Expert Systems With Applications 84 (2017) 74–91

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Utilizing advances in correlation analysis for community structure detection

Lian Duan[a,*], Yanchi Liu[b], W. Nick Street[c], Haibing Lu[d]

[a] Department of Information Systems and Business Analytics, Hofstra University, Hempstead, NY 11549, USA
[b] Department of Management Science and Information Systems, Rutgers University, Newark, NJ 07102, USA
[c] Department of Management Sciences, University of Iowa, Iowa City, IA 52242, USA
[d] Department of Operations Management and Information Systems, Santa Clara University, Santa Clara, CA 95053, USA

ARTICLE INFO

Article history:
Received 27 April 2016
Revised 23 April 2017
Accepted 4 May 2017
Available online 8 May 2017

Keywords:
Correlation analysis
Community detection
Modularity
Leverage
Likelihood ratio
Community simulation

ABSTRACT

In the era of big data, some data records are interrelated with each other in many areas, such as marketing, management, health care, and education. These interrelated data can be more naturally represented as networks with nodes and edges. Inside this type of networks, there is usually a hidden community structure where each community represents a relatively independent functional module. Such hidden community structures are useful for many applications, such as word-of-mouth marketing, promoting decentralized social interactions inside organizations, and searching biological pathways related to various diseases. Therefore, how to detect hidden community structures becomes an important task with wide applications. Currently, modularity-based methods are widely-used among many existing community structure detection methods. They detect communities with more internal edges than expected under the null hypothesis of independence. Since research in correlation analysis also searches for patterns which occur more than expected under the null hypothesis of independence, this paper proposed a framework of changing the original modularity function according to different existing correlation functions in the correlation analysis research area. Such a framework can utilize not only the current but also the future potential research progresses in correlation analysis to advance community detection. In addition, a novel graphical analysis on different modified-modularity functions is conducted to analyze their different preferences, which are also validated by our evaluation on both real life and simulated networks. Our work to connect modularity-based methods with correlation analysis has several significant impacts on the community detection research and its applications to expert and intelligent systems. First, the research progress in correlation analysis can be utilized to define a more effective objective function in community detection for better detection results since different real-life applications might need communities with different resolutions. Second, any existing research progress for the modularity function, such as the Louvain method for speeding up the search and different extensions for overlapping community detection, can be applied in a similar way to the new objective function derived from existing correlation functions, because the new objective function is unified within one framework with the modularity function. Third, our framework opens a large unexplored area for the researchers interested in community detection. For example, what is the best heuristic search method for each different objective function? What are the characteristics of each objective function when applied to overlapping community detection? Among different extensions to overlapping community detection, which extension is better for each objective function?

1. Introduction

In the era of big data, companies and researchers collect a huge amount of data from different sources to gain insights for different applications. Many applications utilize data which only contain independent records, while a significant amount of applications utilize data with interrelated records. Such interrelated records can be more naturally represented as networks with nodes and

[*] Corresponding author.
E-mail addresses: lian.duan@hofstra.edu (L. Duan), yanchi.liu@rutgers.edu (Y. Liu), nick-street@uiowa.edu (W. Nick Street), hlu@scu.edu (H. Lu).

http://dx.doi.org/10.1016/j.eswa.2017.05.010
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

edges, which is widely applied in WWW (Page, Brin, Motwani, & Winograd, 1999), marketing (Bolander, Satornino, Hughes, & Ferris, 2015; Kozinets, De Valck, Wojnicki, & Wilner, 2010), business process (Huang & Kumar, 2012), management (Schilling & Fang, 2014), health care (Kroenke et al., 2013), education (Ahn & Rodkin, 2014), biology (Barabási, Gulbahce, & Loscalzo, 2011; Pilosof et al., 2014), and others (Sasidharan, Santhanam, Brass, & Sambamurthy, 2012). One important feature of networks is their community structures where nodes in the same community are more likely to interact with each other than nodes from different communities. Each community represents a relatively independent functional module on a given network, and community structure information can help us understand complex network systems. For example, the performance of word-of-mouth marketing is related to community-level factors (Kozinets et al., 2010), such as the interpersonal orientation of the word-of-mouth communications and the adaptation of commercial-communal tension. Communities in biological networks correspond to biological pathways related to various diseases (Barabási et al., 2011). In enterprise systems (Sasidharan et al., 2012), community structures can be used to identify influential persons in different communities and create more effective decentralized social interaction opportunities.

