

Missing data in networks:

Exponential random graph (p^*) models for networks with non-respondents ¹

Garry Robins,

Philippa Pattison,

Jodie Woolcock.

Department of Psychology,

University of Melbourne.

15 April 2004

¹ This work has been supported by the Australian Research Council. We are grateful for comments from John Skvoretz and Filip Agneessens, and from two anonymous reviewers.

Abstract

Survey studies of complete social networks often involve non-respondents, whereby certain people within the “boundary” of a network do not complete a sociometric questionnaire – either by their own choice or by the design of the study – yet are still nominated by other respondents as network partners. We develop exponential random graph (p^*) models for network data with non-respondents. We model respondents and non-respondents as two different types of nodes, distinguishing ties between respondents from ties that link respondents to non-respondents. Moreover, if we assume that the non-respondents are missing at random, we invoke homogeneity across certain network configurations to infer effects as applicable to the entire set of network actors. Using an example from a well-known network dataset, we show that treating a sizeable proportion of nodes as non-respondents may still result in estimates, and inferences about structural effects, consistent with those for the entire network.

If, on the other hand, the principal research focus is on the respondent-only structure, with non-respondents clearly not missing at random, we incorporate the information about ties to non-respondents as exogenous. We illustrate this model with an example of a network within and between organizational departments. Because in this second class of models the numbers of non-respondents may be large, values of parameter estimates may not be directly comparable to those for models that exclude non-respondents. In the context of discussing recent technical developments in exponential random graph models, we present a heuristic method based on pseudo-likelihood estimation to infer whether certain structural effects may contribute substantially to the predictive capacity of a model, thereby enabling comparisons of important effects between models with differently sized node sets.

Introduction

A commonly occurring problem in survey-based studies of complete networks is that of non-responding network members. Although methods for handling non-responses in general survey contexts have attracted considerable interest (e.g. Little & Schenker, 1995), over the last two decades there has been relatively little discussion of the problem of non-response in survey network studies. Some recent work suggests that this important issue may now be receiving more sustained attention (for example, Butts, 2003; Kossinets, 2003) but methods for effectively dealing with non-response continue to require further development.

Non-respondents create significant and potentially insidious problems for network analysis. In particular, many network studies are based on the premise that in order to understand some social phenomenon of interest, it is necessary to understand the arrangement of network ties into larger network structures and sub-structures. If this premise is correct and a network tie is missing, then we not only have a limited capacity to describe the network context of those individuals whose ties are missing, but we may also lack significant information on the network context of many other neighbouring actors as well.

For survey studies of complete networks, researchers have to decide on some putative “boundary” to the network in advance of the survey (Laumann, Marsden & Prensky, 1989). In other words, they must make a decision about the individuals or entities considered to constitute the set of actors in the network. Often this “boundary” can be inferred from the research question. For instance, for research on networks in work teams, a natural boundary might be provided by a workgroup of individuals. Of course, there may be no real “boundary” to a network (e.g., White, 1992), but the

practicalities of conducting an empirical research investigation often require that some such decision be taken, at least implicitly.

Having decided on a boundary, researchers may consider several methods to elicit sociometric responses. The definition of the boundary may be utilized directly: for instance, participants may be asked to list those in their workgroup whom they trust. Alternatively, the researcher may obtain a list of individuals “within” the boundary. Each person on the list is asked to select from the list those who are his or her network partners. There are several possible sources of missing data in this survey design: for instance, a list of names provided by the researcher may be incomplete²; secondly, not all individuals “within” the boundary may respond to the sociometric questionnaire.³

An incomplete specification of individuals may often be avoided by a well-justified specification of the boundary and by obtaining sufficient preliminary information about a network of interest. Of course, not every research situation is quite so simple, and given that in all but some extreme examples there are no impermeable boundaries to networks, the potential for ambiguity may be considerable. The problem of boundary specification in network studies has been widely recognised for some time, and the ramifications of various approaches have been well-canvassed (Laumann, et al, 1989). Nonetheless, network analysis has made little progress on dealing with a mis-specified boundary beyond a clear account of its potential problems. Fortunately, this

² When researchers use the definition of the boundary directly to elicit sociometric responses (e.g. “List those of your work colleagues whom you trust”), we do not consider the “forgetting” of names of those within the boundary as missing data. Rather, the data is potentially confounded, with a non-listing signifying, in this case, either “not trusted” or “trusted but forgotten”. Researchers, of course, may choose to assume that it is the “trusted and remembered” category that is most important to an individual’s behavior, so the confounding may not be particularly invidious. For recent work on forgetting in network data collection, see, for instance, Brewer and Webster (1999). These issues relate to *informant accuracy* about underlying network structures, brought into focus by the Cognitive Social Structure approach of Krackhardt (1987). Recent developments open new possibilities in dealing with cognitive social structures (e.g Butts, 2003; Koehly & Pattison, in press).

³ Kossinets (2003) also discusses fixed choice designs, where participants are restricted in the number of network nominations they can make. This is not an issue we consider here.

source of missing data is to some degree within the control of researchers, through careful decisions about the set of individuals relevant to a particular research question, and with some thought to the type of generalizations that might be drawn from the results.

The second source of missing data is more problematic and is the focus of this article. As is the way of data collection in the social sciences, 100% response rates to sociometric questionnaires are rare. Several actors may be included as network partners in other participants' questionnaires, but they themselves may choose not to complete the questionnaire. It is not an unusual situation, then, for network data to include two sets of participants: *respondents*, who complete the sociometric survey and thereby fully participate in the study; and *non-respondents*, who do not complete the survey but who are included in the network in the sense that respondents have identified them as network partners.⁴

A slightly different version of this situation arises with snowball sampling approaches to complete network studies. Here, boundaries to the network may not be clearly defined but key individuals are asked to specify network partners, with some but not necessarily all such partners then in turn asked to specify their own network partners. At some point the snowball stops rolling, with a number of non-respondents a likely outcome (except in rare cases where the network is small and really is constituted from a self-contained set of individuals.) Another important variation occurs if the study involves, for instance, individuals within one department of an organization, each of whom is a respondent, but who may also have formal or informal connections to individuals from other departments not included in the survey. (In such a study, it is

⁴ There are current discussions about the ethics of including non-respondent individuals in network data (e.g. Klov Dahl, 2002). Whatever the outcome of these considerations, presumably these ethical matters do

quite possible that the number of non-respondents may exceed the number of respondents.) In this case, the research focus may be on the original department, and the modeling endeavor may be directed towards understanding the structural patterns among respondents only. But in doing so, it seems sensible to include the additional information about links to other departments, even if that information is not itself explicitly modeled. In a sense, there are two nested “boundaries” operating in such studies: one involving respondents only, the central focus of the research, and the second including (a possibly large number of) non-respondents as well.

It is worth briefly noting that these methods of data collection, and the fact that there may be both respondents and non-respondents, imply that the network is measured as *directed*, even if for conceptual or other reasons researchers may choose to infer non-directed ties from the directed observations. For instance, researchers may be interested in non-directed networks (e.g. mutual friendship ties) and transform the data under some rule (e.g. the observation of reciprocated responses), a technique that may be problematic when there are non-respondents and the observations are inherently directed.

Methods for dealing with non-responses

In general, the network literature provides little guidance on how to proceed when there are non-respondents. It is notable, for instance, that a comprehensive text such as Wasserman and Faust (1994) gives little advice on handling missing data in networks. Various researchers have commented on the difficulty of accommodating non-responses in network studies (e.g., Burt, 1987; Rogers & Kincaid, 1981). One pragmatic approach to the problem of missing network data has been to restrict attention

not arise when the nodes represent entities other than individuals (e.g. corporations) where interlocking information is not complete.

to the subset of individuals for whom network information is complete. This approach to missing data effectively leads to a re-specification of the network “boundary”. As discussed below, this convenient “solution” to the problem will be considerably less than optimal in many circumstances.

Stork and Richards (1992) are among the few to discuss missing data issues in some detail and to provide suggestions about how to analyze network data with non-respondents, as well as how to improve response rates. In preference to the blunt tactic of simply removing non-respondents, Stork and Richards (1992) propose a process that they term *reconstruction*. In reconstructing a network, researchers assume that if a respondent nominates a non-respondent, then the tie between the two exists, so that the respondent’s description of the relationship is accorded to the non-respondent as well. Stork and Richards advise that the validity of this approach should be checked against the data, so that if the ties are to be conceptualized as mutual, and hence non-directed, levels of reciprocity in actual responses should be examined. In directed network studies, questions reflecting both directions of the relationship should be asked (for instance, people should nominate both those to whom they give advice and those from whom they receive advice.) If, among respondents, descriptions of ties tend to match across network partners, then this dyadic reconstruction process may be applied to ties with non-respondents. Stork and Richards note that some ties – specifically, those from one non-respondent to another – cannot be reconstructed by this process, and so remain as missing. If the missing ties are at random and in small numbers, then the development of some type of imputation approach might be considered, but if all of the ties for a potential network member are missing, imputation is unlikely to be very successful.

