Physica A 468 (2017) 53–69

Contents lists available at ScienceDirect

Physica A

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

Exploring triad-rich substructures by graph-theoretic characterizations in complex networks

Songwei Jia [a,b], Lin Gao [a,*], Yong Gao [c], James Nastos [d], Xiao Wen [a], Xindong Zhang [a], Haiyang Wang [a]

[a] *School of Computer Science and Technology, Xidian University, Xi'an 710071, China*
[b] *School of Software, Xidian University, Xi'an 710071, China*
[c] *Department of Computer Science, University of British Columbia, Okanagan, Kelowna, British Columbia, Canada V1V 1V7*
[d] *Department of Computer Science, Okanagan College, Kelowna, British Columbia, Canada BC V1Y 4X8*

HIGHLIGHTS

- The novel triad-rich assumption.
- The definition of 2-club substructure.
- The edge niche centrality.
- The 2-hop overlapping strategy.
- The effective algorithm DIVANC.

ARTICLE INFO

Article history:
Received 8 March 2016
Received in revised form 27 July 2016
Available online 12 October 2016

Keywords:
Complex networks
Community detection
Triad-rich substructure
Graph-theoretic characterizations
Edge niche centrality
Overlapping

ABSTRACT

One of the most important problems in complex networks is how to detect communities accurately. The main challenge lies in the fact that traditional definition about communities does not always capture the intrinsic features of communities. Motivated by the observation that communities in PPI networks tend to consist of an abundance of interacting triad motifs, we define a 2-club substructure with diameter 2 possessing triad-rich property to describe a community. Based on the triad-rich substructure, we design a DIVision Algorithm using our proposed edge Niche Centrality DIVANC to detect communities effectively in complex networks. We also extend DIVANC to detect overlapping communities by proposing a simple 2-hop overlapping strategy. To verify the effectiveness of triad-rich substructures, we compare DIVANC with existing algorithms on PPI networks, LFR synthetic networks and football networks. The experimental results show that DIVANC outperforms most other algorithms significantly and, in particular, can detect sparse communities.

1. Introduction

One of the most important problems in complex networks is how to detect communities accurately [1]. Communities are the subsets of vertices with real physical sense. For example, in biological networks they are referred to as various biological functional modules such as protein complexes, GO terms and pathways; in social networks, communities may be various

[*] Corresponding author.
E-mail address: lgao@mail.xidian.edu.cn (L. Gao).

http://dx.doi.org/10.1016/j.physa.2016.10.021

social circles such as groups of people with common interests, etc. Traditional communities, which are typically described as dense subgraphs (subnetworks) explicitly or implicitly. The underlying assumption is that objects in some communities really tend to interact more frequently than in other regions of the network. Around the issue of how to detect communities, scholars have proposed many popular algorithms based on traditional community definitions which can identify parts of communities successfully at a certain degree. Examples of algorithms that detect communities by dense subnetworks include (i) random-walk based methods such as MCL [2] and INFOMAP [3]; (ii) seed-growing methods such as MCODE [4] and ClusterOne [5]; (iii) algorithms based on clustering, optimization, or statistical techniques such as LinkComm [6], LOUVAIN [7], and OSLOM [8]; and (iv) algorithms based on deeper graph-theoretic features such as EPCA [9–11].

While traditional definitions can offer some insight into some of the structure of communities, more and more recent studies show that these intuitions are unreliable [12–16]. Some of these examples include: overlapping communities have a higher density of links in the overlapping parts than in the non-overlapping ones, which are in contrast with the common picture of communities [16]; there is a paradox that the detection of well-defined communities is more difficult than the identification of ill-defined communities [12]. All of these counterintuitive evidences hint at the necessity of modifying the general defining characteristics of traditional communities. While there is a general consensus on the fact that there is a need for an adjustment of the notion of community or clusters, there is no clear direction to a remedy. Scholars [15] point out that there are two possible scenarios for filling the gaps between traditional definitions and communities. One is to include additional topological features in refining the traditional definitions beyond the standard measures of link density, degree correlations or density of loops, etc.; the other is to add requirements based non-topological knowledge, such as domain-specific background knowledge [17–19] for the detection of communities. However, in the former case, solely adjusting the structural conditions sought for may still not obtain satisfying results as the essence of communities in all contexts may not be characterized by equivalent topology. In the latter case, adding various domain-specific background knowledge may be effective on a limited number of cases, but the reliance on rigid domain-specific knowledge makes the resulting algorithms unlikely to exhibit scalability or transferability to other domains.