Although community structure information is essential to various applications, not all the related applications have a straightforward way to get community structure information. Generally speaking, it is relatively easy to get community structure information for some social network applications as people (nodes) can voluntarily create community labels, such as groups in Facebook, and provide their own community information. However, this method has two potential problems. First, not everyone will provide his/her community information. Therefore, only partial information is directly available. Second, this method is not applicable to other types of networks, such as biological networks, WWW, and telecommunication networks, because nodes in these networks cannot voluntarily provide community structure information. Therefore, an alternative way to get community structure information is highly desired. Besides voluntarily provided information, community structures can also be inferred through node interactions based on the fact that nodes in the same community are more likely to interact with each other than nodes from different communities. This alternative way is more widely used because capturing node interactions and inferring community structure information through these interactions is more practical than asking nodes to voluntarily provide community structure information. Therefore, searching for community structures through network topology becomes a very important research topic as it has wide applications in many areas.

2. Related work

Starting with the first related research (Weiss & Jacobson, 1955) in 1955 to detect communities in a working relationship network among government agency employees, hidden community structures now can be detected through many different methods, such as modularity-based methods (Clauset, Newman, & Moore, 2004; Shiokawa, Fujiwara, & Onizuka, 2013), spectral-based methods (Luxburg, 2007), divisive methods (Girvan & Newman, 2002), label propagation methods (Raghavan, Albert, & Kumara, 2007; Zhou, Lü, Yang, Wang, & Kong, 2015), density-based methods (Mancoridis, Mitchell, & Rorres, 1998), statistical-inference-based methods (Newman, 2013), Louvain methods (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008), local search methods (Kloumann & Kleinberg, 2014; Xin, Xie, & Yang, 2016), etc. Although all the existing methods have very different procedures to detect community structures, all their research progresses can be classified into three different perspectives: (1) feature selection, (2) objective function, and (3) search procedure.

2.1. Feature selection

As original feature spaces contain noisy information, which might reduce community structure detection performance, spectral-based methods (Luxburg, 2007) focuses on transforming data in a high dimensional space into a fewer relevant dimensional space. The l eigenvectors related to the l smallest eigenvalues in the Laplacian matrix (Luxburg, 2007) is used to transform the original adjacency matrix in spectral-based methods. After the change of feature space through the Laplacian matrix, community structures become more clear and any traditional clustering method like K-mean (MacQueen, 1967) can be used to detect communities on the transformed feature space.

2.2. Objective function

Objective functions are functions utilized to describe our goal in a mathematical format. Given a network with n nodes, it can be partitioned into l communities $P = \{G_1, G_2, \ldots, G_l\}$ where $1 \leq l \leq n$ and each node in the given network is assigned to one or many communities in the partition P. If each node is assigned to one and the only one community, the related partition P is a non-overlapping partition. If any node is assigned to more than one communities, the related partition P is an overlapping partition (Palla, Derényi, Farkas, & Vicsek, 2005; Xie, Kelley, & SZYMANSKI, 2013). Although there are many possible partitions in a given network with n nodes, only one of them perfectly matches the ground-truth community structure, and an ideal objective function should assign the highest score to the perfect matching partition. If such an ideal objective function can be defined, it can be utilized to find the partition with the highest value among all the possible partitions to reveal the ground-truth community structure. However, among many factors making the community detection problem hard to solve, one root cause is that there are many different ground-truth community structures for different purposes even in the same network. Take the social interaction network in a company for example. The non-overlapping ground-truth community structure is needed when assigning employees into non-overlapping groups for different full-time tasks, while the overlapping ground-truth community structure is needed for the information diffusion because the employees with multi-memberships serve as the important bridge nodes across different communities for information diffusion. In addition, the hierarchical structure, a special case of overlapping structures in many real-life networks, makes the community detection problem even harder. For example, one student can belong to one department and also one college at the same time in a university. The ground-truth community structure depends on which level of partition is needed given a specific purpose.

Since nodes in the same community are more likely to interact with each other than nodes from different communities, any objective function for community structure detection can be generalized into a balance between two sub-objectives: (1) more possible for nodes in the same detected community to be connected, and (2) less possible for nodes from different detected communities to be connected. As there are numerous ways to describe two sub-objectives through mathematical formulas and strike a balance between two objectives, many research efforts on community structure detection focus on how to create a better objective function. If the objective function is designed for overlapping community detection, an additional belonging factor vector (Xie et al., 2013) for each node will be included for the fuzzy assignment.