Physica A 428 (2015) 154–160

Contents lists available at ScienceDirect

Physica A

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

Defining least community as a homogeneous group in complex networks

Bin Jiang *, Ding Ma

Faculty of Engineering and Sustainable Development, Division of Geomatics, University of Gävle, SE-801 76 Gävle, Sweden

HIGHLIGHTS

- A new community detection algorithm inspired by the head/tail breaks.
- A new way of thinking for community detection or classification in general.
- Far more small communities than large ones in complex networks.
- Simple networks like mechanical watches, while complex networks like human brains.
- Empirical evidence on power laws of the detected communities.

ARTICLE INFO

Article history:
Received 20 June 2014
Received in revised form 7 November 2014
Available online 12 February 2015

Keywords:
Head/tail breaks
ht-index
Scaling
k-means
Natural breaks
Classification

ABSTRACT

This paper introduces a new concept of least community that is as homogeneous as a random graph, and develops a new community detection algorithm from the perspective of homogeneity or heterogeneity. Based on this concept, we adopt head/tail breaks – a newly developed classification scheme for data with a heavy-tailed distribution – and rely on edge betweenness given its heavy-tailed distribution to iteratively partition a network into many heterogeneous and homogeneous communities. Surprisingly, the derived communities for any self-organized and/or self-evolved large networks demonstrate very striking power laws, implying that there are far more small communities than large ones. This notion of far more small things than large ones constitutes a new fundamental way of thinking for community detection.

1. Introduction

A network or graph is simply a set of vertices or nodes joined by edges or links. Sometimes, the edges are given a direction or weight. This paper considers only a binary network, neither a direction nor a weight for all of the edges. A network may be as large as having thousands, millions and even billions of nodes and edges. Large-scale real-world networks are of primary interest because of their large sizes and structural complexity, herein complex networks. Complex networks could be social such as friendships and collaborations, biological such as protein interactions and food webs, technological such as the Internet and streets, and informational such as the World Wide Web (e.g., Refs. [1,2]). Large networks are unlikely to be regular, such as lattices (e.g., crystal in reality) in which each node has the same number of edges, or random (e.g., gas in a container), in which every pair of nodes has the same probability of being linked (both regular and random graphs with homogeneous structures); instead, they are something in between regular and random. Subsequently, real-world networks differ fundamentally from their random counterparts in that they display a very significant heterogeneity. This heterogeneity

* Corresponding author.
E-mail addresses: bin.jiang@hig.se (B. Jiang), ding.ma@hig.se (D. Ma).

