Physica A xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Physica A

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

Density shrinking algorithm for community detection with path based similarity

Q1 Jianshe Wu [a,b,*], Yunting Hou [a], Yang Jiao [a], Yong Li [a], Xiaoxiao Li [a], Licheng Jiao [a]

[a] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an, Shaanxi Province 710071, China

[b] Information and System Laboratory, Department of Electronic Engineering, University of New Orleans, USA

HIGHLIGHTS

- A density shrinking algorithm for community detection is provided.
- The performance is higher than a recent density shrinking algorithm and the runtime is obviously reduced.
- The main process of the proposed density shrinking algorithm is repeatedly finding and merging dense pairs.
- Compared with the recent density shrinking algorithm, analyses show that the improvement promotes the performance and reduces the runtime.
- The analytical results are verified by comparative experiments on large scale networks.

ARTICLE INFO

Article history:
Received 15 April 2014
Received in revised form 3 February 2015
Available online xxxx

ABSTRACT

Community structure is ubiquitous in real world complex networks. Finding the communities is the key to understand the functions of those networks. A lot of works have been done in designing algorithms for community detection, but it remains a challenge in the field. Traditional modularity optimization suffers from the resolution limit problem. Recent researches show that combining the density based technique with the modularity optimization can overcome the resolution limit and an efficient algorithm named DenShrink was provided. The main procedure of DenShrink is repeatedly finding and merging micro-communities (broad sense) into super nodes until they cannot merge. Analyses in this paper show that if the procedure is replaced by finding and merging only dense pairs, both of the detection accuracy and runtime can be obviously improved. Thus an improved density-based algorithm: ImDS is provided. Since the time complexity, path based similarity indexes are difficult to be applied in community detection for high performance. In this paper, the path based Katz index is simplified and used in the ImDS algorithm.

Q2

1. Introduction

Community structure is ubiquitous in many real world networks [1,2]. Finding the community structure is the fundamental of understanding those networked systems [3,4]. For example in social networks, finding the community structure is the key to analyze the relationship between people [5]. Q3

[*] Corresponding author at: Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an, Shaanxi Province 710071, China. Tel.: +86 29 88202279.
E-mail address: jshwu@mail.xidian.edu.cn (J. Wu).

http://dx.doi.org/10.1016/j.physa.2015.03.044
0378-4371/© 2015 Elsevier B.V. All rights reserved.

In the past decade, many algorithms for community detection have been proposed [6–16]. Among them, the most popular method is the modularity optimization presented by Newman and Girvan [17–21]. Higher value of modularity indicates a better partition of the network into communities. But modularity is not a scale independent standard, which indicates that algorithms based on maximizing modularity often lead to the resolution limit problem [22,23]. That is to say, when a community is less than a certain size then it may not be detected. Modularity maximization is an NP-complete problem [24], thus most modularity optimization algorithms are also time-consuming.

Density-based technique has been commonly used in data clustering [25,26]. In fact, the community detection can be also considered as a problem of data clustering. The network clustering algorithm SCAN [27] was extended from the traditional density-based clustering algorithm DBSCAN [26], which can detect meaningful clusters, hubs, and outliers in networks. The recently proposed graph-skeleton-based clustering (gSkeletonClu) algorithm is also a density-based network clustering algorithm, which projects a network to its core-connected maximal spanning tree [28]. Recent research shows that combining the density-based technique with the modularity optimization can overcome the resolution limit problem and an efficient algorithm can be obtained [29].

In Ref. [29], a density-based modularity optimization algorithm called DenShrink was provided, which was derived from the structural network clustering algorithm SCAN [27]. DenShrink is free from the resolution limit that most modularity-based algorithms suffer from. By repeatedly finding micro-communities (broad sense, see definition in Section 2) and merging them into super nodes, DenShrink is much faster than traditional modularity-based algorithms. The time complexity of DenShrink is $m\log(N)$ when using a similarity index of only neighbor information, where N is the number of nodes and m is the number of edges in the network. DenShrink is a promising algorithm for community detection.

But analyses show that DenShrink has two aspects that can be further improved:

(1) When using a similarity index with only neighbor information, e.g., the common neighbor (CN) index or the cosine (COS) index [30,31], there are many micro-communities (narrow sense, see definition in Section 2) to be merged (DenShrink uses the COS index). When using a more accurate index to improve the performance, e.g., the path based similarity indexes [32], there are seldom micro-communities in the network (this will be explained in Section 2). Finding micro-communities becomes inefficient and useless.
(2) Merging micro-community (narrow sense) may degrade the detection accuracy of the algorithm compared with merging only dense pairs (this will be explained in Section 2 in detail).

In this paper, an improved DenShrink algorithm is provided, denoted as ImDS for clarity. The contribution of this is as follows:

(1) The path based Katz index is simplified for computation and used in ImDS, which can improve the detection accuracy of the algorithm.
(2) The procedure of finding and merging micro-communities (broad sense) in DenShrink is replaced by only finding and merging dense pairs in ImDS, which can improve the detection accuracy as well as reduce the runtime.

Overall, the detection accuracy of ImDS is obviously improved and the runtime is reduced compared with the original DenShrink. The rest of this paper is arranged as follows. In Section 2, motivations of this paper are provided. Section 3 describes the details of the ImDS algorithm. Experiments are performed in Section 4. Section 5 is the conclusion of this paper.

2. Motivations

Some definitions and notions are described before the motivations, which are the bases of the proposed algorithm.

2.1. Preliminaries

Let $G(V, E, W)$ be a weighted undirected graph, where V, E, and W are the node, edge, and weight (similarity) sets respectively. $\Gamma(i)$ is the set of neighbors of node i including itself, in formula

$$\Gamma(i) = \{j \in V | \{i, j\} \in E\} \cup \{i\}. \tag{1}$$

The similarity of a pair of nodes is a measure about whether they should be allocated into the same community or not. The definition of similarity has a heavy influence on the performance of the algorithms [26,29]. Commonly used definitions of similarity include the local neighbor information based indexes, e.g. common neighbor (CN) and cosine (COS) [30,31], and the path (global information) based indexes, e.g. the Katz index [32]. Before our motivations are described, the definitions of the CN and COS indexes are introduced. Let $s(i, j)$ denote the similarity between nodes i and j.

Definition 1 (*CN Index [30]*)**.** In an undirected network, the common neighbor based similarity is defined as follows:

$$s_{CN}(i, j) = |\Gamma(i) \cap \Gamma(j)|, \tag{2}$$

where $|\Gamma|$ denotes the number of nodes in the set Γ. According to the definition of neighbor set (see (1)), if there is an edge between nodes i and j, the following formula holds

$$s_{CN}(i, j) = |\Gamma(i) \cap \Gamma(j)| \geq 2.$$

Please cite this article in press as: J. Wu, et al., Density shrinking algorithm for community detection with path based similarity, Physica A (2015), http://dx.doi.org/10.1016/j.physa.2015.03.044