Journal of Mathematical Psychology 57 (2013) 284–296

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Conditional estimation of exponential random graph models from snowball sampling designs

Philippa E. Pattison [a,*], Garry L. Robins [a], Tom A.B. Snijders [b,c,d], Peng Wang [a]

[a] School of Psychological Sciences, University of Melbourne, Australia
[b] Department of Statistics, University of Oxford, United Kingdom
[c] Department of Politics, University of Oxford, United Kingdom
[d] Department of Sociology, University of Groningen, Netherlands

HIGHLIGHTS

- Snowball sampling designs for networks lead to partial observations on network ties.
- Exponential random graph models (ERGM) are a general class of models for networks.
- We propose a conditional estimation approach for ERGM parameters from a snowball sample.
- We demonstrate via simulation the effectiveness of the conditional estimation method.

ARTICLE INFO

Article history:
Available online 15 July 2013

Keywords:
Social networks
Exponential random graph models
Snowball sampling
Conditional Markov chain Monte Carlo
maximum likelihood estimation

ABSTRACT

A complete survey of a network in a large population may be prohibitively difficult and costly. So it is important to estimate models for networks using data from various network sampling designs, such as link-tracing designs. We focus here on snowball sampling designs, designs in which the members of an initial sample of network members are asked to nominate their network partners, their network partners are then traced and asked to nominate their network partners, and so on. We assume an exponential random graph model (ERGM) of a particular parametric form and outline a conditional maximum likelihood estimation procedure for obtaining estimates of ERGM parameters. This procedure is intended to complement the likelihood approach developed by Handcock and Gile (2010) by providing a practical means of estimation when the size of the complete network is unknown and/or the complete network is very large. We report the outcome of a simulation study with a known model designed to assess the impact of initial sample size, population size, and number of sampling waves on properties of the estimates. We conclude with a discussion of the potential applications and further developments of the approach.

1. Introduction

In many empirical network studies, information is obtained about the network contacts of every member of a circumscribed population of actors, leading to a census of the presence or absence of a network tie for every pair of actors in the population. Such an approach is characteristic, for example, when assessing a network in an organization of a moderate size, such as a firm or a school. The resulting network data may be represented in the form of a graph, and in many cases can be visualized and/or analysed in a way that helps to illuminate key structural properties of the network, such as the degree to which network partners of an actor are themselves connected (clustering), the distribution of the number of network partners across actors in the network (the degree distribution), and the presence of actors and network ties whose presence is critical to the connectivity of the network (cutpoints and bridges).

Over the past two decades, significant progress has been made in developing statistical models for such networks that can be estimated from this form of single census of all its ties. In these models, the presence or absence of a tie between any pair of actors in the network is regarded as a random variable, and the model assigns a probability to any possible outcome of the entire observation process. Particularly promising models of this form are *exponential random graph models* (ERGMs) (see, e.g., Frank & Strauss, 1986; Snijders, Pattison, Robins, & Handcock, 2006; Wasserman & Pattison, 1996). These models have been demonstrated in a variety of empirical contexts to be capable of capturing some of the key structural features of networks such as the degree of clustering,

* Corresponding author.
E-mail address: pepatt@unimelb.edu.au (P.E. Pattison).

0022-2496/$ – see front matter
http://dx.doi.org/10.1016/j.jmp.2013.05.004

the degree distribution, and features of network connectivity (see, e.g., Goodreau, 2007; Robins, Snijders, Wang, Handcock, & Pattison, 2007).

Increasingly, though, we are interested in modelling larger networks. This interest is largely driven by the recognition that many processes of interest, such as the diffusion of information or attitudes or the spread of a contact-borne disease, depend crucially on the structure of networks over potentially very large populations. If we are to understand the dynamics of these processes and their outcomes at the population level – such as the distribution of knowledge, attitudes, or disease states – we need to understand the structured nature of the relevant networks. In other words, when a census of the network is not possible, we require methods for estimating population-level network models from sampled observations on the network. This is the problem we address here.

