Information Sciences 314 (2015) 77–99

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Dense community detection in multi-valued attributed networks

Xin Huang [a,b,*], Hong Cheng [a], Jeffrey Xu Yu [a]

[a] The Chinese University of Hong Kong, Hong Kong, China
[b] University of British Columbia, Canada

ARTICLE INFO

Article history:
Received 25 September 2014
Received in revised form 23 February 2015
Accepted 30 March 2015
Available online 4 April 2015

Keywords:
Community detection
Dense subgraph
Attributed graph
Random walk with restart

ABSTRACT

The proliferation of rich information available for real world entities and their relationships gives rise to a general type of graph, namely *multi-valued attributed graph*, where graph vertices are associated with a number of attributes and a vertex may have multiple values on an attribute. It would be useful if we can cluster such graphs into densely connected components with homogeneous attribute values. Recent work has studied graph clustering in attributed graphs considering the full attribute space. However, full space clustering often leads to poor results due to irrelevant attributes.

In this paper, we study subspace clustering in multi-valued attributed graph and propose an algorithm SCMAG for community detection. Our algorithm uses a cell-based subspace clustering approach and identifies cells with dense connectivity in the subspaces. Random walk with restart is used to measure the structural connectivity and attribute similarity. An indexing scheme is designed to support efficiently calculating cell connectivity from random walk scores. We also propose a new cell combining strategy on dimensions of categorical attributes and a novel mechanism to handle multi-valued attributes. Experimental results on IMDB data and bibliographic data demonstrate that SCMAG significantly outperforms the state-of-the-art subspace clustering algorithm and attributed graph clustering algorithm. Case studies show SCMAG can find dense communities with homogeneous properties under different subspaces.

1. Introduction

Graph clustering has attracted a lot of attention recently in the literature. The goal of graph clustering is to group densely connected vertices into a cluster. Graph clustering has found broad applications in community detection, image segmentation, identification of functional modules in large protein–protein interaction networks, etc. Nowadays with rich information available for real world entities and their relationships, graphs can be built in which vertices are associated with a set of attributes describing the properties of the vertices. Such attributes can be numerical or categorical. For some attribute, an entity can have more than one value on that attribute. For example, the *genre* of a movie can be both "*drama*" and "*crime*"; the *research topics* of a researcher can be "*database*", "*data mining*" and "*bioinformatics*". Multi-valued attributes are very common in many real world data. This gives rise to a new and more general type of graph, called *multi-valued attributed graph*.

* Corresponding author at: University of British Columbia, Canada.
E-mail addresses: xhuang@se.cuhk.edu.hk (X. Huang), hcheng@se.cuhk.edu.hk (H. Cheng), yu@se.cuhk.edu.hk (J.X. Yu).

http://dx.doi.org/10.1016/j.ins.2015.03.075
0020-0255/

Traditional graph clustering methods [29,21,34,26,27] mainly focus on the connections in a graph and try to achieve a dense subgraph within a cluster. But these methods do not consider attribute information associated with graph vertices. On the other hand, a graph summarization method [31] partitions a graph according to attribute values, but does not enforce dense connections within a partition.

There have been some recent studies [6,37] on clustering attributed graph, i.e., SA-Cluster and its fast version Inc-Cluster, which partition the graph into several densely connected components with similar attribute values. SA-Cluster finds non-overlapping clusters in the full attribute space. Although SA-Cluster differentiates the importance of attributes with an attribute weighting strategy, it cannot get rid of irrelevant attributes completely. As graph vertices may have many attributes, the high-dimensional clusters are hard to interpret, or there is even no significant cluster with homogeneous attribute values in the full attribute space. If an attributed graph is projected to different attribute subspaces, various interesting clusters embedded in subspaces can be discovered which, however, may not exhibit in the full attribute space. In this paper, we study subspace clustering on multi-valued attributed graph, and discover clusters embedded in subspaces. Such subspace clusters should not only have homogeneous attribute values but also have dense connections, i.e., correspond to communities with homogeneous properties. The detected clusters in multi-valued attributed graphs can overlap, as the nodes may belong to multiple clusters in different subspaces. In contrast, non-overlapping community detection partitions the network into several disjoint clusters. Thus, overlapping communities are more common and natural than non-overlapping communities in attributed graphs. For example, an individual in a social network belongs to many social circles corresponding to different relationships, such as friends, schoolmates, family, research community and so on. Let us look at an example to illustrate the motivation for this subspace clustering problem.