http://dx.doi.org/10.1016/j.physa.2015.02.029

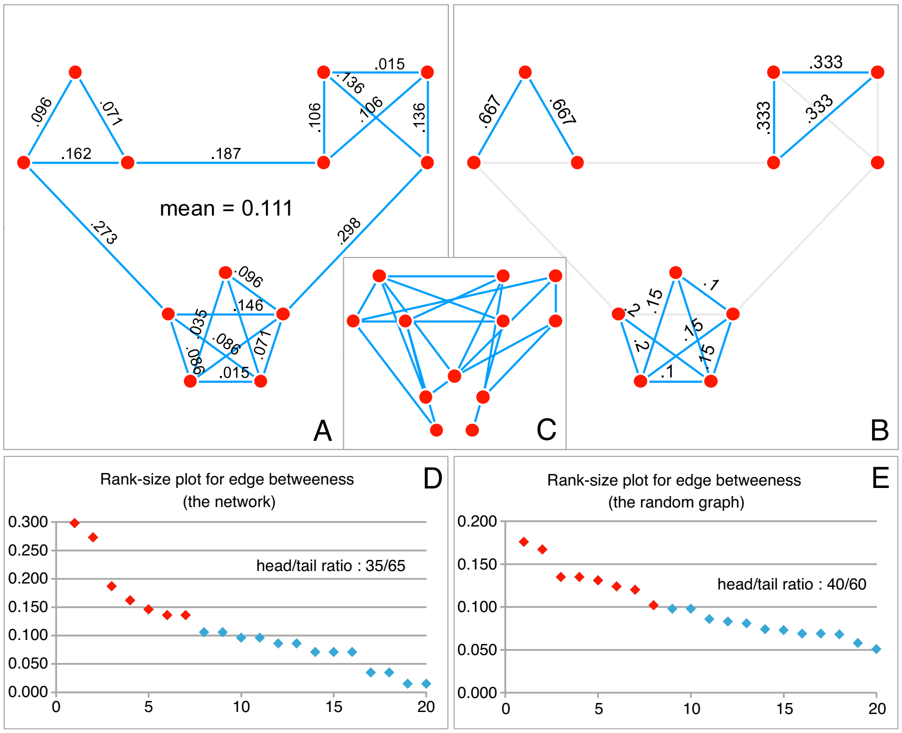

Fig. 1. Illustration of the community detection algorithm. (Note: A fictive social network (Panel A) with 12 vertices linked by 20 edges, its four detected communities with sizes 5, 3, 3, and 1, shown in Panel B, in which the removed edges are shown in gray, and its equivalent random graph (Panel C); the edge betweenness of the network is heterogeneous with a head/tail ratio of 35/65 (Panel D), whereas that of the random network is homogeneous with a head/tail ratio of 40/60 (Panel E). The random graph's head percentage is just a reference to determine whether the fictive network is heterogeneous, i.e., the fictive network's head is smaller than that of the random graph. The actual partition relies on the network's mean or its head percentage rather than that of the random graph.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

implies that complex networks consist of many different relatively independent compartments, leading to the notion of community or community structure.

A community, also called a module or a cluster, is loosely defined as a subset of vertices with many inside edges and a few outside edges. In other words, vertices within a community are densely connected, whereas connections or edges between communities are sparser. This definition sounds very intuitive and straightforward; however, it is hardly operable for community detection, in particular for large complex networks. This paper introduces a new concept of least community as a homogeneous group—as homogeneous as a random graph. Based on this concept, a heterogeneous network is partitioned into many homogeneous communities by referencing its random graph. The random graph is used as a reference because it is considered homogeneous enough and its edges are imposed with the same probability, or it contains only one community. Considering a network as a set of edges characterized by the measure edge betweenness [3], the issue of community detection becomes that of classification, i.e., classifying all edges into different homogeneous groups as homogeneous as a random graph or, more specifically, into inside and outside edges.

The classification relies on edge betweenness to determine different classes or communities. The edge betweenness of real-world networks demonstrates a heavy tail distribution, indicating that conventional methods such as *k*-means [4] and natural breaks [5] could not effectively derive the classes that reflect the underlying scaling pattern. These conventional methods use the mean or the average to characterize individual classes, but the edge betweenness is right skewed or scale free. Given the circumstance, head/tail breaks, a newly developed classification scheme [6], is more appropriate and effective for data with a heavy tail distribution. Head/tail breaks partitions all the edges into the head (those edges with betweenness greater than the mean) and the tail (those edges with betweenness less than the mean), and recursively continues the partition process until the head percentage is as large as that of the random graph (c.f., the next section for illustrations). This ending condition implies that the head and tail are well balanced, and the derived classes or communities are homogeneous enough. During the recursive partition process, some heterogeneous communities are identified as well. Eventually, both homogeneous and heterogeneous communities are derived at different coarse-graining levels. The central argument of this paper is that any self-organized and/or naturally evolved real world network contains far more small communities than large ones, or its communities exhibit a power law or heavy-tailed distribution in general.

Community structure or community detection has received disproportionate attention in the past years, largely because of the availability of rich data from the Internet and social media, and its far-reaching implications for a variety of disciplines (e.g., Refs. [7,8]). Communities could be social groupings in a social network based on interest, related papers in a citation