Using a simulation study, Kossinets (2003) has examined the impact of various types of missing data on the structural properties of social networks. This case study, based on a bipartite graph rather than a unimodal network, suggests that boundary specification inadequacies and fixed choice designs present major problems. In regard to non-response, Kossinets endorses the dyadic reconstruction approach of Stork and Richards (1992), provided that the number of non-respondents is not large. Dyadic reconstruction makes good sense when its assumptions are justified. For directed network studies, survey questions to examine both directions of the relationship need to be incorporated into the design of the study before the data is collected. But there is no guarantee that relationship descriptions will tend to match across network partners. Moreover, as we show in our second example below, there can be strong reciprocity effects in the data, yet the majority of ties may still not be reciprocated. It is not clear what to do in regard to reconstruction in this case.

Even when reconstruction is not appropriate, it may still be useful to retain non-respondents in the data set, but only to analyze those network constructs that can be defined in terms of incoming ties. An adjacency matrix that includes missing data may be constructed to have a row of zeroes for each non-respondent. A non-respondent column, on the other hand, will not comprise zeroes if the non-respondent is nominated at least once. It would be possible to create, for instance, a partition of approximately structurally-equivalent blocks by clustering solely on the columns, as long as the structural equivalence is understood in terms of incoming, and not outgoing, ties. But of course there are many research questions for which an analysis based solely on incoming ties is inadequate.

Missing data are problematic in any context and it is a truism to state that none of the strategies mentioned above is universally successful. Indeed, judgments about

the appropriateness of any strategy will almost certainly depend both on the researchers' beliefs about the underlying processes by which the network data are generated and on the kind of network characteristics that the researchers intend to measure. More broadly, Butts (2003) has written an interesting review of issues relating to network ontology: is a network principally a cognitivist construction, or is it underpinned by a "real" structure of interactions potentially observable by third parties? Butts (2003) notes that different perspectives may result in different approaches to the analysis of networks. These differing perspectives may themselves inform preferred treatment of issues relating to informant inaccuracy and missing data.

In any event, as Kossinets (2003) notes, the proportion of non-respondents is clearly important. If this proportion is sizeable, then the simple strategy of excluding non-respondents is difficult to defend. Furthermore, exclusion provides no evidence as to whether the data are missing data *at random* (i.e., whether non-respondents are different in their network patterning from respondents), nor whether the missing data mask important structural properties. The latter prospect might be realized, for instance, if individuals who are central in an organization do not respond because they are too busy. Stork and Richards (1992) note the importance of similarities between respondents and non-respondents. In cases of dissimilarity at either structural- or individual-level, exclusion distorts conclusions.

Importantly, many network studies have non-respondents for which the missing at random assumption is quite inappropriate. This occurs in the class of studies we noted above, where the research focus is on respondents, for instance, within one organizational department, but with additional information about ties to non-respondents from a quite separate context (such as from other departments). Usually in

such cases there are no grounds to assume that the ties from respondents to non-respondents present with the same structural patterns as those among respondents.

An exponential random graph approach

In this article, we present an approach in which all available data are modeled simultaneously – both the full network data for respondents and the incoming ties for non-respondents – without making reconstruction assumptions. Our models do not require that respondents and non-respondents be similar, either structurally or at the individual-level, although the models permit some exploration of whether respondents and non-respondents are involved in similar structural patterns.

We utilize exponential random graph models – commonly referred to as p^* models – for our purposes (Frank & Strauss, 1986; Pattison & Wasserman, 1999; Robins, Pattison & Wasserman, 1999; Wasserman & Pattison, 1996). The advantage of this class of models is that they model global network structure as the outcome of processes occurring in *local social neighborhoods* of the network. A local social neighborhood can be construed as a set of network tie variables that are hypothesized to be mutually conditionally interdependent (Pattison & Robins, 2002). The form of these local social neighborhoods is determined by an hypothesized dependence structure, that is, by a set of assumptions about which pairs of potential ties are dependent, conditional on the values of all other tie variables (e.g., see Robins & Pattison, in press). A common dependence assumption has been a Markovian one (Frank & Strauss, 1986) in which two tie variables are assumed to be conditionally independent only when they do not have nodes in common. Given a particular set of dependence assumptions, and a consequent specification of the form of local social neighborhoods, the resulting random graph model expresses the probability of a global network structure as a function of

parameters and observed statistics pertaining to certain network configurations (small sub-graphs) that occur within local neighborhoods of the network.

In the case of survey network data with missing data, a respondent can be thought of as having two types of tie: those that are expressed to other respondents; and those that are expressed to non-respondents. The local social neighborhood of any possible tie can then be hypothesized to contain both types of tie. Accordingly, exponential random graph models can be constructed in which the two types of tie are kept distinct and the model for the network is expressed in terms of configurations containing one or both types of tie.

In essence, we propose to model the sub-matrix of the adjacency matrix that includes all columns but only the rows pertaining to respondents. We impose homogeneity across all respondents and also separately across all non-respondents. The resulting model parameters, then, refer to isomorphic network configurations in which respondents and non-respondents are distinguished.

If the researcher has confidence that the non-respondents are missing at random, then we propose a further simplification of the model in which homogeneity is imposed across appropriate configurations pertaining to respondents alone and to both respondents and non-respondents.

When, however, the research focus is principally on the respondents and in circumstances when the missing at random assumption does not apply, we could of course simply ignore non-respondents and use, for instance, a standard Markov graph model (Frank & Strauss, 1986). But it seems sensible to include all available information in attempting to understand the respondent-to-respondent connections. So we propose models in which the ties to non-respondents are exogenous predictors of ties involving only respondents.

Our rationale for distinguishing respondent-to-respondent ties and respondent-to-non-respondent ties in this way is that they differ markedly in the quality of the local network information that we have available for modeling. In the case of respondent-to-respondent ties, the dyadic local neighbourhood is necessarily complete, and there is likely to be a mix of non-missing and missing data in higher-order (e.g., triadic) neighbourhood configurations. For respondent-to-non-respondent ties, on the other hand, even the dyadic neighbourhoods involve missing data, and there is a larger proportion of higher-order neighbourhoods with missing data. As a result, any attempt to model ties involving non-respondents is likely to be less successful than attempts to model ties involving only respondents. Nonetheless, ties involving non-respondents can provide important information about the local neighborhood of a tie linking two respondents (for example, a tie between two respondents may be more likely if both respondents share a tie to a non-responding third party). Consequently, we can adopt the approach of treating respondent-to-non-respondent ties as exogenous in a model for respondent-to-respondent ties.

The use of predictor variables in exponential random graph models has hitherto focused on the development of social influence, social selection and temporal models (Robins, Pattison & Elliott, 2001; Robins, Elliott & Pattison, 2001; Robins & Pattison, 2001, respectively) with both social influence and temporal models involving a set of network ties as predictors. Using the non-respondent ties as predictors, we develop models that can be compared with models for the data with non-respondents ignored altogether. Such comparisons provide useful information about how important non-respondent ties are to an understanding of the major effects in the respondent-only network.

Below, we present both of these modeling approaches: the more general model for both respondent and non-respondent structure, which might be further simplified if the missing at random assumption can be justified; and models for the respondent-only network, using non-respondents ties as exogenous predictors. Comparisons between various models within these classes may also be revealing about the pattern of responses by respondents specifically in regard to non-respondents, especially when an effect seems substantial in one model but not so in another. There is a difficulty here in making such comparisons. The effect of “scaling-up” in these models – that is, the effect of increasing the number of nodes for fixed parameter values – is not yet well-understood (Robins, Pattison & Woolcock, in press). This is likely to be a particular problem when non-respondent numbers are large. Suffice to say that across networks with different numbers of nodes, the *values* of estimates of a parameter for a given neighborhood form may not be directly comparable in terms of indicating, for instance, the size of the neighborhood effect (even if models have the same parameterization.) We need to consider which effects may be considered as the more important in the model, irrespective of the actual estimates.

A difficulty here is that, despite rapid advances in Markov Chain Monte Carlo estimation techniques (Handcock, 2002a, 2002b, 2003; Handcock, Hunter, Butts, Goodreau & Morris, 2004; Snijders, 2002; see also Wasserman & Robins, in press, for a summary), pseudo-likelihood estimation is still at this time the most practicable option for the estimation of more complex models, including for large networks⁵. In this article we use pseudo-likelihood estimation techniques, so that we do not have available accurate standard errors for parameter estimates. Given the desirability of comparing

⁵ although this situation is changing quite rapidly with the development of the *ergm* program (Handcock et al., 2004) and the *Siena* program within the *StOCNET* package (Snijders, 2002).

across networks with different numbers of nodes, however, it is useful to establish criteria to identify the important effects in the various models, criteria that are sensitive to the large number of cases that can arise. But, because the distributional assumptions are not applicable, one needs to resist the temptation of using the standard tests of the logistic regression procedure from which pseudo-likelihood estimates are obtained. The approach we propose here uses a simple heuristic based on criteria related to model fit. We discuss this below in the context of using pseudo-likelihood estimation as an exploratory technique.

In the next section, we describe the general exponential random graph model for a unimodal network and then explain how to model the respondent-only structure, using the non-respondent information as exogenous. We proceed to a section covering certain technical details, including our heuristic for determining the important effects. We then present two empirical examples of our approach. The first example involves a well-known network data set, for which we treat a proportion of nodes as non-respondents to compare our models against the known model for the entire network. The second example is from an organizational study, with the research question related to the effect of external linkages on internal organizational structure. We conclude with a discussion of further extensions of our approach to models that involve multivariate networks and node attributes.

