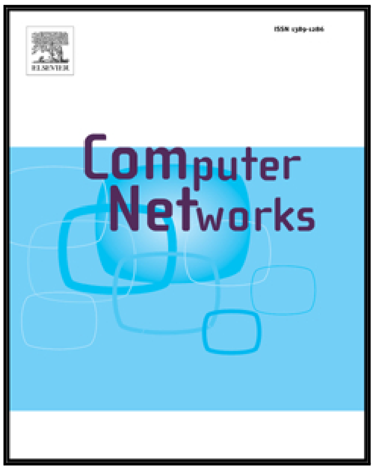
Accepted Manuscript



A social community detection algorithm based on parallel grey label propagation

Qishan Zhang , Qirong Qiu , Wenzhong Guo , Kun Guo ,
Naixue Xiong

PII: S1389-1286(16)30182-7
DOI: 10.1016/j.comnet.2016.06.002
Reference: COMPNW 5920

To appear in: *Computer Networks*

Received date: 30 October 2015
Revised date: 28 May 2016
Accepted date: 2 June 2016

Please cite this article as: Qishan Zhang , Qirong Qiu , Wenzhong Guo , Kun Guo , Naixue Xiong , A social community detection algorithm based on parallel grey label propagation, *Computer Networks* (2016), doi: 10.1016/j.comnet.2016.06.002

A social community detection algorithm based on parallel grey label propagation

Qishan Zhang[a], Qirong Qiu[a], Wenzhong Guo[b,c], Kun Guo[b,c*], Naixue Xiong[d]

[a]*School of Economics and Management, Fuzhou, 350116, China*
[b]*College of Mathematics and Computer Science, Fuzhou, 350116, China*
[c]*Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou 350116, China*
[d]*Department of Business and Computer Science, Southwestern Oklahoma State University, OK 74074, USA*

Abstract

Community detection is one of the important methods for understanding the mechanism behind the function of social networks. The recently developed label propagation algorithm (LPA) has been gaining increasing attention because of its excellent characteristics, such as a succinct framework, linear time and space complexity, easy parallelization, etc. However, several limitations of the LPA algorithm, including random label initialization and greedy label updating, hinder its application to complex networks. A new parallel LPA is proposed in this study. First, grey relational analysis is integrated into the label updating process, which is based on vertex similarity. Second, parallel propagation steps are comprehensively studied to utilize parallel computation power efficiently. Third, randomness in label updating is significantly reduced via automatic label selection and label weight thresholding. Experiments conducted on artificial and real social networks demonstrate that the proposed algorithm is scalable and exhibits high clustering accuracy.

Keywords: Community detection; Parallel computation; Label propagation

1. Introduction

As Web 2.0 applications rapidly develop, social networking has become an important means for people to share information and communicate with each other worldwide. Aside from analyzing network structures, detecting groups of people with common interests or behavior patterns also plays an important role in research on social networks. Unlike traditional networks, social networks are complex and have their own special characteristics. First, social networks are usually extremely large in scale. For example, active Facebook users have already reached one billion in 2014 [1]. The high number of vertices and edges of networks significantly affect processing capability. Second, social networks are typical heterogeneous networks that consist of different types of data from various sources. For example, people can share their favorite writings, pictures, songs, or movies on Facebook or Twitter. Therefore, handling heterogeneous data is a necessary function of modern tools for social network analysis. Finally, community structures hidden in social networks may be highly complex. Communities may overlap or may be hierarchical. For example, the circle of friends of a person is usually combined with those of his/her colleagues, or teachers working in a university can be grouped into different levels, such as colleges and departments.

Many community detection algorithms have been proposed to identify complex community structures in social networks [2]. Community detection appears to be similar to traditional clustering or graph partitioning. Thus, several effective clustering or graph partitioning algorithms have been applied in community detection. The Kernighan–Lin algorithm aims to minimize the difference between intra-edges and inter-edges to detect communities [3]. Spectral methods determine the minimum cut to separate a graph and recursively divide a network into sections where communities emerge naturally [4]. The critical modularity character of networks was discovered when hierarchical clustering was first applied by Grivan and Newman to detect communities [5]. Since its discovery, modularity has been used frequently to design community detection algorithms. For example, it can be utilized as an evaluation measure, an optimization objective, or a stopping criterion [2].

Early community detection algorithms cannot handle large social networks because these algorithms exhibit space complexity and require substantial time [1]. A promising algorithm, called the label propagation algorithm (LPA), was proposed recently [6]. This algorithm is particularly suitable for large social networks with complex and overlapping communities because of various reasons. First, the running time of LPA is linear to network size. Thus,

* corresponding author: gukn123@163.com (Kun Guo), xiongnaixue@gmail.com (Naixue Xiong)