Community Detection in Incomplete Information Networks

Wangqun Lin
National University of Defense Technology
Changsha, China
linwangqun2005@gmail.com

Xiangnan Kong
University of Illinois at Chicago
Chicago, Illinois
xkong4@uic.edu

Philip S. Yu
University of Illinois at Chicago
Chicago, Illinois
psyu@uic.edu

Quanyuan Wu
National University of Defense Technology
Changsha, China
quanyuanwu@nudt.edu.cn

Yan Jia
National University of Defense Technology
Changsha, China
yanjia@nudt.edu.cn

Chuan Li
Sichuan University
Chengdu, China
lcharles@scu.edu.cn

ABSTRACT

With the recent advances in information networks, the problem of community detection has attracted much attention in the last decade. While network community detection has been ubiquitous, the task of collecting complete network data remains challenging in many real-world applications. Usually the collected network is incomplete with most of the edges missing. Commonly, in such networks, all nodes with attributes are available while only the edges within a few local regions of the network can be observed. In this paper, we study the problem of detecting communities in incomplete information networks with missing edges. We first learn a distance metric to reproduce the link-based distance between nodes from the observed edges in the local information regions. We then use the learned distance metric to estimate the distance between any pair of nodes in the network. A hierarchical clustering approach is proposed to detect communities within the incomplete information networks. Empirical studies on real-world information networks demonstrate that our proposed method can effectively detect community structures within incomplete information networks.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Application-Data Mining

General Terms

Algorithms, Experimentation

Keywords

Community detection, incomplete information networks, distance metric learning

1. INTRODUCTION

Information networks arise naturally in a wide range of domains. Examples include biological networks, publication networks and social networks. In these networks, feature vectors are usually available which are associated with nodes. Links represent relationships between the nodes. Identifying communities in information networks is a crucial step to understand the network structures. The community is defined as a group of nodes which are densely connected inside the group, while loosely connected with the nodes outside the group.

Community detection in network data has been extensively studied in the literature [17, 19, 3, 18, 2, 21, 14]. Conventional approaches focus on detecting communities based upon linkage information. They assume that the complete linkage information within the entire network is available. However, in many real-world networks, such as terrorist-attack information networks, the complete linkages are very difficult or even impossible to obtain. Instead, the complete linkage information is only available within a few small local regions. We notice that a similar problem has also been studied in [13]. However, in this paper, we focus on incomplete information networks with local information regions. For example, in work relation networks, it is usually impossible to obtain the complete linkage information among all the people. But usually we can afford to obtain the work relationships within a small number of local regions, such as groups or organizations. These networks are called *incomplete information networks* in this paper. The local regions with complete linkage information are called *local information regions*. An incomplete information network with local information regions is shown in the upper left level of Figure 1. Some real-world examples for community detection in incomplete information networks are listed as follows:

- **Terrorist-attack network**. Let us consider a terrorist attack activity networks within a period in a certain country. Each node in the network represents a terrorist activity. Terrorist attacks committed by the same terrorist organization are linked with each other. Investigating the community structures within these networks is a challenging problem, since most of the connections/links between attacks are not clearly resolved. Detecting the communities in these incomplete information networks is crucial for analyzing the structures of terrorist-attack activities.
- **Food Web** The food web of a large ecosystem is usually a highly complex network. Each node in the network represents a living organism, while the links rep-

WWW 2012, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1229-5/12/04.

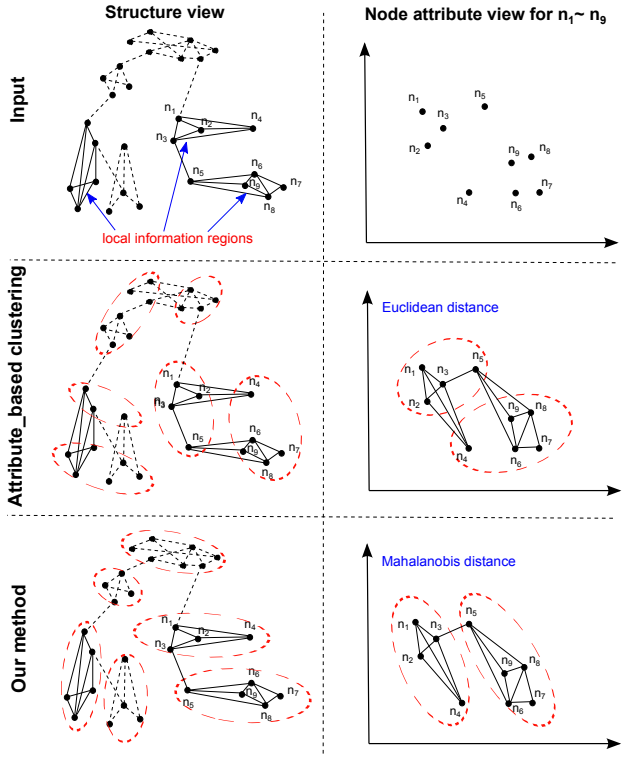


Figure 1: Comparison of different clustering methods on incomplete information networks with missing edges.

resent the relations between them. Usually, it is very difficult to resolve all of the links within a food web. However, it is relatively easier to figure out some local regions within the food web. Discovering communities in these incomplete food webs can help us identify micro ecosystems and the corresponding living organisms of each micro ecosystem.

Finding communities in incomplete information networks is a challenging task. Conventional graph-based clustering methods can not be directly applied to it. The reason is that traditional graph clustering methods, such as normalized cut based methods [24] and modularity based methods [19], mainly focus on the topological structure of the network. Since most of the links are absent in incomplete information networks, it is impossible to cluster the network with this kind of methods. As shown in the middle level of Figure 1, if we cluster the nodes using the traditional attribute based methods such as k-means, the most likely result is that we place nodes with the most similar attributes in the same cluster. However, the nodes which are densely connected in structure may not necessarily mean they have the most similar attributes, i.e., they may be only similar on a subset of the attributes. For example, in the food web networks, a community usually stands for a micro ecosystem and can contain various kinds of living organisms, which can have very different attributes. Recently, some new algorithms [31] which perform clustering based on both structures and the attributes of the network are proposed. However, they can not be applied on incomplete information networks due to the absence of the complete linkage structure.

