented in the books and buyers.

Furthermore, we figure out that each clustering process of the 14 attached nodes contributes to the gain of modularity and the total contribution is up to 22.78%.

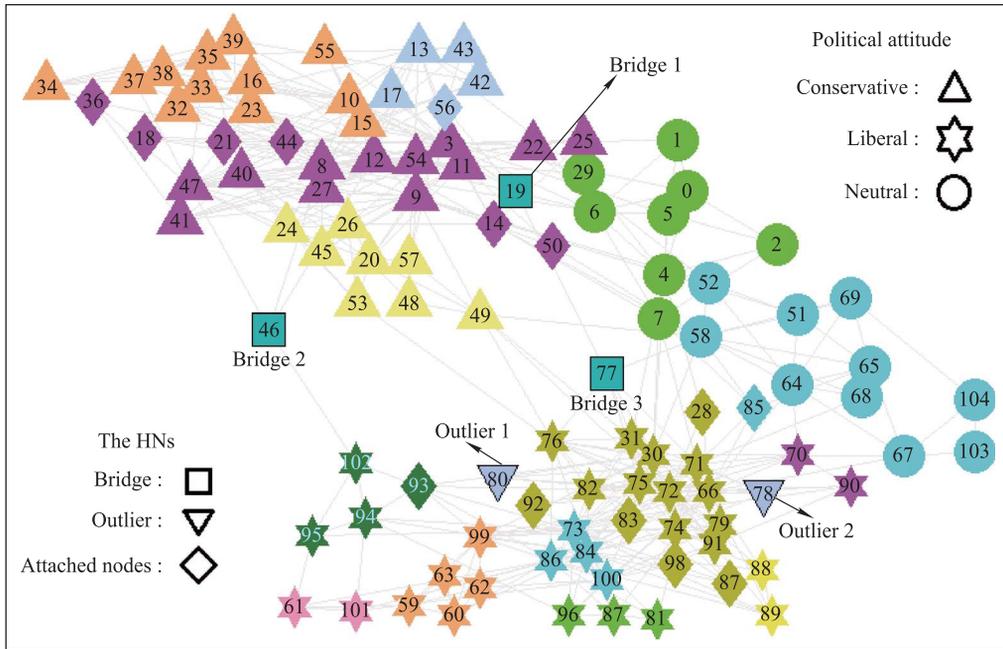


Fig. 4 Experiment on books about US politics

4.1.2 Zachary’s karate club

Zachary’s karate club [33] contains the network of friendships among 34 members in a karate club. And the network is built based on an observation lasting for three years. Afterwards, a conflict between the club president and the instructor then led to the break of the club into two separate groups [6]: one is the set of nodes at the top right of the bold straight line, the other is the set of nodes at the lower

left side.

Various clustering algorithms have been tested in this classical benchmark graph. However, results of previous algorithms did not express the cause of the break clearly, while the application of DBRSM reveals some underlying factors.

As Fig. 5 shows, DBRSM clusters the whole club into five communities marked with different colors.

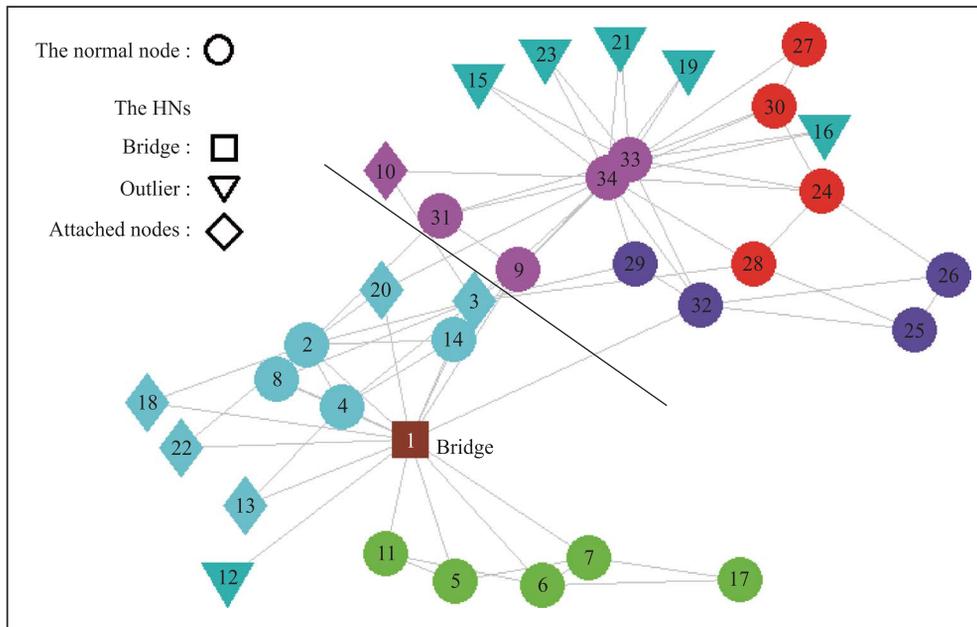


Fig. 5 Experiment on Zachary’s karate club

In Fig. 5 one hub is detected and marked with reddish brown square. Six outliers and six attached nodes are denoted with upside-down triangles and diamonds respectively. Attached nodes are colored the same with the community they are clustered into by DBRSM. We can see that all attached nodes are clustered correctly into the groups which they finally support in real world.

