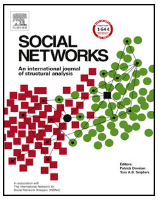
Social Networks 38 (2014) 88–99

Contents lists available at ScienceDirect

Social Networks

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



Statistical power of the social network autocorrelation model

Wei Wang[a,*], Eric J. Neuman[b], Daniel A. Newman[b]

[a] University of Central Florida, United States
[b] University of Illinois at Urbana-Champaign, United States

ARTICLE INFO

Keywords:
Network autocorrelation model
Social network analysis
Statistical power

ABSTRACT

The network autocorrelation model has become an increasingly popular tool for conducting social network analysis. More and more researchers, however, have documented evidence of a systematic negative bias in the estimation of the network effect (ρ). In this paper, we take a different approach to the problem by investigating conditions under which, despite the underestimation bias, a network effect can still be detected by the network autocorrelation model. Using simulations, we find that moderately-sized network effects (e.g., ρ = .3) are still often detectable in modest-sized networks (i.e., 40 or more nodes). Analyses reveal that statistical power is primarily a nonlinear function of network effect size (ρ) and network size (N), although both of these factors can interact with network density and network structure to impair power under certain rare conditions. We conclude by discussing implications of these findings and guidelines for users of the autocorrelation model.

1. Introduction

Identifying and determining network effects are some of the major goals and unique advantages of social network analysis. Of the many models proposed to investigate network effects on individual outcomes, the network autocorrelation model (Anselin, 1988; Cliff and Ord, 1981; Doreian, 1980, 1981; Ord, 1975) is perhaps the dominant approach; it has been recently touted as "a workhorse for modeling network influences on individual behavior" (Fujimoto et al., 2011, p. 231). 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. Because of these advantages, scholars continue to use and build upon the model. For instance, Dow (2007) extended the one-network autocorrelation model to multiple networks and applied it to understand the simultaneous multiple processes of cultural transmission. The standard one-mode network autocorrelation model has also been extended to a two-mode model (i.e., actor × event) by Fujimoto et al. (2011). More importantly, the primary estimation method for network autocorrelation models—maximum likelihood—that was originally elaborated by Doreian (1981) has now been integrated in modern statistical packages such as R (Butts, 2008), Matlab (LeSage, 1999), and Stata (Pisati, 2001). These developments have made the model much more accessible to network researchers.

Despite the obvious benefits of the model and its rising popularity, there has been growing evidence that the maximum likelihood algorithm used to estimate the parameters of autocorrelation models produces estimates of the network effect (ρ) that are negatively biased (Dow et al., 1982; Farber et al., 2009; Mizruchi and Neuman, 2008; Neuman and Mizruchi, 2010; Smith, 2009). This issue potentially leads to two serious problems for users of the network autocorrelation model. First, the model may fail to detect a network effect that truly exists, thus committing a Type II error (i.e., β error). Second, if the model does detect a network effect, the parameter of the network effect may be underestimated. Without further understanding the magnitude of these problems, users may begin to doubt the veracity of *all* network effect results, not just those subject to the conditions in which the bias has been detected.

In this paper, we take a different approach to studying the underestimation problem. Rather than look for more conditions in which ρ is underestimated, we investigate the likelihood of identifying a statistically significant network effect by the network autocorrelation model under various conditions. Specifically, given certain known network properties (e.g., size of network effect ρ, network density, network size, and network structure) what is the likelihood—that is, what is the *statistical power*—of identifying a network effect using the autocorrelation model? While investigating this question, we also attempt to answer a more practical question: *what network size (N) is required in order to obtain decent power (e.g., 80% power) to detect a network effect, given*

[*] Corresponding author at: Department of Psychology, University of Central Florida, 4000 Central Florida Blvd., Psychology Bldg 99 Ste. 320, Orlando, FL 32816, United States. Tel.: +1 407 823 4350; fax: 407-823-5862.
E-mail addresses: wei.wang@ucf.edu, weiglobe@gmail.com (W. Wang).

http://dx.doi.org/10.1016/j.socnet.2014.03.004
0378-8733/© 2014 Elsevier B.V. All rights reserved.

approximate network effect size, density, and structure? We believe such information provides useful guidelines to users of the network autocorrelation model.

By using simulations and manipulating network properties such as network effect ρ, network density, and network structure, we show that for a common network effect size of $\rho = .3$, a network size (N) of 40–80 nodes is sufficient to obtain statistical power of 80% or higher, depending on the network structure. In addition, we find that the Type I error (i.e., the probability of statistically supporting network effects that do not exist) remains acceptably small. We conduct further analyses to reveal that statistical power is primarily a function of network effect size (ρ) and network size (N), although both of these factors can interact with network density and network structure to impair power under certain rare conditions. We conclude by discussing the implications of these findings and offer guidelines for users of the autocorrelation model.