Models for unimodal directed networks with missing data

Models for all the data: respondents and non-respondents

Some terminology and notation

For a set of n persons or *actors*, we represent a *relational tie* between persons i and j as a binary random variable X_{ij} where $X_{ij} = 1$ if person i considers person j as a

partner under the relationship, and where $X_{ij} = 0$, otherwise. In other words, a relational tie is a property of an ordered pair (i, j) of persons, although a tie may not be possible for all ordered pairs. We define a *couple* as an ordered pair of actors (i, j) between whom a relational tie is possible. We regard the network as a random (directed) graph $\mathbf{X} = [X_{ij}]$ with the fixed node set $K = \{1, 2, \dots, n\}$ and with an edge directed from node i to node j if $X_{ij} = 1$ ⁶. We let $\mathbf{x} = [x_{ij}]$ denote the matrix of realizations of the variables X_{ij} .

In the case of missing data, the node set K is the union of two disjoint subsets, R , the subset of respondents, and N , the subset of non-respondents. The set of couples then is defined as $C = \{(i, j): i \in R, j \in K = R \cup N, \text{ and } i \neq j\}$: that is, any ordered pair (i, j) with $i \in N$ is not a couple (a tie from a non-respondent is not considered possible for the purposes of the analysis); nor is a self-tie permitted, as is standard in most network procedures.

Exponential random graph models

We begin by briefly outlining the class of exponential random graph models for networks with only respondents, that is, when $K = R$. These models were first introduced into network analysis through the Markov random graphs of Frank and Strauss (1986). Wasserman and Pattison (1996) discussed p^* models for univariate networks, with further elaborations for multivariate networks provided by Pattison and Wasserman (1999), and for valued networks by Robins, Pattison and Wasserman (1999).

Wasserman and Robins (in press) summarize the basic approach, and the underpinning of the models through dependency structures is presented in Robins and

⁶ If (i, j) is not a couple in \mathbf{X} , X_{ij} is considered a structural zero. In fact this makes no difference at all in the ensuing description. Rather than a matrix (with some structural zeroes), \mathbf{X} can be considered as a set of random variables on the couples.

Pattison (in press). In summary, the models in the exponential random graph class take the form:

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{\kappa} \exp\left(\sum_{T \subseteq C} \theta_T \prod_{(s,t) \in T} x_{st}\right) \quad (1)$$

where C is the set of couples; κ is a normalizing quantity; the parameters θ_T relate to network *configurations* (local subgraphs) of particular types depending on the model, and the network statistic pertaining to a configuration $T - \prod_{(s,t) \in T} x_{st}$ – indicates whether that configuration is actually observed in the network \mathbf{x} ⁷. So each parameter relates to the presence or absence of a particular network configuration in the observed network. By imposing various homogeneity constraints – so that parameters of isomorphic configurations are equated – and by restricting the order of terms in (1), an identifiable model results, with the parameters expressing tendencies for certain classes of network configurations to be observed.

The simplest model of the class is that of Bernoulli random graphs (Erdős & Renyi, 1959; Frank & Nowicki, 1993), where the couples are assumed to be independent of each other, so the only relevant network configurations pertain to single ties. A homogeneity assumption here is equivalent to the assumption that each tie occurs in the graph with equal probability, and the model is simply:

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{\kappa} \exp(\theta L) \quad (2)$$

⁷ More formally, a dependence graph is used to represent the hypothesized dependencies among the couples, with the couples as nodes and the edges representing conditional dependencies between couples (see Robins and Pattison, in press, for details). The parameterization of the models is determined by the structure of the dependence graph, in particular by the cliques of the graph. The parameter θ_T is non-zero if and only if T is a clique in the dependence graph, and there is one and only one parameter for each clique. Different dependence assumptions leads to different types within the class of exponential random graph models.

where θ is a parameter relating to the density of the network and L is the number of observed ties. A slightly more complex version assumes that dyads are independent of each other, with the simplest form of homogeneity resulting in parameters relating to ties and to reciprocated ties:

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{\kappa} \exp(\theta L + \rho M) \quad (3)$$

where θ and L are as before, with ρ a parameter relating to reciprocity and M the number of mutual ties⁸.

Markov random graph models assume that two network couples, (i, j) and (r, s) , are independent unless they share a node. As Frank and Strauss (1986) noted, the resulting configurations for the Markov graph model relate to single ties, mutual ties, various stars, and triadic configurations. In this article we focus on models based on such Markov dependence structures⁹, and with parameters based on the configuration types depicted in the left hand column of Figure 1 below. These include the familiar p^* triadic and two-star configurations, here labeled as τ_9 to τ_{15} (for the full numbering system, which is based on the standard network triad count, see Figure 2 of Robins and Pattison, in press), together with 3-in-star and 3-out-star configurations labeled σ_3^I and σ_3^O , respectively. (We use the same labeling for configurations and for their related parameters in the model.) This Markov random graph model then becomes:

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{\kappa} \exp \left(\sum_{p=9}^{15} \tau_p T_p(\mathbf{x}) + \sigma_3^I S_3^I(\mathbf{x}) + \sigma_3^O S_3^O(\mathbf{x}) \right) \quad (4)$$

⁸ A looser homogeneity assumption results in the well-known dyadic independence p_1 model of Holland and Leinhardt (1981) – see Robins and Pattison (in press).

⁹ More complex models, involving longer paths, higher order configurations, and setting structures, have also been developed – see Pattison and Robins (2002) and Snijders, Pattison, Robins & Handcock (2004).

where $T_p(\mathbf{x})$ is a count of triads of type p in the observed graph \mathbf{x} , $S_3^I(\mathbf{x})$ is the number of 3-in-stars in \mathbf{x} , and $S_3^O(\mathbf{x})$ the number of 3-out-stars. Because this is a model where all nodes are respondents, below we term this an *R model*.¹⁰

Equation (4) expresses a distribution of random graphs, each of which can be construed as arising from an agglomeration of the configurations represented by the parameters. So the parameters can be interpreted as indicating the strength of the local structural effects that produce the graph. It should be noted that interpretation of a parameter is relative to higher or lower order configurations. For instance, in a dyadic model a positive reciprocity parameter (ρ) in the presence of a negative density, or edge, parameter (θ) indicates that there are more reciprocated ties than would be expected by chance, given the number of ties (alternatively, the density) observed in the graph.

Introducing non-respondents

The presentation of the *R* models above makes clear why they can be used to model both respondents and non-respondent data simultaneously. There is nothing in the formulation of (1) that requires that the matrix \mathbf{X} be square, so that it could comprise rows represented by respondents and columns by both respondents and non-respondents. Indeed, (1) is quite general, in that there is no restriction on the set of couples C . Accordingly, for networks with missing data, we can present (1) in an expanded form that makes explicit the inclusion of non-respondents:

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{\mathcal{K}} \exp\left(\sum_{P \subseteq R \times R} \theta_P \prod_{(s,t) \in P} x_{st} + \sum_{\substack{Q \subseteq C: \\ Q \cap (R \times N) \neq \emptyset}} \theta_Q \prod_{(s,t) \in Q} x_{st} \right) \quad (5)$$

where the first summation is over configurations P involving only respondents, and where the second summation is to be taken over configurations Q that include at least

¹⁰ Readers familiar with p^* network models may be puzzled by the inclusion of three-star parameters.

one couple (s, t) with t a non-respondent¹¹. Note that if researchers exclude non-respondents and treat the network as comprising respondents only, in fitting a p^* model they use a model based on only the first, but not the second, summation.

For our general model, we simply utilize (5) but in applying homogeneity constraints we distinguish between nodes that are respondents and those that are non-respondents. So, for instance, the parameter for a two-in-star (τ_{14}) involving three respondents is different from the parameter for a two-in-star with two respondents and one non-respondent (see the second row of Figure 1). Technically, we treat the nodes as *colored* depending on whether they represent respondents or non-respondents, and we impose homogeneity across isomorphic configurations where the isomorphism preserves both edges and colors. A similar homogeneity approach is adopted in exponential random graph social selection and social influence models.

Under a Markov graph dependence assumption, the result is a model with parameters relating to the configurations in Figure 1. We term this an *RN* model. It takes the form:

$$\begin{aligned}
P(\mathbf{X} = \mathbf{x}) = & \frac{1}{\kappa} \exp\left(\sum_{p=9}^{15} \tau_p T_p(\mathbf{x}) + \sigma_3^I S_3^I(\mathbf{x}) + \sigma_3^O S_3^O(\mathbf{x})\right) \\
& + \sum_{p=12}^{15} \tau_{pN} T_{pN}(\mathbf{x}) + \tau_{9N} T_{9N}(\mathbf{x}) + \tau_{12NN} T_{12NN}(\mathbf{x}) \\
& + \sigma_3^I S_{3N}^I(\mathbf{x}) + \sigma_3^O S_{3N}^O(\mathbf{x})
\end{aligned} \tag{6}$$

Obviously, in any configuration there cannot be a tie from a non-respondent. So there are no parameters for reciprocated ties and cyclic triads involving non-respondents. Information on ties to non-respondents is not helpful in determining trends

Hitherto, published applications of these models have typically only included two-stars as the highest order star parameters. We discuss this point below.

