Physica A 392 (2013) 2278–2293

Contents lists available at SciVerse ScienceDirect

Physica A

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

Obtaining communities with a fitness growth process

Mariano G. Beiró [a,b,*], Jorge R. Busch [a,b], Sebastian P. Grynberg [a], J. Ignacio Alvarez-Hamelin [a,b]

[a] Facultad de Ingeniería, Universidad de Buenos Aires, Paseo Colón 850, C1063ACV, Buenos Aires, Argentina
[b] INTECIN (CONICET–U.B.A.), Paseo Colón 850, C1063ACV, Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 31 August 2012
Received in revised form 29 November 2012
Available online 17 January 2013

Keywords:
Community detection
Social networks
Complex systems

ABSTRACT

The study of community structure became an important topic of research over the last years. But, while successfully applied in several areas, the concept lacks of a general and precise notion. Facts like the hierarchical structure and heterogeneity of complex networks make it difficult to unify the idea of community and its evaluation. The global functional known as modularity is probably the most used technique in this area. Nevertheless, its limits have been deeply studied. Local techniques as the one by Lancichinetti et al. (2009) [1] arose as an answer to the resolution limit and degeneracies that modularity has.

Here we propose a unique growth process for a fitness function based on the algorithm by Lancichinetti et al. (2009) [1]. The process is local and finds a community partition that covers the whole network, updating the scale parameter dynamically. We test the quality of our results by using a set of benchmarks of both heterogeneous and homogeneous graphs. We discuss alternative measures for evaluating the community structure and, in the light of them, infer possible explanations for the better performance of local methods compared to global ones in these cases.

1. Introduction

In the last years, community detection became one of the top research topics in the area of Complex Networks. Due in part to the explosion of social networking, and also to its application in diverse areas as ecology and computational biology, an interest arose in defining, detecting, evaluating and comparing community structures. For a thorough, yet not exhaustive, reference of its applications, see the survey by Fortunato [2].

The early research by Newman started with use of *betweenness* to divide the network into modules [3], and the definition of *modularity* to evaluate communities [4]. Then he proposed using the modularity as a functional to be maximized [5]. Different optimization techniques were developed, of which we recall the algorithm by Guimerà et al. based on simulated annealing [6] for its good results, and the Louvain algorithm [7] for its fast convergence within large networks.

Later works questioned the global optimization methods based on modularity, for being prone to resolution limits (Fortunato [8]) and extreme degeneracies (Good et al. [9]). Local techniques were proposed, as the Clique Percolation Method (CPM) (Palla et al. [10]), and the algorithm in Lancichinetti et al. [1], based on a fitness function. Both of them find overlapping communities, and in the latter, the notion of *natural community* arose. The *natural community* of a vertex is a locally-computed set, and its size depends on a resolution parameter α.

It has also been observed that the resolution limits for modularity found in Fortunato et al. [8] are particularly common in heterogeneous graphs with heavy-tailed community sizes and vertex degree distributions (see Ref. [2], Section VI.C).

* Corresponding author at: Facultad de Ingeniería, Universidad de Buenos Aires, Paseo Colón 850, C1063ACV, Buenos Aires, Argentina. Tel.: +54 1143430891x258.
E-mail addresses: mbeiro@fi.uba.ar (M.G. Beiró), jbusch@fi.uba.ar (J.R. Busch), sebgryn@fi.uba.ar (S.P. Grynberg), ignacio.alvarez-hamelin@cnet.fi.uba.ar (J.I. Alvarez-Hamelin).

0378-4371/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2013.01.031

In these graphs, small communities will often be masked into larger ones by modularity maximization techniques when they are interconnected just by a few links.

In this work we detect communities based on a *fitness function* analogous to the one defined by Lancichinetti et al. in Ref. [1]. After analyzing the role of the resolution parameter α in these functions, we propose a *uniform fitness growth process* which scans the whole graph and whose resolution parameter is updated dynamically. Then, we extract a community partition from the output of this process. The details of our method are described in Sections 2 and 3, and the algorithmic complexity is discussed in Section 4.

In Section 5 we show the results. We use a benchmark developed in Ref. [11] to build a dataset of heterogeneous and homogeneous networks. We observe an important improvement using our fitness growth process when compared to the global modularity maximization techniques, which suggests that local methods may outperform global ones in these cases. In order to discuss this conjecture, we analyze the consequences of the resolution limits and give a possible explanation to the differences in performance between the two methods.

As a measure for comparing community structures, Danon et al. [12] proposed using the *normalized mutual information*. We shall use it in order to make comparisons with global methods and with community structures known *a priori*. We also apply the algorithm to real networks and show the results. Finally, we discuss the robustness (repeatability of the results) of our process.

2. Our method

Given a graph $G = (V, E)$, the work by Lancichinetti et al. [1] define a process based on a fitness function with a resolution parameter α such that, given a set $C \subset V$:

$$f(C) = \frac{k_{in}}{(k_{in} + k_{out})^{\alpha}}$$

where k_{in} is the number of edges that join vertices in C, and k_{out} is the number of edges that join some vertex in C to some vertex not in C. This notion of community is related to the one proposed by Radicchi et al. [13]. In fact, a choice of $\alpha = 1$ corresponds to the definition of weak community introduced in that paper.

The process starts with a community made up by the seed vertex v and proceeds by stages, where in each stage the steps are: (1) select a vertex whose addition increments the fitness function, and add it to the present community; (2) delete from the present community all the vertices whose deletion increments the fitness function. The algorithm stops when, being in stage 1, it finds no vertex to add. Step 2 is time consuming, and usually very few vertices are deleted, but it is necessary due to the local, vertex-by-vertex nature of the analysis. The authors called the final result of the algorithm the *natural community of* v. The resolution parameter α is related to the natural community size.

In order to obtain a covering by overlapping communities, they select a vertex at random, obtain its natural community, select a vertex not yet covered at random, obtain its natural community, and so on until they cover the whole graph.

In this process, the resolution parameter α of the fitness function is kept fixed. The authors perform an analysis in order to find the significant values of α.

Our contribution extends that work to define a *uniform growth process*. This process covers the whole graph by making a course throughout its communities. We modify the *fitness function* $f(C)$ and analyze the role of α in the termination criteria for the process. Then we propose an algorithm for increasing the fitness function monotonically while traversing the graph, dynamically updating the parameter. Finally, a *cutting technique* divides the sequence of vertices obtained by the process, in order to get a partition into communities.

2.1. Previous definitions

We shall deal with simple undirected graphs $G = (V, E)$, with $n = |V|$ vertices and m edges (here $|\cdot|$ denotes the cardinal of a set). To avoid unnecessary details, we assume that $E \subset V \times V$ is such that $(v, w) \in E$ implies that $(w, v) \in E$.

We set $\delta_E(v, w) = 1$ if $(v, w) \in E$, $\delta_E(v, w) = 0$ in the other case. We then have the following expression for the degree of a vertex v

$$\deg(v) = \sum_{w \in V} \delta_E(v, w).$$

Thus, $|E| = \sum_{w \in V} \deg(w) = 2m$. We shall use two measures, m_V and m_E, the first one on V and the second one on $V \times V$. Given $C \subset V$,

$$m_V(C) = \sum_{v \in C} \deg(v)/|E|$$

is the normalized sum of the degrees of the vertices in C. Given $D \subset V \times V$,

$$m_E(D) = \sum_{(v,w) \in D} \delta_E(v, w)/|E|.$$