Physica A 416 (2014) 173–182

Contents lists available at ScienceDirect

Physica A

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

# Semi-supervised community detection based on discrete potential theory

Dong Liu [a,b,c], Xiao Liu [a,b], Wenjun Wang [a,b,*], Hongyu Bai [a,b]

[a] School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
[b] Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin 300072, China
[c] School of Computer and Information Technology, Henan Normal University, Xinxiang 453007, China

## HIGHLIGHTS

- A novel semi-supervised algorithm for community detection is proposed based on the discrete potential theory.
- The proposed algorithm is particularly suitable for networks with obscure community structure.
- The proposed algorithm owns near linear time complexity and can effectively find communities both in real-world networks and artificial networks.

## ARTICLE INFO

*Article history:*
Received 15 November 2013
Received in revised form 17 July 2014
Available online 29 August 2014

*Keywords:*
Semi-supervised
Community detection
Potential theory

## ABSTRACT

In recent studies of the complex network, most of the community detection methods only consider the network topological structure without background information. This leads to a relatively low accuracy. In this paper, a novel semi-supervised community detection algorithm is proposed based on the discrete potential theory. It effectively incorporates individual labels, the labels of corresponding communities, to guide the community detection process for achieving better accuracy. Specifically, a number of vertices with user-defined labels are first identified to act as unit elementary charges which can generate different electrostatic fields. Then, community detection can be translated into a potential transmission problem. By formulating the problem using combinational Dirichlet, labels of those unlabeled vertices can be determined by the labels for which the greatest potential is calculated. Finally, a better community partition can be obtained. Our extensive numerical experiments in both artificial and real networks lead to two key observations: first, individual labels play an important role in community detection; and second, our proposed semi-supervised community detection algorithm outperforms existing counterparts in both accuracy and time complexity, especially for obscure networks.



## 1. Introduction

In recent years network study has attracted a considerable amount of attention in various fields ranging from computer science, physics, information science, to mathematics and statistics [1]. Complex networks, such as biological network, social network, technological network and information network, are usually effectively modeled as graphs by regarding each entity as a vertex and each connection as an edge. A lot of evidence has shown that the network commonly owns the community structure, which is a significant feature of the complex network. Although no definition of community is universally accepted at present, we usually consider the community as a sub-network with many edges joining vertices of the same community

* Corresponding author at: School of Computer Science and Technology, Tianjin University, Tianjin 300072, China. Tel.: +86 15822908976.
*E-mail address:* xiaoxiao@tju.edu.cn (W. Wang).

http://dx.doi.org/10.1016/j.physa.2014.08.051
0378-4371/

while comparatively few edges joining vertices of different communities [2]. Such a community can be regarded as a fairly independent functional unit of a network, playing a similar role like, e.g., the protein in the cell of human body. Therefore, community detection becomes a heated topic and the community structure plays an important role in revealing the organization and function of the networks. Generally speaking, there are two types of communities: overlapping community and non-overlapping community. Many researchers have devoted great concerns and made great contributions to the overlapping community detection [3–5]. However, we mainly focus on non-overlapping community detection in this paper.

A lot of models and algorithms have been proposed for community detection, such as betweenness-based algorithms [2,6], modularity-based methods [7–11], Potts model based on the Markov process [12,13], spin model [14] and stochastic block-models [15]. However, these methods are a kind of unsupervised community detection methods. These work merely using the topological information of the network and ignoring its background knowledge. Actually, some prior information is of great value in identifying the community structure. For instance, a few proteins have been known to belong to certain functional classes in protein–protein interaction networks [16]. Therefore, how to combine the prior information with topological structure to guide the community detection is an interesting problem worth researching. Recently, a few semi-supervised community detection methods have been proposed. Ma et al. [17] proposed a semi-supervised method based on symmetric nonnegative matrix factorization, which incorporates pairwise constraints (via must-links and cannot-link) on the cluster assignments of vertices for identifying the community structure in a network. Eaton et al. [18] presented a semi-supervised algorithm based on the spin-glass model, which can incorporate prior knowledge in the forms of individual labels (via known cluster assignments for a fraction of vertices) and pairwise constraints into the process of extracting the community structure. Zhang [19,20] developed a method that implicitly encodes the pairwise constraints by modifying the adjacency matrix of the network, which can also be regarded as the de-noising process of the consensus matrix of the community structures. Although these existing semi-supervised community detection methods can improve the community identification accuracy, some of which have limitations in high time complexity or unable apply to networks with very obscure community structure.

The discrete potential theory on graph has been applied to various fields. Grady [21] exploited it for image segmentation. Zhang and Zhou [22] applied it to directed networks; they presented a new mechanism for the local organization of directed networks and designed the corresponding link prediction algorithm. Wang [23] came up with a semi-supervised clustering method based on generalized point charge models for text data classification. However, there are some differences among image data, text data and complex network. Thus, it will be of great value, applying the potential theory to detect the community structure of the complex network.

In this paper, we propose a novel semi-supervised community detection algorithm based on the discrete potential theory (SDPT). A few number of vertices with user-defined labels are first specified and acted as unit point charges, while the remaining unlabeled vertices are placed in the electrostatic fields generated by these charges. Then calculate the labels of the remaining vertices through solving a system of sparse linear equations, as described in Section 2.2. Finally, compared with the existing algorithm of Girvan and Newman [2], Markov clustering algorithm [24], Clauset's algorithm (CNM) [25] and Infomap algorithm [26], the experimental results on both real-world networks and synthetic LFR benchmark networks demonstrate the effectiveness of our approach, especially under the condition of obscure community structure.

The remainder of the paper is structured as follows. Section 2 provides a detailed description of our new semi-supervised community detection algorithm. Experimental results on real-world networks and artificial networks are presented in Section 3 and the paper ends with a conclusion in Section 4.

## 2. Semi-supervised community detection based on discrete potential

In this section, the specific process of our proposed SDPT algorithm is presented. This section describes three aspects of our proposed algorithm: restating problem and notions, establishing the system of equations to solve the problem and discussing the complexity of the algorithm.

### 2.1. Problem restatement and notions

We begin with defining a precise notion for a graph. A graph is mathematically defined as the pair $G = (V, E)$ where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of vertices and $E \subseteq V \times V$ represents the set of edges. It is frequently used to denote an unweighted and undirected network. A key point must be mentioned here that multiple edges and self-connections are not taken into account. Generally, a network can be expressed by its adjacent matrix $A$, where the element $A_{ij}$ of $A$ is equal to 1 if there is an edge between vertices $v_i$ and $v_j$ and 0 otherwise. The degree matrix $D$, which is a diagonal matrix containing the vertex degree $d(v_i)$ $(i = 1, 2, \ldots, n)$ of a graph on the diagonal, is also a common matrix used in the network analysis, that is

$$D = \begin{pmatrix} d(v_1) & 0 & \cdots & 0 \\ 0 & d(v_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d(v_n) \end{pmatrix}.$$