¹¹ More formally, the summations are over dependence graph cliques P that contain only respondent-respondent couples, and over cliques Q that contain at least one respondent-non-respondent couple.

towards reciprocation or cyclicity, so that RN models do not assist in estimating such trends.

(Figure 1 about here)

Homogeneous RN models

Suppose a researcher has confidence that the non-respondents are missing at random. Then we can attempt to fit a model imposing homogeneity across appropriate configurations presented in Figure 1. This amounts to equating parameters across the rows of the Figure. More particularly, we equate the following parameters: $\tau_{15} = \tau_{15N}$; $\tau_{14} = \tau_{14N}$; $\tau_{13} = \tau_{13N}$; $\tau_{12} = \tau_{12N} = \tau_{12NN}$; $\tau_9 = \tau_{9N}$; $\sigma_3^I = \sigma_{3N}^I$; and $\sigma_3^O = \sigma_{3N}^O$. To be clear in what follows, we will label such parameters with an ‘H’ subscript (referring to “homogeneity”), and we will term this an *RH* model. Inferences based on this model might be taken to apply to the network as a whole, irrespective of the non-respondents. Clearly such inferences will only be viable if the missing at random assumption is reasonable. Comparisons between the two models may be useful in deciding the validity of this assumption.

Models for respondents with non-respondent ties as exogenous

If the central focus of the research is on the respondents, and if the missing at random assumption is not viable, it may be helpful to fit a model predicting only respondent ties, but using non-respondent ties as exogenous predictors. We term this an *R+* model, in that we are fitting a model only for the network of respondents but taking into account information relating to non-respondents. A typical research question here might be whether ties external to the respondents’ organization are important in determining structure among respondents.

Using Markov dependence assumptions, the resulting model is relatively straightforward with the parameters a subset of those for the more general model (6).¹² The essential difference is that we are only modeling ties in configurations that involve respondents. For instance, in Figure 1 consider the τ_{12N} configuration, an out-star that contains two ties, one to a respondent and one to a non-respondent. We can incorporate the τ_{12N} configuration into our model, but in this case only as a predictor of the tie between the two respondents. Any configuration that does not involve a tie among respondents, then, will not enter the model. Accordingly, the model includes configurations τ_{12N} , τ_{13N} and τ_{9N} , as well as the standard τ_9 to τ_{15} Markov configurations, σ_3^I and σ_3^O . The model does not include parameters for τ_{14N} , τ_{12NN} , σ_{3N}^I and σ_{3N}^O , as these configurations, if observed, would not include any respondent-to-respondent ties.

In summary then, we are dealing with four models. The *R* model is simply a Markov random graph model with non-respondents excluded. The *RN* model is the fullest model and treats non-respondents as a different type of node, retaining all the resultant homogeneous effects. The *RH* model derives from the *RN* model by equating various of the *RN* effects to produce parameters akin to those of standard Markov graph models; it may be applicable if the missing at random assumption holds. The *R+* model, on the other hand, treats ties to non-respondents as exogenous, and may be appropriate when the missing at random assumption is not applicable. With these four models in place, we present two examples: one from a well-known network data set from which we treat a certain number of nodes as non-respondents, derive an *RN* and an *RH* model, and compare against the *R* model for the entire network. Our second example comes

¹² Technically, to develop models with network ties as predictors, we derive a dependence graph from a

from an organizational study, with the research question pertaining to network structure among respondents within the one organizational department. The data, however, included network ties from respondents to non-respondents in other departments and organizations. We use an R^+ model to investigate whether there are any important “boundary-spanning” effects involving ties outside the department that influence the shape of the network within the department.

Before we present these examples, however, we discuss some technical details related to model comparisons, parameter estimation, and the behavior of “near degenerate” models.

Estimation and model comparisons

In this article we use pseudo-likelihood estimation, suggested for this class of models by Strauss and Ikeda (1990). There have been recent promising developments in Monte Carlo maximum likelihood estimation for Markov random graph models (Handcock, 2002a, 2002b, 2003; Snijders, 2002), based on algorithms for long-run simulations, but these methods have yet to be implemented for more complex models, nor in practical terms are they yet available for very large graphs. As our immediate purpose is simply to illustrate the new models, we use the more convenient pseudo-likelihood estimation procedure with the warning that the parameter estimates need to be seen as approximate.

One issue with any estimation procedure for these models relates to degeneracy, first discussed in the context of Markov random graphs by Strauss (1986). Near degenerate models occur when particular combinations of parameter values lead to the simulation procedure behaving unusually, either being “trapped” in a particular region of parameter space, or possibly oscillating between different regions, as illustrated by

some of Snijders' (2002) examples. For such parameter values, any estimation procedure is likely to be problematic. Handcock (2002a, 2002b, 2003) showed that this behavior is to be expected in certain regions of the parameter space, but that for non-degenerate regions, Monte Carlo estimation can proceed satisfactorily. Our own work on large-scale simulations of Markov random graph models (Robins, Pattison & Woolcock, in press) suggests that models that include negative parameters for three-stars can give rise to plausible model properties in some circumstances. This is a preliminary finding – and indeed recent work suggests that the inclusion of certain higher order non-Markov parameters will also be useful in avoiding near degenerate models (Snijders et al, 2004) – but our current recommendation for Markov models is that they should at least include parameters pertaining to 3-in- and out-stars, as above.

The combination of 2- and 3-stars in the one model enables a better modeling of the degree distribution. A positive 2-star parameter suggests a tendency for actors to have multiple network partners, while a negative 3-star parameter suggests a ceiling on this effect. In other words, if the magnitude of the negative 3-star parameter were sufficiently large, many actors would have multiple network partners but there would be few with very many. (For further details of 3-star interpretation, see Robins et al, in press.)

Pseudo-likelihood

Pseudo-likelihood estimation can be implemented through standard logistic regression procedures. Each couple represents a case, with the observation of a tie predicted from statistics associated with each parameter. Each statistic is the number of configurations pertaining to the parameter that the tie would complete, if it were observed. (For example, in an R model, if there are three two-paths from i to j , then a tie from i to j would complete one τ_{15} configuration, three 2-in-stars, τ_{14} , three 2-out-stars,

τ_{12} , and three transitive triads, τ_9 .) In Appendix 1, we briefly outline how to calculate the statistics for each of the configurations in Figure 1.

Any logistic regression output typically produces standard errors and a deviance statistic, which is a measure of fit. In pseudo-likelihood estimation, the standard errors are unreliable and may often be too small (Snijders, 2002), although they might be taken as a rough indicator of scale. In standard logistic regression models for independent observations, the deviance is asymptotically distributed as chi-squared. The pseudo-likelihood deviance is still a useful measure of fit but its distribution is unclear¹³. Moreover, the pseudo-likelihood data file for a sizeable network will contain a very large number of cases, and it would normally be desirable to take this into account in determining an appropriate alpha level. When there is substantial power in the presence of many cases, blind adherence to chi-squared approaches with standard alphas will result all too readily in the retention of many small effects that achieve notional significance. For the same reasons, the use of the Wald statistic from logistic regression procedures is an uncertain guide to indicate important parameters in these models. So, pseudo-likelihood estimation needs to be seen as an exploratory technique, not one that is appropriate for formal null hypothesis significance testing. This is not always a problem for network analytic studies which often rely on exploratory techniques in any case.

Moreover, the scale-up properties of these models – that is, the behavior of the same model with increased numbers of nodes – are not yet well understood (Robins, Pattison & Woolcock, in press). Hence it is difficult to make direct comparisons of parameter estimates (irrespective of the estimation procedure): for instance, in

¹³ Indeed, given that the sampling space of random graphs on a fixed number of nodes is finite, albeit large, it is not certain that the use of asymptotic results is appropriate in any case.

comparing across graphs with different numbers of nodes, parameter estimates should not necessarily be seen as representing effect sizes on comparable scales. Nevertheless, conclusions can be drawn about the importance of various effects in a model, and on that basis comparisons might be made. The advantage of Monte Carlo estimation techniques is the availability of reliable standard errors, from which confidence intervals for estimates can be produced. In the more exploratory world deriving from pseudo-likelihood estimation, the absence of reliable standard errors makes this approach problematic.

A heuristic for model simplification

As a means to making decisions about important parameters, we provide here a non-distributional heuristic based on the pseudo-likelihood deviance as a measure of fit. The idea is that parameters that are not important would not affect model interpretation grossly if they were removed, so the basis of the heuristic is to ensure that the conditional probabilities of a tie being present, as estimated from the models, do not vary substantially for too many cases if a parameter were to be removed. We may still retain the parameter in the model, but then treat it as “unimportant” in the sense that it does not greatly affect interpretation.