Given the assumption that the structure of the network has a close relation with attributes of each object in the information network, in this paper, we propose a novel approach for community detection in incomplete information networks. To the best of our knowledge, this is the first attempt to formulate and address the incomplete information network problem. The main idea of our approach is that, since the structure of the network has a strong relation with the attributes of the objects in the network, we can learn a global distance metric from the local information regions with complete linkage information. Then, we use the global metric to measure the distance between any pair of nodes in the network. Because the metric is learned from the structure of the network, the distance will reflect the hidden linkage structure in the network. Finally, we propose a distance-based clustering algorithm to cluster the nodes in the incomplete information network. The different clustering results are shown in Figure 1. To summarize, this work contributes on the following aspects:

- We identify and define the problem of community detection in incomplete information networks with local information regions, i.e., an incomplete information network that still has a few tiny local regions where the complete linkage information is available.
- In order to find a measurement, which can reflect the structural relation between the nodes in incomplete information networks, we cast the side information of the network into an optimization problem. Then a metric, which can be used to measure the distance between any pair of nodes, is learned.
- Based on the learned metric, we devise a distance-based modularity function to evaluate the quality of the communities.
- Finally, we propose a distance-based algorithm DSHRINK which can discover the hierarchical and overlapped communities. Moreover, in order to speedup the clustering process, an effective strategy is also taken.

This paper is organized as follows. We introduce the related work in Section 2. The formal definition of our problem is presented in Section 3. In Section 4, we introduce how to make use of the side information to learn a global metric. In Section 5, we explain the distance-based clustering algorithm. The experimental results are presented in Section 6. Finally, we conclude in Section 7.

2. RELATED WORK

Community detection in networks and graphs has been widely studied in recently years[16, 4]. Many approaches mainly focused on the topological structures based on various criteria including modularity [19], normalized cut [24], structural density [30] and partition density [3]. Given a graph, which is clustered into k communities, the modularity function Q is defined as:

$$Q = \sum_{i=1}^{k} \left[\frac{l_i}{L} - \left(\frac{d_i}{2L} \right)^2 \right] \quad (1)$$

where L is the number of edges in the graph, l_i is the number of edges between nodes within community i, and d_i is the sum of the degrees of the nodes in community i. The optimal

clustering result is achieved by maximizing the modularity value which ranges from 0 to 1. In general, maximizing Q is a NP-hard problem. Hence, many heuristic approaches, which try to approximate the optimal modularity value, were proposed [10]. Such approaches include greedy agglomeration [19, 28], mathematical programming [1], spectral methods [25], simulated annealing [11], sampling techniques [22], etc. However, modularity is not a scale-invariant measure, and therefor, by relying on its maximization, can not detect communities smaller than a certain size [8]. Besides, Palla et. al. [20] proposed a clique percolation method, which can detect overlapped communities, but is not suitable for detecting hierarchical structures. Huang et. al. [12] proposed a parameter-free algorithm SHRINK, which can not only discover overlapped and hierarchical communities but also the hub nodes and outliers among them. Rosvall et. al. [21] tried to compress the information of the graph by optimizing the minimum description length of the random walk and proposed a highly accurate algorithm namely Infomap. Ahn et al. [3] insisted that link communities are fundamental building blocks, and the overlapped and hierarchical communities in networks are two aspects of the same phenomenon. They proposed a link-based approach which reveals the real world communities effectively. Other link-based methods were also devised by [6].

There are also some graph clustering methods which based on attributes. Tian et al. [26] proposed an OLAP-style aggregation approach to summarize large graphs by grouping nodes based on user-selected attributes and relationships. This method achieves homogeneous attribute values within clusters but ignores the intra-cluster topological structures. Tsai et al. [27] proposed a feature weight self-adjustment mechanism for k-means clustering. In that study, finding the appropriate weight is modeled as an optimization problem which tries to minimize the separations within clusters and maximize the separations between clusters. Since most of the attributes based methods mainly focus on the homogeneity of the clusters, the cohesive internal structure of the clusters can not be guaranteed.

Recently, some clustering methods based on both links and attributes were also proposed. [9] introduced the connected k-center(CkC) problem, which checks whether an attributed graph can be partitioned or not by considering both attributes and the links. Since the CkC problem is NP-complete, the authors proposed a constant factor approximation algorithm and a heuristic algorithm for the large data sets. [31] proposed SA-Cluster, which is based on both structural and attribute similarities through a unified distance measure. In that study, a graph is partitioned into k clusters so that each cluster contains a densely connected subgraph with homogeneous attribute values. Then, in order to learn the degree of contributions of structural similarity and attribute similarity automatically, an effective method was proposed.

3. PROBLEM DEFINITION

In this section, we formally define our problem and introduce several related concepts.

Definition 1 (Information Network) *An information network is denoted as $G = (V, E, A)$, where V is the set of vertices, $E \subseteq V \times V$ is the set of edges, and $A = \{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_{|V|}\}$ is the set of node attributes which describe the properties of vertices in V.*

For convenience, we use $A(v)$ to denote the attribute vector of node v and $E(U)$ to represent the edges among nodes in $U (U \subseteq V)$.

