Physica A 421 (2015) 510–521

Contents lists available at ScienceDirect

Physica A

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

Overlapping community detection using neighborhood ratio matrix

Justine Eustace, Xingyuan Wang *, Yaozu Cui

Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China

HIGHLIGHTS

- Proposes vertex neighborhood ratio matrix which is used to represent relation between nodes.
- Proposes NRATIO (Neighborhood RATIO) algorithm.
- Detects overlapping communities.
- Experiments show that; proposed algorithm gives more accurate results compare to the existing close related algorithms.

ARTICLE INFO

Article history:
Received 23 September 2014
Received in revised form 24 October 2014
Available online 24 November 2014

Keywords:
Neighboring ratio
Complex networks
Community detection algorithms
Data mining
Peron clusters

ABSTRACT

The participation of a node in more than one community is a common phenomenon in complex networks. However most existing methods, fail to identify nodes with multiple community affiliation, correctly. In this paper, a unique method to define overlapping community in complex networks is proposed, using the overlapping neighborhood ratio to represent relations between nodes. Matrix factorization is then utilized to assign nodes into their corresponding community structures. Moreover, the proposed method demonstrates the use of Perron clusters to estimate the number of overlapping communities in a network. Experimental results in real and artificial networks show, with great accuracy, that the proposed method succeeds to recover most of the overlapping communities existing in the network.

1. Introduction

In network science, a community can be considered as a sub-graph, which is internally more densely connected than its outside connections. The detection of these modular structures corresponds with identifying many community related features in a networked system. For example, communities in biological networks can range from a set of frequently interacting proteins (protein–protein interaction) to predator–prey relations [1]. In communication networks a community may exist as a set of devices which more frequently transmit signals to each other than other connected devices [2]. Another example of a community related structure can be found when considering linguistic attributes, whereby a community may be a set of consonants sharing the same phonetic features [3].

Complex network communities, can be categorized as overlapping or non-overlapping community structures. Of particular interest to this work, is the overlapping community which is a notable feature in many networked systems. For instance, overlapping features can be observed in scientific collaboration networks in which scientists participate in multiple disciplines [4]. Similarly, in social networks, an individual can belong to multiple social groups. Hence the detection of overlapping communities in complex networks is a problem of significant importance.

* Corresponding author.
E-mail addresses: justineustace@yahoo.com (J. Eustace), wangxy@dlut.edu.cn (X. Wang), cyz3471@sina.com (Y. Cui).

http://dx.doi.org/10.1016/j.physa.2014.11.039
0378-4371/© 2014 Elsevier B.V. All rights reserved.

In the past decade, a number of overlapping community detection methods have been proposed. These include modularity based methods [5–7], spectral based methods [8–11] and matrix factorization based methods [12–14]. Matrix factorization methods such as Non-Negative Matrix Factorization (NMF) [15], can be used to classify nodes into corresponding communities. For example, Wang et al. [12] propose various NMF frameworks that can be used in overlapping community detection. Also, Zarei et al. [16], proposed a NMF-based method to detect overlapping communities using Laplacian matrix of a given network. NMF can also be used to detect communities on large networks [17]. However, the paramount drawback of such methods is, the number of communities must be known in advance, which is often not feasible.

To overcome the above mentioned challenge, several NMF-based method like, Bayesian NMF [13], Bounded Non-Negative Matrix Tri-Factorization [18] and Binary matrix factorization [19,20] have been proposed. Nodes in Bayesian NMF are classified into corresponding communities using Bayesian NMF and the number of communities present in the network is defined as the inner rank of network relation graph. Bounded Non-Negative Matrix Tri-factorization [18], uses the stated method to detect overlapped communities. Binary matrix factorization, such as Symmetric Binary Matrix Factorization (SBMF) [19,20] uses optimized NMF methods on binary matrices to detect communities in the network. For instance, Zhang et al. [20] proposed an overlapping community detection method using SBMF. In SBMF [20], partition density [21] is used to compute the number of communities present in the network. Although these methods can be extended in link communities [22,23], they are still characterized by limited resolution and high computation complexity.

This paper addresses the above mentioned challenges. The accuracy of NMF-based method is enhanced by the use of a vertex Neighborhood Ratio Matrix. This matrix is a modified adjacency matrix of a given network in which two nodes are connected only if they possess more than the average number of neighbors of all nodes present in the network. This matrix reduces the influence of unrelated nodes in community structures during community detection. This paper also uses Perron clusters to determine the number of communities present in a network. Unlike other methods, the number of communities detected using Perron clusters, can be determined in one iteration. This significantly reduces computational complexity of the method herein. Finally, the performance of the proposed method is realized by conducting experiments on real, artificial and random networks.

The proposed algorithm detects crispy partition in overlapped and non-overlapped networks. As a results, lack of preliminary knowledge about the structure of the network [24,25] and in Ref. [26] does not affect the performance of the proposed algorithm.

The upcoming 2 sections of this paper are organized as follows; Section 2.1, formalizes the symbols used and introduces the problem presented in this paper. Sections 2.2 and 2.3, present the methods of Neighborhood Ratio and Non-Negative Matrix Factorization for estimating the number of communities present in the network. The algorithm proposed in this paper is presented in Section 2.4. Experiments and results are presented in Section 3. Finally, the work is concluded in Section 4.

2. Model formulation

2.1. Motivation

Considering the network model in Fig. 1, the 7 gray nodes share more neighbors between each other, than the rest of the unshaded nodes. If node $c(v_0)$ is chosen as a reference node, gray nodes would represent the closer related neighbors of node $c(v_0)$ than the unshaded nodes. Each gray node shares more neighbors with other gray nodes than unshaded nodes. By using a transitive property, this effect can be explored by propagation throughout the network whereby each node is considered as a reference vertex.

From the general definition of a community structure, nodes that have a large number of neighbors in common, are more likely to belong to the same community. Therefore we can assume that; all gray nodes belong to the same community structure. As such, this can be taken as the refined adjacency list of node $c(v_0)$. Though node 12 is adjacent to node $c(v_0)$, it is, in fact unrelated node to $c(v_0)$. A similar case is observed with node 5, which is connected to only two gray nodes (4 and 6) and other 3 connectivity is gray color unfilled nodes. Hence node 5, does not belong to the same group as the other gray nodes. From Ref. [27], unrelated nodes can adversely effect the accuracy of the detected community. Therefore, refining the adjacency list serves as an effort to minimize this effect.

Suppose that $G = (V, E)$ is an undirected network, where $V = \{v_1, v_2 \dots, v_n\}$ is a non-empty set of n vertices, $\boldsymbol{E}$, is the set of edges $e_{ij} \in \boldsymbol{E}$, such that each edge connects vertices v_i and v_j. The value of $n = |V|$ and $m = |E|$ is the total number of vertices and edges respectively, that are present in a network. For each node $v_i \in V$, $N(v_i)$ is a set of all vertices adjacent to v_i. In other words, $N(v_i)$ is the Neighborhood Set, of vertex v_i. The value of $\delta(v_i)$ denotes the degree of the vertex v_i. Adjacency matrix, $\boldsymbol{A}$, of a graph, G, represents a relation between nodes where, $\boldsymbol{A}_{ij} = 1$, if there is an edge between v_i and v_j and $\boldsymbol{A}_{ij} = 0$ otherwise.

2.2. Neighborhood ratio matrix

The relation of nodes is organized by using the principle of overlapping neighborhood ratio. This ratio measures the connectivity between nodes that are expected to be in the same community. It is an intersection ratio between the neighborhood