



Treatment of non-response in longitudinal network studies

Mark Huisman^{a,*}, Christian Steglich^{b,1}

^a Department of Psychology, University of Groningen, Grote Kruisstraat 2/1, 9724 TS Groningen, The Netherlands

^b Department of Sociology, University of Groningen, Grote Rozenstraat 31, 9712 TG Groningen, The Netherlands

ARTICLE INFO

Keywords:

Missing data
Wave non-response
Imputation
Network evolution

ABSTRACT

The collection of longitudinal data on complete social networks often faces the problem of actor non-response. The resulting incomplete data pose a challenge to statistical analysis, as there typically is no natural way to treat the missing cases. This paper examines the problems caused by actors missing as nominators, but still occurring as nominees, in complete, directed networks measured in a panel design. In the framework of stochastic actor-driven models for network change ("SIENA models"), different methods to cope with such incomplete data are investigated. Data on a friendship network among female high school students are used to illustrate the procedures. Missing data problems related to early panel exit and late panel entry are not addressed.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Data analysis in social sciences is often hampered by non-response. In the analysis of social networks, non-response results in missing network information. This means that ties from one actor to another are not observed and/or information on actor attributes is not available. According to Burt "missing data are doubly a curse to survey network analysis" (Burt, 1987, p. 63), compared to other types of analyses (see also Borgatti and Molina, 2003). First, the complexity of items in network surveys is more likely to generate missing data (e.g., Marsden, 2005), and second, network analysis is especially sensitive to missing data because of the dependence structure of the data. If a network tie, or worse, an actor is missing, there is limited capacity to describe the network context of the actors whose ties are missing and there is lack of information on the context of neighboring actors (Robins et al., 2004).

The effects of non-response and missing data on the structural properties of networks are investigated in several studies (Burt, 1987; Costenbader and Valente, 2003; Kossinets, 2006). The general conclusion is that missing data have a negative effect on network mapping (Borgatti and Molina, 2003) and estimating structural network properties: the strength of relationships is underestimated (Burt, 1987), centrality measures become unstable as well as degree measures (Costenbader and Valente, 2003; Kossinets, 2006), and clustering coefficients are underestimated (Kossinets, 2006). Still,

Costenbader and Valente (2003) find that measures based on indegrees are reasonably robust for small proportions of missing data when the observed incoming ties of non-respondents are used in the analysis. This latter result shows a unique property of social networks: non-participation by respondents does not necessarily mean that they are not included in the study (Borgatti and Molina, 2003). Respondents report ties to non-respondents, that is, the incoming ties of non-respondents are available.

Missing data treatment methods can use the information on non-respondents from the nominations of observed actors. Stork and Richards (1992) propose using these partially described ties between respondents and non-respondents to reconstruct the missing outgoing ties: substitute the missing ties by the value of the tie in the opposite direction. This imputation method is appropriate if ties tend to match across actors, for instance, in undirected networks. For directed networks, this (ad hoc) imputation method seems less suitable. Another imputation method is suggested by Burt (1987), who finds that missing relations are strongly associated with weak relations and therefore can be replaced with values indicating such weak relations.

More recent missing data methods are proposed by Robins et al. (2004), Gile and Handcock (2006), Handcock and Gile (2007), and Koskinen (2007). These methods are also based on all available data, including the incoming ties of non-respondents. The proposed methods are model-based treatment methods within the framework of exponential random graph models (ERGMs). Robins et al. (2004) model the ties from respondents to non-respondents separately from the fully described ties, which allows exploring the structural effects for the entire network. The model is especially helpful when the non-respondents systematically differ from the respondents with respect to ties. Gile and Handcock (2006),

* Corresponding author. Tel.: +31 50 3636345; fax: +31 50 3636304.

E-mail address: j.m.e.huisman@rug.nl (M. Huisman).

¹ The author was funded by the Netherlands Organization for Scientific Research (NWO) under grant 401-01-550.

Handcock and Gile (2007), and Koskinen (2007) use Markov chain Monte Carlo methods to fit ERGMs to incomplete network data. This is a more traditional missing data approach based on the marginal distribution of the observed data (e.g., see Schafer and Graham, 2002), allowing for proper inferences for network properties for both respondents and non-respondents. As the methods repeatedly sample from the conditional distribution of the missing data, they can also be used to impute the data sets.

All these methods are designed for modeling *single*, incomplete observations of a network. Moreover, possible treatments are either simple ad hoc procedures (the imputation methods of Stork and Richards, 1992, and Burt, 1987), or are embedded within ERGMs (Robins et al., 2004; Gile and Handcock, 2006). For the case of longitudinal network data, studies on the effect of non-response or the effect of treatment procedures are lacking. In this paper we examine the effects of non-response and missing data techniques on longitudinal network data. The effects of missing data treatments are investigated within the framework of the actor-driven models for network evolution proposed by Snijders (2001, 2005), using simulations under a known evolution model. The missing data treatments that are used in the simulation study are the analysis of complete cases, two ad hoc imputation methods based on reconstruction (Stork and Richards, 1992) and preferential attachment (Barabasi and Albert, 1999), respectively, and a hybrid imputation procedure based on simulating networks with the actor-driven network evolution models (Snijders, 2005).