Definition 2 (IIN) *Incomplete Information Networks with Local Information Regions (IIN) are defined as follows: given an information network $G = (V, E, A)$ and a network $G' = (V', E', A')$, network G' is called an incomplete information network with local information regions of G iff (1) $V' = V, E' \subset E$ (2) $\forall v' \in V', \forall v \in V$, if $v = v'$, then $A'(v') = A(v)$. (3) $\exists V'' \subseteq V'$, for $\forall e \in E(V'')$, then $e \in E'(V'')$. Specifically, we call the local subnetwork $g(V'', E'', A'')$ as the* local information region *denoted by L, where $E'' = E(V'')$ and $A'' = A(V'')$.*

From Definition 2, we know that an incomplete information network with local information regions is a network $G' = (V', E', A')$ with a small set of connected regions $V'' \subset V'$, where the edges in $E' - E(V'')$ are missing. There can be many different types of incomplete information networks. In this paper, we focus on the incomplete information network that has some local information regions, where the linkage structures are completely preserved as in Definition 2. For the remaining of the paper, we will just refer to this type of "incomplete information network with local information regions" as incomplete information network. Obviously, there can be more than one local information regions in an incomplete information network G'. Furthermore, when we are talking about an incomplete information network, we assume that there is a corresponding information network, potentially. A typical incomplete information network is shown in the upper left level in Figure 1.

Definition 3 (Dissimilar Node Pair) *Given an information network $G = (V, E, A)$ and its k clusters $C_1, C_2, ..., C_k$, where $\bigcup_{i=1}^{k} V(C_i) = V(G)$, any pair of nodes (v_i, v_j) is a dissimilar node pair iff (1) $v_i \in C_m \wedge v_j \in C_n \wedge m \neq n (1 \leq m, n \leq k)$; (2) $E(\{v_i, v_j\}) = \emptyset$. We denote the dissimilar node-pair set as $\mathcal{D}$.*

For an information network G, if $C = C_1, C_2, ..., C_k$ is the set of clusters which are based on the linkage structures of G, we formalize our community detection problem as: given an incomplete information network G' of G and the dissimilar node-pair set $\mathcal{D}$, based on some similar criteria, the objective is to find the set C' which should be as similar as possible to C.

4. OPTIMIZATION FRAMEWORK

In this section, we address the problem of how to learn a global metric which is used to measure the distance between any pair of nodes in an incomplete information network G'. This goal is achieved by solving an optimization problem, which makes use of the side information getting from the link relations of G' and the dissimilar node-pair set $\mathcal{D}$.

Definition 4 (Structure Similarity) *Given a network $G = (V, E)$, for any pair of nodes $v_i, v_j \in V$, the structure similarity between node v_i and v_j is defined as*

$$s(v_i, v_j) = \frac{|\Gamma(v_i) \cap \Gamma(v_j)|}{\sqrt{|\Gamma(v_i)||\Gamma(v_j)|}} \quad (2)$$

where $\Gamma(v)$ is the set containing v and its neighbors.

We compute the structure similarity between any pair of nodes u_i, v_i in local information region L_i by Equation (2). Then, we define the similar node-pair set $\mathcal{S}$ as a 3-tuple set about the structure similarity as follows:

$$\mathcal{S} = \{(u_i, v_i, s_i) | s_i = s(u_i, v_i), u_i, v_i \in V(L_i)\} \quad (3)$$

Based on the similar node-pair set $\mathcal{S}$ and the dissimilar node-pair set $\mathcal{D}$, our objective is to find a metric by which, the similar nodes should be close together and the dissimilar nodes should be far away from each other. Moreover, the extent of closeness between any pair of similar nodes should be based on the structure similarity between them. Inspired by [29], this objective can be achieved by learning a distance metric. Let the matrix $\mathcal{M} \in R^{m \times m}$ represent the distance metric. Then, the distance between any two nodes $u_i, v_i \in V$ is defined by

$$d_{\mathcal{M}}(u_i, v_i) = \|u_i - v_i\|_{\mathcal{M}} = \sqrt{(u_i - v_i)^T \mathcal{M}(u_i - v_i)} \quad (4)$$

In order to make sure the distance metric defined by Equation (4) satisfies non-negativity and the triangle inequality, we constraint $\mathcal{M}$ to be positive semi-definite. Now, we can formalize our objective as an optimization problem as follows:

$$\begin{aligned} \min_{\mathcal{M}} \quad & \sum_{(u_i,v_i)\in\mathcal{S}} (s_i \|u_i - v_i\|_{\mathcal{M}})^2 \\ \text{s.t.} \quad & \sum_{u_i,v_i\in\mathcal{D}} \|u_i - v_i\|_{\mathcal{M}} \geq w \\ & \mathcal{M} \succeq 0 \end{aligned} \quad (5)$$

where w is a constant.

We notice that our objective function (5) is a linear function of $\mathcal{M}$. Further more, both of the constraints given in Equations (5) are convex. Hence, our optimization problem is convex, which enables us to compute the global optimal resolution.

Despite our optimization problem falls into the category of convex programming, it does not fall into any special class of convex programming, e.g., quadratic programming and semi-definite programming. Hence, the global solution can only be solved by a generic approach. We also notice that the learned optimal $\mathcal{M}$ can appear in two forms, which are diagonal matrix and full matrix.

In order to get the diagonal form of $\mathcal{M}$, we give the equivalent Equations (5) similar to [29]:

$$\begin{aligned} & f(\mathcal{M}) = f(\mathcal{M}_{11}, ..., \mathcal{M}_{nn}) \\ = & \sum_{u_i,v_i\in\mathcal{S}} s_i^2 \|u_i - v_i\|_{\mathcal{M}}^2 - \log\left(\sum_{(u_i,v_i)\in\mathcal{D}} \|u_i, -v_i\|_{\mathcal{M}}\right) \end{aligned} \quad (6)$$

Minimizing Equation (6) can be resolved by using the Newton-Raphson method. Furthermore, in order to keep the semi-definite characteristics of $\mathcal{M}$, we replace the Newton update $H^{-1}\nabla_f$ by $\alpha H^{-1}\nabla_f$, where α is the a step-size parameter optimized via a line-search which gives the largest downhill steps subject to $\mathcal{M}_{ii} \geq 0$ [29].

