Future Generation Computer Systems 89 (2018) 265–285

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

# A parallel self-organizing overlapping community detection algorithm based on swarm intelligence for large scale complex networks

Hanlin Sun [a,b], Wei Jie [c], Jonathan Loo [c], Lizhe Wang [d,e,*], Sugang Ma [a,b], Gang Han [f], Zhongmin Wang [a,b], Wei Xing [a,b]

[a] School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, China
[b] Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, China
[c] School of Computing and Engineering, University of West London, UK
[d] Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, China
[e] School of Computer Science, China University of Geosciences, China
[f] Department of Electronic Science and Technology, Northwestern Polytechnical University, China

## HIGHLIGHTS

- Swarm intelligence is a feasible framework for community evolution simulating.
- A distributed parallel overlapping community detection algorithm is proposed.
- Extended modularity is not practicable for algorithm performance comparison.
- Extended modularity density is not practicable, either.
- The structural communities of three large real-world networks are analyzed.

## ARTICLE INFO

*Article history:*
Received 24 January 2018
Received in revised form 25 April 2018
Accepted 27 May 2018
Available online 2 July 2018

*Keywords:*
Overlapping community detection
Community structure analysis
Complex network analysis
Swarm intelligence
Parallel network analysis

## ABSTRACT

Community detection is a critical task for complex network analysis. It helps us to understand the properties of the system that a complex network represents and has significance to a wide range of applications. Though a large number of algorithms have been developed, the detection of overlapping communities from large scale and (or) dynamic networks still remains challenging. In this paper, a Parallel Self-organizing Overlapping Community Detection (PSOCD) algorithm ground on the idea of swarm intelligence is proposed. The PSOCD is designed based on the concept of swarm intelligence system where an analyzed network is treated as a decentralized, self-organized, and self-evolving systems, in which each vertex acts iteratively to join to or leave from communities based on a set of predefined simple vertex action rules. The algorithm is implemented on a distributed graph processing platform named Giraph++; therefore it is capable of analyzing large scale networks. The algorithm is also able to handle overlapping community detection well because a vertex can naturally joins to multiple communities simultaneously. Moreover, if some vertexes and edges are added to or deleted from the analyzed network, the algorithm only needs to adjust community assignments of affected vertexes in the same way as its finding joining communities for a vertex, i.e., it inherently supports dynamic network analysis. The proposed PSOCD is evaluated using a number of variety large scale synthesized and real world networks. Experimental results indicate that the proposed algorithm can effectively discover overlapping communities on large-scale network and the quality of its detected overlapping community structures is superior to two state-of-the-art algorithms, namely Speaker Listener Label Propagation Algorithm (SLPA) and Order Statistics Local Optimization Method (OSLOM), especially on high overlapping density networks and (or) high overlapping diversity networks.



* Corresponding author.
*E-mail addresses:* sunhanlin@xupt.edu.cn (H. Sun), wei.jie@uwl.ac.uk (W. Jie), Jonathan.Loo@uwl.ac.uk (J. Loo), LZWang@ceode.ac.cn (L. Wang), msg@xupt.edu.cn (S. Ma), hangang668866@163.com (G. Han), zmwang@xupt.edu.cn (Z. Wang), xingwei@xupt.edu.cn (W. Xing).

https://doi.org/10.1016/j.future.2018.05.071
0167-739X/

## 1. Introduction

A lot of complex systems, such as the World Wide Web, mobile communication networks, online social networks, power grids, traffic road networks and so on, are often modeled as complex networks to investigate. A complex network usually shows some interesting properties such as high network transitivity, power-law degree distribution, small world, scale free, the existence of community structures, and much more. The study of community structures can help us to understand those systems at a mesoscopic level, just between the macroscopic level in which the whole system is considered and the microscopic level in which each node is analyzed individually. In addition, the analysis of community structures has significance to many applications. For example, community structure analysis can be used in social networks (e.g. Facebook) which presents relationships between members. The analysis of such networks will help to design reliable friend recommendation systems. As another example, community structure analysis can be used in detecting communities of customers with similar purchasing interest in e-business networks. This can lead to setting up efficient product recommendation systems and thus improving business opportunities for product retailers.

An exact definition of a community depends on the underlying problem and its application, thus there is no a unanimous definition. For example, the definition could be based on degrees of vertexes [1], k-cliques [2], k-clans [2], k-clubs [2], etc. Filippo et al. [1] gave out the definition of strong sense community and weak sense community according to member connection strength. Michele et al. [3] proposed a number of meta definitions. Intuitively, a community is a group of vertexes in a network that has more edges (connections) among its members but comparatively has less edges between its members and the rest of the network vertexes. This simple concept is the core of nearly all community definitions.

