Information Sciences 367–368 (2016) 874–889

Contents lists available at ScienceDirect

Information Sciences

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

Quantitative function and algorithm for community detection in bipartite networks

Zhenping Li[a], Rui-Sheng Wang[b], Shihua Zhang[c,*], Xiang-Sun Zhang[c,*]

[a] School of Information, Beijing Wuzi University, Beijing, China
[b] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA
[c] National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, CAS, Beijing, China

ARTICLE INFO

Article history:
Received 17 September 2015
Revised 6 May 2016
Accepted 6 July 2016
Available online 7 July 2016

Keywords:
Bipartite network
Community
Quantitative function
Integer programming
Label propagation algorithm

ABSTRACT

Community detection in complex networks is a topic of high interest in many scientific fields. A bipartite network is a special type of complex network whose nodes are decomposed into two disjoint sets such that no two nodes within the same set are adjacent. Many relationships in real-world systems can be represented by a bipartite network, such as predator-prey relationships, plant-pollinator interactions, and drug-target interactions. While community detection in unipartite networks has been extensively studied in the past decade, identification of modules or communities in bipartite networks is still in its early stage. Several quantitative functions have been developed for evaluating the quality of bipartite network divisions, however, these functions were designed based on null model comparisons and thus are subject to certain resolution limits. In this paper, we propose a new quantitative function called bipartite partition density for community detection in bipartite networks, and use some network examples to demonstrate that this quantitative function is superior to the widely used Barber's bipartite modularity and other functions. Based on the bipartite partition density, the bipartite network community detection problem is formulated into an integer nonlinear programming model in which a bipartite network can be partitioned into reasonable overlapping communities by maximizing the quantitative function. We further develop a heuristic adapted label propagation algorithm (BiLPA) to optimize the bipartite partition density in large-scale bipartite networks. BiLPA is efficient and does not require any prior knowledge about the number of communities in the networks. We conduct extensive experiments on simulated and real-world networks and demonstrate that BiLPA can successfully identify the community structures of bipartite networks.

1. Introduction

Many real-world systems can be modeled as complex networks [2], such as Internet [22], social networks [40], biological networks [53], and sensor networks [32]. Although these networks come from different fields, they all have modular structure or community structure [14,16,19,21,23,24,28,29,32,41,43,53–55,57]. A community in a complex network is often defined as a set of nodes which are densely connected with each other, but sparsely connected with nodes outside of the

[*] Corresponding authors.
E-mail addresses: zsh@amss.ac.cn (S. Zhang), zxs@amt.ac.cn (X.-S. Zhang).

http://dx.doi.org/10.1016/j.ins.2016.07.024
0020-0255/© 2016 Elsevier Inc. All rights reserved.

set. Studies have shown that community structure of complex networks is highly relevant to the organization and functions of the corresponding systems [58]. For example, the communities in World Wide Web correspond to webpages with similar topics. People in the same community of social networks often share common hobbies or interests. In biological networks, biomolecules from the same community tend to have similar functions. Nowadays, community structure information has been used in designing sensor networks to provide better services [32]. Moreover, community structure information can also be used for general data mining tasks [15] and network controlling [49]. Therefore, identification of community structure is critical not only for discovering what makes entities come together, but also for understanding the structural and functional properties of the whole network [13].

In the past decades, several quantitative functions have been proposed to evaluate the quality of a network partition or community detection. Based on these quantitative functions, a wide range of successful algorithms have been developed to discover the community structures in complex networks. Although most attention has focused on community detection in unipartite networks (see [16] and references therein), many real-world relations, such as paper-author, movie-actor, plant-pollinator, order-item, and event-attendee, can be better represented as bipartite networks [20]. In bipartite networks, nodes are divided into two disjoint sets (i.e., a bipartite network is composed of two types of nodes) such that no two nodes within the same set are adjacent. For example, in paper-author networks, there are two types of nodes: papers and authors. Edges only exist between papers and authors. Each paper connects to all of its authors. In movie-actor networks, each actor is connected to the films where he or she starred. Many biological networks such as protein-protein networks are naturally bipartite, where the nodes can be partitioned into two types: bait proteins and prey proteins.

So far there has been no agreement on a standard definition for community detection in bipartite networks. Some studies of bipartite networks depend on the one-mode projection of the original network into two unipartite networks. But some information of the original bipartite network will be lost in the projected networks. To overcome this drawback, several bipartite modularity functions have been proposed based on the assumption that a community is a bipartite subgraph composed of nodes of both types [4,5,37,50]. Since these definitions of bipartite modularity were designed based on the null model used in the Newman–Girvan modularity function, they all have the resolution limit issue [17,37] and fail to detect communities smaller than a detectable scale that depends on the size of a network and the interconnectivity of its communities. When the network size is sufficiently large, optimizing the above mentioned bipartite modularity functions favors network divisions with groups of small communities merged into larger communities, which may lead to ambiguities [33].

In this paper, we propose a normalized quantitative function – bipartite partition density for evaluating community partitions in bipartite networks. We formulate the community detection problem for a bipartite network into an integer nonlinear programming model in which a bipartite network can be partitioned into reasonable overlapping communities by maximizing its bipartite partition density. We show that the bipartite partition density function can overcome the resolution limits in bipartite community detection through theoretical and numerical analyses on some network examples. Then we design a heuristic algorithm (BiLPA) based on bipartite partition density for efficient community detection in large-scale bipartite networks. We demonstrate the effectiveness and efficiency of BiLPA by conducting extensive computational experiments on both simulated and real-world bipartite networks.

The rest of this paper is organized as follows: in Section 2, we mainly review some related work on bipartite community detection. Section 3 describes the new bipartite partition density function we developed; in Section 4, some examples are provided to show that the bipartite partition density function can improve resolution limits. The bipartite community detection problem is formulated into an integer nonlinear programming model in Section 5. Section 6 describes the heuristic algorithm BiLPA for bipartite community detection; Experimental results and analyses are presented in Section 7; and we conclude the paper in Section 8.

2. Related work

To evaluate the quality of a network partition or community structure, Newman and Girvan [42] introduced a modularity function Q, which measures the number of edges within communities as compared to a null model. In other words, a partition with high modularity should be the one that the number of edges within communities is significantly higher than random expectation. Modularity optimization has become a very popular method for community detection in the past decade. However, Q has been shown to have serious resolution limit issues [17,45]. It contains an intrinsic scale that depends on the total number of edges in a network which makes it fail to detect dense communities smaller than this scale [17]. Recently, Bagrow [3] reported that trees and treelike networks can have arbitrarily high values of modularity Q which contradicts the notion of communities as groups of nodes that are densely interconnected.

To overcome the above issues, several other quantitative functions have been defined [1,30,31,48,56]. For example, Li et al. [30] developed modularity density D based on the concept of average degree of a community. Sun et al. [48] presented a variation of modularity density D for evaluating the cohesiveness of a community. Ahn et al. [1] proposed the concept of link community detection and defined a quantitative function based on the average link density of link communities. Li et al. [31] improved Ahn's quantitative function for link community detection. Although these quantitative functions [1,30,31,56] show certain advantages over modularity function Q, they are designed for detecting communities in unipartite networks.

In the past several years, community detection in bipartite networks has attracted great interests as well [7,10–12,20,27,33–37,47,51,52]. Several quantitative functions and algorithms for this problem have been developed