Besides, we find that there is no qualified hub between any two communities coming from the two divided groups. That is a sign of fission. If knowing this earlier, we might try to build a qualified hub deliberately and prevent the conflict. However, in DenShrink, regarding node 10 and node 20 as hubs would lead to a overly optimistic viewpoint that communities are well connected. Besides, the clustering of the six attached nodes increases the modularity by 6.41%. The increase of modularity here is relatively low because the low number (only six) of attached nodes.

## 4.2 Experiment on synthetic data

We also apply the DBRSM algorithm to synthetic data. They are six 2-dimension spatial spherical clusters generated based on Gaussian distribution with the following means of each cluster:  $\mu_1 = (1, 1)$ ,  $\mu_2 = (3, 2)$ ,  $\mu_3 = (1, 3)$ ,  $\mu_4 = (2.5, 4.5)$ ,  $\mu_5 = (4, 4)$ ,  $\mu_6 = (4.5, 1.0)$ . Their covariance matrices are:

$$\delta_1 = \dots = \delta_6 = \begin{bmatrix} 0.3^2 & 0 \\ 0 & 0.3^2 \end{bmatrix}.$$

DenShrink tends to cluster the random data into many small communities and leave behind many HNs not being clustered. Then DBRSM helps to detect the relatively close communities of the attached nodes among the HNs. To evaluate the performance of DBRSM in community detection of different size, we use multiple sets of spherical clusters. The size of each cluster in different test sets ranges from 30 to 150 with the step 30. We list the clustering results in Table 1.

**Table 1** Community detection result of different sizes of test sets

Test set	Cluster size	Average community size by DenShrink	Percentage of attached nodes detected by DBRSM/%	Percentage of hubs detected by DBRSM/%	Modularity gain by clustering attached nodes/%
1	30	11.8	6.7	1.10	21.68
2	60	9.0	7.2	1.70	31.65
3	90	8.3	8.0	1.30	40.27
4	120	9.1	6.0	0.69	33.92
5	150	9.7	6.7	0.89	39.61

In Table 1, we can see that HNs devote a stable percentage of the statistical sample while DenShrink tends to miss the clustering of them. Though the amount is not large, clustering them by DBRSM gains big improvement of modularity, which would promote the stability of the community structure.

In summary, the experiments on both real world and synthetic datasets show that HNs clustering gets practical significance and DBRSM performs properly in the management of HNs and is also able to improve the modularity optimization compared with DenShrink.

## 5. Conclusions

In overlapping communities, HNs get realistic meaning. Some of them are potential hubs; some are likely to be addicted by a certain community. Their community structure is often implicit thus current approaches have difficulties in clustering HNs.

In this paper, we propose a DBRSM for the clustering of HNs by combining the density-based clustering algorithm with the rough set theory. It adds flexibility to the density-base algorithm for overlapping community detec-

tion and takes advantage of the rough set theory to make use of the local information denoted by HNs. The strength of the trust path proves its rationality in theory and our experiments show that DBRSM also promotes modularity maximization.

In the future, the compatibility of the clustering method DBRSM with the optimization of modularity requires more investigation, which will promote the application of DBRSM in complex networks analysis. Besides, one of the basic notions in DBRSM, the neighborhood of HN, can be extended and become more flexible.

## References

- [1] L. Duan, L. Xu, F. Guo, et al. A local-density based spatial clustering algorithm with noise. *Information Systems*, 2007, 32(7): 978–986.
- [2] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967: 281–297.
- [3] M. E. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 2004, 69(6): 066133.
- [4] G. Palla, I. Derényi, I. Farkas, et al. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, 435(7043): 814–818.
- [5] J. Baumes, M. Goldberg, M. Magdon-Ismail. Finding commu-

