Physica A 445 (2016) 295–309

Contents lists available at ScienceDirect

Physica A

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

A scanning method for detecting clustering pattern of both attribute and structure in social networks

Tai-Chi Wang, Frederick Kin Hing Phoa *

Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 11529, Taiwan

HIGHLIGHTS

- We propose a new community detection method for network data via scan statistics.
- Both the network structure and node attribute are taken into account in this method.
- A frequentist likelihood ratio test is proposed for the significance of cluster existence.
- This method outperforms the existing methods (like CESNA) in several common criteria.
- We apply this method to three real-life demonstrated examples.

ARTICLE INFO

Article history:
Received 18 November 2014
Received in revised form 1 June 2015
Available online 10 November 2015

Keywords:
Social networks
Community/cluster detection
Scanning window
Scan statistic
Structure and attribute cluster
Statistical significance

ABSTRACT

Community/cluster is one of the most important features in social networks. Many cluster detection methods were proposed to identify such an important pattern, but few were able to identify the statistical significance of the clusters by considering the likelihood of network structure and its attributes. Based on the definition of clustering, we propose a scanning method, originated from analyzing spatial data, for identifying clusters in social networks. Since the properties of network data are more complicated than those of spatial data, we verify our method's feasibility via simulation studies. The results show that the detection powers are affected by cluster sizes and connection probabilities. According to our simulation results, the detection accuracy of structure clusters and both structure and attribute clusters detected by our proposed method is better than that of other methods in most of our simulation cases. In addition, we apply our proposed method to some empirical data to identify statistically significant clusters.

1. Introduction

Networks are widely applied to explain relationship among the nodes or actors. In general, networks can be represented as *graphs* to visually understand their structures and complexities. More explicitly, the visualizations of networks are used to unmask the relationship among the members of a club [1], to realize the protein interaction in a neural system [2], and to explore the food chain web in an ecological system [3]. Networks in different fields have different features, but a common feature in various types of networks is the cluster structure.

Clusters or communities are defined as groups of vertices that share common properties or play similar roles in a graph or network [4]. Cluster detection drew much attention due to its importance in applications. For examples, finding the purchase pattern of customers in a shop can facilitate customers' shopping and enhance the commercial profit [5], and determining

* Corresponding author. Tel.: +886 2 6614 5611x405; fax: +886 2 2783 1523.
E-mail addresses: taichi43@stat.sinica.edu.tw (T.-C. Wang), fredphoa@stat.sinica.edu.tw (F.K.H. Phoa).

http://dx.doi.org/10.1016/j.physa.2015.10.009
0378-4371/

the clustering relationship in a social network can efficiently deliver information among a group or an organization [6]. Therefore, approaches of cluster/community detection have been well developed recently [4].

Cluster detection methods are generally designed by comparing the similarity within the groups and difference between inside and outside the groups. One of the most popular methods is the modularity-based method [7], which considers a modularity measure to evaluate the similarity of groups and use a spectral optimization method to classify a network. There are many extended methods developed from this criterion, such as greedy techniques [8] and annealing methods [9]. However, the modularity optimization suffers from a severe limitation, that prevents it from finding clusters smaller than a given scale, even when they are very pronounced and easy to find [10]. Besides, a large value of modularity determined in a network does not guarantee the existence of a cluster [4], and most modularity methods do not have a statistical testing procedure to verify cluster existence. To determine the statistical significance, Lancichinetti et al. [11], [12] provided statistical testing procedures to test communities in networks.

Recent studies further paid attentions on networks with attributes. For example, a tendency called "homophily" [13] suggested that people usually interact with others whom are similar to themselves with some attributes [14]. Some studies also discussed how homophily affects network integration [15]. Zhou et al. [16] developed a distance-based transition probability, based on similarities of both structure and attribute, to construct a clustering algorithm. Yang et al. [17] modeled the links of network and node attributes to provide a probability regime to detect community memberships. Due to the network complexity, some heuristic algorithms are created to detect communities [18–20]. On the other hand, Handcock et al. [21] and Heard et al. [22] proposed Bayesian models, called latent position cluster model, to consider network transitivity and similarity on attributes and clustering simultaneously. In addition, Latent Dirichlet Allocation (LDA) models [23–25], which combined graph and content information, were provided for improving topic models and community discovery. Similar to other Bayesian models in other fields, Bayesian network models are also judged by their prior selections and the computing loadings [26]. Unfortunately, the numbers of nodes and parameters of a social network are usually too huge to efficiently construct a Bayesian network model.

Compared with the Bayesian models, frequentist testing methods that considered likelihood of network structure and attribute drew fewer attentions. Since the model of network structure and distributions of attributes are presumed, frequentist methods can avoid the problems addressed in Bayesian methods. Due to the similarity of data structure in spatial statistics and social networks, some frequentist methods originated from spatial statistics were applied for analysis of social networks in recent years [27,28]. Furthermore, Wang et al. [29] provided a scan statistic, which is generalized from the spatial scan statistic [30], to detect clusters in social networks. Nevertheless, only the statistic for testing structure clusters was provided and the testing performance is not discussed. Since most attributes have their own distributions, considering the distance or probability between attributes of nodes is not appropriate when the distributions of attributes considered as random variables are known. In this study, we extend the scan statistics to consider both structural network and its attribute.

We provide a whole picture of how to apply the scan statistics to identify clusters in social networks in Section 2, including an introduction of the relationship between scan statistics and social networks, and a derivation of testing statistics and testing procedures. In addition, a toy example for understanding the whole testing procedures is demonstrated in the end of this section. In Section 3, simulation studies are conducted to perform the powers of our proposed method for testing clusters of structure only (S-cluster in abbreviation), attribute only (A-cluster in abbreviation), and both of structure and attribute (SA-cluster in abbreviation) respectively. Our proposed method and the modularity method proposed by Newman and Girvan [7] are compared in Section 4. In Section 5, three different types of data, structure only, attribute only, and both of them, are used to illustrate our proposed method. Finally, a brief conclusion and discussions are reported in Section 6.

2. Scan statistic in social networks

2.1. Scan statistic and scanning window

In order to detect clusters in networks, we apply the idea of clustering in social networks. A measure of clustering usually refers to "clustering coefficient", which is defined as the degree which nodes in a graph tend to cluster to each other, that is, a pair of nodes with a common neighbor tend to connect to each other [31]. However, this measure cannot provide a statistical significance to tell which nodes shall be put together. A scan statistic is a useful tool to distinguish the difference when data are mixed with different components.

A scan statistic was first proposed for detecting clusters in time domain [32,33], and became popular in spatial domain [30]. However, there are few studies and methods [29] applied this approach to detect clusters in social networks. According to our experience and knowledge, networks and spatial statistics have many features in common, such as neighborhood structure and distances between neighbors. If fixed coordinates are given in a network and the edges are drawn in a graph, it will look like a spatial map. For example, Wong et al. [27] applied spatial models to describe network structure.

A scan statistic is provided for the purpose of comparing two disjoint subsets. In general, a scan statistic is a likelihood-based test statistic. Suppose a parameter of interest is θ and a subset is selected as Z. We can separately estimate θ in the disjoint subsets Z and Z^c, which are defined as θ_z and θ_c respectively, under the independent assumption between Z and Z^c.