Statistics and Probability Letters 121 (2017) 54–60

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A note on jointly modeling edges and node attributes of a network

Haiyan Cai

Department of Mathematics and Computer Science, University of Missouri - St. Louis, St. Louis, MO 63121, United States

ARTICLE INFO

Article history:
Received 28 November 2015
Received in revised form 11 October 2016
Accepted 12 October 2016
Available online 15 October 2016

MSC:
60
62
82

Keywords:
Network model
Dynamical network
Random graph
Limiting graph

ABSTRACT

We are interested in modeling networks in which the connectivity among the nodes and node attributes are random variables and interact with each other. We propose a probabilistic model that allows one to formulate jointly a probability distribution for these variables. This model can be described as a combination of a latent space model and a Gaussian graphical model: given the node variables, the edges will follow independent logistic distributions, with the node variables as covariates; given edges, the node variables will be distributed jointly as multivariate Gaussian, with their conditional covariance matrix depending on the graph induced by the edges. We will present some basic properties of this model, including a connection between this model and a dynamical network process involving both edges and node variables, the marginal distribution of the model for edges as a random graph model, its one-edge conditional distributions, the FKG inequality, and the existence of a limiting distribution for the edges in an infinite graph.

1. Introduction

In modeling networks (Goldenberg et al., 2010; Kolaczyk, 2009; Newman, 2010), the usual focus is on the network topologies. A network is typically modeled as a random graph (Bollobas, 2001; Durrett, 2010) defined in terms of a probability distribution of the edge status. However, networks in many applied problems are not always just about links or edges. More extensive data for certain networks, containing information not only for edges but also for some node variables or attributes, are becoming available. For such data and for some important problems in network study, a limitation of the random graph model is the absence of information from the node variables. Such a model is incapable of catching interactions between edges and nodes. In a social network problem, for example, one might be interested in studying users' behaviors (or some dynamical attribute in the user-profiles) as a function of the network topology, or vice versa (McAuley and Leskovec, 2012; Mislove et al., 2010), or in a gene network problem, one might be interested in inferring gene expression levels as a function of an underlying regulatory network, or vice versa (Wang and Huang, 2014). When it comes to analyzing behaviors of the nodes in a network or the influence of node behaviors on network topologies, the utility of the random graph models becomes limited.

The latent space model (Handcock and Raftery, 2007; Hoff et al., 2002; Kolaczyk, 2009) is another popular network model. It does assume the dependence of the edge probabilities on some node variables. The model however treats these variables as latent variables, paying little attention to the inference on these variables. On the other hand, a Gaussian graphical model describes a distribution for node variables on a network with built-in edge information of the network (through the inverse covariance matrix). It however treats the network topology as a static parameter which remains constant regardless how the node variables will change.

E-mail address: haiyan_cai@umsl.edu.

http://dx.doi.org/10.1016/j.spl.2016.10.014
0167-7152/© 2016 Elsevier B.V. All rights reserved.

In this paper, we propose a joint probability distribution for both edges and node variables (Section 2). A study of such a model can shed lights on how edges and nodes interact with each other in a network so that information for both edges and nodes can be utilized in studying the networks. In a way, our model can be described as a combination of the latent space model and the Gaussian graphical model: given the node variables, the edges will follow independent logistic distributions, with the node variables as covariates in the logistic function; given edges, the node variables will be distributed jointly as multivariate Gaussian, with their conditional covariance matrix depending on the graph induced by the edges. In terms of the marginal distribution for the node variables, our model generalizes the Gaussian graphical model to allow for the underlying graphical structure to be random. In other words, it is now a mixture of Gaussian graphical models over all the possible edge configurations of the network. Our model also leads to a non-trivial and interesting random graphical model when we take the marginal distribution for the edges.

A reason that motivates us to propose such a model is that it provides a sensible framework for modeling network dynamics in which the edge status and the node variables change their values over time, as we will explain in the end of Section 2. We will see that the dynamical system updates edge status and node variables alternatively according to the conditional distributions between edges and nodes. The equilibrium (stable) distribution of the dynamical system is then exactly the joint distribution we propose here. In other words, our model can be viewed as the stable probability law of a dynamical network process.