The idea of using the network structure itself when sampling from a network to discover some of its structural properties is not new. There is a long history of studies in which personal networks or egonets, i.e., collections of other individuals linked to the respondents, are sampled from a network in order to throw light on individuals' immediate social environments within larger networks (see, e.g., Bott, 1957), and the well-known study by Milgram (1967) introduced a novel means of sampling paths in networks. The technique of snowball sampling was also devised to observe parts of a network that might be important to social processes of particular interest. Barton (2001) attributed an early interest in snowball sampling to Paul Lazarsfeld, who had found in studies of voting behaviour that individuals reported high levels of interpersonal influence from opinion leaders in their immediate social environment. Lazarsfeld recognized that random samples of voters provided an ineffective means of studying such processes:

> "The opinion leaders of a community could be best identified and studied by asking people to whom they turn for advice on the issue at hand and then investigating the interaction between the advisors and advisees. It is obvious that in a study involving a sample, like the present one, that procedure would be difficult if not impossible since few of the related leaders and 'followers' would happen to be included within the sample (Lazarsfeld, Berelson, & Gaudet, 1944, 49:50; cited in Barton (2001, 254–255))."

Barton noted a study conducted in 1943 and reported by Katz and Lazarsfeld (1955) and Merton (1957) as a very early instance of a snowball design.

Coleman (1958) is now recognized as introducing snowball sampling to the methodological literature. Coleman's work inspired Goodman's (1961) now classic analysis of snowball sampling which set out inferential procedures for determining the number of mutual ties and cycles in a network for various snowball sampling schemes and sampling assumptions.

More generally, the problem of understanding properties of a network from some sampling of network components is one that has received a small though steady stream of attention for a number of years. Frank (2005) and Handcock and Gile (2007, 2010) have summarized much of this work. An important body of work in the domain has used network-based sampling designs, including snowball sampling methods, to obtain population estimates of individual characteristics, often in difficult-to-find populations such as heroin addicts (Frank & Snijders, 1994), the homeless (Dávid & Snijders, 2002), or cocaine users (Bieleman, Diaz, Merlo, & Kaplan, 1993). In these snowball sampling designs, and in more general adaptive (Thompson & Collins, 2002; Thompson & Frank, 2000) or respondent-driven (Heckathorn, 1997) sampling designs, the goal is often to estimate characteristics of a population of individuals in a way that utilizes network links to trace members of a rare population but that also takes this sampling design into account in computing the required estimates.

Another strand of work has, like Goodman's (1961), attempted to identify characteristics of the network itself. As both Frank (2005) and Handcock and Gile (2010) have observed, this can be done in one of two primary ways. The first, a *design-based* approach, attempts to assess specific population-level network characteristics in a way that makes no particular assumptions about the form of the network. The second, a *model-based* approach, takes a particular model form as its starting point, often with the task of building population-level network models in mind.

As Handcock and Gile (2010) demonstrated, the capacity for progress within a design-based framework is inherently limited in many link-tracing designs, but they also established that the model-based approach can be used successfully to estimate ERGMs from a number of different link-tracing designs, including snowball sampling designs. In particular, they showed that likelihood-based inference could be based on evaluation of the full data likelihood by enumeration over all possible values for the unobserved data, or by Monte Carlo simulation. They demonstrated the implementation of this approach for two-wave samples from a 36-actor network, showing the low bias and high efficiency of the approach.

We also adopt a model-based approach here, and consider the problem of estimation in cases where Handcock and Gile's (2010) approach would be difficult to apply in practice because the number of nodes in the total network is large or unknown. We believe that our method has a practical relevance for parameter estimation of ERGMs in large networks, and a theoretical relevance because it elucidates a connection between specific dependence assumptions that define specifications of the ERGM and snowball sampling designs.

The paper is organized as follows. After describing snowball sampling, we present the general modelling framework within which this work is set, and describe the particular class of models that we assume. We then propose a conditional estimation approach that permits estimation of models within the class for particular snowball sampling designs. We present the results of several simulation studies designed to assess the effectiveness of the approach. Finally, we discuss the way in which the proposed scheme complements the likelihood-based method for sampled data proposed by Handcock and Gile (2010), and set out its potential application to different sampling processes and model classes. In doing so, we illustrate the connection between properties of the snowball sampling process and the form of the exponential random graph model assumed for the network.

2. Snowball sampling

As we have already observed, it is generally not feasible to observe the state of every potential tie in the network when the node set of the network is large. Snowball sampling was devised to sample partial network data in the vicinity of one or more selected nodes, and it can be described using the concepts of zones and neighbourhoods in networks. We begin with some notation and essential definitions.

