Social Networks 54 (2018) 209–227

Contents lists available at ScienceDirect

Social Networks

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

Detecting node propensity changes in the dynamic degree corrected stochastic block model

Lisha Yu [a,*], William H. Woodall [b], Kwok-Leung Tsui [a]

[a] Department of System Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong
[b] Department of Statistics, Virginia Tech, Blacksburg, VA, USA

ARTICLE INFO

Article history:
Received 23 May 2017
Received in revised form 12 March 2018
Accepted 17 March 2018

Keywords:
Dynamic networks
Multivariate control charts
Network surveillance
Statistical process monitoring

ABSTRACT

Many applications involve dynamic networks for which a sequence of snapshots of network structure is available over time. Studying the evolution of node propensity over time can be important in exploring and analyzing these networks. In this paper, we propose a multivariate surveillance plan to monitor node propensity in the dynamic degree corrected stochastic block model. The method is flexible enough to detect anomalous nodes that arise from different mechanisms, including individual change, individuals switch, and global change. Experiments on simulated and case study social network data streams demonstrate that our surveillance strategy can efficiently detect node propensity changes in dynamic networks.

1. Introduction

Social networks consist of interactions between connected individuals or organizations often involving relationships such as communication, friendship or collaboration. With the rapid development of the Internet, social networks such as Facebook and Twitter form large dynamic networks. As dynamic social networks are constantly undergoing changes, analyzing these networks is crucial to understanding their structure and behavior. Our focus is on the detection of changes in individuals' propensities to communicate within a network that has community structure.

Monitoring to detect anomalous behavior within dynamic networks is known as network surveillance. The main tasks in network surveillance are to detect change-points in time at which a subset of the network deviates from normal behavior, and to identify the particular parts of the network that are responsible for the change. A considerable amount of research specifically designed for anomaly detection in the dynamic network environment has been proposed. Methods have been successfully applied to a number of domains, including fraud detection (Akoglu et al., 2010; Hassanzadeh et al., 2012), threat detection (Eberle et al., 2010; Chen et al., 2011), review spam detection (Jindal et al., 2010; Fire et al., 2012), financial trade fraud detection (Li et al., 2012) and auction fraud detection (Chau et al., 2006). An overview of methods was given in recent review papers by Savage et al. (2014), Ranshous et al. (2015), Bindu and Thilagam (2016) and Woodall et al. (2017).

Social network surveillance for anomaly detection has attracted much recent attention. It is a relatively new research area as pointed out by McCulloh and Carley (2011), Savage et al. (2014) and Woodall et al. (2017). The social network surveillance literature seems to have been developed somewhat independently of the computer network surveillance literature. The structure of social network data is different from that of other types of networks and the objectives of monitoring are typically different. The goal of social network surveillance is to prospectively monitor the interactions among individuals so as to detect the unexpected behavior of a particular individual or detect sudden changes in the behavior of groups of individuals. A general framework for guiding social network surveillance is statistical process monitoring (Wilson et al., 2017). The philosophy behind statistical process monitoring is to distinguish between common-cause variation that is attributable to a relatively stable underlying process, and special-cause variation that is unusual for the underlying process. In general, statistical process monitoring provides a methodology for the real-time monitoring of any characteristic of interest.

Some centrality metrics have been proposed to evaluate node importance. A common approach is to monitor these extracted metrics through time. As an example, McCulloh and Carley (2011) constructed control charts based on different network centrality measures. In the work of Priebe et al. (2005), Marchette (2012) and Neil et al. (2013), scan-based network monitoring schemes

* Corresponding author at: City University of Hong Kong, 83 Tat Chee Ave, Kowloon, Hong Kong.
E-mail addresses: lishayu2-c@my.cityu.edu.hk (L. Yu), bwoodall@vt.edu (W.H. Woodall), kltsui@cityu.edu.hk (K.-L. Tsui).

https://doi.org/10.1016/j.socnet.2018.03.004
0378-8733/

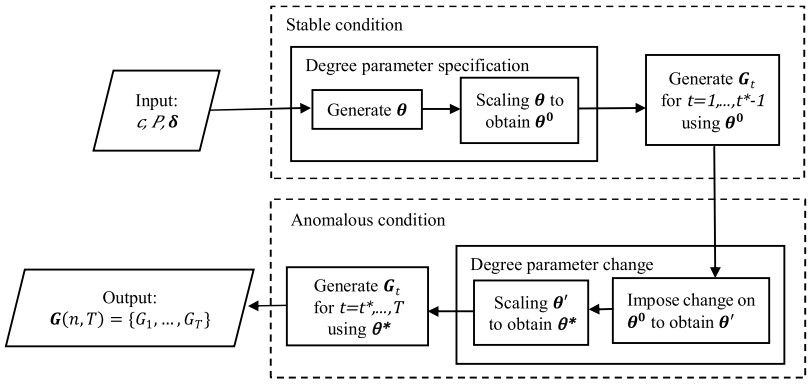

Fig. 1. Flowchart for generating dynamic DCSBM.

were proposed. Similarly, Park et al. (2013) used a fusion of network statistics to detect changes in a stream of networks. By taking advantage of node attributes, Azarnoush et al. (2016) modeled the probability of edge existence as a function of node attributes and applied likelihood ratio methods to detect changes arising from different network edge formation mechanisms.

