Physica A 421 (2015) 233–248

Contents lists available at ScienceDirect

Physica A

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

Overlapping community identification approach in online social networks

Xuewu Zhang [a,*], Huangbin You [a], William Zhu [a,b], Shaojie Qiao [c], Jianwu Li [d], Louis Alberto Gutierrez [e], Zhuo Zhang [a], Xinnan Fan [a]

[a] College of Internet of Things Engineering, Hohai University, Changzhou 213022, China
[b] Minjiang Chair Professor of Computer Science, Minnan Normal University, Zhangzhou 363000, China
[c] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China
[d] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[e] Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA

HIGHLIGHTS

- A method for detecting overlapping communities in online social networks is proposed.
- We present a novel index (CLA) to evaluate the quality of detecting communities.
- We confirm that the overlapping nodes are biased.
- The improved closeness centrality is useful for dividing overlapping nodes.
- The proposed method has lower time complexity and higher division accuracy.

ARTICLE INFO

Article history:
Received 14 July 2014
Received in revised form 2 October 2014
Available online 13 November 2014

Keywords:
Online social network
Overlapping community
Closeness centrality

ABSTRACT

Online social networks have become embedded in our everyday lives so much that we cannot ignore it. One specific area of increased interest in social networks is that of detecting overlapping communities: instead of considering online communities as autonomous islands acting independently, communities are more like sprawling cities bleeding into each other. The assumption that online communities behave more like complex networks creates new challenges, specifically in the area of size and complexity. Algorithms for detecting these overlapping communities need to be fast and accurate. This research proposes method for detecting non-overlapping communities by using a CNM algorithm, which in turn allows us to extrapolate the overlapping networks. In addition, an improved index for closeness centrality is given to classify overlapping nodes. The methods used in this research demonstrate a high classification accuracy in detecting overlapping communities, with a time complexity of $O(n^2)$.

1. Introduction

Systems and relationships in the real world can be described by complex networks [1–3], and finding structures in these networks can play an important role in understanding the communities and the relationships. A fundamental observation that can be made in these networks, is that if connections between nodes are dense, then that may indicate the existence of a community, and sparse connections indicate otherwise. If we extrapolate the structure of complex networks to online

* Corresponding author.
E-mail address: lab_112@126.com (X. Zhang).

http://dx.doi.org/10.1016/j.physa.2014.10.095

Social Networks, where nodes represent individuals, then there is the potential to mine meaningful data about users and their communities.

Recent research has produced a large number of community detection algorithms with varied results [4–7]. Early research focused on non-overlapping communities, which worked on the assumption that a node belongs to a single community. One specific example is the Spectral Bisection Method which is based on the Laplace Matrix [8]. Another example is the Kernighan–Lin algorithm [9], which employs a greedy method. There are also many algorithms in this domain that use modularity, such as the GN algorithm, proposed by Girvan and Newman [10], detects community structures using an iterative removal of edges with high betweenness scores. One disadvantage of this method, however, is a high time complexity, $O(n^3)$ on a network with n nodes. Newman et al. proposed the FastNewman algorithm [11] and CNM algorithm [12]. The two algorithms are implemented similarly, but with the former having a runtime of $O(n(n+m))$, where n and m represent the number of nodes and edges respectively, and the latter is $O(n\log^2 n)$. The CNM algorithm ranks as one of the algorithms with the lowest time complexity as of current research.

The issue with many of these algorithms is that they are predicated on the assumption that a node can only belong to a single community. This assumption may not transfer over to real world applications, given that many users in social networks may belong to numerous communities. In response to this, there has been significant research proposing a number of overlapping community detection algorithms. Palla et al. [13] proposed the k-clique algorithm, where the adjacent k-cliques are combined, and the nodes belonging to several k-cliques represent an overlapping part of the community. Shen et al. [14] proposed a new metric for quantifying overlapping community structures, and a method to follow that communities can be detected through optimization of the index. Chen et al. [15] proposed an overlapping community detecting algorithm which searched an initial partial community from a node with maximal node strength, then adding tighter nodes so as to expand the partial community. Xie et al. [16] proposed Speaker–listener Label Propagation Algorithm (SLPA), it is an extension of the Label Propagation Algorithm (LPA) [17]. The approach applies the underlying network structure to detect both individual overlapping nodes and the entire overlapping communities.

Although social networks share many similar properties with complex networks, there are some aspects in which they differ. The network size is one of the main aspects. As the online social networks maintain a vast number of nodes generally. This necessitates a fast algorithm with a low time complexity. Accuracy is also important in overlapping community detection in social networks, this is because given the huge scale of networks, small errors can propagate quickly.

In this paper we propose a novel overlapping community detection algorithm for online social networks. The crux of the method implements a counter intuitive approach, using a CNM algorithm for detecting non-overlapping communities initially, for the purposes of analyzing nodes and their relationships to other communities. This step adds the node to the corresponding community based on given rules, which can in turn give us the overlapping communities. In order to mitigate the risk of biased results, we implement an index (improved closeness centrality) [18] to divide the overlapping nodes, which maximizes the potential of receiving more accurate results. In addition, we propose an intuitive method for evaluating the quality of the partition.

2. Related works

2.1. Problem statement

The ever increasing size of social networks necessitates algorithms for community detection that are both fast and accurate. Recent work has shown advances in detecting overlapping networks, but it remains an open research area. In this paper we propose an overlapping community detection algorithm which maintains a higher classification accuracy and a lower time complexity than the previously proposed algorithms for overlapping community detection. In addition, we use an index of closeness to divide the overlapping nodes, which helps to improve accuracy.

2.2. Basic concepts

In this section we formalize a few concepts which are necessary for realizing the process of overlapping community detection.

2.2.1. Overlapping community

Previously, we expressed the real-world social construct of a community as a collection of nodes and edges: where densely connected nodes represent communities and sparse otherwise. In an overlapping community there are some nodes which belong to multiple communities. Fig. 1 demonstrates this case, in which node 6 belongs to both communities $C1$ and $C2$.

2.2.2. Degree

Given a network $G(V, E)$ with n nodes and m edges, if $e_{uv} = 1$, then u and v are connected, otherwise, $e_{uv} = 0$, and there is no connection. So we can define the degree k_u of node u as:

$$k_u = \sum_{v \in V} e_{uv}. \qquad (1)$$