In order to get the full matrix of $\mathcal{M}$, we give the equivalent Equations (5) similar to [29]:

$$\begin{aligned} \max_{\mathcal{M}} \quad & g(\mathcal{M}) = \sum_{(u_i,v_i)\in\mathcal{D}} \|u_i, v_i\|_{\mathcal{M}} \\ \text{s.t.} \quad & h(\mathcal{M}) = \sum_{(u_i,v_i)\in\mathcal{S}} (s_i \|u_i, v_i\|_{\mathcal{M}})^2 \leq w \\ & \mathcal{M} \succeq 0 \end{aligned} \quad (7)$$

The reason for giving the transformation of the original optimization problem is for efficiently finding the global optimal full matrix $\mathcal{M}$ by using gradient descent and the idea of iterative projections [5]. We first use a gradient ascent on $g(\mathcal{M})$ to optimize (7). Then, we project the intermediate results to hold the constraints (7). The similar tricks are also used in [29].

Besides, we notice that in Equations (5) and (7), w is a constant whose value is not important. This is because the distance between any pair of nodes in network G' is a relative variable. Changing the value of w only makes the distance between any pair of nodes u_i and v_i change from $\|u_i - v_i\|_{\mathcal{M}}$ to $w^2\|u_i - v_i\|_{\mathcal{M}}$. Hence, we choose $w = 1$ in this paper. For the convenience of discussion, we denote the incomplete information network as $G = (V, E, A, \mathcal{M})$ in the rest of the paper.

5. DISTANCE-BASED CLUSTERING

By optimizing our objective function, we have learned a matrix $\mathcal{M}$ in section 4. In other words, we have gotten a metric which can be used to measure the distance between any two nodes in graph G. Inspired by the density-based clustering approaches, e.g., [30, 12], which cluster nodes from the higher density to lower density, in this section, we propose a distance-based clustering approach DSHRINK which can detect the overlapped and hierarchical communities hidden in the graph.

5.1 Distanced-based Modularity

The distance-based clustering approach DSHRINK places the nodes which have the shorter distance with each other into the same cluster, and the nodes which have the longer distance between them into different clusters. In order to evaluate the quality of clusters, we define the distance-based modularity as follows:

Definition 5 (Distance-based Modularity) *Given an incomplete information network $G = (V, E, A, \mathcal{M})$ and its cluster $C = \{C_1, C_2, ..., C_k\}$, the distance-based modularity Q_d is defined as*

$$Q_d = \sum_{i=1}^{k} \left[\frac{D_i^I}{D^T} - \left(\frac{D_i^C}{D^T}\right)^2\right] \quad (8)$$

where k is the number of clusters, $D_i^I = \sum_{u,v\in C_i} d_{\mathcal{M}}(u, v)$ is the sum of distance between any pair of nodes within cluster C_i, $D_i^C = \sum_{u\in C_i, v\in V} d_{\mathcal{M}}(u, v)$ is the sum of distance between any node in cluster C_i and any node in the network G, and $D^T = \sum_{u,v\in V} d_{\mathcal{M}}(u, v)$ is the sum of distance between any two nodes in the network G.

Obviously, in contrast to the original modularity defined by Newman [19], the value range of Distance-based Modularity is $[-1, 0]$. If $Q_d = 0$, it means all the nodes are either

placed into one cluster or placed into different clusters randomly. The smaller value of Q_d means the better quality of clustering.

Similar to [7, 12], if we combine any two modules C_s and C_t, the distance-based modularity gain $\triangle Q_d$ achieved from the combination can be computed by

$$\triangle Q_d = Q_d^{C_s \cup C_t} - Q^{C_s} - Q^{C_t} = \frac{2D_{st}^U}{D^T} - \frac{2D_s^C D_t^C}{(D^T)^2} \quad (9)$$

where $D_{st}^U = \sum_{u \in C_s, v \in C_t} d_{\mathcal{M}}(u, v)$ is the sum of distance between any two nodes in modules C_s and C_t respectively.

According to Equation (9), we compute the gain of distance-based modularity $\triangle Q_d$ for combing j clusters $C_1, C_2, ..., C_j$ into a new community by

$$\triangle Q_d(j) = \frac{\sum_{s,t \in \{1,...,j\}, s \neq t} 2D_{st}^U}{D^T} - \frac{\sum_{s,t \in \{1,...,j\}, s \neq t} 2D_s^C D_t^C}{(D^T)^2} \quad (10)$$

5.2 Clustering Algorithm

Before addressing our algorithm in detail, we give the following definitions.

Definition 6 (Nearest Neighbor) *Given an incomplete information network $G = (V, E, A, \mathcal{M})$, the nearest neighbor set for $\forall v \in V$ is defined as*

$$NN(v) = \{y | y = \arg\min_x d_{\mathcal{M}}(v, x), x \in V \wedge x \neq v\}. \quad (11)$$

Definition 7 (Mutual Nearest Neighbor) *Given an incomplete information network $G = (V, E, A, \mathcal{M})$, any pair of nodes $u, v \in V$ is said to be mutual nearest neighbor, denoted by $u \overset{\gamma}{\leftrightarrow} v$, i $iff v \in NN(u) \wedge u \in NN(v) \wedge d_{\mathcal{M}}(u, v) = \gamma$, where $\gamma \in \mathcal{R}^+$.*

Definition 8 (Local Community) *Given an incomplete information network $G = (V, E, A, \mathcal{M})$, we call the subgraph $C(v) = (V', E', A', \mathcal{M}, \gamma)$ of G as a local community iff* (1) *$v \in V'$;* (2) *$\forall u \in V', \exists v \in V' \wedge (u \overset{\gamma}{\leftrightarrow} v)$;* (3) *$\{u | u \in V' \wedge u \overset{\gamma}{\leftrightarrow} v \wedge v \notin V'\} = \emptyset$. $\gamma \in \mathcal{R}^+$ is the radius of the local community $C(v)$.*

The distance-based shrinking approach DSHRINK is presented in Figure 3. Our approach can be divided into two phases. At the first phase, we compute the distance between any pair of nodes in graph G and store the distance as a 3-tuple $(v_i, v_j, d_{\mathcal{M}}(v_i, v_j))$ into a map structure (see Figure 2). For any $v_i \in V$, the sum of the distance between node v_i and any other node $v_j \in V$ is saved in S_i^T. Since $D_s^C = \sum_{v_i \in C_s, v_j \in V} d_{\mathcal{M}}(v_i, v_j) = \sum_{v_i \in C_s} S_i^T$, computing S_i^T in advance can speed up computing D_s^C when computing Q_d. For the same purpose, the total distance D^T between any pair of node is also be computed.