We will pay particular attention to the marginal distribution for edges of our model. An explicit formula for the conditional probability of one edge given all other edges is given in Section 3. We will show in Section 4 that the probability distribution for edges is positively associated in the sense that it satisfies the FKG inequality (Grimmett, 2006; Holley, 1974), a property that is shared by many well-known models in statistical mechanics. We then give a weak convergence result for the edge distribution based on the FKG inequality. To ensure consistent results in statistical analysis for very large networks, it is essential that the model, as a probability law, has a limiting distribution. The concept of the limit for random graphs we use here is that of the infinite-volume Gibbs distributions (Georgii, 1988) on graphs, involving both nodes and edges (see Grimmett, 2006 for an example). We also note that there is a close similarity between our model and the random-cluster model derived in the statistical mechanics (Grimmett, 2006): what our model is to the Gaussian graphical model is in some sense similar to what the random-cluster model is to Ising or Potts models. This is indeed another reason that motivated us to propose the model in this note.

2. The random Gaussian graphical model

We will call our model the random Gaussian graphical model and formulate it in this section. Let $G = (V, E)$ be a finite simple graph (undirected, unweighted, no loops, no multiple edges) with E being a subset of $V \times V$ which is fixed. Suppose $|V| = m$ and $|E| = n$. For convenience, we identify V as the integer set $V = \{1, \dots, m\}$. Suppose associated with each node $i \in V$ there is a random variable X_i, representing an attribute of node i. Let $X = \{X_1, \dots, X_m\}$. We will use $x \in \mathcal{R}^m$ to denote a generic value of X. We write (i, j) for the edge in E which is incident with the nodes $i, j \in V$.

We will consider random sub-graphs of G in which V remains the same and E is reduced randomly to some subset of itself. Such a random graph can be represented by a random adjacency matrix $A = \{A_{ij},\ i, j \in V\}$ in which all the diagonal elements $A_{ii} = 0$, and for each edge $(i, j) \in E$, $A_{ij} = A_{ji} = 1$ if the edge is present in the random graph, and $A_{ij} = A_{ji} = 0$ if otherwise. It is always understood that $A_{ij} \equiv 0$ for all $(i, j) \notin E$. We will call A_{ij} an edge variable. With a slight abuse of notation, we let $\mathcal{A} = \{0, 1\}^E$ be the set of all possible values of A. We will use $a = a^T \in \mathcal{A}$ to denote a generic value of the adjacency matrix A.

By “random Gaussian graphical model” we mean the following joint probability density for variables A and X, defined on the space $\mathcal{A} \times \mathcal{R}^m$,

$$\mu(a, x) \equiv \frac{1}{Z} \exp\left\{-\frac{1}{2} H(a, x)\right\}, \quad (a, x) \in \mathcal{A} \times \mathcal{R}^m, \tag{1}$$

where

$$H(a, x) = \alpha \sum_i x_i^2 + \beta \sum_{(i,j)\in E} a_{ij}(x_i - x_j)^2 \tag{2}$$

for some parameters $\alpha > 0$ and $\beta \ge 0$, and Z is the normalizing constant. It is clear that this Z is always finite.

We note that in this model, if all $a_{ij} = 1$ it becomes an usual Gaussian graphical model (as we will see below). Therefore, we can consider the Gaussian graphical model as a “full model” relative to the given edge set E while model (1) as a model that allows us to “turn off” some edges in E at random according to the values of a_{ij}’s. In particular, if all $a_{ij} = 0$ and therefore there are no connections among the nodes in the graph, X_i’s are independent $N(0, 1/\alpha)$ random variables. The joint distribution of a_{ij}’s in turn depends on X_i’s. In general, the likelihood of connectivity among the nodes is determined by the magnitudes of the differences between the corresponding node variables and the value of β. On the other hand, the connectivity of the nodes will, in turn, affect the distribution of the node variables.