Computational Statistics and Data Analysis 82 (2015) 35–46

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Using retrospective sampling to estimate models of relationship status in large longitudinal social networks

A. James O'Malley [a,*], Sudeshna Paul [b]

[a] The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Lebanon, NH 03766, USA

[b] Nell Hodgson Woodruff School of Nursing, Emory University, Atlanta, GA 30322, USA

HIGHLIGHTS

- Estimation of statistical models for social networks is challenging.
- Dyads with no relationship ("null-dyads") are common in large social networks.
- Propose to subsample the "always-null" dyads.
- Develop weighted likelihood Bayesian estimation method.
- Method enables large social networks to be analyzed feasibly and accurately.

ARTICLE INFO

Article history:
Received 22 August 2012
Received in revised form 30 July 2014
Accepted 1 August 2014
Available online 13 August 2014

Keywords:
Conditional independence
Longitudinal
Retrospective sampling
Social network
Sociocentric design
Sparse data
Weighting

ABSTRACT

Estimation of longitudinal models of relationship status between all pairs of individuals (dyads) in social networks is challenging due to the complex inter-dependencies among observations and lengthy computation times. To reduce the computational burden of model estimation, a method is developed that subsamples the "always-null" dyads in which no relationships develop throughout the period of observation. The informative sampling process is accounted for by weighting the likelihood contributions of the observations by the inverses of the sampling probabilities. This weighted-likelihood estimation method is implemented using Bayesian computation and evaluated in terms of its bias, efficiency, and speed of computation under various settings. Comparisons are also made to a full information likelihood-based procedure that is only feasible to compute when limited follow-up observations are available. Calculations are performed on two real social networks of very different sizes. The easily computed weighted-likelihood procedure closely approximates the corresponding estimates for the full network, even when using low sub-sampling fractions. The fast computation times make the weighted-likelihood approach practical and able to be applied to networks of any size.

1. Introduction

In this paper we develop, apply, and evaluate a new method of estimating a dynamic model of the relationship status of all *dyads* (pairs of individuals) in a social network, where both the number of individuals (N) and the number of observation times (T) can be large. Analyses of complete lattices of dyadic data (referred to as *sociocentric network data*) in general

[*] Corresponding author. Tel.: +1 603 653 0854; fax: +1 603 653 0896.
E-mail addresses: Alistair.J.O'Malley@Dartmouth.edu, James.OMalley@Dartmouth.edu (A.J. O'Malley), sudeshna.paul@emory.edu (S. Paul).

http://dx.doi.org/10.1016/j.csda.2014.08.001

seek to identify the important determinants of dyadic relationships and gain insights into properties or determinants of the network. For example, one phenomena that is often thought responsible for the formation of relationships is *homophily* – commonly described as "birds of a feather flock together" – whereby individuals with similar attributes are more likely to form or maintain relationships, leading to clusters of individuals with similar traits within the network. However, the primary objective of this paper is demonstrating that the new estimation method is feasible to implement on networks of any N and T, overcoming the practical limitations of existing methods. The following two publicly-available social network datasets, judiciously chosen due to the difference in N and T between them, will be analyzed and used to appraise our method of computing estimates.

The smaller network is from the excerpt of 50 schoolgirls in the Teenage Friends and Lifestyle Study (TFLS) described in Snijders (2014). Students in the study named up to 12 close friends at three surveys conducted during 1995–1997 (Michell and Amos, 1997; West and Sweeting, 1995). After dropping the two girls who did not nominate and were not nominated by anyone, the final network comprised $N = 48$ girls (1128 dyads) observed on $T = 3$ occasions (two relationship change opportunities). The number of friends named by each schoolgirl (*out-degree*) could range from 0 to 12 while the number of times a girl could be named by others as their friend (*in-degree*) had a range from 0 to 47. The students were also asked about substance use and adolescent behavior associated with lifestyle, sporting behavior and tobacco, alcohol and cannabis consumption. A particular question of importance is whether homophily of smoking behavior exists; were girls who were both smokers or both non-smokers more likely to become friends.