We suppose that $N = \{1, 2, \ldots, n\}$ is a fixed set of nodes, and we let $Y(i, j)$ denote a nondirected *tie variable* for the pair of nodes i and j, with $i, j \in N$. The variable $Y(i, j)$ takes the value 1 if there is a tie between node i and node j; $Y(i, j) = 0$, otherwise. We denote by $Y = [Y(i, j)]$ the $n \times n$ array of tie variables, and by $y(i, j)$ and y a realization of $Y(i, j)$ and Y, respectively. For the main part of the paper, we assume nondirected tie variables (that is, we do not distinguish $Y(i, j)$ and $Y(j, i)$), but we touch briefly on the case of directed networks in the final section. By $N^{(2)}$ we denote the collection of unordered pairs $\{\{i, j\} | i, j \in N, i \neq j\}$ that label the tie variables.

For a network y on the node set N, consider a particular node, termed a *seed* node, $s \in N$. Define a *path* of *length* k from actor s to

Table 1
Adjacency matrix of the network y with rows and columns ordered according to zones of some seed set A.

	Z_0	Z_1	Z_2	…	Z_k	Z_{k+1}	Z_{k+2}	…
Z_0	Y_{00}	Y_{01}	0	…	0	0	0	0
Z_1	Y_{10}	Y_{11}	Y_{12}	…	0	0	0	0
Z_2	0	Y_{21}	Y_{22}	…	0	0	0	0
…	…	…	…	…	…	…	…	…
Z_k	0	0	0	…	Y_{kk}	$Y_{k,k+1}$	0	0
Z_{k+1}	0	0	0	…	$Y_{k+1,k}$	$Y_{k+1,k+1}$	$Y_{k+1,k+2}$	…
Z_{k+2}	0	0	0	…	0	$Y_{k+2,k+1}$	$Y_{k+2,k+2}$	…
…	0	0	0	…	0	…	…	…

another actor t to be a sequence $s = s_0, s_1, \ldots, s_k = t$ of distinct nodes in N such that $y(s_{m-1}, s_m) = 1$ for $m = 1, 2, \ldots, k$. A path from node s to node t is a *geodesic* if it is path of minimum length; the length of such a path is defined to be the *geodesic distance* d_{st} between s and t. Define $Z_k(s)$, the *zone of order k of node s in the network y*, to be the set of nodes at geodesic distance k from s in y. We also define $Z_k(A)$, the *zone of order k of a set A of nodes in the network y* as follows. Let $Z_0(A) = A$, and define $Z_k(A) = \{t : t \in Z_k(s)$ for some $s \in A$ and $t \notin Z_h(u)$ for any $h < k$ and $u \in A\}$. The zone of order k of the set A therefore comprises those nodes whose minimum distance to any node in A is k. Where dependence of a zone on the seed set A is clear, we drop reference to A in the notation and refer simply to Z_k as the zone of order k. The subarray of tie variables between members of Z_k is denoted by y_{kk}, and the subarray of tie variables from Z_k to Z_l by y_{kl}. We also use $y_{0+1+\cdots+k,0+1+\cdots+k}$ to refer to the subarray of tie variables in y between members of $Z_0 \cup Z_1 \cup \cdots \cup Z_k$.

The *k-wave snowball sample* with *seed set A* can be described as follows. Assume that the set A of seed nodes is given. In the initial wave (wave 0) of snowball sampling, we identify the presence or absence of ties among members of $Z_0 = A$; that is, we identify values in the subarray y_{00}. We also identify in the initial wave the set Z_1 of network partners of members of A who are not included in A itself. The presence or absence of ties from members of Z_0 to members of Z_1 is therefore also identified; that is, the subarray y_{01} is observed. In the next wave of sampling, we observe values in the subarray y_{11}; that is, we observe ties among members of Z_1, and we also identify the set Z_2 of nodes who are tied to one or more node in Z_1 but not to any node in Z_0. In addition, we identify the subarray y_{12}, and hence all ties from nodes in Z_1 to nodes in Z_2. More generally, in wave k, we identify the network partners of nodes in Z_k that have not already been included in Z_h for any $h < k$, and hence the members of Z_{k+1}. We also observe the subarray y_{kk} corresponding to tie variables between nodes in Z_k and the subarray $y_{k,k+1}$ referring to tie variables between nodes in Z_k and nodes in Z_{k+1}. By construction, the ties in y_{hm} for $|h-m| \geq 2$ all have values of 0. The general structure of the adjacency matrix of the network y with rows and columns organized according to the zones of A is shown in Table 1.