2. Network autocorrelation model and its applications in social science

The network autocorrelation model was initially proposed by geographers to remedy the dependence problem in the error terms of regression analysis for geographic proximity data (Cliff and Ord, 1981; Ord, 1975). Spatial dependence is quite common in geographic data. For example, the average real estate prices of two proximal areas are closer than those of two distant areas. If this spatial dependence is not acknowledged and accounted for in the ordinary least-squares (OLS) regression model (i.e., $Y = X\beta + \varepsilon$), then the model residuals of proximal areas are more similar than the residuals of distant areas. Such an error term ε thus violates a fundamental assumption for the conventional regression model: The error terms should be independent with zero mean and a constant variance and should follow a Gaussian distribution. To solve the assumption violation problem and to remove the spatial dependence of the disturbance, geographic researchers proposed two autocorrelation models (Cliff and Ord, 1981; Ord, 1975). The first model, termed the spatial disturbances model or the spatial error model (Anselin and Hudak, 1992), decomposes the problematic spatially dependent error term ε into $\varepsilon = \rho W\varepsilon + \upsilon$, where W is an $N \times N$ adjacency matrix of the spatial distances among the observations (e.g., W is a social network matrix), ρ is the parameter representing the correlation strength of spatial dependence in the residuals of ε, and υ is now the vector of Gaussian-distributed residuals. The second model, which is more straightforward than the first model and is the model we focus on in the current paper, models the spatial dependence directly on the dependent variable Y instead of on the model residuals. This second model was termed the spatial effect model (Doreian, 1980), the network effect model (Doreian et al., 1984), or the spatial lag model (Anselin and Hudak, 1992), and is $Y = \rho WY + X\beta + \varepsilon$, where W is the same $N \times N$ matrix of spatial distances among the observations as specified for the first model (e.g., W is the social network matrix). However the error term ε in this model follows a Gaussian distribution $N(0, \sigma^2 H)$ and the parameter ρ represents the strength of spatial dependence in the dependent variable Y. Because of its versatility, this model was soon adopted by social scientists (Doreian, 1990; White et al., 1981) who used it to model social influence. Now these models of spatial and network autocorrelation have been applied in many social sciences such as political science (Cho, 2003; Franzese and Hays, 2007; Franzese et al., 2012), sociology (Crowder and South, 2008; Loftin and Ward, 1983), cultural psychology and anthropology (Dow, 2007; Dow and Eff, 2008), and organizational studies (Ibarra and Andrews, 1993; Mizruchi et al., 2006).

3. The estimation challenge

Despite the many advantages of the network autocorrelation model, one serious problem has emerged. In numerous simulation studies as far back as the early 1980s, researchers have shown that maximum likelihood estimation of the network effect ρ can be negatively biased under several conditions. The earliest known evidence of an estimation bias was identified by Dow et al. (1982). In a study of the disturbances model using small networks ($N = 20$, 30, and 40), Dow and colleagues found that ρ was underestimated across a variety of target ρ's (.2, .4, .6, and .8) for a random W with density of .1, and also for a "language"-structured W with higher density. For the random networks, the magnitude of the bias increased as the target ρ increased, and network size had little effect on the bias. For the structured networks, the bias was less severe than that found with the random networks, decreased in magnitude as the target ρ increased, and was less severe for the largest networks.

Despite the authors' bold claim that "estimates of the significance of $\hat{\rho}$ are unreliable from ML [maximum likelihood] procedures" (Dow et al., 1982:198), the problem of a potential bias in properly-specified models was ignored for over two decades. Interest has been renewed in this area over the past five years as computing power makes it possible to carry out in-depth simulations over a host of different network conditions.

Perhaps the first such systematic investigation of bias was conducted by Mizruchi and Neuman (2008). Using random networks of sizes 40, 50, and 100 and across network densities from .05 to .95, they reported strong evidence of a negative bias in the estimation of ρ using the network effects model. Regardless of network size or whether W was row-standardized, they consistently found that the underestimation of ρ increased with increasing density of W and that the relationship between network density and negative bias in the estimate of ρ became stronger at higher levels of target ρ. Only when the noise in the residual term, ε, of the autocorrelation model was reduced to unrealistically low levels or when the number of exogenous variables (X's) in the model was increased to unrealistically high levels was much of the underestimation bias of ρ attenuated—though it was never entirely eliminated.

In follow-up work (Neuman and Mizruchi, 2010) the authors extended their previous study to examine whether the estimation bias held for non-random networks. Using larger networks than before ($N = \sim 400$), they ran simulations of star, caveman, and small-world networks along with random networks (all network densities $\leq .5$) at target ρ's of 0, .2, and .5. The pattern of findings was the same as before: a negative bias in the estimation of ρ with a magnitude that increased with increasing network density. Yet they also identified a negative bias in the estimation of ρ for low-density star networks. At a minimum, this underestimation of ρ for low-density star networks suggests that high density is not the sole source of the underestimation bias. More strongly, it might suggest that high density itself does not directly cause the bias but that high density networks create a condition that leads to the bias, and that this condition could also be caused by other network configurations (e.g., low-density star models).

Consistent with Mizruchi and Neuman's simulation findings for the effects model, Smith (2009) analytically showed that for maximally-connected networks (that is, W's with density = 1) maximum likelihood estimates of ρ exhibit a negative bias in both the spatial effects and spatial disturbances models. Smith then conducted simulations of both effects and disturbances autocorrelation models to examine cases where W is not maximally-connected. Using 50-node, randomly-connected networks with target $\rho = .5$ and with densities of .3, .5, .8, .9, .95, and .99, he replicated Mizruchi and Neuman's finding that ρ is seriously underestimated as the density of W increases. Smith's results further showed that for his