At the second phase, (1) we first begin at an arbitrary node and span the node to a local community based on Definition 8. All the nodes, which are in the local community, will be tagged as "visited". Then, we choose the next unvisited node in graph G and repeat the above step. This process will not stop until all the nodes are visited. (2) Secondly, for each local community discovered by the first step, we view each single node in it as a community. Then $\triangle Q_d$ is computed according to Equation (10). If $\triangle Q_d < 0$,

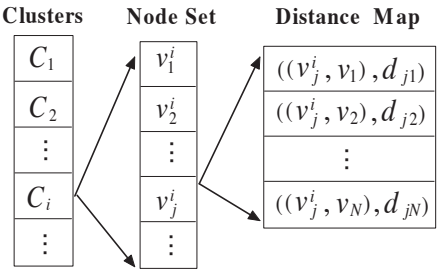


Figure 2: Data structure used in DSHRINK

which means the combination of the communities can decrease the total distance-based modularity Q_d, we shrink the local community as a super node. Otherwise, the local community will not be shrunk. (3) Thirdly, we tag all the nodes including super nodes and the common nodes as "unvisited" and repeat the first and second steps. The above steps will be repeated many times until shrinking any local maximal community can not decrease the Q_d any more. Finally, the nodes condensed in a super node form a community, and different super nodes stand for different communities.

According to the definition of local community, we know that the order of traversing the nodes in the incomplete information network G does not change the final members of the local community. Moreover, from the clustering process, we know that the local community, in which nodes have a shorter distance, will be shrunk at the prior or the same iteration than the local community, in which nodes have a longer distance. Single nodes, which have not been shrunk to any other super nodes, are viewed as hubs or outliers depending on how many communities they are close to. If we want to form the overlapped communities, the hub nodes will be placed into more than one communities. Otherwise, each hub node will only be placed into the community which makes the most decrease of Q_d by adding the hub node. If we view the distance between each pair of nodes as the structure similarity, the above shrinking process is similar to [12].

5.3 Speeding up the Clustering Process with Approximation

It is possible to speed up the clustering process by allowing some approximation in the determination of the local community. We define ϵ-approximate mutual nearest neighbor and ϵ-approximate local community as follows:

Definition 9 (ϵ-approximation Mutual Nearest Neighbor) *Given an incomplete information network $G = (V, E, A, \mathcal{M})$, any pair of nodes $u, v \in V$ is said to be ϵ-approximation mutual nearest neighbor in G, denoted by $u \overset{\epsilon}{\leftrightarrow} v$, iff $(v \in NN(u) \wedge |d_{\mathcal{M}}(u, v) - d_{\mathcal{M}}(v, x)| \leq \epsilon) \vee (u \in NN(v) \wedge |d_{\mathcal{M}}(u, v) - d_{\mathcal{M}}(u, y)| \leq \epsilon)$, where $x \in NN(v)$, $y \in NN(u), \epsilon \in \mathcal{R}^+$.*

Obviously, ϵ-approximate mutual nearest neighbor is an extended version of mutual nearest neighbor.

Definition 10 (ϵ-approximation Local Community) *Given an incomplete information network $G = (V, E, A, \mathcal{M})$, $C(v) = (V', E', A', \mathcal{M}, \epsilon)$ is a subgraph of network G. $C(v)$ is said to be a ϵ-approximation local community of G iff* (1)

DSHRINK($G = (V, E, A, \mathcal{M})$)

Input:
$G = (V, E, A, \mathcal{M})$: Incomplete information network.
Output:
$C = \{C_1, C_2, ..., C_k\}$: Cluster set.
HO : Hubs and outliers.
Process:

```
1  Initialize each v_i ∈ V as a community and put it in C;
2  for each v_i ∈ V do
3    for each v_j ∈ V ∧ v_i ≠ v_j do
4      Compute d_M(v_i, v_j) according to Equation 4;
       Store (key(v_i, v_j), value(d_M(v_i, v_j))) with
       ascending order into the distance map.
5      S_i^T += d_M(v_i, v_j);
6      D^T += d_M(v_i, v_j);
7    end
8  end
9  while true do
10   for each v_i ∈ V do
11     if v_i.visited then continue
12     Span a local community C(v_i) according to
       Definition 8;
13     for each v_j ∈ V(C(v_i)) do
14       v_j.visited = ture;
15     end
16     C ← C ∪ C(v_i);
17   end
18   Q_d.descrease = false;
19   for each C_j ∈ C do
20     Compute △Q_d according to Equation 10;
21     if △Q_d < 0 then
22       v_s ← V(C_j);
23       C ← (C − C_j) ∪ v_s;
24       Q_d += △Q_d;
25       Q_d.descrease = true;
26       v_s.visited = false;
27     end
28   end
29   if !(Q_d.descrease) then break;
30 end
31 Get single nodes from C and put them into HO
32 return C, HO;
```

Figure 3: The Description of DSHRINK

$v \in V'$; (2) $\forall u \in V', \exists v \in V' \wedge (u \underset{\epsilon}{\leftrightarrow} v)$; (3) $\{u | u \in V' \wedge u \underset{\epsilon}{\leftrightarrow} v \wedge v \notin V'\} = \emptyset$; (4) *let* $f(r) = \{r | r = d_{\mathcal{M}}(s,t), s \underset{\epsilon}{\leftrightarrow} t \wedge s \in V' \wedge t \in V'\}$, $|Max(f(r)) - Min(f(r))| \leq \epsilon$; (5)*when* (3) *and* (4) *can not be held at the same time,* (4) *is prior to* (3) *to be guaranteed.* $\epsilon, r \in \mathcal{R}^+$.