For the discussion below, we suppose that we have first obtained either a one-wave or higher-wave snowball sample based on some seed set A. We assume that each party to a tie reports it accurately, and hence that $y(i, j)$ has the same value whether reported by i or j. It is evident that, in general, a k-wave snowball sample leads to the observation of a larger and more complete portion of the network in the vicinity of the seed nodes than a $(k-1)$-wave sample. Snowball sampling can be contrasted with other network sampling methods in this respect, such as *random sampling of egonets* (randomly sampled nodes and their associated ties) or a *random walk* on a network, namely, a sequence of randomly selected edges, in which a partner of a seed node is identified at random, a partner of that partner is identified at random, and so on. In both of these latter cases, information about the network structure in the vicinity of the seed nodes is only partial. As we see below, it is the local 'completeness' of the data observed through the snowball sampling process that supports the conditional estimation procedure that we develop.

3. Exponential random graph models

A random graph model for Y assigns a probability to each possible realization y of Y. For a nondirected network on n nodes, the set Ω of possible realizations of Y contains $2^{n(n-1)/2}$ distinct graphs, since each of the $n(n-1)/2$ pairs of nodes may or may not be linked by an edge.

Building on Frank and Strauss (1986), Holland and Leinhardt (1981) observed that a general framework for the development of probability models for graphs could be developed by considering the collection of tie variables as an interactive system of variables (Besag, 1974). Their insight led to the development of models for networks that avoided the need to assume independence of dyads, an assumption characteristic of earlier approaches to statistical network modelling (see, e.g., Holland & Leinhardt, 1981). Frank and Strauss proposed a *Markov dependence* assumption for network tie variables in which two network tie variables are assumed to be conditionally independent, given the values of all other network tie variables, unless they have a node in common.

In their most general form (Frank & Strauss, 1986), exponential random graph models are probability distributions of the form

$$\Pr(Y = y) = \exp(\Sigma_P \theta_P z_P(y))/\kappa(\theta), \tag{1}$$

where

- P is a variable ranging over a collection of subsets of $N^{(2)}$, referring to possible ties (and each value of P thus defines a potential network configuration);
- θ_P is a parameter associated with the configuration P and is nonzero if and only if every pair of variables in the subset P are dependent conditionally on all tie variables (i, j) for (i, j) outside P;
- $z_P(y) = \Pi_{Y(i,j)\in P} y(i, j)$ is the network statistic corresponding to the configuration P, and indicates whether or not all ties in the configuration are present in the network y; and
- $\kappa(\theta)$ is a normalizing quantity.

To reduce the number of model parameters, Frank and Strauss (1986) introduced a homogeneity constraint that $\theta_P = \theta_{P'}$ whenever P and P' are isomorphic configurations (that is, whenever there is a one-to-one mapping ϕ on N so that, for all $Y(i, j)$ in the set P, $Y(i, j) \in P$ if and only if $Y(\phi(i), \phi(j)) \in P'$). With this constraint, there is a single parameter $\theta_{[P]}$ for each class $[P]$ of isomorphic configurations. The statistic in the model corresponding to the class $[P]$ is then

$$z_{[P]}(y) = \Sigma_{P\in[P]} \Pi_{Y(i,j)\in P} y(i, j), \tag{2}$$

that is, a count of all observed configurations in the graph y that are isomorphic to the configuration corresponding to P. Such homogeneous models can be extended by dependence on covariates.

A fundamental example is the case of a homogeneous Markov random graph, i.e., a random graph satisfying the Markov dependence condition as defined by Frank and Strauss (1986). These authors proved that the model then takes the form

$$\Pr(Y = y) = \exp(\theta L(y) + \Sigma_k \sigma_k S_k(y) + \tau T(y))/\kappa,$$

where $L(y)$, $S_k(y)$, and $T(y)$ are the number of edges, k-stars $(2 \leq k \leq n-1)$, and triangles in the network y, and θ, $\sigma_k (2 \leq k \leq n-1)$, and τ are corresponding parameters. An *edge* in the network is simply a subgraph comprising two connected nodes, that is, a pair $\{i, j\}$ of nodes for which $y(i, j) = 1$. A *k-star* is a subgraph comprising a node i and distinct nodes $m_1, m_2, \ldots, m_k$, to each

of which i is connected by a tie; that is, $y(i, m_h) = 1$ for $h = 1, 2, \ldots, k$. A *triangle* is a completely connected subgraph of three nodes, that is, a triple $\{i, j, k\}$ of nodes for which $y(i, j) = 1 = y(i, k) = y(j, k)$.