The second and larger longitudinal friendship network is from the offspring cohort of the Framingham Heart Study (FHS). Since the offspring cohort's inception in 1971, its members have been followed from 1971–2008 through eight periodic health exams, at which an extensive array of personal and medical information (e.g., height, weight, age, smoking status) was collected. Friendship ties at each exam were ingeniously obtained from the nomination of close-friends who might be in a position to know where the study member would be in two to four years (Christakis and Fowler, 2007, 2008). Subjects were not restricted from naming multiple friends but on most occasions only named a single friend, resulting in a sparsely-connected network. Out-degrees were typically 0 or 1 while in-degrees were more widespread with values ≥ 2 relatively common. Emulating Paul and O'Malley (2013), all FHS offspring members who named or were named by another offspring cohort member over any two consecutive exams were included in the analysis, yielding $N = 831$ individuals observed at up to $T = 8$ exams (7 relationship change opportunities). A plethora of personal characteristics (gender, age, BMI, smoking status, various medical quantities) are available although herein we focus on age. More details of both the FHS and TFLS networks appear in Paul and O'Malley (2013).

In these networks relationship status (close friendships between schoolgirls or between study members) is presumed known for all $N(N-1)/2$ dyads, yielding complete sociocentric data. Close friendship is represented as a binary random variable ($1 = \text{yes}$, $0 = \text{no}$) with the presence thereof referred to as a *tie*. Because there is no constraint that a tie from one individual to another implies that a tie exists in the reverse direction the networks are *directional*.

To identify the presence of homophily or some other relationship feature (e.g., reciprocity) in the network, other possible explanations for the formation and dissolution of ties need to be statistically adjusted for or controlled. Finding the important determinants of a network is aided by longitudinal data. However, such data has historically been elusive. Not surprisingly, methods for longitudinal analysis of sociocentric data are scarce and those that do exist are confronted by computational challenges. For example, we previously developed a novel model for a longitudinally-observed sociocentric network that allowed homophily effects and other network phenomena to be estimated. Although the methodology was sound, implementation was restricted to small- to mid-sized networks by CPU and time constraints (Paul and O'Malley, 2013). One of the reasons for the challenging computations is that the number of dyads in a sociocentric dataset has order N^2 as opposed to the order N number of observations in individual level analyses. Because large networks with $N \geq 1000$ are becoming commonplace, the development of methods of estimating models of networks for any N and T is timely.

The method proposed herein adapts ideas from survey sampling methodology to accurately approximate estimates of the full network in minimal computational time. The genesis of the method is the observation that as N increases the number of dyads that remain null (no ties) over time increases. Therefore, as long as the sampling design is accounted for in the analysis, in large networks only a small fraction of the always-null dyads may be needed to accurately approximate the estimates computed on the full network. To account for the dependences introduced by sampling, we develop a novel *weighted likelihood* (WL) estimation procedure that weights the observations for each dyad by the inverse of the probability of sampling that dyad. The proposal to subsample null-dyads is not without precedent (Raftery et al., 2012; Kleinbaum, 2012). However, to our knowledge we are the first to consider subsampling in the context of longitudinal sociocentric networks.

In Section 2 we define notation and specify models for longitudinal analysis of sociocentric data. In Section 3 we describe our proposed sampling design and develop associated WL estimation and implementation procedures. To evaluate the efficacy of the WL estimation procedure, we compare it to a full information *observed data likelihood* (ODL) procedure on the smaller TFLS network data for which estimation of the ODL procedure is feasible and discuss the limitations of ODL methods on larger or more intensely observed networks. The estimation methods are applied to the two longitudinal sociocentric network datasets described above in Section 4 with comparisons between the methods and other results reported in Section 5. Section 6 reviews the primary findings and discusses limitations.