Social Networks 49 (2017) 27–36

Contents lists available at ScienceDirect

Social Networks

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

Penalized component hub models

Charles Weko [a], Yunpeng Zhao [b,*,1]

[a] Headquarters, Department of the Army, United States Army, United States
[b] Department of Statistics, George Mason University, United States

ARTICLE INFO

Article history:
Available online 5 December 2016

Keywords:
Social network analysis
Regularization method
Finite mixture model
Hub model

ABSTRACT

Social network analysis presupposes that observed social behavior is influenced by an unobserved network. Traditional approaches to inferring the latent network use pairwise descriptive statistics that rely on a variety of measures of co-occurrence. While these techniques have proven useful in a wide range of applications, the literature does not describe the generating mechanism of the observed data from the network.

In a previous article, the authors presented a technique which used a finite mixture model as the connection between the unobserved network and the observed social behavior. This model assumed that each group was the result of a star graph on a subset of the population. Thus, each group was the result of a leader who selected members of the population to be in the group. They called these *hub models*.

This approach treats the network values as parameters of a model. However, this leads to a general challenge in estimating parameters which must be addressed. For small datasets there can be far more parameters to estimate than there are observations. Under these conditions, the estimated network can be unstable.

In this article, we propose a solution which penalizes the number of nodes which can exert a leadership role. We implement this as a *pseudo-Expectation Maximization* algorithm.

We demonstrate this technique through a series of simulations which show that when the number of leaders is sparse, parameter estimation is improved. Further, we apply this technique to a dataset of animal behavior and an example of recommender systems.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Networks consist of discrete *nodes* or *vertices* which are connected by *links* or *edges*. These pairwise connections are frequently represented by a square matrix called an *adjacency matrix*. Network analysis has drawn attention in a wide variety of scientific and engineering disciplines because of the practicality of the network structure. The applications of networks include concrete problems such as finding the shortest path through a transportation system or determining the maximum flow through a n electrical transmission system (Hiller and Lieberman, 2001). The generality of networks allows for their application to more abstract problems such as the propagation of disease or information through a population (Jackson, 2008). Applications further extend to identifying key nodes in social networks (Koschützki et al., 2005), community detection among weblogs on the World Wide Web (Karrer and Newman, 2011), link prediction in social and biological networks (Liben-Nowell and Kleinberg, 2007; Zhao et al., 2013), as well as many others (Kolaczyk, 2009; Goldenberg et al., 2010; Newman, 2011).

Traditionally, statistical network analysis focuses on modeling the random generation of observed or *explicit* network structure. For physical networks, like communication systems or railway networks, the nodes are clearly defined and the links between nodes can be directly observed (Hiller and Lieberman, 2001; Kolaczyk, 2009; Newman, 2011).

In other fields of research, explicit network structure may not be observable. This is especially true in the social sciences where the observed raw data is usually the social behavior instead of an explicit network structure (Freeman et al., 1989; Whitehead, 2008). This situation may also occur in the analysis of protein–protein interaction or gene regulatory networks. In these situations, the

* Corresponding author at: Department of Statistics, George Mason University, Volgenau School of Engineering, 4400 University Drive, MS 4A7, Fairfax, VA 22030-4444, United States.
E-mail addresses: charles.w.weko.mil@mail.mil (C. Weko), yzhao15@gmu.edu (Y. Zhao).
[1] This work is supported by NSF Grant DMS 1513004..

http://dx.doi.org/10.1016/j.socnet.2016.09.003
0378-8733/© 2016 Elsevier B.V. All rights reserved.

Table 1
Dataset for six children and three birthday parties, Adapted from (Wasserman and Faust, 1994).

Party	Child					
	Allison	Drew	Eliot	Keith	Ross	Sarah
1	1	0	0	0	1	1
2	0	1	1	0	1	1
3	1	0	1	1	1	0

observed behavior is presumed to result from a latent network structure. For instance, researchers may not directly observe "friendships" within a population; instead, they may observe some social behavior (e.g., four people gather together with a certain frequency or they visited each other's house at least once in a month).

The notion that there is a connection between observable behavior and network structure can be traced to the so-called *social network perspective* proposed by Moreno (1934). Wasserman and Faust (1994) also gave a detailed explanation of this concept. The central principle of the social network perspective is that a network model governs the action of individual nodes and makes them behave interdependently. This relationship between behavior and network structure suggests that the network may be inferred from such observed behavior. In a previous article, Zhao and Weko (2016) developed a model which used the network as a parameter for the random generation of observed behavior.