As a first step, comparison of means of absolute residuals is useful. In addition, we propose a more defensive strategy when removing a parameter, a step that will lead to a worsening of residuals (i.e. a worsening in the model’s predictive capacity). Larger changes in estimated probabilities of a tie being correctly observed (or not observed) might be tolerated for cases when the model is already successful but only smaller changes might be accepted where the model is weakly predictive. For this purpose, the pseudo-likelihood (PL) deviance statistic turns out to be valuable. Decisions are required about the level of deviations in predicted probabilities that are regarded as

tolerable. This gives an indication of the level of overall deviation that is acceptable, and the change in the PL deviance is a useful summary statistic for that purpose.

We provide detail of our approach in Appendix 2. In summary, we suggest removing a parameter from the model if the resulting change of deviance is less than $-2N \log(1-\delta)$ where N is the number of cases (in a unimodal binary network, the number of couples) and where δ (defined in the Appendix as an acceptable level for the proportional change in predicted probabilities) is a small number, possibly 0.001 or 0.005. For instance, suppose we have a network of 50 actors, all of whom are respondents, so that there are 2450 couples, and suppose we set $\delta = 0.005$. Then we would remove parameters from the model if they did not diminish the PL deviance by at least 24.6. With $\delta = 0.001$, we would remove parameters only if the PL deviance was not diminished by less than 4.9. Clearly the researcher has a choice here, with a smaller δ being a more rigorous criterion¹⁴.

This approach can be used to simplify models by parameter removal (see for instance, Robins et al, 1999), or simply to indicate the parameters that are not important to a model's predictive capacity, which is what we do in this article. If the approach is used to simplify models, there is one possible proviso to the removal of a parameter that is not a substantial contributor to the PL deviance. Configurations contain within them various other sub-configurations. Accordingly, in many circumstances it is desirable to keep models hierarchical, so that parameters that relate to lower order configurations are retained in the model in the presence of substantial higher order parameters. Admittedly, there are times when a non-hierarchical model may be pragmatically helpful in terms of

¹⁴ In the sense that for smaller δ two models that differ by one parameter are considered "equivalent" if the difference in their PL deviances is smaller. In other words, it is easier to consider a parameter "unimportant" if the δ is larger.

simplifying interpretation (see Robins and Pattison, 2001, for an example), but a decision to use a non-hierarchical models should be made with care.

Example 1: The Kapferer tailor shop data

As an example we use the Kapferer (1972) tailor shop data for instrumental interactions (work and assistance-related) in a Zambian tailor shop. Binary directed observations were made on 39 actors at two time points. We use the data from the first time point here. The dataset is available in UCINET5 (Borgatti, Everett & Freeman, 1999).

We begin by fitting a Markov random graph model for the entire network. The data comprises observations on 1,482 couples (39×38). If we use a stringent $\delta = 0.001$, then $-2N \log(1 - \delta) = 3.0$, so that effects that diminish the PL deviance by 3 or more are considered important. The results are in top panel of Table 1 with unimportant parameters marked with ‘#’. We retain these parameters in the model but do not interpret them. Perhaps not surprisingly, the strongest effect is for reciprocity (the τ_{11} estimate). There are fewer two-paths across the network than expected (a negative τ_{13} estimate) unless those paths are closed into transitive triads (the positive τ_9 estimate). Interestingly there is a strong negative effect against 3-cycles (τ_{10} estimate). So, apart from reciprocity, the major effects are for transitivity and against cycles, indicating hierarchical network closure. There is also a positive τ_{12} estimate, suggesting variation in expansiveness; so there are likely to be some actors who have relatively high outdegree. The three-star parameters are not important in this model.

Table 1 about here.

To examine our non-respondent models in comparison with the results in Figure 1, we arbitrarily chose the last 19 of the 39 actors as non-respondents. It cannot be confidently said that these non-respondents are missing at random. A blockmodel of the

entire network, with separate blocks for the first 20 and the last 19 actors, indicates relatively high density in blocks (0.14 in the first and 0.10 in the second) yet lower density between blocks (0.05 for block 1 selecting block 2, and 0.01 for block 2 selecting block 1). Nevertheless, the non-respondent ties are not exogenous here, so that *RN* models – and possibly *RH* models depending on how poorly the missing at random assumption holds – are more appropriate than *R+* models with exogenous non-respondent ties.

We begin with parameter estimates for the respondent-only (*R*) model, presented in the bottom panel of Figure 1. Comparison between the two models in Figure 1 indicates that many of the effects in the entire data are captured quite well by the respondent *R* model, but there are some important differences, notably a change in the sign of the 2-path (τ_{13}) estimate, a strengthening of the positive 2-out-star (τ_{12}) estimate and of the negative 3-cycle (τ_{10}) estimate, and an important negative 3-out-star effect. These results are consistent with the blockmodel – there are relatively more 2-paths within the denser respondent block, so that, relative to the increased 2-path effect, there are even fewer 3-cycles (recall that a 3-cycle is “built up” of three 2-paths). Moreover, within the denser respondent block there is a greater tendency for multiple choices of network partners, but there is nevertheless a ceiling on this effect (hence the negative 3-star effect.) The pseudo-likelihood deviances here are not comparable, given that they relate to different numbers of cases, but the mean absolute residual suggests that the *R* model does not do anywhere near as well as the full model in predicting ties and non-ties.

In Table 2, we present estimates for the *RN* model. From the mean absolute residual, we see that this model does appreciably better than the *R* model. Here we briefly interpret the important effects relating to non-respondents. There is an important

popularity effect for non-respondents (the positive τ_{14N} estimate), so that if a non-respondent is chosen at all there is a tendency for them to be chosen multiple times. But the negative 3-in-star effect (the σ_{3N}^I estimate) places a ceiling on popularity levels. This is consistent with the blockmodel: the indegree distribution for the second block chosen by the first indicates that of the 11 non-respondents selected by respondents, six have indegrees 2 or 3, but none with indegrees higher than 3.

As might be expected given the lower density of selections of non-respondents by respondents, there is an effect against 2-paths ending in the second block (the negative τ_{13N} estimate). The previously observed outdegree effect is even more pronounced when a non-respondent is selected (the positive τ_{12N} and τ_{12NN} estimates, counterbalanced by the negative 3-outstar σ_{3N}^O estimate), reflecting the fact that a couple of respondents select a high number of 5 or 6 non-respondents.

Table 2 about here.

In Table 3, we present the *RH* model. Because the numbers of cases are identical, the pseudo-likelihood deviances are directly comparable between the *RH* and *RN* models. We see that the *RN* model yields better predictions, with a difference of 34.9 in PL deviances but at the cost of an additional 8 parameters, an average of 4.4 per variable. In the Tables, to determine importance of variables, we have used a stringent $\delta = 0.001$, so that for these *RN* and *RH* models, only variables that contribute 1.5 or less to the PL deviance are considered unimportant. If we had used the less stringent $\delta = 0.005$, variables contributing less than 7.6 to the deviance would be considered unimportant. This suggests that *RN* model is indeed better but perhaps not dramatically so, a conclusion confirmed by the mean absolute residuals. So we might infer that although there is structure in the blockmodel, the missing at random assumption for these non-respondents has some plausibility.

Table 3 about here

If we used the *RH* model to infer effects for the entire network, then a comparison of Table 3 with the first panel of Table 1 indicates that we would not draw vastly different conclusions. This is in a situation where we have almost half the original actors specified as non-respondents, so the *RH* model is performing well in coping with the missing data.

We have chosen this example to illustrate the potential of this approach, even when the proportion of non-respondents is as high as 50%, and to permit others to replicate the results. But as pointed out earlier, we have not chosen non-respondents at random. Our experience of choosing 19 non-respondents at random for this network suggests that for the majority of occasions the *RH* estimates are consistent with the full network estimates as in Figure 1. There are occasions when this is not so, particularly when the 3-cycle (τ_{10}) estimate becomes extremely negative to the point of suggesting overfitting of the parameter. Our original inference was that there were fewer 3-cycles than expected in the network. Creating a large number of non-respondents, and thereby removing a large number of ties, may result in there being very few or no 3-cycles in the data at all. In these circumstances the estimation procedure obviously cannot function well and the parameter will appear as overfitted. If this occurs in actual data, we suggest an investigation of the presence of the configuration in the data to confirm its relative absence (which might be a useful conclusion in its own right), and then the removal of the parameter from the model.

The problem obviously does not apply to the same extent if the proportion of non-respondents is lower. We have investigated the Kapferer data further, producing an *RH* model for each of 15 selections of 9 non-respondents at random. The mean and

standard deviation of the parameter estimates are produced in Table 4. It will be seen that these results are strikingly consistent with the model for the full data in Table 1.

Table 4 about here.

Example 2: An organizational department with external non-respondents.

We fit the models to a data set from an Australian government organization (Rogers, Kempt & Bergin, 2002). A number of networks were measured for 60 respondents in one department of the organization. Respondents were free to choose their network partners from other individuals within that department or from other departments. The network we model here is a binary work frequency network, signifying frequent work partners. Nominations for frequent work partners included the original 60 respondents, together with an additional 171 non-respondents. In total there were 577 observed ties, of which 245 were among respondents and 332 from respondents to non-respondents. The question we ask here is whether external connections beyond the department are important to an understanding of departmental structure.

