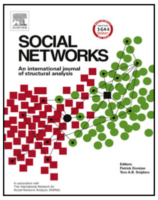
Social Networks 45 (2016) 89–98

Contents lists available at ScienceDirect

Social Networks

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



Multiple imputation for missing edge data: A predictive evaluation method with application to Add Health

Cheng Wang[a,*], Carter T. Butts[b], John R. Hipp[c], Rupa Jose[d], Cynthia M. Lakon[e]

[a] Department of Sociology, University of Notre Dame, United States
[b] Departments of Sociology and Statistics, University of California, Irvine, United States
[c] Departments of Criminology, Law and Society, and Sociology, University of California, Irvine, United States
[d] Department of Psychology and Social Behavior, University of California, Irvine, United States
[e] Program in Public Health, University of California, Irvine, United States

ARTICLE INFO

Article history:
Available online 7 January 2016

Keywords:
Missing edge data
ERGM-based imputation
Held-Out Predictive Evaluation (HOPE)

ABSTRACT

Recent developments have made model-based imputation of network data feasible in principle, but the extant literature provides few practical examples of its use. In this paper, we consider 14 schools from the widely used In-School Survey of Add Health (Harris et al., 2009), applying an ERGM-based estimation and simulation approach to impute the network missing data for each school. Add Health's complex study design leads to multiple types of missingness, and we introduce practical techniques for handing each. We also develop a cross-validation based method – Held-Out Predictive Evaluation (HOPE) – for assessing this approach. Our results suggest that ERGM-based imputation of edge variables is a viable approach to the analysis of complex studies such as Add Health, provided that care is used in understanding and accounting for the study design.

1. Introduction

Missing edge variable data – i.e. edge variables in an observed network whose states are unknown – has long been recognized to be a serious problem for social network analysis (Burt, 1987). Network analytic concepts and measures are generally defined with respect to a completely observed graph (Wasserman and Faust, 1994) and the non-extensive nature of many network properties makes them difficult or impossible to estimate by e.g. simply averaging observed local network information. While ad-hoc methods such as treating missing edges as absent, dropping vertices with missing edge information, etc. have been employed, these can produce misleading or incorrect estimates (see Ghani et al., 1998; Huisman and Snijders, 2003; Kossinets, 2006; Huisman and Steglich, 2008; Huisman, 2009; Almquist, 2012); methods for handling missingness from one source by integrating measurements from other sources (e.g. Butts, 2003) can work well, but require data unavailable to most network researchers. Unfortunately, missingness is sometimes impossible to avoid, or arises from flaws in study design that are unrecognized until after data collection. Given the importance and scope of this problem, finding practical and principled ways to deal with it has been an important priority in network research.

A significant development in this regard has been the emergence of techniques for fitting exponential family random graph models (ERGMs) in the presence of missing data. The core insight (introduced by Handcock in 2002) is that the latent missing data framework developed by Rubin (1976) in a non-network context can also be applied to edge variables: given a parametric model, and appropriate assumptions regarding the nature of missingness, one can derive the likelihood of the observed data as a marginalization of the complete-data likelihood over the possible states of the missing variables (in some cases weighted by a factor related to the probability of the observed pattern of missingness). Techniques for performing maximum likelihood estimation (MLE) under these conditions (and theory regarding the nature of the assumptions required) have been developed by Robins et al. (2004) and Handcock and Gile (2010), with recent Bayesian extensions by Koskinen et al. (2010, 2013).

The current state of the art may be briefly summarized as follows. First, it is usually assumed that the pattern of missingness is ignorable (i.e., that any unknown parameters governing the observation process are distinct from those being estimated, and the probability of the pattern of missingness depends only on the values of the observed data and/or covariates). Ignorability can in some

[*] Corresponding author at: Department of Sociology, University of Notre Dame, 810 Flanner Hall, Notre Dame, IN 46556, United States. Tel.: +1 6073196561.
E-mail address: cwang3@nd.edu (C. Wang).

http://dx.doi.org/10.1016/j.socnet.2015.12.003

cases be relaxed (albeit not without altering the likelihood calculation), but is satisfied exactly or approximately for many real-world designs [see, e.g., Handcock and Gile (2010) for a discussion]. Second, a model is posited for the graph as a whole (here, a parametric model in ERGM form). Finally, the likelihood for a given parameter vector is then calculated by marginalizing the ERGM likelihood for the full network over all possible complete networks that are compatible with the observed data. This observed data likelihood is then employed for purposes of inference.

Although the emphasis of these techniques is on ERGM inference, it is clear that they also provide an approach to the more general problem of network imputation: given an adjacency matrix Y with realization y of which portion y^{obs} is observed and y^{mis} is missing, y^{mis} can be modeled via conditional prediction from an ERGM fit to y^{obs}. Specifically, let θ' be an estimate (e.g., an MLE) for the parameter vector of an ERGM with sufficient statistic t given data y^{obs}. Then we generate draws $Y^{mis} \sim \mathrm{ERGM}(\theta\text{'}|y^{obs})$, where $\mathrm{ERGM}(\theta\text{'}|z)$ denotes the ERGM distribution with statistic t and parameter θ conditional on z (i.e., with the elements of Y contained in z held fixed). Such draws may be taken using standard Markov chain Monte Carlo (MCMC) methods (see Snijders, 2002; Snijders et al., 2006; Wasserman and Robins, 2005), and indeed simulations of this sort are used as part of the latent missing data estimation process described above. Draws from Y^{mis} can then be used to estimate various features of y^{mis} (the true missing data) or $y = y^{obs} \cup y^{mis}$ (the true state of the graph).