The properties of very complex networks induce three staple challenges for a community detection algorithm: (1) overlapping community structure detection, especially from a high overlapping density network of which a large percent of vertexes are overlapping vertexes, and (or) from a high overlapping diversity network of which an overlapping vertex belongs to a great number of communities; (2) large scale network analysis, e.g., the number of vertexes and edges could reach the scale of several millions and even more; and (3) dynamic changing of the analyzed networks topology, i.e., a number of vertexes and edges could appear or disappear frequently. The problem about large scale and dynamic networks is how to find community structures within them quickly with as less effort as possible. Designing efficient algorithms to meet these problems remains challenging.

In this paper, we develop the Parallel Self-Organizing Community Detection (PSOCD) algorithm based on the idea of swarm intelligence (SI) to further near to a final solution. Swarm intelligence is the collective behavior of decentralized and self-organized systems, either natural or artificial. An SI system generally consists of a large number of simple individuals who can only perform simple actions and interact with nearby neighbors as well as with the system existing environment. Intelligence will emerge as a consequence of the sum of these simple actions and interactions. The main innovations and contributions of this paper are:

(1) We proposed the PSOCD algorithm applying concepts of SI. In PSOCD, an analyzed network is modeled as a SI system in which each vertex as an individual decides its own actions, i.e. leaving its original communities or joining into new communities, depending on a set of predefined simple action rules. Eventually an optimal community structure will emerge whilst each vertex acts iteratively. A vertex is naturally allowed to join to multiple communities, thus it is able to find overlapping community structures. The algorithm inherently supports dynamic network analysis very well; in that, if new vertexes and edges are added to the analyzed network, or existing vertexes and edges are deleted, it only needs to adjust community associations of affected vertexes in the same way as finding their previous joining communities.

(2) We implemented the PSOCD in a distributed manner on the parallel graph processing platform Giraph++ [4], leveraging the properties of SI, namely distributed, self-organizing, and self-evolving. As a result, it is capable of handling large scale network analysis.

(3) We found that the extended modularity or modularity density [5] for evaluating overlapping community structure quality should not be used as a metric for performance comparison of different algorithms. The reason is that the incorporation of overlapping properties in the metrics has great impact on its values, thus may lead to a quite opposite conclusion.

(4) We applied the PSOCD to analyze structural communities of three large scale real world networks and got reasonable results, which are closer to the functional communities reported in previous studies than those discovered by the two compared state-of-art algorithms.

The remainder of this paper is structured as follows: In Section 2, some most related overlapping community detection algorithms are briefly reviewed. In Section 3, the design of the PSOCD algorithm is outlined. A current implementation of the proposed algorithm on the platform Giraph++ and its computational complexity are described in Section 4. In Section 5, the evaluation results of the algorithm for a number of synthesized and real large complex networks are presented, and the limits of the current implementation are discussed. Finally, Section 6 concludes the paper.

## 2. Related works

Community structure analysis of complex networks has attracted much interest, and a number of algorithms originating from different fields, such as physics, statistics, data mining, evolution computation and many more have been proposed. There are many different strategies behind these community detection algorithms, such as divisive hierarchy, agglomerative hierarchy, random walking, information diffusion, spectrum analysis, statistical inference and so on. Several comprehensive reviews of these methods have been conducted, for example, a survey of community discovery methods was provided with a special focus on techniques designed by statistical physicists [6]. Meta definitions of a community in complex network were given and majority community discovery methods were summed up based on their own definitions [3]. Overlapping community structure analysis algorithms were reviewed in [7,8], while those for social network analysis were reviewed in [9]. The performance of a number of algorithms were compared in [8,10]. In this section, we briefly review some algorithms most related to our work.

### 2.1. LPA

The label propagation algorithm (LPA) is currently the fastest algorithm for community structure analysis, with a near-linear computational complexity. The idea is that, as information propagates on a network with a community structure, it will have a high probability flowing within a community. At first, each vertex is assigned a label, indicating the community to which it belongs, then each vertex sends its label to its neighbors and selects a label received from neighbors, e.g., the label observed most frequently, as its new label. By iteratively propagating labels among neighboring vertexes, the community structure will gradually emerge. Assuming that a vertex is able to hold more than one label, LPA