The paper is organized as follows. Section 2 addresses the problem of non-response in longitudinal network data, defining the missing data patterns that are considered in this study. In Section 3 the family of actor-driven models for network evolution of Snijders (2001, 2005) is briefly described. Section 4 presents the missing data treatments, of which the performance (i.e., the effects of the treatments on modeling the data with actor-driven models) is investigated in a simulation study. The design of this study is presented in Section 5 and in Section 6 the results in terms of convergence of the estimation procedure and the absolute and relative bias in the parameter estimates. The paper ends with a discussion of the results and some general recommendations.

2. Non-response in longitudinal network studies

In missing data research usually two types of non-response are distinguished: unit non-response, where complete cases are missing, and item non-response, where the unit participated but data on particular items are missing. For social network data, *unit non-response* means that an actor does not participate in the study and therefore all his or her outgoing ties are unavailable for analysis. *Item non-response* means that only particular (outgoing) ties are unavailable in the analyses.

In the case of longitudinal data where respondents are repeatedly contacted at successive time points, non-response patterns can be further distinguished by including partial non-response (De Leeuw et al., 2003), or wave non-response (Lepkowski, 1989). *Wave non-response* is characterized by time dependency and means that only at certain time points data are available. This is often due to panel mortality or attrition, which results in completely missing cases after a certain time point. For social network data, we define wave non-response as complete non-response at one or more time points, which results in completely missing outgoing ties of some actors for these time points.

The present study restricts analyses to unit and wave non-response, and considers the case of longitudinal network data with two observations moments and completely missing actors at one or both time points. This means we assume that at a certain observation moment non-response results in completely missing outgoing

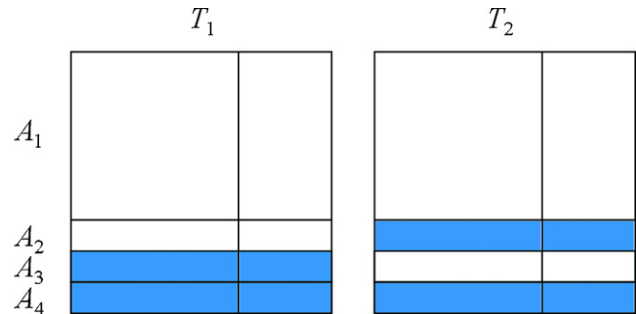


Fig. 1. Partial non-response in a network observed at two time points (T_1 , T_2), identifying four subsets of actors (A_1 – A_4). Gray areas indicate missing outgoing ties.

ties of some actors. The incoming ties of these non-responding actors from responding actors are observed. Moreover, occasionally missing ties (item non-response) are not considered. This results in four kinds of missingness patterns, which are presented in Fig. 1. For more than two observations of the network the procedures are generally the same. The main difference is in the number of possible missing data patterns.

Fig. 1 shows the sociomatrixes of a network at two observations moments. As the rows and columns of the matrices represent the actors, the subsets of actors distinguish mutually exclusive subsets of rows and columns. The observed missingness patterns consist of four set of actors: A_1 , actors observed at both time points, A_2 , actors observed at T_1 and missing at T_2 (wave non-response), A_3 , actors missing at T_1 and observed at T_2 (wave non-response), and A_4 , actors missing at both time points (unit non-response). At observation moment T_1 , the set of observed actors is $A_1 \cup A_2$ and the data set consists of the corresponding rows in the sociomatrix. These are the white areas in first adjacency matrix in Fig. 1, consisting of incoming and outgoing ties of observed actors in $A_1 \cup A_2$, and incoming ties of non-respondents in $A_3 \cup A_4$. At observation time T_2 the set of observed actors is $A_1 \cup A_3$.

Although the non-response patterns play an important role in treating the missingness, the most important question is whether the non-response causes systematic bias in the analysis due to systematic differences between respondents and non-respondents. This leads to the difficult task of assessing whether the data are missing at random and, as a result, the missingness mechanism can be ignored (Rubin, 1976). Data are called *missing at random* (MAR) if the missingness is unrelated to the (unknown) value of the missing item itself. For network analysis this means that missingness is unrelated to the value of the tie. In this case the missingness may be related to completely observed actor characteristics, but not to network properties.

If the missingness is related to the value of the missing tie itself, the data are *missing not at random* (MNAR). In this case, network properties are biased because they are computed from the network in which tie values are systematically missing. The extent in which the structural properties are affected depends on the property itself. Measures based on indegrees, for instance, are found to be less affected than other measures, because incoming ties are only partially missing for all respondents.

3. Actor-driven models for network evolution

The prominent tool for modeling and analyzing longitudinal, complete network data is the family of stochastic, actor-driven models introduced by Snijders (1996, 2001, 2005). Estimation of these models is implemented in the SIENA software (shorthand for S imulation I nvestigation for E mpirical N etwork Analysis; Snijders

Download English Version:

<https://daneshyari.com/en/article/1129549>

Download Persian Version:

<https://daneshyari.com/article/1129549>

[Daneshyari.com](https://daneshyari.com)