Physica A 416 (2014) 386–399

Contents lists available at ScienceDirect

Physica A

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

CK-LPA: Efficient community detection algorithm based on label propagation with community kernel

Zhen Lin [a,b,*], Xiaolin Zheng [a,c], Nan Xin [a], Deren Chen [a]

[a] College of Computer Science & Technology, Zhejiang University, Hangzhou 310027, China
[b] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[c] School of Medicine, Stanford University, CA 94305, USA

HIGHLIGHTS

- Existing shortages of LPA-based algorithms are studied.
- A novel label propagation-based model with community kernel is proposed.
- Several strategies are adopted to solve existing defects.
- Analyses show CK-LPA approach performs in nearly linear time complexity.
- Performance of CK-LPA is better than recent related algorithms during experiments.

ARTICLE INFO

Article history:
Received 20 July 2014
Received in revised form 13 August 2014
Available online 16 September 2014

Keywords:
Community detection
Online social network
Label propagation
Community kernel
Graph mining

ABSTRACT

With the rapid development of Web 2.0 and the rise of online social networks, finding community structures from user data has become a hot topic in network analysis. Although research achievements are numerous at present, most of these achievements cannot be adopted in large-scale social networks because of heavy computation. Previous studies have shown that label propagation is an efficient means to detect communities in social networks and is easy to implement; however, some drawbacks, such as low accuracy, high randomness, and the formation of a "monster" community, have been found.

In this study, we propose an efficient community detection method based on the label propagation algorithm (LPA) with community kernel (CK-LPA). We assign a corresponding weight to each node according to node importance in the whole network and update node labels in sequence based on weight. Then, we discuss the composition of weights, the label updating strategy, the label propagation strategy, and the convergence conditions. Compared with the primitive LPA, existing drawbacks are solved by CK-LPA. Experiments and benchmarks reveal that our proposed method sustains nearly linear time complexity and exhibits significant improvements in the quality aspect of static community detection. Hence, the algorithm can be applied in large-scale social networks.

1. Introduction

With the rapid development of Web 2.0 and the rise of online social networks, finding community structures in large-scale user data has become a hot topic in network analysis. In networks, communities with the same structure, i.e., groups

* Corresponding author at: College of Computer Science & Technology, Zhejiang University, Hangzhou 310027, China. Tel.: +86 571 87951245.
E-mail addresses: linz@illinois.edu, nblinz@gmail.com (Z. Lin), xlzheng@zju.edu.cn (X. Zheng), xinnannan@gmail.com (N. Xin), drc@zju.edu.cn (D. Chen).

http://dx.doi.org/10.1016/j.physa.2014.09.023
0378-4371/

that consist of people with common interests, students in the same school, or researchers in a particular academic field, may exist. Such communities or small circles can reveal a number of hidden social relationships and explain many social phenomena.

The newly emerged social network has led to the rapid increase of research on this field. Compared with traditional networks, social networks present a clearer structure. Moreover, connections among individuals, such as friendships in social networks, reference relationships in paper networks, and links among blog networks, are evidently exposed. With regard to community structures in social networks, no universally recognized definition is yet available. At present, community structures in academic research typically refer to divisions based on nodes. Such reference causes the link density within communities to exceed that among communities [1].

However, social networks differ sharply from traditional networks. First, they exhibit more heterogeneous characteristics. Nodes and edges in social networks are typically of different types; therefore, processing these kinds of data is important. Moreover, the scale of social networks is significantly larger than that of traditional networks. For example, by the end of March 2014, active monthly Facebook users exceeded 1.3 billion,[1] whereas active daily users of Sina Weibo (a microblog service similar to Twitter that is popular in China) exceeded 66.6 million.[2] Therefore, processing technologies used in social networks are far more demanding in terms of efficiency. Lastly, social networks are more dynamic, i.e., their nodes and edges are constantly changing. Such dynamic evolution of social networks poses many challenges, such as timelines and robustness in processing technologies.

The label propagation algorithm (LPA) was first applied in the community detection field in 2007 [2]. This algorithm, which has become one of the most widely used algorithms in community detection, is practical and simple. In addition, it has been widely used in other fields, such as in text information classification, information retrieval in multimedia, and community detection.

However, some drawbacks, such as the formation of a "monster" community, weak robustness, and high randomness, remain in traditional LPA [3]. When community labels in a large community stop spreading, while labels in a small community near it have not yet spread, the labels of the large community will likely affect the labels of the small community, and the small community will be swallowed. After several rounds of such phenomenon, a monster community may be formed. Consequently, the quality of community detection becomes significantly low. Furthermore, the start node is randomly chosen, and the label is randomly selected when the label number of the nodes of a neighbor is the same. Therefore, the results exhibit high randomness, which considerably affects their stability. In addition, the number of iterations in traditional LPA is difficult to predict. The algorithm often iterates more times than expected.

In this study, we propose an LPA based on a community kernel that can rapidly detect communities to solve the aforementioned problems. We validate the algorithms on benchmark datasets and real social networks. The experiments show that the proposed algorithm is close to linear time complexity and can be applied to detect large-scale network communities.

The structure of this paper is as follows. Section 1 specifies the background and significance of fast community detection in social networks. Section 2 mainly introduces the research background of community detection algorithms in networks. Section 3 proposes the fast community detection algorithm based on the label propagation concept. The proposed algorithm is performed on benchmark and real datasets to verify its efficiency and accuracy in Section 4. Finally, Section 5 summarizes the paper.

2. Related works

Static community detection has always been a relatively popular research direction. To date, many relevant articles have applied traditional graph-related algorithms to community detection, which can be divided into many categories, such as graph partitioning [4–6], hierarchical clustering [1], and spatial clustering [7]. In recent years, as social networks develop, community detection has received increasing attention. Many algorithms, including dividing algorithm [8], clustering algorithm [9,10], modularity-based algorithm [8], semantic network-based algorithm [11], information cascade-based model [12], user interaction based algorithm [13], and dynamic algorithm [14], have been proposed.

Nearly all of these algorithms detect community structures by partitioning networks and fail to consider propagation characteristics. Moreover, the time complexity of most algorithms is high. Therefore, a community detection algorithm based on label propagation was first proposed in Ref. [2]. The proposed algorithm considers the network structure and propagation properties of social networks. Furthermore, it has a nearly linear time complexity. However, the proposed algorithm also has some drawbacks [3], such as the formation of a monster community, weak robustness, and high randomness, which have been improved by many experts [15–18]. Nevertheless, the time complexity of the algorithm increases rapidly.

With regard to the problems existing in the original LPA, several improved algorithms were proposed based on the original algorithm. To find overlapping communities, COPRA algorithm [19] and SLPA algorithm [20] were proposed. To partition massive graphs, many approaches, such as balanced label propagation [21], multivariate graph-based method [22], and parallel SLPA [23], were presented. To avoid losing network topology information, the weighted-CNP algorithm [24] was put forward. Moreover, a modularity-specialized label propagation algorithm (LPAm) was proposed to avoid the formation

1 http://www.statisticbrain.com/facebook-statistics/.

2 http://www.chinainternetwatch.com/tag/sina-weibo/.