The construction of latent networks often relies on data structures generated from surveys in which individuals or researchers report relationships (Sampson, 1969; Zachary, 1977). In this article we focus on an alternative type of dataset which is frequently collected in the social sciences and which can be generalized to other areas of research. Wasserman and Faust (1994) introduce such a dataset using the example of children attending birthday parties. In Table 1, the value 1 indicates that a specific child attended a party, and 0 indicates otherwise. For example, Allison attended Parties 1 and 3 but did not attend Party 2. Whitehead (2008) refers to each party as a *group* and Table 1 as a *group-by-individual matrix*. Zhao and Weko (2016) referred to this type of data as *grouped data*.

The existing methods for network inference from grouped data are essentially descriptive statistics. The most common approach is to use the frequency of co-occurrence between two nodes to estimate the strength of the link between individuals (Zachary, 1977; Freeman et al., 1989; Wasserman and Faust, 1994; Kolaczyk, 2009). We refer to this measure as the *co-occurrence matrix*. As an alternative, the *half weight index* (Dice, 1945; Cairns and Schwager, 1987; Bejder et al., 1998; Whitehead, 2008) estimates the strength of the link by the frequency that two nodes co-occur given that one of them is observed.

One shortcoming of these techniques is that they do not define how the observed data is generated from the estimated statistics. A particular challenge is that the probability of co-occurrence is not equivalent to the probability of connection. For example, in Table 1 it is possible that two children who do not know each other attended the same party because they are invited by a mutual friend. It remains unclear what model assumption justifies the network structure inferred by these measures.

Zhao and Weko (2016) proposed a simplistic generating mechanism for grouped data based on a network structure. The *hub model* (HM) assumes that each observed group is the result of a leader bringing together a subset of the population. That is, every group is brought together by a central node (often referred to as the *leader*). The other members of the group are present based on their relationship to this leader. Thus, the hub model parameters have an interpretation which can be easily applied to relevant research questions.

Despite the fact that hub models assume an intuitive generating mechanism and perform well with sufficient observations, the number of parameters in the model presents a challenge. If we let n be the number of nodes in the network, the network contains $O(n^2)$ parameters. Therefore, a moderate-sized network (e.g., $n = 50$) would in principle require a large number of observations to accurately estimate the network.

Moreover, in most practical situations, the central node of each group is unobserved. Without any prior information, it is possible for any node in the group to be the central node for that group. Zhao and Weko (2016) use an Expectation-Maximization (EM) algorithm to identify the central node for each group. As n increases, the possibility for larger and larger groups also increases. Thus identifying the central node of such groups can be difficult because there are many nodes which could be central and the probability of each node being central can be small.

In practice, it is not necessary to model every node in the population as a potential leader. For example, there may be low ranking members of the population who do not have the authority or influence to initiate a group. This is especially true when the number of observations is small. Therefore, we propose a *penalized component hub model* (PCHM) to reduce the hub model's complexity. Using a penalized likelihood of hub models, the probability that a node is a leader is shrunk towards 0 when that probability is small.

The PCHM assumes *sparse* parameters. That is, only a small proportion of the nodes have a non-zero probability of forming a group. Since the hub model is an example of a finite mixture model, we essentially penalize the number of components in the mixture model.

This penalization technique belongs to the class of regularization methods which have been extensively studied in the statistical literature. For example, least absolute shrinkage and selection operator (LASSO) introduced by Tibshirani (1996) is a famous L_1 regularization method for variable selection in linear regression. Ridge regression (Hoerl and Kennard, 1970) applies L_2 regularization to reduce the variance of the coefficients estimates and hence obtains smaller mean square error than least square estimates. Similarly in the PCHM case, regularization on the probabilities of nodes being leaders increases the stability of the estimated networks and yields better performance when the sample size is limited.

Regularization techniques have been widely used in graphical models and covariance estimation to obtain a "sparse" estimated adjacency matrix (Bickel and Levina, 2008; Friedman et al., 2008; Guo et al., 2010). However, the definition of "sparse" in these techniques is different from the definition we will use. Traditional techniques define the network structure solely based on an adjacency matrix. Thus a "sparse" network is one where the adjacency matrix contains many elements which are equal to zero. In this case, regularization of the network is achieved by penalizing the elements of the adjacency matrix of the network.

Hub models define the network structure using two parameters (a mixing distribution and an adjacency matrix). Under PCHM, sparsity is defined on the mixing distribution. Thus PCHM penalizes the probability of nodes being centers. The detailed explanation motivating this approach will be given in Section 2 and further elaborated in Section 3.

The rest of this article is organized as follows. We start with a brief review of hub models to motivate our approach in Section 2. We propose PCHM and the algorithm for solving the penalized likelihood in Section 3. In Section 4, we discuss the application of the Bayesian Information Criterion (BIC) for tuning parameter selection. Simulation studies are provided in Section 5. In Section 6, we apply the PCHM to a dataset of Hector's dolphins (Bejder et al., 1998) and a recommender system in supplemental materials.

Download English Version:

https://daneshyari.com/en/article/5126786

Download Persian Version:

https://daneshyari.com/article/5126786

Daneshyari.com