Social Networks 49 (2017) 113–123

Contents lists available at ScienceDirect

Social Networks

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

Network autocorrelation models with egocentric data

Daniel K. Sewell

Department of Biostatistics, University of Iowa, Iowa City, IA 52242, United States

ARTICLE INFO

Article history:
Available online 23 January 2017

Keywords:
Actor attributes
Bayesian estimation
Social influence
Spatial autoregressive model

ABSTRACT

Network autocorrelation models have been widely used for decades to model the joint distribution of the attributes of a network's actors. This class of models can estimate both the effect of individual characteristics as well as the network effect, or social influence, on some actor attribute of interest. Collecting data on the entire network, however, is very often infeasible or impossible if the network boundary is unknown or difficult to define. Obtaining egocentric network data overcomes these obstacles, but as of yet there has been no clear way to model this type of data and still appropriately capture the network effect on the actor attributes in a way that is compatible with a joint distribution on the full network data. This paper adapts the class of network autocorrelation models to handle egocentric data. The proposed methods thus incorporate the complex dependence structure of the data induced by the network rather than simply using ad hoc measures of the egos' networks to model the mean structure, and can estimate the network effect on the actor attribute of interest. The vast quantities of unknown information about the network can be succinctly represented in such a way that only depends on the number of alters in the egocentric network data and not on the total number of actors in the network. Estimation is done within a Bayesian framework. A simulation study is performed to evaluate the estimation performance, and an egocentric data set is analyzed where the aim is to determine if there is a network effect on environmental mastery, an important aspect of psychological well-being.

1. Introduction

Network autocorrelation models can help capture complex dependencies in individual level data and can also estimate how and to what extent an individual's network influences that individual's attributes or behaviors. Fujimoto et al. (2011) describes the network autocorrelation model as "a workhorse for modeling network influences on individual behavior." Wang et al. (2014) states "The network autocorrelation model has some clear advantages over other conventional approaches (e.g., egocentric or dyadic) in that it simultaneously accommodates network effects and individual attributes." This class of models has been used for decades in a variety of contexts, such as determining the network effect on gender roles in labor (White et al., 1981), educational and occupational aspirations (Duke, 1991), U.S. interstate commodity flows (Chun et al., 2012), policy influence (Carpenter et al., 1998), task performance (Carr and Zube, 2015), and phylogenetics (Björklund, 1990).

Network autocorrelation models describe stochastic data generating processes using the joint distribution of all actor attributes given the network structure. This is both a benefit and a curse. The positive aspect of this, and indeed the motivation for employing such an approach, is that by jointly modeling all actors in the network, the complex dependence structure is explicitly modeled, and social influence can be directly quantified and estimated. The downside is that to utilize such a model, one needs to collect data on all actors in the network. This can be a problem for (at least) four reasons. First, often times the network is simply too large to sample (Granovetter, 1976), or there are monetary constraints to obtaining data on all the actors of the network. Second, the actors of the network may not be easily accessible to the researchers, especially if the network is defined by controversial or illegal behaviors, or will not give consent in human subject research. Third, there may be nonresponses in the data collected. This has been a widely noted problem, and has the potential to be particularly damaging to network analyses (Stork and Richards, 1992). Fourth, the boundary of the network may not be identifiable. For example, suppose one wishes to know the network effect of peers on adolescent behaviors. Is the network of interest defined by all adolescents in a particular class or school? Or perhaps it can be defined by some on-line social media platform? Or perhaps it is all adolescents in a particular city, state, or country? Doreian (1989) makes the statement, which still holds true today, "locating boundaries remains a persistent and vexing problem."

E-mail address: daniel-sewell@uiowa.edu

http://dx.doi.org/10.1016/j.socnet.2017.01.001
0378-8733/

Researchers often avoid the difficulty of collecting data on all actors of the network by obtaining a subsample of the actors and focusing on the ties involving the sampled actors. The resulting data is referred to as egocentric network data. This type of data can be collected in a variety of ways, such as a simple random sample, targeted sampling, snowball sampling, respondent-driven sampling, etc (see, e.g., Heckathorn, 1997). Egocentric network analyses have been used to study interorganizational collaborations (Ahuja, 2000), health behaviors (O'Malley et al., 2012), personal and group communication (Fisher, 2005), contraceptive use (Behrman et al., 2003), support network after cancer diagnoses (Ashida et al., 2009), and many others.

The use of egocentric data has been limited primarily to the study of either dyadic relationships or structural/positional measures of the entire network (Provan et al., 2007). Recently there has been work on utilizing a subclass of exponential random graph models to estimate homophily and network structure (Krivitsky and Morris, 2015). Methods to study actor attributes using egocentric data are more limited in scope; this type of analysis is often done in an ad hoc manner by using as a covariate some summary statistic of the egos' personal networks such as density, network size, or an average of some alter attribute. Doing so ignores the network effect on the covariance matrix of the dependent variable; that is, ignoring the network effect naively implies homoscedastic independent errors.

This paper proposes a method that allows researchers to perform linear regression on egocentric network data that accounts for the heteroscedasticity and correlation that exist in such data while simultaneously estimating the effect that social influence has on the response variable of interest. To the author's knowledge there is a lack of methodology that appropriately accounts for the complex dependence structure sure to exist within egocentric data, and no mechanism for egocentric data to be used to learn the effects of social influence. Very purposefully, the proposed methodology adapts the widely used network autocorrelation models rather than inventing some new model to handle egocentric data; thus we do not require researchers to assume a different data generating mechanism based on how much data has been observed. Further, as the proposed method is derived directly from the joint distribution of all actors in the network, one can study network data even if the boundary of the network is unknown or ill-defined.