The Markov model has rarely provided a good fit to empirically observed social networks except in cases where the number of nodes is low (e.g., less than 20), or the average number of edges per node is very low (e.g., less than 2), To achieve better fitting models, Snijders et al. (2006; see also Hunter & Handcock, 2006) proposed an alternative *realization-dependent* (Baddeley & Möller, 1989) model form. Following Pattison and Robins (2002), they argued that conditional dependencies may emerge from the network processes themselves, with new dependencies created as network ties are generated. In particular, they assumed that, in addition to Markov dependencies, two network ties $Y(i, j)$ and $Y(k, l)$ might be conditionally dependent in the case that there is an observed tie between, say, j and k and between l and i, that is, if the presence of a tie from i to j and from k to l would create a 4-cycle in the graph.

Snijders et al. (2006) showed that the consequence of this assumption was a set of additionally permitted nonzero parameters in an exponential random graph model, including those referring to collections of 2-paths with common starting and ending nodes, and collections of triangles with a common base. These configurations have been very useful in providing model specifications that achieve a good model fit for many empirical data sets, when implemented in the following way. We define a *k-2-path* to be a subgraph comprising two nodes, i and j, and a set of k paths of length 2 through distinct intermediate nodes $m_1, m_2, \ldots, m_k$. A *k-triangle* is a subgraph comprising two *connected* nodes, i and j, and a set of k paths of length 2 through distinct intermediate nodes $m_1, m_2, \ldots, m_k$.

Following Snijders et al. (2006), if we let σ_k, ν_k, and τ_k be the model parameters associated with a k-star, a k-2-path, and a k-triangle, respectively, then we can entertain assumptions about relationships among related parameters, such as

$$\sigma_{k+1} = -\sigma_k/\lambda \ (k \geq 2), \quad \text{and}$$
$$\nu_{k+1} = -\nu_k/\lambda \quad \text{and} \quad \tau_{k+1} = -\tau_k/\lambda \ (k \geq 1),$$

for $\lambda \geq 1$ a (fixed) constant. This is just a hypothesis, and its adequacy needs to be assessed empirically. Under these assumption, the statistics

$$S^{[\lambda]}(y) = \Sigma_k(-1)^k S_k(y)/\lambda^{k\text{-}2},$$
$$U^{[\lambda]}(y) = \Sigma_k(-1)^k U_k(y)/\lambda^{k\text{-}2},$$

and

$$T^{[\lambda]}(y) = \Sigma_k(-1)^k T_k(y)/\lambda^{k\text{-}2},$$

where $S_k(y)$, $U_k(y)$, and $T_k(y)$ are, respectively, the number of k-stars, k-2-paths, and k-triangles in the network y, become single statistics associated with the parameters σ_2, ν_1, and τ_1. These statistics may be termed respectively the *alternating star*, the *alternating 2-path statistic*, and the *alternating triangle statistic*. It should be noted that the value of λ need not be the same for each statistic.

Hunter (2007) and Hunter and Handcock (2006) referred to the latter two statistics $U^{[\lambda]}(y)$ and $T^{[\lambda]}(y)$ as GWDSP (*geometrically weighted dyad-wise shared partners*) and GWESP (*geometrically weighted edge-wise shared partners*), respectively. The first is equivalent to a geometrically weighted aggregate over k of the number of dyads with exactly k shared partners and the second to a geometrically weighted aggregate over k of *connected* dyads with exactly k shared partners.

Robins et al. (2007) coined the term *social circuit dependence* to characterize those ERGMs (1) whose statistics $z_p(y)$ are of the form (2) with each P corresponding to a subgraph with the property that any pair of nonadjacent edges in P is part of a 4-cycle. In other words, for any P with a nonzero parameter, if $\{i, j\}, \{t, s\} \in P$ for four distinct vertices i, j, s, t, then also either $\{i, t\}, \{j, s\} \in P$ or $\{i, s\}, \{j, t\} \in P$. The definitions of k-triangles and k-2-paths clearly satisfy this definition, as well as many other configurations.