An ideal paradigm for fully analyzing communities would include identification of communities via certain specific intrinsic features, combined with a method for capturing deeper domain-specific structure in a general topological framework on which one can further develop algorithms. Here, we develop such a new framework by incorporating a novel and more subtle assumption based on graph-theoretic properties of communities and design efficient computing procedures to detect non-overlapping and overlapping substructures that have the desired properties. As shown in Fig. 1(a) and (b), both of the communities 'nuclear origin of replication recognition complex' [20] (dense) and 'GID complex' [21] (sparse) consist of abundantly interacting triad motifs, for instance. More details about the two complexes can be found in Appendix A. Motivated by the observation that communities in a PPI network are either quite dense or quite sparse and tend to consist of an abundance of interacting triad motifs [22–25], we define a 2-club substructure with diameter 2 possessing triad-rich property to describe a community. Based on the triad-rich substructure, we design a DIVision Algorithm using our proposed edge Niche Centrality DIVANC to detect communities effectively in complex networks. We also extend DIVANC to detect overlapping communities by proposing a simple 2-hop overlapping strategy. To verify the effectiveness of triad-rich substructures, we compare DIVANC with existing algorithms on PPI networks, LFR synthetic networks and football networks. The experimental results show that DIVANC outperforms most other algorithms significantly and, in particular, can detect sparse communities.

The rest of the paper is organized as follows. In Section 2, we present our framework for detecting communities. After discussing our datasets and providing statistical evidence that motivates and supports our triad-rich assumption about communities in Sections 2.1 and 2.2, we give a formal definition of a 2-club substructure in Section 2.3. In Section 2.4, we discuss the details of our algorithm for 2-club substructure detection, including a new edge-centrality measure specifically designed for 2-club substructures as well as a 2-hop-based strategy for extracting overlapping 2-club substructures. In Section 3, we report and discuss our experimental results. In Section 4, we conclude the paper and give closing discussion.

2. Methods

2.1. The datasets

We apply our framework on PPI networks [26,27], LFR synthetic networks [28,29] and football networks [30,31]. In the following, we give details about the relative networks and six golden standard sets of communities in PPI networks, respectively.

S. cerevisiae PPI networks (*Sce*DIP) are obtained from DIP [26] and *H. sapiens* PPI networks (*Hsa*HPRD) are extracted from HPRD [27]. For *Sce*DIP, we use the sets from the Munich Information Center for Protein Sequences (MIPS) [32], Saccharomyces Genome Database (SGD) [33] and *S. cerevisiae* GO terms (*Sce* GO term) as golden standards [34,35]. For *Hsa*HPRD, we use the sets of Human Protein Complex Database with a Complex Quality Index (PCDq) [36], Comprehensive Resource of Mammalian Protein Complexes (CORUM) [37] and *H. sapiens* GO terms (*Hsa* GO term) [34,35] as golden standards. *Sce*DIP consists of 4980 proteins and 22 076 interactions; *Hsa*HPRD consists of 9269 proteins and 36 917 interactions. The GO terms are not composed of all the terms but the high-level GO terms whose information content is more than 2 [34,35]. The definition of the information content (IC) of a GO term g is $IC = -\log(|g| / |root|)$ as given in the literature [34], where '*root*' is the corresponding root GO terms across the three aspects of molecular function (MF), biological process (BP) or cellular