- nities by clustering a graph into overlapping subgraphs. *IADIS AC*, 2005, 5: 97–104.
- [6] S. Fortunato. Community detection in graphs. *Physics Reports*, 2010, 486(3): 75–174.
- [7] J. Huang, H. Sun, J. Han, et al. Density-based shrinkage for revealing hierarchical and overlapping community structure in networks. *Physica A: Statistical Mechanics and Its Applications*, 2011, 390(11): 2160–2171.
- [8] C. Guo, Y. Zang. Clustering algorithm based on density function and niche PSO. *Journal of Systems Engineering and Electronics*, 2012, 23(3): 445–452.
- [9] M. Ester, H. P. Kriegel, J. Sander, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996: 226–231.
- [10] M. Ankerst, M. M. Breunig, H. P. Kriegel, et al. OPTICS: ordering points to identify the clustering structure. *ACM SIGMOD Record*, 1999, 28(2): 49–60.
- [11] X. Xu, N. Yuruk, Z. Feng, et al. SCAN: a structural clustering algorithm for networks. *Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007: 824–833.
- [12] M. E. Newman, M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 2004, 69(2): 026113.
- [13] S. Gelareh, S. Nickel. Hub location problems in transportation networks. *Transportation Research Part E: Logistics and Transportation Review*, 2011, 47(6): 1092–1111.
- [14] A. Nocera, D. Ursino. PHIS: a system for scouting potential hubs and for favoring their “growth” in a social internetworking scenario. *Knowledge-Based Systems*, 2012.
- [15] S. H. Askarabadi, R. Valizadeh, M. Zarghami. Comparison amount of depression, anxiety and obsession between active and inactive men students of university. *Procedia-Social and Behavioral Sciences*, 2011, 30: 2401–2404.
- [16] S. Merler, M. Ajelli. Human mobility and population heterogeneity in the spread of an epidemic. *Procedia Computer Science*, 2010, 1(1): 2237–2244.
- [17] C. Kumar, J. B. Norris, Y. Sun. Location and time do matter: a long tail study of website requests. *Decision Support Systems*, 2009, 47(4): 500–507.
- [18] C. Anderson, M. P. Andersson. *The long tail*. Stockholm: Bonnier Fakta, 2007.
- [19] A. Enders, H. Hungenberg, H. P. Denker, et al. The long tail of social networking: Revenue models of social networking sites. *European Management Journal*, 2008, 26(3): 199–211.
- [20] B. H. Chou, E. Suzuki. *Discovering community-oriented roles of nodes in a social network*, in *data warehousing and knowledge discovery*. Berlin: Springer Verlag, 2010: 52–64.
- [21] Z. Pawlak. Rough sets. *International Journal of Computer & Information Sciences*, 1982, 11(5): 341–356.
- [22] Y. Kim, H. S. Song. Strategies for predicting local trust based on trust propagation in social networks. *Knowledge-Based Systems*, 2011, 24(8): 1360–1371.
- [23] M. E. Newman. Power laws, pareto distributions and Zipf’s law. *Contemporary Physics*, 2005, 46(5): 323–351.
- [24] E. Leicht, P. Holme, M. Newman. Vertex similarity in networks. *Physical Review E*, 2006, 73(2): 026120.
- [25] S. P. Borgatti. Centrality and network flow. *Social Networks*, 2005, 27(1): 55–71.
- [26] D. Chen, L. Lü, M. S. Shang, et al. Identifying influential nodes in complex networks. *Physica A: Statistical Mechanics and Its Applications*, 2012, 391(4): 1777–1787.
- [27] G. Yan, T. Zhou, B. Hu, et al. Efficient routing on complex networks. *Physical Review E*, 2006, 73(4): 046108.
- [28] M. Kryszkiewicz. Rough set approach to incomplete information systems. *Information Sciences*, 1998, 112(1): 39–49.
- [29] J. Golbeck, J. Hendler. *Engineering knowledge in the age of the semantic web*. Berlin: Springer Verlag, 2004: 116–131.
- [30] T. D. Huynh, N. R. Jennings, N. Shadbolt. Developing an integrated trust and reputation model for open multi-agent systems. *Proc. of the 7th International Workshop on Trust in Agent Societies*, 2004: 65–74.
- [31] J. A. Golbeck. *Computing and applying trust in web-based social networks*. Washington, USA: University of Maryland, 2005.
- [32] V. Krebs. Books about US politics. <http://www.orgnet.com/>.
- [33] W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 1977, 33(4): 452–473.

## Biographies



**Jun Wang** was born in 1969. He is currently a professor in the Department of Information Systems, Beihang University, Beijing, China. His current research interests include knowledge management, knowledge systems engineering and business intelligence.  
E-mail: king.wang@buaa.edu.cn



**Jiaxu Peng** was born in 1992. She is a graduate student of Beihang University. Her research interests include knowledge management, data mining and complex networks.  
E-mail: peng.jia.xu@gmail.com



**Ou Liu** was born in 1976. He is an assistant professor of the Hong Kong Polytechnic University. His research interests include business intelligence, virtual communities, knowledge management, ontology engineering and evolutionary computation.  
E-mail: affiuou@polyu.edu.hk