We note that this relaxation of the definition of local community can greatly speed up the clustering process. In order to take advantage of ϵ-approximation local community, the only difference in DSHRINK is to span a ϵ-approximation local community instead of a local community in step (12). When we span the ϵ-approximation local community ($\epsilon > 0$), the final clustering result may rely on the visiting sequence of the nodes. In this paper, we give priority to the shorter distance nodes among all of the ϵ-approximation neighbours when spanning the ϵ-approximation local communities. Our experimental results show that the final clustering effect is almost not affected by the order of the visiting sequence of nodes by taking the above strategy. Furthermore, given an appropriate parameter ϵ, we find that this relaxation does not affect the practical quality of the communities obtained.

Table 1: Summary of experimental data sets

Dataset	# Nodes	# Links	# Attributes	# Classes
DBLP-A	4638	16,447	102	6
DBLP-B	4559	14,407	102	6

6. EXPERIMENTS

In this section, we use two real-world data sets to validate the effectiveness and efficiency of our approach. All the experiments are conducted on a machine with Intel 8-core 2.7 GHz processors and 28GB memory.

6.1 Data Sets

DBLP-A Dataset: DBLP-A is the data set extracted from DBLP database[1] which provides bibliographic information on computer science journals and proceeding. We extract paper information from 16 top conferences which cover 6 research fields including *Artificial Intelligence*, *Information Retrieval*, *Computer Vision*, etc. We create the coauthor network by choosing authors, who published at least 2 papers during $2000 - 2010$, as the nodes of the network. Any pair of authors who have coauthored are linked in this coauthor network. This coauthor network contains 4638 nodes and 16447 links in total. Each node is attached with a bag-of-words which extracted from the paper titles published by him/her. We first apply the standard text preprocessing such as stemming, stop words removal. Then we reduce the dimension of the bag-of-words to 100 by PCA and use them as the features of the corresponding node. In addition, the number of co-authors and publications are also used as features of the nodes.

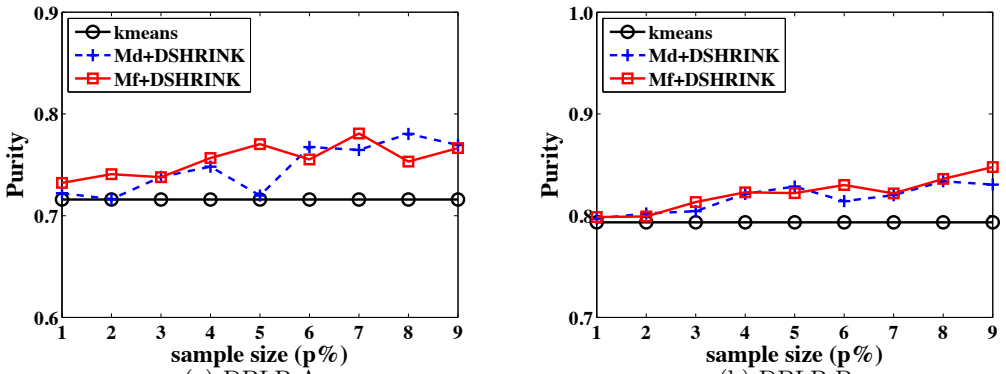
DBLP-B Dataset: We also extract paper information from 16 top conferences of 6 research fields such as *Algorithms & Theory*, *Natural Language Processing*, *Bioinformatics*, etc. The same setups with DBLP-A are also used here to build the coauthor network as our second data set, called DBLP-B.

We summarize our data sets in Table 1.

6.2 Incomplete Information Network Generation

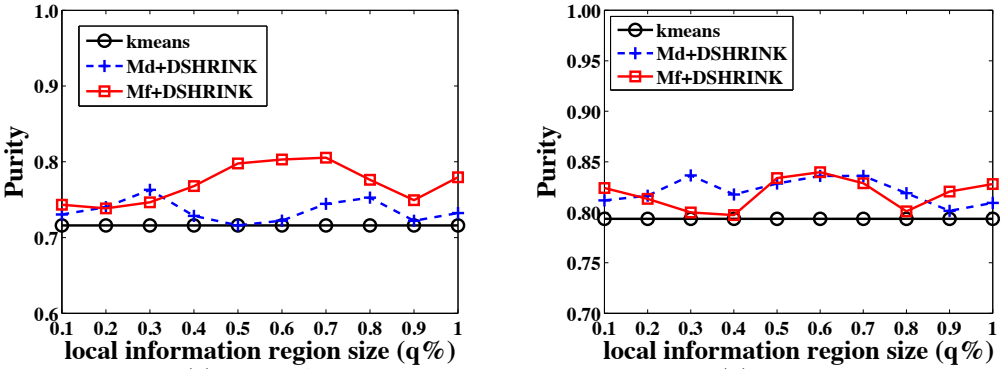
In order to simulate the incomplete information networks with local information regions, we use the following experiment setting. If we perform random sampling on the nodes, the sampled network usually ends up being sparsely connected, without local information regions. In this paper, we use the snowball sampling [23] to sample a group of connected local region at a time. We randomly sample one node and use BFS to include its neighboring nodes into the sampled region until a fixed number of nodes are sampled. We repeat this process until a number of local regions are sampled. Then we assume the links within the local informative regions are available to the algorithms, while the remaining links in the network are removed. In order to control the total number of nodes being sampled, we introduce a parameter p, called *sample ratio*, i.e., the ratio of the nodes in the network being sampled into the local region. In addition, we introduce another parameter q, called *local information region size*, to control the size of each local information region. In detail, we first randomly choose a node

[1]http://www.informatik.uni-trier.de/∼ley/db/



(a) DBLP-A (b) DBLP-B

Figure 4: Accuracy comparison between different methods ($q\% = 0.3\%$).



