Expert Systems With Applications 42 (2015) 9634–9646

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An approach for overlapping and hierarchical community detection in social networks based on coalition formation game theory

Lihua Zhou [a], Kevin Lü [b,*], Peizhong Yang [a], Lizheng Wang [a], Bing Kong [a]

[a] Department of Computer Science, Yunnan University, Kunming 650091, China
[b] Brunel University, Uxbridge UB8 3PH, UK

ARTICLE INFO

Keywords:
Social network
Community detection
Coalition formation game theory

ABSTRACT

With greater availability of data and increasing interaction activities taking place on social media, detecting overlapping and hierarchical communities has become an important issue and one that is essential to social media analysis. In this paper, we propose a coalition formation game theory-based approach to identify overlapping and hierarchical communities. We model community detection as a coalition formation game in which individuals in a social network are modelled as rational players aiming to improve the group's utilities by cooperating with other players to form coalitions. Each player is allowed to join multiple coalitions, and those coalitions with fewer players can merge into a larger coalition as long as the merge operation is beneficial to the utilities of the merged coalitions, thus overlapping and hierarchical communities can be revealed simultaneously. The utility function of each coalition is defined as the combination of a gain function and a cost function. The gain function measures the degree of interactions amongst the players inside a coalition, while the cost function instead represents the degree of the interactions between the players of the coalition and the rest of the network. As game theory provides a formal analytical framework with a set of mathematical tools to study the complex interactions among rational players, applying game theory for detecting communities helps to identify communities more rationally. Some desirable properties of the utility function, such as the non-resolution limit and the non-scaling behavior, have been examined theoretically. To solve the issue of pre-setting the number and size for communities and to improve the efficiency of the detection process, we have developed a *greedy agglomerative manner* to identify communities. Extensive experiments have been conducted on synthetic and real networks to evaluate the effectiveness and efficiency of the proposed approach which can be applied for real world applications.

1. Introduction

With social networks gaining popularity, social network analysis has become an important research issue, with a significant impact on the society (Fortunato, 2010; Li, Pan, Xiao, & Huang, 2014a). One major and fundamental topic in social network analysis is community detection, i.e. to identify groups of vertices in a network such that the vertices within a group are much more connected to each other than to the rest of the network (Fortunato, 2010; Newman & Girvan, 2004). Because individuals belonging to the same community are more likely to have common features, such as social functions, interests on some topics, viewpoints, etc. (Zhao et al., 2012), the identified communities can be used in the improvement of services (Krishnamurthy & Wang, 2000), knowledge sharing (Liu, Raahemi, & Benyoucef, 2010), collaborative recommendation (Yuan, Guan, Lee, Lee, & Hur, 2010), information spreading (Wu, Huberman, Adamic, & Tyler, 2004), structure visualizing (Wu & Li, 2011), and other applications. In recent years, community detection has received a great deal of attention as it has significance relating to online influence analysis, online marketing and ebusiness (Bagrow, 2012; Francesco and Clara 2014; Li, Ng, & Ye, 2014b; Papadopoulous, Kompatsiaris, Vakali, & Spyridonos, 2012; Zhou & Lü, 2014).

However, community detection is not a straightforward task, because in real networks communities can be overlapped or hierarchical, and these features often occur simultaneously. The overlap of communities implies that vertices simultaneously belong to more than one group, for instance, people belong to different social groups, depending on their activities, interests, etc. (Palla, Derenyi, Farkas, & Vicsek, 2005). This breaks the assumption that a community should have more internal than external connections (because highly over-

* Corresponding author. Room 104, ESGW Building, Brunel University, Uxbeidge UB8 3PH, UK. Tel.: +1 44 1895 265254.
E-mail addresses: lhzhou@ynu.edu.cn (L. Zhou), kevin.lu@brunel.ac.uk (K. Lü), pzyang@ynu.edu.cn (P. Yang), lzhwang@ynu.edu.cn (L. Wang), bkong@ynu.edu.cn (B. Kong).

http://dx.doi.org/10.1016/j.eswa.2015.07.023
0957-4174/

lapping communities can have many more external than internal connections), and demands a method that is able to detect either overlapping or non-overlapping communities (Lancichinetti, Fortunato, & Kertesz, 2009). The hierarchical form of communities implies that the communities are recursively grouped into a hierarchical structure, i.e. small communities can form larger ones, which in turn can group more communities together to form even larger ones, etc. In the presence of hierarchies, the concept of community structure becomes richer, and demands a method that is able to detect communities at different levels, not just within a single level (Lancichinetti et al., 2009). Another two essential challenges in community detection are the efficiency of algorithms and the priori knowledge on the number and size of communities, because the presence of many vertices and links in a large network results in heavy computation, and the number and size of communities are usually unknown beforehand. At present, these issues have not been solved satisfactorily. In existing community detection algorithms, some require a priori knowledge on the number and size of communities before performing the task of detecting communities, some are not able to detect overlapping and hierarchical communities, and some are not applicable to large-scale networks due to the low efficiency.

