Physica A 436 (2015) 878–896

Contents lists available at ScienceDirect

Physica A

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

Discovering natural communities in networks☆

Angsheng Li[a,*], Jiankou Li[a,b], Yicheng Pan[a,c]

[a] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, PR China
[b] University of Chinese Academy of Sciences, Beijing, PR China
[c] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, PR China

HIGHLIGHTS

- We proposed an information theoretical measure of complexity of networks, namely, the structure entropy of networks.
- We proposed a novel algorithm for detecting communities of networks by structure entropy minimization.
- We verified that our algorithm identifies or approximates the natural communities of networks both by models and nature evolving.
- We found that the communities found by our algorithm are balanced, with modularity comparable or larger than that by existing algorithms.

ARTICLE INFO

Article history:
Received 5 January 2015
Received in revised form 9 February 2015
Available online 12 May 2015

Keywords:
Community detection
Networks
Modularity
Structure entropy
Natural community

ABSTRACT

Understanding and detecting natural communities in networks have been a fundamental challenge in networks, and in science generally. Recently, we proposed a hypothesis that homophyly/kinship is the principle of natural communities based on real network experiments, proposed a model of networks to explore the principle of natural selection in nature evolving, and proposed the measure of structure entropy of networks. Here we proposed a community finding algorithm by our measure of structure entropy of networks. We found that our community finding algorithm exactly identifies almost all natural communities of networks generated by natural selection, if any, and that the algorithm exactly identifies or precisely approximates almost all the communities planted in the networks of the existing models. We verified that our algorithm identifies or very well approximates the ground-truth communities of some real world networks, if the ground-truth communities are semantically well-defined, that our algorithm naturally finds the balanced communities, and that the communities found by our algorithm may have larger modularity than that by the algorithms based on modularity, for some networks. Our algorithm provides for the first time an approach to detecting and analyzing natural or true communities in real world networks. Our results demonstrate that structure entropy minimization is the principle of detecting the natural or true communities in large-scale networks.

☆ All authors are partially supported by the Grand Project "Network Algorithms and Digital Information" of the Institute of Software, Chinese Academy of Sciences, by an NSFC Grant No. 61161130530, and by a China Basic Research Program (973) Grant No. 2014CB340302. Yicheng Pan was partially supported by China Postdoctoral Science Foundation Grant No. 2014M550870.

* Corresponding author.
E-mail address: angsheng@ios.ac.cn (A. Li).

http://dx.doi.org/10.1016/j.physa.2015.05.039
0378-4371/

1. Introduction

Natural or true communities are basic to many interacting systems in nature, society and networks. Identifying and analyzing natural communities of real world networks are essential to understanding the networks, with potential applications in understanding, for instance: the roles and functions of the modules of social and technological systems, the roles and mechanisms of social groups in nature and society, the mechanisms of group intelligence, the mechanisms of interacting learning and games among social groups, diagnosing and curing of complex diseases, and designing of new medicines etc. Our algorithm provides for the first time a method which may exactly identify or precisely approximate the natural or true communities of many real world networks and interacting systems in nature and society. Equally important, our algorithm also provides an efficient method for cyberspace information compression, allowing efficient access of knowledge in large-scaled cyberspace.

Real world networks reflect the organizations of order and regulations in the real world. The order and regulations give rise to some common phenomena of real graphs such as the phenomenon of the heavy tail degree distribution, the property of the high clustering coefficients and the property of the small average shortest paths etc. [1–3].

Topological properties of networks are usually characterized by homogeneity and heterogeneity, which are measured by the degree centrality, the closeness centrality, and the betweenness centrality, etc., which have already been extensively studied [4–9].

Community detection is an important tool to explore the network structure, and to extract useful information from a network [10,11]. Leskovec et al. [12] analyzed community structure in large real networks and tried to find the "best" communities at various sizes, showing that the "best" communities seem to have sizes no more than 100 nodes, and that large subsets of a graph are not well-defined communities. Andersen, Chung and Lang [13] proposed an efficient algorithm to compute an approximate pagerank.

Li and Peng [14] proposed a definition of communities. Given a graph $G = (V, E)$, and a set $S \subset V$, we say that S is a community of G, if the induced subgraph of S in G is connected, and the conductance of S in G is (bounded by a number) inversely proportional to a power of the size $|S|$ of S, that is, $\Phi(S) = O(\frac{1}{|S|^{\beta}})$ for some constant β. This means that if S is large, then it is unlikely that $\Phi(S) = O(\frac{1}{|S|^{\beta}})$ due to the fact that $O(\frac{1}{|S|^{\beta}})$ is extremely small, so that S is unlikely to from a well-defined community. It has been shown that for each of the classical models, $\mathcal{M}$ say, either networks generated by $\mathcal{M}$ are rich in such communities, or are free of the communities [14,15].

Kumar et al. [16] found that the WWW graph is rich in bipartite cliques. Newman [7,17] studied the patterns of cooperations from the co-authorship graphs, showing that most authors cooperate with a few, one or two, long term collaborators, instead of with many authors evenly distributed. In addition, distributions of the sizes of communities by certain community detection algorithms, on some networks, have been found to be skewed [18,19].

Newman and Girvan [20] defined the notion of modularity to measure the quality of community structure of a network. The notion of modularity has become the major measure of quality of community structures of networks, leading to numerous extensions and modifications of algorithms for finding communities in graphs [11].

The authors of this paper [21] proposed a community finding algorithm based on fitness of networks, which discovers interpretable communities in real world networks. By comparing the experimental discoveries and Darwinian theory of natural selection [22], the authors proposed a homophyly/kinship hypothesis that homophyly is the extension of kinship, and is the controlling principle of nature evolving from random variation [21]. To explore the principle of natural selection in networks, the authors proposed a homophyly/kinship model of networks.

The homophyly/kinship model explores that homophyly/kinship is the principle of natural communities in nature, society and networks, and that natural communities of networks satisfy a number of properties, including, interpretability, robustness, stability, leadership property, internal centrality, external de-centrality, inclusiveness and exclusiveness, reciprocity, and king node properties etc. Our theory implies that natural communities exist in real world networks, for which natural selection is the underlying principle. A fundamental question is hence: Can we really find the natural or true communities of networks?

In the present paper, we proposed the measure of structure entropy of networks, and proposed a new community finding algorithm, written by $\mathcal{E}$. We verified that our algorithm $\mathcal{E}$ exactly finds almost all or most natural communities of networks generated by our homophyly/kinship model, and by existing models with planted communities, provided that the natural communities exist in the networks, and that the well-known algorithm based on modularity maximization fail to find nontrivial natural communities of networks. Our algorithm $\mathcal{E}$ provides for the first time an approach to identifying and analyzing natural or true communities in networks.

We organize the paper as follows. In Section 2, we introduce the notion of structure entropy of graphs. In Section 3, we analyze the principles of the notion of structure entropy. In Section 4, we describe our new community finding algorithm by minimization of structure entropy of networks. In Section 5, we introduce our homophyly/kinship model of networks. In Section 6, we verify that our algorithm exactly identifies or precisely approximates the natural communities of networks generated by our homophyly/kinship model and the planted l-partitioning model. In Section 7, we verified that our algorithm exactly identifies or very well approximates the ground-truth communities of some real world networks, if the semantically defined ground-truth communities are well-defined. In Section 8, we summarize the conclusions of the paper.