Eb&D: A new clustering approach for signed social networks based on both edge-betweenness centrality and density of subgraphs

Xingqin Qi [a],*, Huimin Song [a], Jianliang Wu [b], Edgar Fuller [c], Rong Luo [c], Cun-Quan Zhang [c]

[a] *School of Mathematics and Statistics, Shandong University (Weihai), Weihai, 264209, China*
[b] *School of Mathematics, Shandong University, Jinan, 250100, China*
[c] *Department of Mathematics, West Virginia University, Morgantown, WV, 26506, USA*

HIGHLIGHTS

- A clustering method is presented for signed networks.
- This method is a graph theory based approach.
- It does not need the number of clusters in advance.
- It performs better than other existing method with applying on several data sets.

ARTICLE INFO

ABSTRACT

Clustering algorithms for unsigned social networks which have only positive edges have been studied intensively. However, when a network has like/dislike, love/hate, respect/disrespect, or trust/distrust relationships, unsigned social networks with only positive edges are inadequate. Thus we model such kind of networks as signed networks which can have both negative and positive edges. Detecting the cluster structures of signed networks is much harder than for unsigned networks, because it not only requires that positive edges within clusters are as many as possible, but also requires that negative edges between clusters are as many as possible. Currently, we have few clustering algorithms for signed networks, and most of them requires the number of final clusters as an input while it is actually hard to predict beforehand. In this paper, we will propose a novel clustering algorithm called Eb&D for signed networks, where both the betweenness of edges and the density of subgraphs are used to detect cluster structures. A hierarchically nested system will be constructed to illustrate the inclusion relationships of clusters. To show the validity and efficiency of Eb&D, we test it on several classical social networks and also hundreds of synthetic data sets, and all obtain better results compared with other methods. The biggest advantage of Eb&D compared with other methods is that the number of clusters do not need to be known prior.

© 2017 Elsevier B.V. All rights reserved.

---

* Corresponding author.
   E-mail address: qixingqin@163.com (X. Qi).

1. Introduction

Social network analysis has gained great attention in recent years. Commonly, vertices in a network represent people or individuals while the edges show relationships between them. In the usual setting, only positive relationships which denote proximity or similarity are considered. However, when a network has like/dislike, love/hate, respect/disrespect, or trust/distrust relationships, such a representation with only positive edges is inadequate as it fails to encode the sign of a relationship. This kind of networks can be modeled as signed networks, where edge weights can be either positive or negative, representing positive or negative relationships, respectively. Early uses of signed networks can be found in anthropology, where negative edges have been used to denote antagonistic relationships between tribes [1].

The most dominant theory related to signed networks is "social balance", whose intuition can be interpreted as "a friend of my friend is my friend", "an enemy of my friend is my enemy", and "an enemy of my enemy is my friend". A graph is *strongly balanced* if no part of the graph violates this intuition.

Cartwright and Harary [2,3] showed that a network which follows the strong balance notion can be clustered into **two** perfect antagonistic groups such that the edges within groups are positive and the edges between groups are negative. However, in real world not every network has the structure of *strongly balanced*, they may have more than two antagonistic groups. Davis [4] proposed the notion of "weak balance" as a generalization of social balance. They relax the balanced relationships by allowing "an enemy of one's enemy can still be an enemy".

Theorem 1 (*[4]*)**.** *A network is "weakly balanced" if and only if either* (i) *all of its edges are positive, or* (ii) *the nodes can be clustered into k groups such that the edges within groups are all positive and the edges between groups are all negative. This kind of network is also called* **perfect** *k-weakly balanced (where $k > 2$).*

But real signed networks are seldom perfect k-weakly balanced, thus for a given signed network, we want to partition it into small groups such that (1) positive edges within clusters are as many as possible and (2) negative edges between clusters are as many as possible. This problem is called *clustering for signed networks*.

Clustering algorithms for unsigned networks with only positive edges have been studied intensively. The most widely used algorithms are spectral clustering method [5], the KL (Kernighan–Lin) algorithm [6], GN (Girvan–Newman) algorithm [7], Newman algorithm [8], and HITS (hyperlink induced topic search) algorithm [9] etc. Clustering for unsigned networks also can be seen as a process to detect dense subgraphs in G. In [10], an approach named Quasi-clique Merger Algorithm was proposed to detect all dense subgraphs in G with various levels and construct a hierarchically nested system to illustrate their inclusion relationships. Yang et al. [11] gave a comprehensive survey of the clustering algorithm for unsigned networks, where clustering algorithms are partitioned into two classes: optimization based methods and heuristic methods. Particularly, Malliaros and Vazirgiannis [12] gave a survey for clustering directed networks.

But clustering algorithms for signed networks cannot be successfully carried out by merely extending the theory and algorithms for unsigned networks in a straightforward manner. Many notions and algorithms for unsigned networks break down when edge weights are allowed to be negative. There are fewer algorithms for signed networks than for unsigned networks. Doreian et al. [13] proposed a local search strategy which is similar to the Kernighan–Lin algorithm. Nikihil et al. [14] also considered the clustering problem for signed graphs, though their formulation is motivated by correlation clustering. Yang et al. [15] proposed an agent-based method which essentially conducts a random walk on the graph. Anchuri et al. [16] proposed hierarchical iterative methods that solve 2-way frustration and signed modularity objectives using spectral relaxation at each hierarchy. Another framework for clustering signed graphs was proposed by Chiang et al. [17] who introduced a new criterion called balance normalized cut, and they developed a multilevel clustering framework "S-Graclus" for signed networks, which "*can be viewed as a generalization of Graclus [18] for the signed network setting*"(see Ref. [17]).

However, most of these existing algorithms (especially spectral clustering methods) required the number of final clusters k as an input, that is, k should be known in advance before clustering process. But in practice we cannot initially predict the number of communities beforehand. As said in SAS/STAT 9.2 User's Guide "*There are no completely satisfactory algorithms that can be used for determining the number of population clusters for many type of cluster analysis*". Thus, clustering methods for signed network which do not need the number of clusters in advance are in urgent need.

In this paper, we will propose a novel clustering algorithm for signed networks which avoids using k as an input. Motivated by the GN (Girvan–Newman) algorithm [7] which constructs clusters by progressively removing edges which are least central (i.e., the edges which are most "between" clusters), we will look for the "least central" edges as well in a signed network, but we will also take the density of subgraphs into consideration. A hierarchically nested system are constructed to illustrate the inclusion relationships of clusters. This paper is organized as follows. Some terminologies and notations are introduced in Section 2; In Section 3, a new clustering algorithm for signed networks named Eb&D is presented; Its validity is shown by applying it on some classical real data sets and hundreds of synthetic data sets in Section 4; Conclusions are made in Section 5.

2. Terminologies and notations

A signed network $G = (V, E)$ can be represented as an adjacency matrix that describes relationships between entities. Formally, let $G = (V, E)$ be a graph with weight $w(e) \in \{-1, 0, 1\}$ on each edge, the adjacency matrix A is defined as