Motivated by the need for developing an algorithm that can detect both overlapping and hierarchical communities without prior knowledge on the number and size of communities in large-scale networks, we develop an approach by applying cooperative game theory (Zlotkin & Rosenschein, 1994) to detect communities in this study. Cooperative game theory (Zlotkin & Rosenschein, 1994) studies the cooperative behaviors of groups of rational players, where players cooperate with each other for improving the group's utility, such a group of players is called a coalition. One class of cooperative games is coalition formation games (Saad, Han, Debbah, Hjørungnes, & Basar, 2009), whose main objective is to analyze the formation of coalitional structures through players' interaction. Coalition formation games are generally not superadditive due to the presence of costs that reduce the gains from forming the coalition. In social network environments, the behaviors of individuals are not independent (Zacharias, MacMillan, & Hemel, 2008), and joining a community provides one with tremendous benefits, such as members feeling rewarded in some ways for their participation in the community, and gaining honor and status for being members (Sarason, 1974). In which case, every individual has an incentive to join communities; however, in real-world cases not only does each individual receive benefit(s) from the communities it belongs to, but the individual must also pay a certain price to maintain its membership within these communities (Chen, Liu, Sun, & Wang, 2010). These characteristics make coalition formation game theory applicable to community detection.

In this study, we first model the process of community detection as a coalition formation game, in which individuals in a social network are modelled as rational players aiming to achieve the maximal group's utility by cooperating with other players to form coalitions. A coalition is a subset of players. Each player is allowed to join multiple coalitions, which reflects the concept of "overlapping communities". Meanwhile, coalitions with fewer players can merge into a larger coalition as long as such merge operations could improve the utilities of the coalitions. This process reveals, in fact, the hierarchical structure of communities. A coalition is regarded as a *stable community* if it cannot further improve its utility by merging with other coalitions. If no coalition can further improve its utility by merging with other coalitions, the game achieves an *equilibrium state of coalitions*, and the configuration of communities at this state is called *the stable community structure*.

Next, we introduce the utility function for each coalition, which is the combination of a gain function and a cost function. The gain function measures the degree of the interaction amongst the players inside a coalition, while the cost function represents the degree of the interaction between the players of that coalition and the rest of the network. Based on the defined utility function, two coalitions without any link between them cannot improve their utilities by merging into a larger coalition, thus whether a coalition is merged with others it can be decided by looking only at its neighbors (coalitions that have at least one link between them), rather than necessitating the performance of an exhaustive search over the entire network. This can speed up the computation considerably.

Then, we develop a *greedy agglomerative manner* to identify communities, which starts from the vertices as separate coalitions (singletons); coalitions are iteratively merged to improve the group's utilities until no further merging of coalitions is needed. This *greedy agglomerative manner* does not require a priori knowledge on the number and size of the communities, and it matches the real-world scenario, in which communities are formed gradually from bottom to top.

Finally, we conduct extensive experiments on different networks to assess the performance of our approach. Meanwhile, we also compare our results with other related studies. The experimental results show that our algorithm is effective and efficient in identifying overlapping and hierarchical communities.

The main contributions of this study can be summarized as follows:

- The coalition formation game theory is applied to address the community detection problem. This approach considers community formation as the result of the group behaviors of rational players who cooperate with each other to form coalitions for achieving and improving a group's utilities.
- A utility function for modeling the benefit and cost of each coalition is introduced, and the properties of the utility function, such as the non-resolution limit and the non-scaling behavior, have been examined theoretically.
- An algorithm based on the *greedy agglomerative manner* is proposed to identify communities. The proposed algorithm does not require a priori knowledge on the number and size of communities, and it can detect the overlapping and hierarchical communities simultaneously.
- Extensive experiments on synthetic and real networks have been conducted to evaluate the effectiveness and efficiency of the proposed approach.

The rest of this paper is organized as follows: Section 2 introduces related work; Section 3 presents a coalition formation game theory-based framework for community detection; Section 4 provides an algorithm that uses the *greedy agglomerative manner* to identify communities. The experimental results on the synthetic and real networks are presented in Section 5, and Section 6 concludes this paper.

2. Related work

A well-known method for detecting non-overlapping and non-hierarchical communities is the use of modularity-based methods (Newman & Girvan, 2004), which is based on the idea that a random graph is not expected to have a cluster structure, so the possible existence of clusters is revealed by the comparison between the actual density of edges in a subgraph and the density one would expect to have in the subgraph, if the vertices of the graph were attached regardless of community structure (Fortunato, 2010). However, modularity-based methods implicitly assume that communities do not intersect with one another, which is usually not the case for real-world communities (Chen et al., 2010). Fortunato and Barthélemy (2006) found that modularity optimization may fail to identify communities smaller than a scale which depends on the total number of links of the network and on the degree of interconnectedness of the communities, even in cases where communities are unambiguously defined. Brandes et al. (2008) also identified counterintuitive properties of modularity, such as non-locality and sensitivity to satellites.