Information Sciences 181 (2011) 1060–1071

Contents lists available at ScienceDirect

Information Sciences

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

Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks

Peng Gang Sun [a,*], Lin Gao [a,*], Shan Shan Han [b]

[a] School of Computer Science and Technology, Xidian University, Xi'an 710071, China
[b] Faculty of Science, University of Copenhagen, Copenhagen 1307K, Denmark

ARTICLE INFO

Article history:
Received 14 December 2009
Received in revised form 18 November 2010
Accepted 19 November 2010

Keywords:
Community structure
Fuzzy clustering
Fuzzy equivalence relation
Equivalence class

ABSTRACT

This paper proposes a novel method based on fuzzy clustering to detect community structure in complex networks. In contrast to previous studies, our method does not focus on a graph model, but rather on a fuzzy relation model, which uses the operations of fuzzy relation to replace a traversal search of the graph for identifying community structure. In our method, we first use a fuzzy relation to describe the relation between vertices as well as the similarity in network topology to determine the membership grade of the relation. Then, we transform this fuzzy relation into a fuzzy equivalence relation. Finally, we map the non-overlapping communities as equivalence classes that satisfy a certain equivalence relation. Because most real-world networks are made of overlapping communities (e.g., in social networks, people may belong to multiple communities), we can consider the equivalence classes above as the skeletons of overlapping communities and extend our method by adding vertices to the skeletons to identify overlapping communities. We evaluated our method on artificial networks with built-in communities and real-world networks with known and unknown communities. The experimental results show that our method works well for detecting these communities and gives a new understanding of network division and community formation.

1. Introduction

In the past several years, understanding the properties of networks, such as the World Wide Web networks, citation networks, transportation networks, social networks and biochemical networks, has grown in importance [2,11,30]. One significant feature in these networks is the "community structure/module," in which groups of vertices, within which connections are dense, but between which they are sparse. Identifying these communities is of great importance because such structures often correspond to important functional organizations. For example, social networks often include communities based on common locations, interests and occupations. Metabolic networks have communities based on functional groupings. Citation networks form communities based on similar research topics [11,22]. Communities (modules) in protein–protein interaction (PPI) networks may correspond to known or even unknown functional modules/protein complexes that play special roles in cellular systems [12]. Therefore, the detection of the community structure in a network has important practical applications and can help us understand the complex network system.

Many methods have been developed to detect community structure in complex networks [1,8,10–12,14–17,20,22–24,26–29,31,32], i.e., the betweenness-based method [11] and the spectral clustering algorithm [24,32]. Although the notion of community structure is easy to understand, construction of an efficient algorithm for community structure detection is

* Corresponding authors. Tel.: +86 29 88207344.
E-mail addresses: sun198268@163.com (P.G. Sun), lgao@mai.xidian.edu.cn (L. Gao).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2010.11.022

highly nontrivial. There are two main difficulties in detecting community structure. The first is that the number of communities is unknown for a given network beforehand. The second is that some vertices in a network may belong to more than one community. This feature is the overlapping community structure that is familiar to us in real-world networks. Overlapping vertices play a special role in a complex network system. For example, in biological networks such as a PPI network, one vertex (protein or gene) belonging to two functional modules/protein complexes may act as a bridge between them, which transfers biological information or acts as multiple functional units [12,34].

In terms of community structure, because the vertices within a community may have certain similar relations (e.g., people in a social network may have similar relations in interests or hobbies), we can use it to classify vertices into different communities. However, most existing methods for community structure detection treat this similar relation in a graph as a deterministic relation, which leads to the division of networks being insignificant. Actually, in real-world networks such as social networks, this similar relation between people is fuzzy and indeterminate. It is not a good strategy for us to use a deterministic relation to describe it. The deterministic relation addresses that a vertex either belongs to or does not belongs to a community, but the fuzzy relation emphasizes that a vertex belongs to more than one community with a membership degree valued in the real unit interval $[0,1]$. Therefore, a generalization is to treat this similar relation as a fuzzy, indeterminate relation [7].

Since the fuzzy set theory was proposed in 1965 by Zadeh [36,37], fuzzy clustering as a fundamental tool has been applied to many fields extensively [3–5,9,19]. In this paper, we provide a novel method based on fuzzy clustering to detect community structure in complex networks. The existing methods often treat the complex network system as a graph model and mine community structure based on this graph. In contrast to previous methods, our method is no longer focused on a graph model, but on a fuzzy relation model; it uses the operations of fuzzy relation to replace a traversal search of the graph for identifying community structure. By the composition of fuzzy relation, the information between vertices is transmitted with different paths from one node to another. Ultimately, each community corresponds to an equivalence class because they share the same properties (reflexivity, symmetricity and transitivity), and the details can be seen in Section 2.

According to the discussion above, in our method, we first use a fuzzy relation to describe the relation between vertices and the similarity in network topology to determine the membership grade of the relation. A graph is denoted by a fuzzy relation. Then, we transform this fuzzy relation into a fuzzy equivalence relation by transitive closure. Finally, we map the non-overlapping communities as equivalence classes that satisfy a certain equivalence relation. Because most real-world networks are made of overlapping communities (e.g., people may belong to multiple communities in social networks), we can consider the equivalence classes above as the skeletons of overlapping communities, and then extends the skeletons by adding vertices to them when detecting the overlapping communities. Our method can address all the issues discussed above.

The rest of our paper is organized as follows. In Section 2, we describe the theoretical basis about fuzzy clustering and similarity in network topology. Section 3 presents our method for detecting overlapping and non-overlapping community structure in complex networks. In Section 4, we test our method using artificial networks and real-world networks and give an analysis on its performance. The conclusion is provided in Section 5.

2. Theoretical basis

2.1. The definitions of fuzzy clustering

In this paper, let $G = \{V,E\}$ be a graph, where V is a set of vertices and E is a set of pairs (unordered) of distinct vertices called edges. In this subsection, we will give the details of fuzzy clustering [3,7,36,37].

Definition 1 (*Fuzzy relation*). Let U and V be nonempty sets. A fuzzy relation R, $R \in F(U \times V)$, $F(U \times V)$ is the set of all the fuzzy relations of $U \times V$. $\forall (u,v) \in U \times V$, $R(u,v)$ can be interpreted as the grade of membership of the ordered pair (u,v) in R. If $U = V$, then we can say that R is a binary fuzzy relation in V [7].

Definition 2 (*Inclusion relation*). Let $R, Q \in F(U \times V)$, and if $\forall (u,v) \in U \times V$, $R(u,v) \leqslant Q(u,v)$, then $R \subseteq Q$ [7].

Definition 3 (*Composition of fuzzy relations*). Let $R \in F(U \times V)$ and $Q \in F(V \times W)$, then $R \circ Q \in F(U \times W)$ can be defined in the following way [7]:

$$(R \circ Q)(u,w) = \vee_{v \in V}(R(u,v) \wedge Q(v,w)), \quad (u,w) \in U \times W.$$

If $R \in F(V \times V)$, then $R^0 = I$, $R^n = R^{n-1} \circ R$, $n = 1, 2, \ldots$, and I is the identity relation.

Definition 4 (*Fuzzy equivalence relation*). Let $R \in F(V \times V)$, and R is a fuzzy equivalence relation if it satisfies the following conditions [7]:

(1) Reflexivity : $\forall v_i \in V$, $R(v_i, v_i) = 1$.
(2) Symmetricity: $\forall (v_i, v_j) \in V \times V$, $R(v_i, v_j) = R(v_j, v_i)$.
(3) Transitivity : $R \circ R \subseteq R$.