It is worth noting that this data is not particularly suitable for reconstruction along the lines of Stork and Richards (1992), as the 245 ties among respondents included only 39 mutual dyads¹⁵. If we were to apply reconstruction to the non-respondent ties, we would likely be wrong more often than right. This is not to say that there is no reciprocity effect in the data: given that the density of the network is some 7%, the fact that 78 out of 245 ties (32%) are reciprocated suggests quite high levels of

¹⁵ This result is less puzzling when other aspects of the data set are taken into account. The network responses were quite strongly influenced by the formal hierarchy of the organization, so that in choosing network partners, individuals seemed tempted to choose those on similar or higher levels of the hierarchy. In other words, in surveying their perceived working environment, respondents tended to look “upwards”, rather than “downwards”.

reciprocity. This example illustrates that, even when there is a tendency towards reciprocity, reconstruction assumptions may not apply particularly well¹⁶.

R-models

We begin by fitting the respondent-only R model. The respondent-only data comprises observations on 3,450 couples (60×59). If we use a $\delta = 0.001$, then $-2N \log(1 - \delta) = 6.9$, so that effects that diminish the PL deviance by 7 or more are considered important. The results are presented in Table 5,

The model suggests reciprocity effects (τ_{11}), and some not surprising evidence for hierarchy, with transitivity effects (τ_9 – suggesting that the global structure has some tendency towards clustering of nodes based on hierarchically ordering), and an effect against cyclicity (τ_{10} – suggesting tendencies against the clustering of nodes based on cyclic exchange).

Table 5 about here

R+ model

The $R+$ model is presented in Table 6. The first point to note is that none of the three additional non-respondent parameters are important in this model. Nor does the PL deviance, nor the mean absolute residual, differ greatly for that of the Markov R model in Table 5. We conclude that external ties beyond the department are not important in understanding the internal departmental work patterns¹⁷. Note that this conclusion may not apply if the work frequency network were included in a broader multivariate network analysis, or if actor attributes were included – see the discussion below.

¹⁶ Examination of such statistics for all parameters also gives confidence that the exploratory pseudo-likelihood estimates are successfully indicating strong effects in the data.

¹⁷ The presence of the additional parameters does permit the 3-in-star parameter to reach the criterion of importance, but this does not seem a major change.

Table 6 about here

Conclusions

In this article, we have modeled networks with missing data in the form of non-respondents to a sociometric questionnaire, using exponential random graph models. Our approach is deliberately pragmatic: we attempt to use whatever information is available to strengthen interpretations, rather than to discard data when it is incomplete (e.g. when there is information on non-respondents in terms of indegrees but not outdegrees.) As with any missing data in social sciences, there is no foolproof way to infer from what is not observed, and we underline the importance of obtaining the best and most complete data that one reasonably can. Nevertheless, network data is difficult to collect, and there are times when there are many non-respondents. In these all-too-common circumstances, we need to use the data as best we can.

The exponential random graph approach works with missing data because the models essentially adduce the global structure of a network as the aggregation of local sub-structures. So even though there may be many gaps in the data, sufficient numbers of local social neighborhoods may still be observed from which to make reasonable inferences. Other network techniques that rely more on observed global network structures, and not so explicitly on local neighborhoods, are not particularly adequate in the face of substantial missing data.

We see this article as in the tradition of exploratory analysis, rather than as presenting procedures that demand formal statistical testing (even though the models do have a statistical basis). This exploratory approach is forced on us to a degree by the approximate pseudo-likelihood estimation technique we have used, although new developments in Monte Carlo maximum likelihood estimation open the prospect of

using these models in a more formal way. But we see our primary goal as the obtaining of insight into the data, rather than in the reaching of definitive statistical conclusions (although the latter are important if available). It is in this light that we have presented our various proposals for heuristics in model selection and interpretation.

The techniques in this article can be readily extended to the analysis of multivariate and valued networks in a natural way based on the multivariate and valued extensions of p^* models (Pattison and Wasserman, 1999; Robins et al, 1999). More complex generalizations would involve actor attributes, for instance, in social selection models. If attribute information is available on non-respondents, then again the extension follows naturally with a rather direct application of these techniques to social selection models. If attribute information is not available for non-respondents, however, the non-respondents then in effect create a new category of attribute. For instance, if information on a binary attribute such as sex is not available for non-respondents, then there are in effect three “colors” on the nodes of the network: non-respondents, male respondents, female respondents. Homogeneity constraints could then be imposed based on isomorphic configurations, where the isomorphism preserves the three colors, in analogy with the approach above for two colors.

References

- Borgatti, S., Everett, M., & Freeman, L. (1999). *UCINET 5 for windows: Software for social network analysis*. Harvard, MA: Analytic technologies.
- Brewer, D.D., & Webster, C.M. (1999). Forgetting of friends and its effects on measuring friendship networks. *Social Networks*, 21, 361-373.
- Burt, R.S. (1987). A note on missing social network data in the General Social Survey. *Social Networks*, 9, 63-73.
- Butts, C. (2003). Network inference, error, and informant (in)accuracy: a Bayesian approach. *Social Networks*, 25, 103-140.
- Erdős, P., & Renyi, A. (1959). On random graphs. I. *Publicationes Mathematicae (Debrecen)*, 6, 290-297.
- Frank, O., & Nowicki, K. (1993). Exploratory statistical analysis of networks. In J. Gimbel, J.W. Kennedy & L.V. Quintas (Eds.), *Quo Vadis, Graph Theory? Annals of Discrete Mathematics*, 55, 349-366.
- Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81, 832-842.
- Handcock, M. S. (2002a). *Degeneracy and inference for social network models*. Paper presented at Sunbelt XXII International Social Network Conference, New Orleans, February 2002.
- Handcock, M. S. (2002b). *Assessing degeneracy in statistical models for social networks*. Paper presented at Workshop on Dynamic Social Network Analysis, Washington, Nov 7-9, 2002.

- Handcock, M.S. (2003). Statistical models for social networks: Inference and degeneracy. In R. Breiger, K. Carley, & P. Pattison (Eds.), *Dynamic social network modeling and analysis* (pp. 229-240).
- Handcock, M. S., Hunter, D., Butts, C. T., Goodreau, S., & Morris, M. (2004). *ergm*: an R package for the statistical modeling of social networks. Center for Statistics in the Social Sciences, University of Washington, Working Paper No. 40.
- Holland, P.W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76, 33-65.
- Kapferer, B. (1972). *Strategy and transaction in an African factory*. Manchester: Manchester University Press.
- Krackhardt, D. (1987). Cognitive social structures. *Social Networks*, 9, 109-134.
- Koehly, L., & Pattison, P. (in press). Random graph models for social networks: multiple relations or multiple raters. In P. Carrington, J. Scott, & S Wasserman (Eds.) *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.
- Kossinets, G. (2003). Effects of missing data in social networks. M.Phil. qualifying paper, Columbia University. (<http://arxiv.org/abs/cond-mat/0306335>)
- Klov Dahl, A. (2002). *Roundtable on IRB refusals to allow network research*. Discussion session, Sunbelt XXII International Social Network Conference, New Orleans, February 2002.
- Laumann, E.O., Marsden, P.V., & Prensky, D. (1989). The boundary specification problem in network analysis. In Freeman, L.C., White, D.R., & Romney, A.K.

- (eds.), *Research Methods in Social Network Analysis* (pp.61-87). Fairfax, VA: George Mason University Press.
- Little, R.J.A., & Schenker, N. (1995). Missing data. In G.Arminger, C.C. Clogg, & M.E.Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences*. NY: Plenum Press.
- Pattison, P.E. & Robins, G.L. (2002). Neighbourhood-based models for social networks. *Sociological Methodology*, 32, 301-337.
- Pattison, P., Robins, G., & Snijders, T. (2003). *Neighbourhood-based models for social networks: Model specification issues*. Invited address to Institute of Mathematics and its Applications workshop, "Networks and the population dynamics of disease transmission". Minneapolis, 17-21 November 2003.
- Pattison, P. E., & Wasserman, S. (1999). Logit models and logistic regressions for social networks, II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52, 169-194.
- Robins, G.L., Elliott, P., & Pattison, P.E. (2001). Network models for social selection processes. *Social Networks*, 23, 1-30.
- Robins, G.L., & Pattison, P.E. (2001). Random graph models for temporal processes in social networks. *Journal of Mathematical Sociology*, 25, 5-41.
- Robins, G.L., & Pattison, P.E. (in press). *Interdependencies and social processes: Generalized dependence structures*. In P. Carrington, J. Scott, & S. Wasserman (Eds.) *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.
- Robins, G.L., Pattison, P.E., & Elliott, P. (2001). Network models for social influence processes. *Psychometrika*, 64, 371-394.

- Robins, G., Pattison, P.E., & Wasserman, S. (1999) Logit models and logistic regressions for social networks, III. Valued relations. *Psychometrika* 64: 371-394.
- Robins, G.L., Pattison, P., & Woolcock, J. (In press). Small and other worlds: Global network structures from local processes. *American Journal of Sociology*.
- Rogers, P., Kempt, N., & Bergin, S. (2002). Personal communication.
- Rogers, E., & Kincaid, L. (1981). *Communication networks: Toward a new paradigm for research*. NY: Free Press.
- Snijders, T.A.B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3, 2.
- Snijders, T.A.B., Pattison, P., Robins, G., & Handcock, M. (2004). *New specifications for exponential random graph models*. Working Paper, University of Groningen.
- Stork, D., & Richards, W. D. (1992). Nonrespondents in communication network studies: Problems and possibilities. *Group & Organization Management*, 17, 193-210.
- Strauss, D. (1986). On a general class of models for interaction. *SIAM Review*, 28, 513-527.
- Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85, 204-212.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge, UK: Cambridge University Press.
- Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^* . *Psychometrika*, 61, 401-425.
- Wasserman, S., & Robins, G.L. (in press). *An Introduction to Random Graphs, Dependence Graphs, and p^** . In P. Carrington, J. Scott, & S. Wasserman (Eds.)