(a) DBLP-A (b) DBLP-B

Figure 5: Accuracy comparison between different methods ($p\% = 10\%$).

in the network. We then include $q\%$ nodes from its neighbors using BFS search. Common neighbors of any pair of nodes in the sampled region are further included into the sampled local region. The above sampling process continues until we sample $p\%$ of the nodes in the network. In addition to the local regions, we sample the same number of nodes and use them to generate dissimilar pairwise constraints. In the sampled group, the pairs of nodes that are in different classes are then used as the dissimilar node-pair set $\mathcal{D}$. More concretely, for DBLP-A and DBLP-B datasets, we choose the pair of authors, whose research fields are not overlapped as the dissimilar node pair.

6.3 Evaluation Measures

In order to measure the effectiveness of our approach, we adopt *Purity* to evaluate the quality of the communities generated by different approaches. The definition of purity is as follows: each cluster is first assigned with the most frequent class in the cluster, and then the purity is measured by computing the number of the instances assigned with the same labels in all clusters. Formally:

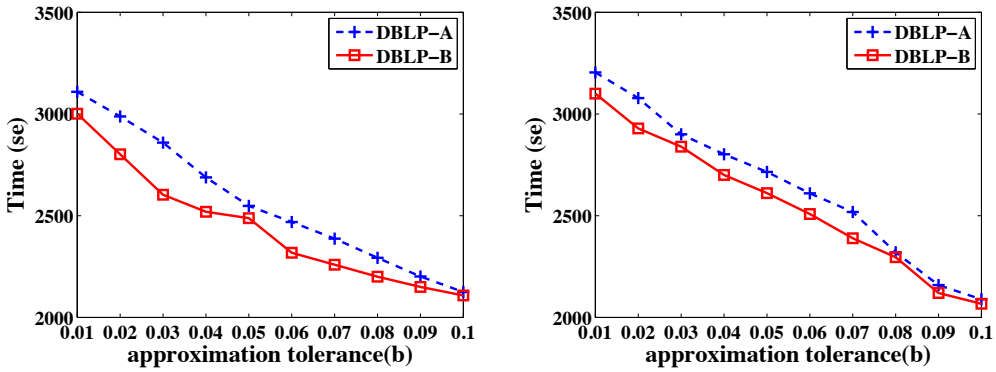
$$\text{Purity} = \frac{1}{n}\sum_{i=1}^{k}\max_{j}|C_i \cap l_j| \tag{12}$$

where $\{C_1, \cdots, C_k\}$ is the set of clusters, l_j is the j-th class label. The value of purity ranges from 0 to 1. The community structure generated by each compared method will be evaluated using the true label of each node such that the higher purity value means the higher accuracy of the method. Since each author can have multiple research areas as its class labels. We computed the purity of the clustering results based on each label separately, and the average results over 6 labels are reported.

6.4 Compared Methods

In order to demonstrate the effectiveness and efficiency of our approach, we compare our approach with the following methods:

- Kmeans: We use the default Euclidean metric to measure the distance between any node x_i and the centroid x_k. The K value used in the dataset of DBLP-A and DBLP-B is 6, which is the same number of clusters with the ground truth.
- $\mathcal{M}_d$+ DSHRINK: We learn a diagonal Mahalanobis matrix $\mathcal{M}_d$ and use it as the input of $\mathcal{M}$ for DSHRINK.
- $\mathcal{M}_f$+ DSHRINK: We learn a full Mahalanobis matrix $\mathcal{M}_f$ and use it as the input of $\mathcal{M}$ for DSHRINK.



(a) $\mathcal{M}_f$+DSHRINK (b) $\mathcal{M}_d$+DSHRINK

Figure 6: The computation time with the different values of b.

6.5 Effectiveness Results

The variation of purity scores under different values of $p\%$ is given in Figure 4. In this experiments, $q\% = 0.3\%$ is used. Since the accuracy of KMeans is not affected by the number of sampling nodes, the purity value of the KMeans is a horizontal line in all cases. We notice that the purity scores of $\mathcal{M}_f$+DSHRINK and $\mathcal{M}_d$+DSHRINK ascend quickly with the increasing number of local information regions sampled. Especially, when $p\% > 2\%$, the purity values of $\mathcal{M}_f$+DSHRINK and $\mathcal{M}_d$+DSHRINK exceed kmeans over all data sets. That is because, firstly, the learned Mahalanobis matrix $\mathcal{M}$ rescales all of the nodes into a new feature space, where the similar nodes are closer, and the dissimilar nodes are further away than the original Euclidean space. Secondly, DSHRINK can automatically detect the most appropriate number of communities by minimizing the distance-based modularity. Since the number of communities in the networks is unknown, $\mathcal{M}_f$+DSHRINK and $\mathcal{M}_d$+DSHRINK have more advantage for discovering the most appropriate community structures than Kmeans. Another observation is that, in most of the cases, with the same value of $p\%$, the purity scores of $\mathcal{M}_f$+DSHRINK are a little higher than $\mathcal{M}_d$+DSHRINK in both DBLP-A and DBLP-B data sets. This demonstrates that the full Mahalanobis matrix performs better rescaling function for separating the similar nodes from the dissimilar nodes than diagonal Mahalanobis matrix in DBLP-A and DBLP-B data sets. However, this principle dose not not always hold. For instance, in DBLP-A data set, the purity score of $\mathcal{M}_f$+DSHRINK is less than $\mathcal{M}_d$+DSHRINK when $p\% = 8\%$.

In Figure 5, given a specified value of $p\% = 10\%$, we also present the changes of purity scores with the different value of $q\%$. We notice that, on the one hand, for a specified $p\% = 10\%$, the larger value of $q\%$ makes more similar node pairs be captured in each local information region, but fewer local information regions get chosen in the whole incomplete information network. On the other hand, with the smaller value of $q\%$, fewer similar node pairs can be captured in each local information region, but more local information regions can be sampled. Since both the number of local information regions and the number of similar node pairs can affect the learning of the metric, finding the balance point of $q\%$ is critical for achieving a better clustering result. From Figures 5 (a) to (b), we know that the balance point of $q\%$ can be gotten between 0.5% to 0.7%.