While the basic logic of ERGM-based network imputation is straightforward, there are to date few published use cases (to our knowledge, e.g., Handcock and Gile, 2010; Koskinen et al., 2010, 2013). Likewise, the existing literature gives little guidance on assessing the quality of network imputation (an important practical consideration in everyday use). In this paper, we attempt to rectify this latter deficit by introducing a simple cross-validation based method – what we term Held-Out Predictive Evaluation (HOPE) – to assess the accuracy of imputed draws from an observed-data ERGM.

We apply the ERGM-based network imputation method to model the missingness and error inherent in the Add Health data set (Harris et al., 2009). This provides for a useful demonstration given that this is a widely used study in the literature, and that it has a high level of missingness making it a very complex and challenging case. The Add Health case is also useful for demonstrating the use of multiple sources of information (particularly, marginal constraints on degree and group-specific mixing) in aiding estimation (something not explored in most published work to date). For our study, we use the friendship networks from 14 schools in the saturated sample of Add Health. As we are using a real-world data set (rather than simulated data), our focus is on technique illustration rather than method evaluation per se; however, as we will demonstrate, one feature of our approach is that it provides some basis for evaluation on available data. As we show, ERGM-based imputation can produce reasonable results in a real-world setting (although careful attention must be paid to the complexities of one's study design).

As a complement to the above-mentioned methods of imputation, we introduce a simple strategy we call Held-Out Predictive Evaluation (HOPE) for evaluating the quality of imputation in real-world settings. As discussed below, HOPE involves holding out a stratified sample of edge variables from the graph prior to model estimation and imputation, and using the predictive accuracy of the model on the imputed data as an indicator of imputation quality. It is worth emphasizing that the HOPE method sets a relatively high bar for accuracy, compared to common methods of assessing the latent missing data imputation framework developed by Rubin (1976) outside of the network context. In that context, the typical approach for assessing the quality of imputation is to assess how well the *estimated parameters* for the imputed data compare to the true parameters. Thus, the question posed is whether the imputed data are able to accurately capture the proper coefficients for a specific model. By contrast, HOPE directly assesses the ability of an imputation model to *correctly identify present and absent* edges in the (unknown) true network. This tough but general standard is useful when imputation is being performed without knowledge of what analyses will need to be subsequently conducted on the resulting graph (e.g., when the imputed draws will be shared with other researchers, or at the early stages of a multi-stage investigation), and/or when the same imputed draws will be used for several different purposes (rendering any single model-based evaluation problematic). HOPE may also be useful as an easily interpretable adjunct to other quality measures, and can serve as the basis for a wider range of predictive evaluation measures employing specific graph properties.[1]

2. Data source, multiple types of network data missingness, and treatments

2.1. Data source

Our data comes from the first wave of the National Longitudinal Study of Adolescent to Adult Health (Add Health), a longitudinal study of a stratified sample of US schools from 7th to 12th grades (see Harris et al., 2009). (A "school" in this case consists of a high school, in some cases united with a "feeder" school whose students ultimately attend it. We use the singular "school" to refer to such high school/feeder school pairs.) All participants were invited to take the In-School Survey ($n = 90{,}118$) during 1994 and 1995. A random sample of 20,745 students selected from the In-School Survey respondents completed a wave 1 In-Home Survey, which was administered between April and December, 1995. Approximately one year later, participants who had not yet graduated from high school were asked to take a Wave 2 In-Home Survey ($n = 14{,}738$) between April and December, 1996. Information on social and demographic characteristics (i.e., gender and grade) of the respondents, attending classes and grades, extracurricular activities (i.e., club and sport-team participation), education and occupation of parents, household structure, risk behaviors including tobacco and alcohol use, expectations for the future, self-esteem, and health status were collected. Each student was also asked to nominate up to five best female friends and five best male friends.[2] In this paper we focus on the saturated sample of 4431 students collected from 14 out of 132 participating schools.[3] As shown in Table 1, the roster size of our 14 schools range from 30 to 2104.

[1] For example, a variant of the HOPE technique could be used to assess the ability to reproduce structurally selected subsets of edge variables (e.g., those known to be embedded in two-paths), rather than randomly selected edge variables.

[2] The friendship network dataset from Add Health has considerable complexity. Respondents (egos) were asked to nominate friends (alters) by entering numbers from a roster listing students at the school (and, in some cases, a feeder school with which it was paired). Because of enrollment changes, some students were not listed on the roster; these "off-rosteršstudents could participate (and hence their outgoing ties are observed) but could not be uniquely identified as alters by other participants. "Off-rosterälters are identified in the data by a generic code, and hence only the total number of ties to such persons (by gender) is observable. Further, the nominees were not limited to participants in the sample: respondents could also nominate persons outside the school. Ties to those outside the school are likewise identified by a generic code, and only the number of such alters (by gender) for each observed ego is known. (Since the survey was administered only to students within the sampled schools, incoming nominations from those outside the school are unobserved.)

[3] Add Health contains a saturated sample of 16 schools (Harris et al., 2009). Among the 16 schools, there is a special education school with constant student turnover, and another school suffering from an administrative error in which the students' IDs at the earlier wave could not be matched with those at later waves. Thus these two schools are not included in this paper.