Physica A 397 (2014) 144–156

Contents lists available at ScienceDirect

Physica A

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

Group detection in complex networks: An algorithm and comparison of the state of the art

Lovro Šubelj *, Marko Bajec

University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, SI-1000 Ljubljana, Slovenia

HIGHLIGHTS

- We propose a propagation-based algorithm for group detection in complex networks.
- The main novelty is a hierarchical refinement procedure for discovery of different groups.
- The algorithm is comparable to the state of the art and has near ideal complexity.
- We consider group detection, hierarchy discovery and link prediction tasks.

ARTICLE INFO

Article history:
Received 19 July 2013
Received in revised form 29 October 2013
Available online 9 December 2013

Keywords:
Complex networks
Group detection
Hierarchy discovery
Label propagation
Clustering

ABSTRACT

Complex real-world networks commonly reveal characteristic groups of nodes like communities and modules. These are of value in various applications, especially in the case of large social and information networks. However, while numerous community detection techniques have been presented in the literature, approaches for other groups of nodes are relatively rare and often limited in some way. We present a simple propagation-based algorithm for general group detection that requires no a priori knowledge and has near ideal complexity. The main novelty here is that different types of groups are revealed through an adequate hierarchical group refinement procedure. The proposed algorithm is validated on various synthetic and real-world networks, and rigorously compared against twelve other state-of-the-art approaches on group detection, hierarchy discovery and link prediction tasks. The algorithm is comparable to the state of the art in community detection, while superior in general group detection and link prediction. Based on the comparison, we also discuss some prominent directions for future work on group detection in complex networks.

1. Introduction

Complex networks of real-world systems commonly reveal groups of nodes with characteristic connection patterns [1] (e.g., densely connected groups known as communities [2]). These correspond to people with common interests in social networks [3] or classes with the same information signature in software networks [4]. Characteristic groups of nodes provide an important insight into the structure and function of real-world networks [5], while group detection also has numerous practical applications, including epidemic outbreak prevention [6], viral marketing [7], software package prediction [8] and compression [9].

Despite an outburst of community detection algorithms in the last decade [10,1], approaches for other groups of nodes are relatively rare and often limited (e.g., demand some a priori knowledge about the network structure). Thus, we here propose a general group detection algorithm based on the label propagation framework in Refs. [4,11] that requires no a

* Corresponding author. Tel.: +386 1 476 81 86.
E-mail addresses: lovro.subelj@fri.uni-lj.si (L. Šubelj), marko.bajec@fri.uni-lj.si (M. Bajec).

0378-4371/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.physa.2013.12.003

priori knowledge. Analysis in the paper confirms that the proposed algorithm is at least comparable to the current state of the art, while its complexity is near ideal. The main novelty of the paper is otherwise a simple hierarchical refinement procedure that enables straightforward discovery of different types of groups. The paper also includes a detailed empirical comparison of a larger number of state-of-the-art approaches for group detection that may be of an independent interest.

The rest of the paper is structured as follows. First, Section 2 gives preliminary discussion on groups of nodes in real-world networks. Next, we introduce the group detection algorithm proposed in the paper in Section 3. Rigorous analysis on synthetic and real-world networks appears in Section 4, whereas detailed comparison with the state of the art on group detection, hierarchy discovery and link prediction tasks is presented in Section 5. Section 6 concludes the paper and gives some prominent directions for future work.

2. Background

Let the network be represented by a simple undirected graph $G(V, L)$, where V is a set of nodes in the network, $|V| = n$, and L is the set of links, $|L| = m$. Next, let Γ_i be the set of neighbors of node v_i, $v_i \in V$, and Δ_i the number of links between the nodes in Γ_i. Last, let k_i be the degree of v_i, $|\Gamma_i| = k_i$, and let $\langle k \rangle$ be the average degree in the network.

2.1. Groups in real-world networks

The present paper is concerned with groups of nodes with characteristic connection pattern that appear in complex real-world networks [12]. While these could be defined in various ways, we adopt two types of groups that have been most popular in the recent literature [10,1].

First, we consider communities [2] (also link-density community [13]) that, e.g., represent groups of people with common interests in social networks [3]. Community is defined as a (connected) group of nodes with more links towards the nodes in the group than to the rest of the network [14].

Second, we consider sparse modules [15] (also link-pattern community [13] and other [16]) that in, e.g., software networks correspond to classes with the same function [4]. A module is defined as a (possibly) disconnected group of nodes with more links towards common neighbors than to the rest of the network [4].

The definition of a module is rather similar to the concept of structural or regular equivalence [17,18] (i.e., blockmodels), although not equivalent. Note that communities can in fact be considered under the definition of modules (most authors have indeed adopted this stance [12,16]), however, there exist important differences between the two [4] (e.g., connectedness).

2.2. Node and network clustering

Based on the above, a group detection approach could gain from differentiating between communities and modules. Note that the definition of a community implies locally dense structure, while no such tendency appears in the definition of a module. The density of a node's neighborhood is usually measured by the node clustering coefficient c [19] defined as

$$c_i = \frac{\Delta_i}{\binom{k_i}{2}}, \tag{1}$$

while $C = \frac{1}{n}\sum_i c_i$ is the network clustering coefficient [19,20].

However, the denominator in Eq. (1) introduces biases into the definition of clustering, since $\binom{k_i}{2}$ often cannot be reached due to fixed degree distribution [21]. The latter is particularly apparent in degree disassortative networks [22], which most real-world networks in fact are. Thus, an alternative definition denoted degree-corrected clustering coefficient d [21] has been proposed as

$$d_i = \frac{\Delta_i}{\omega_i}, \tag{2}$$

where ω_i is the maximum possible number of links among Γ_i and $D = \frac{1}{n}\sum_i d_i$.

Thus, in the presence of communities, one can expect dense network structure with $D \gg 0$. On the other hand, different configurations of modules (e.g., bipartite, multi-partite or star-like structures) imply sparser networks with $D \approx 0$. For example, most prominent modules are found in two-mode networks, where $D = 0$. It should, however, be stressed that a module structure does not necessarily mean $D = 0$, since a tripartite configuration of modules obviously has $D > 0$.

3. The algorithm

The proposed algorithm for group detection is based on the label propagation framework [4,11] that we introduce next. For simplicity, the framework is presented for the case of simple undirected graphs, although a generalization to weighted multigraphs is straightforward.