Models and Methods in Social Network Analysis. New York: Cambridge University Press.

White, H. C. (1992). *Identity and Control.* Princeton, NJ: Princeton University Press.

Appendix 1: Calculation of statistics for pseudo-likelihood estimation

The data can be decomposed into \mathbf{R} , an $R \times R$ matrix of ties among respondents (with zeroes on the diagonal) and \mathbf{N} , an $R \times N$ matrix of ties from respondents to non-respondents. Each cell in these matrices represents a case in the logistic regression that is conducted to obtain pseudo-likelihood estimates. Accordingly for each of the parameters (configurations in Figure 1), there is a matrix of statistics corresponding to the relevant couple. Define some additional matrices: $\mathbf{u1}$ is an $R \times R$ matrix containing only 1's, except for 0's on the diagonal; $\mathbf{u2}$ is an $R \times N$ matrix containing only 1's; $\mathbf{u3}$ is an $N \times N$ matrix containing only 1's, except for 0's on the diagonal; and $\mathbf{0}$ is a matrix of the relevant dimension containing only zeroes. Then the statistics can be calculated via the following matrix arithmetic.

	$R \times R$ matrix	$R \times N$ matrix
τ_{15}	$\mathbf{u1}$	$\mathbf{0}$
τ_{15N}	$\mathbf{0}$	$\mathbf{u2}$
τ_{14}	$\mathbf{u1} * \mathbf{R}$	$\mathbf{0}$
τ_{14N}	$\mathbf{0}$	$\mathbf{u1} * \mathbf{N}$
τ_{13}	$\mathbf{u1} * t(\mathbf{R}) + t(\mathbf{R}) * \mathbf{u1}$	$\mathbf{0}$
τ_{13N}	$\mathbf{u2} * t(\mathbf{N})$	$t(\mathbf{R}) * \mathbf{u2}$
τ_{12}	$\mathbf{R} * \mathbf{u1}$	$\mathbf{0}$
τ_{12N}	$\mathbf{N} * t(\mathbf{u2})$	$\mathbf{R} * \mathbf{u2}$
τ_{12NN}	$\mathbf{0}$	$\mathbf{N} * \mathbf{u3}$
τ_{11}	$t(\mathbf{R})$	$\mathbf{0}$
τ_{10}	$t(\mathbf{R}) * t(\mathbf{R})$	$\mathbf{0}$
τ_9	$\mathbf{R} * \mathbf{R} + t(\mathbf{R}) * \mathbf{R} + \mathbf{R} * t(\mathbf{R})$	$\mathbf{0}$
τ_{9N}	$\mathbf{N} * t(\mathbf{N})$	$\mathbf{R} * \mathbf{N} + t(\mathbf{R}) * \mathbf{N}$

Here ‘*’ indicates matrix multiplication and ‘t(**R**)’ indicates the transpose of **R**. Having calculated these statistics, it is simply a matter of reshaping the data and statistics into the one logistic regression file. It is perhaps simplest to calculate the three-star statistics by computing new variables in the logistic regression file: $s_3^I = t_{14}(t_{14} - 1)/2$;

$$s_{3N}^I = t_{14N}(t_{14N} - 1)/2; s_3^O = t_{12}(t_{12} - 1)/2; \text{ and } s_{3N}^O = t_{12NN}(t_{12NN} - 1)/2.$$

Appendix 2: Proposal for model simplification

Deviations in predicted probabilities

Suppose that pseudo-likelihood estimation is used for models with N cases, where Y_i is a dichotomous variable for the i -th case, where P_{0i} is the probability that $Y_i = 1$ conditional on all observed values of Y_j for $j \neq i$, as predicted under the original model, and where P_{1i} is the analogous conditional probability predicted under a model with one parameter removed. Let π_{0i} indicate the original model's predicted conditional probability of assigning the correct value of Y to case i , that is:

$$\begin{aligned}\pi_{0i} &= P_{0i} && \text{when } Y_i = 1 \\ &= 1 - P_{0i} && \text{when } Y_i = 0.\end{aligned}$$

If we consider the model as predicting probabilities of classifying each case to categories, $Y_i = 1$ and $Y_i = 0$, then π_{0i} is the probability of correct classification. Let π_{1i} be the analogous probability for the revised model. Because the revised model has one less parameter, we would expect that $\pi_{0i} \geq \pi_{1i}$, or equivalently that $P_{0i} \geq P_{1i}$ when $Y_i = 1$ and that $P_{1i} \geq P_{0i}$ when $Y_i = 0$.

The Mean of Absolute Residuals

Denote as R_{0i} the absolute residual in the original model for case i , i.e. $R_{0i} = |Y_i - P_{0i}|$. It follows that for $Y_i = 1$, $R_{0i} = 1 - P_{0i} = 1 - \pi_{0i}$, and that for $Y_i = 0$, $R_{0i} = P_{0i} = 1 - \pi_{0i}$. In other words, the absolute residual for case i is simply the model's predicted probability for incorrect classification. Let \overline{R}_0 signify the mean of the absolute residuals for model 0 and $\Delta \overline{R}_{01}$ the change in the mean absolute residuals from model 0 to model 1. We then have:

$$N\Delta\overline{R}_{01} = N(\overline{R}_0 - \overline{R}_1) = \sum_i (\pi_{1i} - \pi_{0i}) = \sum_{Y_i=1} (P_{0i} - P_{1i}) + \sum_{Y_i=0} (P_{1i} - P_{0i})$$

We would ideally like $|P_{0i} - P_{1i}| = |\pi_{0i} - \pi_{1i}|$ to be suitably small for all i (with “suitably small” to be determined by the level of deviations that are regarded as tolerable). In that case, $\Delta\overline{R}_{01}$ is clearly a simple and useful measure. What is more likely, of course, is that even if $\Delta\overline{R}_{01}$ is small, $|P_{0i} - P_{1i}|$ will be large for some i . Provided that the number of cases where this occurs is not substantial, then we might accept that overall the two models can be treated as approximately equivalent in their abilities to predict conditional probabilities, even though there may be a small proportion of individual cases where this is not so.

Despite the simplicity of $\Delta\overline{R}_{01}$, it is not on a logarithmic scale unlike the logits from which the model’s estimates are derived. Secondly, and relatedly, a particular difference $|P_{0i} - P_{1i}|$ can lead to different interpretations between models, depending on π_{0i} .

For instance, suppose $\pi_{0i} - \pi_{1i} = 0.2$. If the original model performs very well in classifying case i , say $\pi_{0i} = 0.9$, the revised model does not do quite as well but still accords the correct classification a relatively high probability of 0.7. In these circumstances, the overall interpretation of the model might not be substantially changed and this fall in the probability of correctly classifying case i may be considered as not invidious. But if the original model does poorly for case i , say $\pi_{0i} = 0.23$, the revised model will do very poorly to the point that interpretation of the model will have correct classification as quite a rare event. Here, the change in probability does change interpretation substantially.

Proportional changes in probabilities

Accordingly, we might rather accept changes in probability that are relatively large for cases where the original model fits well, but smaller for cases where it fits poorly. A means to this end is - rather than to seek small values of $|\pi_{0i} - \pi_{1i}|$ - to aim for small values of $|\pi_{0i} - \pi_{1i}|/\pi_{0i}$. For instance, in comparison with the example in the previous paragraph, if $(\pi_{0i} - \pi_{1i})/\pi_{0i} = 0.2$, and if $\pi_{0i} = 0.9$, then $\pi_{1i} = 0.78$; whereas if $\pi_{0i} = 0.23$, then $\pi_{1i} = 0.18$.

Moreover, if we revert to the logarithmic scale we can utilize the PL deviance.

The deviance statistic

The deviance is -2 times the logarithm of the pseudo-likelihood function, that is:

$$\begin{aligned} G_{PL}^2 &= -2 \log \left[\prod_i P_i^{Y_i} (1 - P_i)^{(1 - Y_i)} \right] \\ &= -2 \sum_i [Y_i \log P_i + (1 - Y_i) \log(1 - P_i)] \\ &= -2 \sum_{Y_i=1} \log P_i - 2 \sum_{Y_i=0} \log(1 - P_i) \\ &= -2 \sum_i \log \pi_i \end{aligned}$$

so that $G_{PL1}^2 - G_{PL0}^2 = -2 \sum_i \log \left(\frac{\pi_{1i}}{\pi_{0i}} \right) = -2 \sum_i \log \left(1 - \frac{\pi_{0i} - \pi_{1i}}{\pi_{0i}} \right) = -2 \sum_i \log(1 - \delta_i)$

where $\delta_i = (\pi_{0i} - \pi_{1i})/\pi_{0i}$.