6.6 Efficiency Results

We notice that, computing the optimal Mahalanobis matrix and the distance between any pair of nodes can be accomplished in advance before clustering process. In this part, we mainly focus on the clustering process and test how ϵ-approximation local community speeds up the clustering process and affects the quality of clusters. The distance here is relative and changeable according to different values of w in Equation (5). Hence, discussing the value of ϵ is meaningless with a special value of w. Fortunately, we find the top k nearest nodes of each node is a good base for us to compute the appropriate ϵ value. In order to compute an appropriate ϵ value, we average the sum of distance between each node and its corresponding top k nearest nodes as follows:

$$d = \frac{\sum_{i=1}^{N} \sum_{j \in TopK(i)} d_{\mathcal{M}}(v_i, v_j)}{k|N|} \tag{13}$$

where $TopK(i)$ is the set of node index, whose distance to v_i ranks in the top k among all of the nodes to v_i, and $|N|$ is the total number of nodes in the incomplete information network. In this paper, we choose $k = 10$ and give the value of ϵ as $\epsilon = d \times b$. For a specified incomplete information network G and a metric $\mathcal{M}$, the value of d is a constant. Therefor, changing the value of b is equivalent to change the value of ϵ.

In Figures 6(a) to (b), we have illustrated the variation in the efficiency of different b values for $\mathcal{M}_f$+DSHRINK and $\mathcal{M}_d$+DSHRINK. We observe that the computation time decreases quickly with the increasing value of b. It is because relaxing the definition of local community to a certain extent can decrease the iteration times in the clustering process. We also find that the purity score are not changed dramatically with differenct b values.

7. CONCLUSION

In this paper, we presented the first approach for community detection in incomplete information networks with local information regions. While the traditional community detection algorithms make the assumption of the full knowledge of linkage information, they can not solve the problem

of community detection in incomplete information networks. In order to resolve this problem, we explored the metric learning idea and learned a global metric from the side information of the incomplete information network. Moreover, we proposed the distance-based modularity function. Based on this function, we further devised a distance-based clustering algorithm DSHRINK. In order to speed up the clustering process, some helpful approximation strategies were also proposed. Experimental results illustrated the effectiveness and efficiency of our approach.

8. ACKNOWLEDGMENTS

The first author is supported by National Natural Science Foundation of China through grants 60933005, 60873204 and the National High-Tech Program through grant 2010AA012505. The third author is supported in part by US NSF through grants IIS 0905215, DBI-0960443, and Google Mobile 2014 Program.

9. REFERENCES

[1] G. Agarwal and D. Kempe. Modularity-maximizing graph communities via mathematical programming. *European Physical Journal B*, 66:409–418, 2008.

[2] C. Aggarwal, Y. Xie, and P. Yu. Towards community detection in locally heterogeneous networks. SDM, pages 391–402, 2011.

[3] Y. Ahn, J. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466:761–764, 2010.

[4] H. Alani, S. Dasmahapatra, K. O'Hara, and N. Shadbolt. Identifying communities of practice through ontology network analysis. *Intelligent Systems*, 18(2):18–25, 2003.

[5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[6] T. Evans and R. Lambiotte. Line graphs, link partitions and overlapping communities. *Physical Review E*, 80:016105, 2009.

[7] Z. Feng, X. Xu, N. Yuruk, and T. A. J. Schweiger. A novel similarity-based modularity function for graph partitioning. In *DaWak*, pages 385–396, 2007.

[8] S. Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of The National Academy of Sciences*, 104(1):36–41, 2007.

[9] R. Ge, M. Ester, B. J. Gao, Z. Hu, B. K. Bhattacharya, and B. Ben-moshe. Joint cluster analysis of attribute data and relationship data: The connected k-center problem, algorithms and applications. *ACM Transactions on Knowledge Discovery From Data*, 2:1–35, 2008.

[10] B. H. Good, Y. A. de Montjoye, and A. Clauset. The performance of modularity maximization in practical contexts. *Physical Review E*, 81:046106, 2010.

[11] R. Guimera and L. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[12] J. H. H. D. Y. S. Y. L. J. Huang, H. Sun. SHRINK: a structural clustering algorithm for detecting hierarchical communities in networks. In *CIKM*, pages 219–228, 2010.

[13] M. Kim and J. Leskovec. The network completion problem: Inferring missing nodes and edges in networks. In *SDM*, pages 47–58, 2011.

[14] J. Z. Z. N. L. Tang, H. Liu. Community evolution in dynamic multi-mode networks. KDD, pages 677–685, 2008.

[15] A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5):056117, 2009.

[16] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical properties of community structure in large social and information networks. WWW, pages 695–704, 2008.

[17] J. Leskovec, K. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. WWW, pages 631–640, 2010.

[18] Y. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher. Metafac: community discovery via relational hypergraph factorization. KDD, pages 527–536, 2009.

[19] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.

[20] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814, 2005.

[21] M. Rosvall and C. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105:1118, 2008.

[22] M. Sales-Pardo, R. Guimerà, A. Moreira, and L. Amaral. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104(39):15224–15229, 2007.

[23] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

[24] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[25] M. Shiga, I. Takigawa, and H. Mamitsuka. A spectral clustering approach to optimally combining numerical vectors with a modular network. KDD, pages 647–656, 2007.

[26] Y. Tian, R. Hankins, and J. Patel. Efficient aggregation for graph summarization. In *SIGMOD*, pages 567–580, 2008.

[27] C. Tsai and C. Chiu. Developing a feature weight self-adjustment mechanism for a k-means clustering algorithm. *Computational Statistics Data Analysis*, 52:4658–4672, 2008.

[28] K. Wakita and T. Tsurumi. Finding community structure in mega-scale social networks. In *WWW*, pages 1275–1276, 2007.

[29] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *NIPS*, pages 505–512, 2002.

[30] X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger. Scan: A structural clustering algorithm for networks. In *KDD*, pages 824–833, 2007.

[31] Y. Zhou, H. Cheng, and J. Yu. Graph clustering based on structural/attribute similarities. *VLDB Endowment*, 2:718–729, 2009.