Simulated social networks are required to evaluate fully the performance of a proposed method instead of solely applying it to a well-studied network that perhaps best fits the criteria of the method itself (Savage et al., 2014; Woodall et al., 2017; Zhao, 2017). This is required to understand how and when a particular method performs well. Motivated by this, Wilson et al. (2017) proposed the use of the degree-corrected stochastic block model (DCSBM) of Karrer and Newman (2011) to model and monitor dynamic networks that undergo a significant change. We note that the DCSBM is a generalization of the stochastic block model of Nowicki and Snijders (2001). The DCSBM provides a probability distribution for undirected graphs with Poisson-distributed edge weights. The DCSBM specifies the propensity of connection between nodes and captures two important aspects of social networks, heterogeneous connectivity and community structure. The use of the dynamic DCSBM proposed in Wilson et al. (2017) provides a means to model and monitor a social network to detect certain types of change.

Different structural changes can be introduced into the dynamic DCSBM. Individuals could change their behavior over time. For example, one faculty member in an academic department may significantly increase his interaction with other faculty members before a promotion to the role of department head. In this example, the change of interaction will affect that member's propensity to communicate. Wilson et al. (2017) focused on modeling and detecting only increases in the variation of the propensity to communicate for nodes from the same community. Their approach can not be used to detect changes of the propensity for particular nodes. In addition, in their model, they chose the parameter that quantifies the individual propensity as a random variable. At each time step, a different propensity value was generated for each individual. Motivated by the limitations of their approach, we put our focus on node propensity change detection in dynamic social networks using the DCSBM.

We propose a method to detect quickly node propensity changes over time based on the dynamic DCSBM. The following three types of node propensity changes are considered: (i) change of an individual's propensity; (ii) switch between two individuals' propensities; and (iii) change of variability of propensity. The node propensity values in each community of the DCSBM are represented as a multivariate vector that contains estimated individual node propensities as its components. Various multivariate statistical process monitoring methods can be applied for monitoring these vectors. We propose to monitor for anomalous nodes in each community using the compositional T^2 control chart. A joint monitoring scheme at the community level is then proposed. With this approach, the number of statistics we monitor depends on the number of communities instead of the number of nodes in the network.

The remainder of this paper is organized as follows. In the next section, we define the DCSBM, followed by a detailed explanation of the proposed method in Section 3. In Section 4 we present a numerical study to investigate the performance of the proposed method as well as case studies on the MIT Reality Mining personal mobility dataset and the well-studied Enron email network to illustrate the method. Finally, in Section 5 we give some concluding remarks and directions for future work.

2. The degree corrected stochastic block model

2.1. The model

In many applications, nodes can be naturally divided into different communities. The block model is a widely used model for networks with communities (Holland et al., 1983; Nowicki and Snijders, 2001). Consider an undirected graph $G=(V,E)$, where V is the set of n nodes and E is the set of weighted edges. The network can be mathematically represented by its adjacency matrix, an $n\times n$ symmetric matrix $\boldsymbol{A}=\left[A_{ij}\right]$, where A_{ij} is equal to the weight of the edge between node i and j when $i\neq j$. Since self-loops are not allowed, the adjacency matrix has zeros on its diagonal. Let a represent the number of communities, and $\boldsymbol{c}=(c_1,\ldots,c_n)$ contains the community labels where c_i is the community corresponding to node i. To handle variation in the degree distribution, each node i is also assigned an additional propensity parameter θ_i, $i=1,\ldots,n$, which reflects the propensity of the node to connect. Following the DCSBM introduced in Karrer and Newman (2011), the edge variables A_{ij} are independent Poisson random variables with mean $\theta_i\theta_j P_{C_iC_j}$ where $\boldsymbol{P}$ is a $a\times a$ symmetric matrix which contains the propensity of connection between nodes in communities c_i and c_j, i.e.,

$$A_{ij}\sim Poisson\left(\theta_i\theta_j P_{C_iC_j}\right) \tag{1}$$

Sometimes, we call θ_i the "degree parameter" associated with node i, reflecting its individual propensity to form ties.

The parameters $\boldsymbol{\theta}=\left(\theta_1,\ldots,\theta_n\right)$ are arbitrary to within a multiplicative constant which is absorbed into the $\boldsymbol{P}$ parameters. Thus identification requires constraints, and convenient ones force the θ_i values to sum to 1 within each community, i.e.,

$$\sum_{i:C_i=r}\theta_i=1 \tag{2}$$