Now if it so happens that δ_i is suitably small for all i , then $\Delta G_{PL}^2 = G_{PL1}^2 - G_{PL0}^2$ will be small. Suppose we are in the very fortunate position where the parameter being investigated is uniformly irrelevant across all cases, so that, say, $0 \leq \delta_i \leq 0.005$ for all i . Then for all i , $-2 \log(1 - \delta_i) \leq 0.01$ and $\Delta G_{PL}^2 \leq 0.01N$.

This simple calculation suggests a guideline for determining the limits of ΔG_{PL}^2 wherein the two models might be considered equivalent. If we consider δ an acceptable level for the proportional change in predicted probabilities, and assume that the models are well behaved in the sense of the previous paragraph whereby the parameter is uniformly irrelevant, then for all i , $0 \leq \delta_i \leq \delta$, and a change in the deviance of $\Delta G_{PL}^2 = -2N \log(1 - \delta)$ can be considered a tolerable variation. Of course, the assumption of uniform irrelevance is unlikely to hold, but we still might utilize the resulting limit as a necessary condition for considering the two models equivalent.

In general, as the calculation of the limit is dependent on N , for a set δ the limit will increase with increasing N . For maximum likelihood models, this would be tantamount to decreasing the α -level for testing significance, which is in fact a sensible practice as N becomes large.

TABLE 1

Model estimates for Kapferer network

(NB: # indicates parameters whose absence does not change the PL deviance substantially)

Parameter	Estimate
<i>(a) Markov graph model entire network</i>	
(PL deviance = 434.9, MAR = .078)	
τ_{15}	- 4.38
τ_{14}	0.03#
τ_{13}	- 0.16
τ_{12}	0.40
τ_{11}	4.49
τ_{10}	- 0.89
τ_9	0.50
σ_3^I	- 0.04#
σ_3^O	0.03#
<i>(b) Markov graph R model with first 20 actors as respondents</i>	
(PL deviance = 164.2, MAR = .124)	
τ_{15}	- 6.36
τ_{14}	0.11#
τ_{13}	0.37
τ_{12}	1.10
τ_{11}	3.31
τ_{10}	- 1.44
τ_9	0.33
σ_3^I	- 0.01#
σ_3^O	- 0.27

TABLE 2

Model estimates for Kapferer network:
RN model with first 20 actors treated as respondents

Parameter		Estimate	
<i>RN model</i> (PL deviance = 267.4, MAR = .097)			
τ_{15}	- 6.74	τ_{11}	3.71
τ_{15N}	- 5.08	τ_{10}	- 1.40
τ_{14}	0.01#	τ_9	0.32
τ_{14N}	1.07	τ_{9N}	0.23#
τ_{13}	0.58	σ_3^I	- 0.08#
τ_{13N}	- 0.35	σ_{3N}^I	- 1.40
τ_{12}	0.79	σ_3^O	- 0.24
τ_{12N}	0.31	σ_{3N}^O	- 0.54
τ_{12NN}	1.57		

TABLE 3

Model estimates for Kapferer network:
RH model with first 20 actors treated as respondents

Parameter	Estimate
<i>RH model</i> (PL deviance = 302.3, MAR = .109)	
τ_{15H}	-4.74
τ_{14H}	0.12#
τ_{13H}	-0.12
τ_{12H}	0.58
τ_{11H}	3.68
τ_{10H}	-0.52
τ_{9H}	0.26
σ_{3H}^I	0.04#
σ_{3H}^O	-0.06

TABLE 4

Means and standard deviations of parameter estimates
for 15 Kapferer network *RH* models with nine non-respondents selected at random

Parameter	Mean	Standard deviation
τ_{15H}	-4.04	0.26
τ_{14H}	0.01	0.22
τ_{13H}	-0.17	0.08
τ_{12H}	0.34	0.09
τ_{11H}	4.41	0.50
τ_{10H}	-1.03	0.56
τ_{9H}	0.55	0.14
σ_{3H}^I	0.03	0.06
σ_{3H}^O	-0.03	0.02
PL Deviance	368.4	41.6
Mean absolute residual	0.084	0.012

TABLE 5

Models for work frequency network among respondents (*R* model)
 (NB: # indicates parameters that do not change the PL deviance substantially)

Parameter	Estimate
<i>Markov graph model</i> (PL deviance = 1175, MAR = .090)	
τ_{15}	-4.31
τ_{14}	0.20
τ_{13}	-0.14
τ_{12}	0.23
τ_{11}	2.11
τ_{10}	-0.41
τ_9	0.54
σ_3^I	-0.03#
σ_3^O	-0.02#

TABLE 6

Models for work frequency network among respondents
with non-respondent ties exogenous (R^+ model)

(NB: # indicates parameters that do not change the PL deviance substantially.)

Parameter	Estimate	Parameter	Estimate
<i>Markov graph model</i> (PL deviance = 1163, MAR = .090)			
τ_{15}	-4.50	τ_{11}	2.03
τ_{14}	0.23	τ_{10}	-0.39
τ_{13}	-0.16	τ_9	0.51
τ_{13N}	0.03#	τ_{9N}	0.16#
τ_{12}	0.22	σ_3^I	-0.03
τ_{12N}	0.03#	σ_3^O	-0.02#

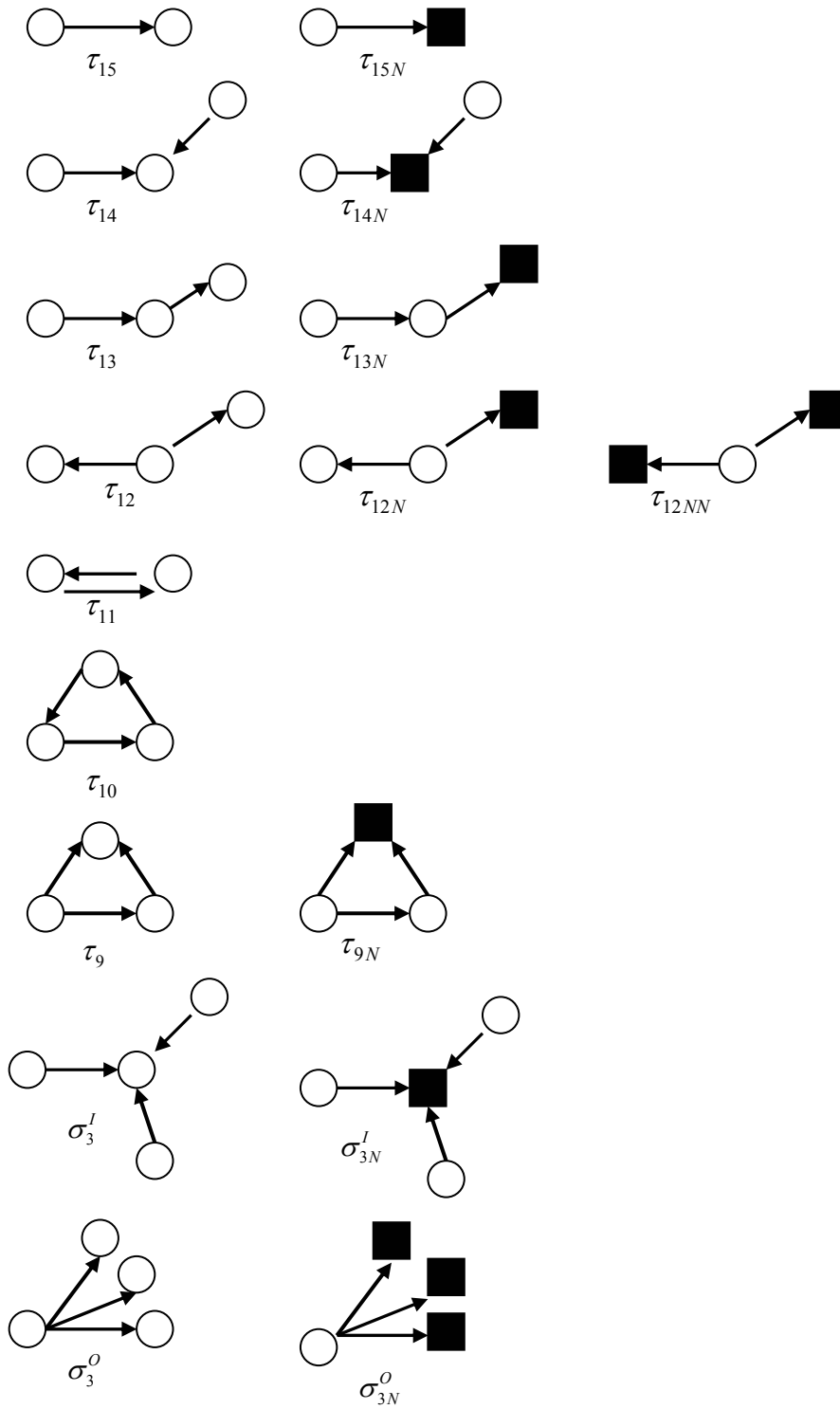


Figure 1
 Configurations for Markov graph model with non-respondents
 (Empty circles indicate respondents; black squares indicate non-respondents)