Section 2 describes the proposed methodology and Bayesian estimation. Section 3 describes a simulation study that compares the performance of the estimation as the true model parameters and underlying network itself varies. Section 4 shows the results from applying the proposed method to an egocentric data set of adults in a rural southeastern Iowa town, with the goal of determining if there is a network effect on environmental mastery. Section 5 provides a brief discussion.

2. Methods

Suppose that we wish to make inference regarding a graph augmented with actor attributes. We may view this as a triple $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{C})$; $\mathcal{V}$ is the set of vertices, or actors, of the network, $\mathcal{E}$ is the set of edges, or relations, between the vertices, and $\mathcal{C}$ is the set of actor attributes, or characteristics, on $\mathcal{V}$. We will denote $|\mathcal{V}|$, the number of actors, by n. Typically one may represent $\mathcal{E}$ by an adjacency matrix A, where the ith row jth column entry of A is 1 if there is an edge between actors i and j and 0 otherwise. The actor attributes $\mathcal{C}$ can be partitioned into the $n \times 1$ response variable vector $\boldsymbol{y}$ and the $n \times p$ matrix of covariates X. The goal is to try to determine how the covariates X and the network affect the response $\boldsymbol{y}$. This is typically accomplished via network autocorrelation models.

The network autocorrelation model has its genesis in spatial statistics (e.g., Ord, 1975; Doreian, 1980). It was soon borrowed by researchers studying complex network data to great effect (e.g., Dow et al., 1982). There are two variations on a theme, namely, (borrowing nomenclature from Doreian, 1980), the network effects model, given by

$$\boldsymbol{y} = X\boldsymbol{\beta} + \rho A\boldsymbol{y} + \boldsymbol{\epsilon}, \tag{1}$$

and the network disturbances model, given by

$$\begin{aligned} \boldsymbol{y} &= X\boldsymbol{\beta} + \boldsymbol{\nu}, \\ \boldsymbol{\nu} &= \rho A\boldsymbol{\nu} + \boldsymbol{\epsilon}, \end{aligned} \tag{2}$$

where $\boldsymbol{\beta}$ is the parameter vector of coefficients, ρ is the coefficient which captures the network effect, and $\boldsymbol{\epsilon}$ is a vector of zero mean independent normal random variables with homogeneous variance σ^2. Note that an equivalent but more concise form of (2) is

$$\boldsymbol{y} = X\boldsymbol{\beta} + \rho A(\boldsymbol{y} - X\boldsymbol{\beta}) + \boldsymbol{\epsilon}. \tag{3}$$

For egocentric data, $\mathcal{G}$ is only partially observed. Fig. 1 illustrates an egocentric network for a small toy data set. The set of actors $\mathcal{V}$ can be partitioned into the sampled egos $\mathcal{V}_e$, the egos' alters $\mathcal{V}_a$ (those unsampled actors with whom the egos have ties), and all other actors in the network $\mathcal{V}_o$, so that $\mathcal{V} = \mathcal{V}_e \cup \mathcal{V}_a \cup \mathcal{V}_o$. We should clarify here that egos may claim ties with other egos. So long as an actor has been sampled we will, in this paper, refer to that actor as an ego. Thus it may be that egos have ties amongst other egos as well as with individuals who have not been sampled (alters). Let n_e, n_a, and n_o denote the number of egos, the number of the egos' alters, and the number of remaining actors in the network respectively, so that $n = n_e + n_a + n_o$. We can partition $\mathcal{E}$ by focusing on the adjacency matrix A, specifically

$$A = \begin{pmatrix} A_e & A_{ea} & \mathbf{0} \\ A'_{ea} & A_a & A_{ao} \\ \mathbf{0} & A'_{ao} & A_o \end{pmatrix},$$

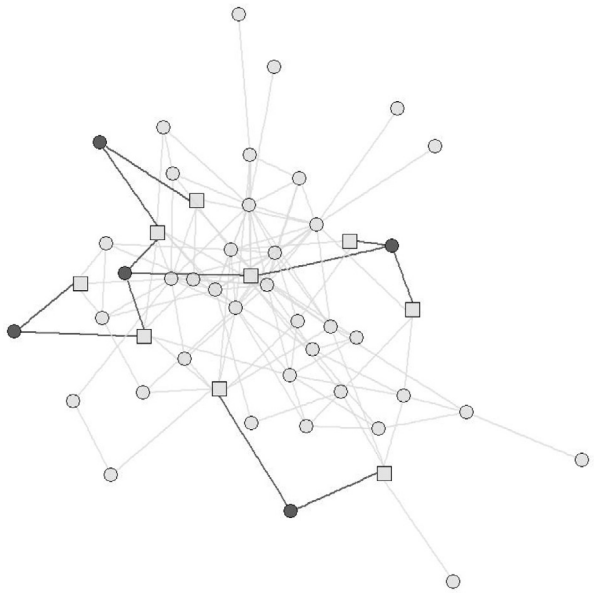

Fig. 1. A toy example of an egocentric network, where the dark circles are the sampled egos, the dark edges are the observed egos' edges, the squares are the alters, and the light gray circles and lines are the unobserved actors and edges respectively.