Fig. 1(a) shows a coauthor network with three attributes {*Topic, Affiliation, Country*}. A vertex represents an author and an edge represents the coauthor relationship between two authors. The attributes and their possible values are listed in Table 1. *Topic* is a multi-valued attribute and an author can have one or more topics, e.g., author v_4. The problem is how to partition the coauthor network into clusters with close collaborations and homogeneous attribute values.

If we apply SA-Cluster [6], the coauthor network is partitioned into 4 clusters, as shown in Fig. 1(b). But this clustering result is not satisfactory, because (1) the attribute values in the same cluster are still quite different, if considering the full attribute space; and (2) v_9 is disconnected from his coauthors and forms a single-node cluster. This is not reasonable. In a high-dimensional attributed graph this problem may become even worse, as it is hard to find clusters with dense connectivity and homogeneous attribute values in the full space.

If we consider subspace clustering, then we can find two clusters under the subspace {*Affiliation, Country*} in Fig. 2(a), and another two clusters under the subspace {*Topic, Affiliation*} in Fig. 2(b). These subspace clusters make much more sense because they not only have homogeneous attribute values in the respective subspace, but also have a cohesive structure within a cluster. Note that, if traditional subspace clustering methods, e.g., CLIQUE [2] and ENCLUS [5] are applied, under the subspace {*Affiliation, Country*} we will have two more clusters $\{v_1, v_4, v_6\}, \{v_2, v_3, v_5, v_7\}$ respectively with attribute values [*Univ., AU*] and [*Univ., CN*]. However, we can easily find these two clusters have very sparse connections, thus not closely collaborating groups.

This example shows neither the recent attributed graph clustering algorithms [6,37] which consider the full attribute space, nor the traditional subspace clustering algorithms [2,5] which completely ignore the structural connectivity are suitable to solve the problem of multi-valued attributed graph clustering. The unique challenges in this problem include the following:

1. how to discover subspaces under which densely connected clusters with homogeneous attribute values are embedded? For example, there are meaningful clusters under the two subspaces shown in Fig. 2(a) and (b), but there is no cluster with dense structural connectivity under the subspace {*Topic, Country*};
2. how to properly handle categorical attributes and multi-valued attributes. Traditional subspace clustering algorithms, e.g., CLIQUE [2] and ENCLUS [5], only handle numerical attributes. But in real data categorical and multi-valued attributes are very common, for example, v_4 has two topics as *DB* and *DM*. Which cluster should v_4 belong to in Fig. 2(b)?
3. how to enforce both structural connectivity and attribute similarity requirements, like in the clusters in Fig. 2 (a) and (b)?

There have been some recent studies which combine attribute subspace clustering and dense subgraph mining on a graph with feature vectors, e.g., CoPaM [19], GAMer [12,13], DB-CSC [11]. CoPaM and GAMer follow a cell-based subspace clustering approach to find clusters that show high similarity in a feature subspace and are densely connected in the given graph, while DB-CSC takes a density-based approach to find subspace clusters. They have the different limitations: (1) CoPaM and GAMer do not have a cell merging strategy for adjacent cells, and the formation of a cell depends on the initial vertex to start with; (2) CoPaM strictly requires all nodes in a cluster have the same value on each attribute in the subspace. GAMer and DB-CSC use a single parameter *maximal width w* to control the value difference on *all the attributes* in the concerned subspace within a cell. But for different attributes, which can be categorical or numerical ones, it is hard to set a uniform threshold to control the attribute value differences; (3) The time complexity of GAMer and DB-CSC increases exponentially with the number of vertices in the input graph, which can hardly scale with large graphs. Even though adjusting by various parameter settings, GAMer is proven to be too slow to generate results in our experiments; and (4) According to the experimental results reported in [12,13,11], only clusters with small size are found by GAMer and DB-CSC on real datasets, i.e., the average size of the found clusters is only around 10 in a graph with 10k vertices. Similarly, according to [19], the largest found subgraph