It is worth reiterating that adoption of a particular ERGM form assumes not only that the number of nodes in the network is fixed, but also that the ERGM model specification is the correct one: the analysis is model based. This means assuming that the network Y is an outcome of a probability model of the ERGM form (1),

$$\Pr(Y = y) = \exp(\Sigma_P \theta_P z_P(y))/\kappa\ (\theta),$$

for which the functions z_P (y) are such that social circuit dependence holds. As noted above, this set of assumptions has proved realistic in a number of social network modelling scenarios (see, e.g., Goodreau, 2007; Robins et al., 2007; Snijders et al., 2006) and, as we discuss in the final section of the paper, it is possible to adapt the procedure developed here to other model specifications. Nonetheless, the selection of an appropriate model specification remains a difficult problem within the ERGM framework (see, e.g., Pattison & Snijders, 2013).

ERGM model parameters may be interpreted by observing that, if a parameter has a large positive (or negative) value, then graphs with high (or low) values of the corresponding statistics have higher (or lower) probability, net of other effects. Nonetheless, it is important to observe that there are complex constraints among the values of the statistics. The model is well understood in the case where only the edge parameter θ is nonzero, this being the well-known Erdös and Rényi (1959) model with uniform tie probability $p = \exp(\theta)/[1 + \exp(\theta)]$. However, a dyad-dependent instantiation of model (1) can be seen as a model for a self-organizing network process, one that is often characterized by highly nonlinear relationships between model parameters and network properties (Handcock, 2003; Robins, Pattison, & Woolcock, 2005).

In many settings, primary interest lies in estimating the parameters of exponential random graph models from observed network data. Early attempts to apply Markov Chain Monte Carlo maximum likelihood estimation (MCMCMLE) approaches were not always successful, because the properties of the models under consideration were not always fully appreciated, as Snijders (2002) and Handcock (2004) demonstrated. However, with a growing understanding of model properties, and more careful attention to model adequacy, MCMCMLE approaches have been successfully implemented (Hunter & Handcock, 2006; Robins et al., 2007; Snijders et al., 2006).

4. A conditional estimation strategy

This paper presents a new conditional estimation strategy for the exponential random graph models just described in the case of multi-wave snowball samples. The one-wave sample approach is presented first because of its conceptual simplicity, but its practical application may be restricted to cases in which the initial seed set comprises a relatively large sample of nodes in N (and, in this case, the method of Handcock and Gile (2010) is likely to be preferred as long as the size of N is known). The approach for two-wave or higher-wave samples is intended to be more practical and more useful for the estimation of large networks. In both cases, we assume a model of the social circuit form described above, and we write the model as

$$\Pr(Y = y) = (1/c)\exp(\Sigma_p \theta_p z_p(y)). \tag{3}$$

We also assume that the seed set A and the graph Y are mutually independent; for example, the seed set may be chosen in advance, without knowledge of Y.

4.1. Conditional estimation based on a one-wave sample

For the one-wave snowball sample design, the conditional estimation conditions on the composition of the seed set Z_0 and the

first-order zone Z_1, as well as the ties between zones Z_0 and Z_1 and those within zone Z_1, and uses as outcomes only the array of tie variables y_{00} within zone Z_0. Conditionality on the node set Z_1 is equivalent to requiring that for all nodes i in Z_1 there is at least one connecting tie to i from some node in Z_0. We let Y_{00} refer to the set of variables for which y_{00} is a realization and Y_{00}^c refer to all variables in Y except those in Y_{00}; y_{00}^c refers to the realization y omitting the values in y_{00}. In this notation, the conditional estimation strategy is based on the probability model $\Pr(Y_{00} = y_{00} \mid Y_{00}^c = y_{00}^c)$, in which y_{00}^c is considered fixed. Note that assumption (1) and the fact that the subsets y_{00} and y_{00}^c of the outcome y do not overlap imply that these conditional probabilities are always positive. Conditioning here does not reflect any assumption about these tie variables being fixed by design, but is merely an estimation device.

Using the outcome $Y_{00} = 0$ as an arbitrary point of reference, we can write

$$\log(\Pr(Y_{00} = y_{00} \mid Y_{00}^c = y_{00}^c)/\Pr(Y_{00} = 0 \mid Y_{00}^c = y_{00}^c)) = \Sigma_p \theta_p [z_p(y) - z_p(y^{00})],$$

where y^{00} is equal to y but with all entries in y_{00} set to 0. Hence

$$\log(\Pr(Y_{00} = y_{00} \mid Y_{00}^c = y_{00}^c)) = C + \Sigma_p \theta_p z_p(y),$$

where $C = \log(\Pr(Y_{00} = 0 \mid Y_{00}^c = y_{00}^c)) - \Sigma_p \theta_p z_p(y^{00})$, which does not depend on y_{00}.

The conditional estimation strategy exploits the characterization of the social circuit model given above. This characterization implies for the one-wave snowball sample that, if $i,\ j \in Z_0$ and the variables $Y(i, j)$ and $Y(k, l)$ are conditionally dependent, where k and l are distinct from i and j, then either $y(i, k) = 1 = y(j, l)$ or $y(i, l) = 1 = y(j, k)$, and, in either case, $k,\ l \in Z_0 \cup Z_1$. Hence, if $Y(i, j)$ is a variable in Y_{00}, then $Y(i, j)$ is conditionally independent of any variable that is not in Y_{00}, Y_{01}, or Y_{11}, so that any nonzero statistics $z_p(y)$ depending on values in y_{00} involve only configurations on some subset of $Z_0 \cup Z_1$. Ties outside $Z_0 \cup Z_1$ therefore cannot contribute to $z_p(y)$. This proves the following result.

Proposition 1. *For a one-wave snowball sample from an ERGM satisfying social circuit dependence, with seed set Z_0 and first-order zone Z_1, denoting by Y_{00} the set of tie variables among nodes in Z_0 and by $Y_{[1,1]}$ the set of tie variables among nodes in $Z_0 \cup Z_1$, it holds that*

$$\Pr(Y_{00} = y_{00} \mid Y_{00}^c = y_{00}^c, Z_0, Z_1) = (1/C') \exp(\Sigma_p \theta_p z_p(y_{[1,1]})) \quad (4)$$

for a constant C' independent of y_{00}, for outcomes y_{00} that yield Z_1 as the first-order zone when starting with seed set Z_0, whereas this probability is 0 for all other values of y_{00}.

The proposition implies that this conditional distribution does not depend on all of y_{00}^c but only on that part of y_{00}^c which is observable from a one-wave snowball sample. The conditioning is on y_{00}^c, and therefore also on the choice of the initial set of nodes Z_0. It follows that the parameters of the conditional probability model for $\Pr(Y_{00} = y_{00} \mid Y_{00}^c = y_{00}^c)$ can be estimated from the observed data on $y_{[1,1]}$ without needing any further information about the rest of the network, such as the total number of nodes. Note that, depending on y_{00}^c, it is in theory possible that (4) is insensitive to some parameters in θ, and these parameters are then unidentified under this conditional probability model. If the seed set Z_0 is sufficiently large, however, this has a vanishingly small probability.

The essential insight underlying the procedure is that a conditional probability model for the collection Y_{00} of ties among nodes in Z_0 has the same parameters as the ERGM for the network as a whole, but can be estimated conditionally from the one-wave snowball sample with seed set Z_0. This result is based on the observation that a potential tie Y_{ij} linking two nodes i and j in Z_0 may depend, given the assumed model, on any tie between node i or node j and another partner k (in which case, k is in N_1) or on any tie Y_{kl} for nodes k and l, each of whom is directly connected to either i or j (and therefore both k and l are in Z_1). Either way, the relevant observations are included in the one-wave snowball sample with seed set Z_0, and each Y_{ij} in Y_{00} depends only on potential ties whose values have been observed, and not on any unobserved ties.

Estimates of θ obtained as a result of MCMC maximization of (4) may be called conditional Markov chain Monte Carlo maximum likelihood estimates (CMCMCMLEs). The estimation proceeds as a modification of the MCMCMLE of the network $y_{[1,1]}$. Estimation algorithms for ERGMs (see Handcock, Hunter, Butts, Goodreau, & Morris, 2008, Snijders, 2002) have two components: simulation of the ERGM for a known parameter vector by Gibbs sampling or a Metropolis–Hastings algorithm, and inserting the results of such simulations in a procedure for solving the likelihood equation. For the CMCMCMLE, the second component is identical to the procedures for the MCMCMLE discussed in the literature (the Robbins–Monro algorithm as in Snijders (2002), or the Geyer–Thompson algorithm as in Handcock et al. (2008)). The first component, however, is different: in the Gibbs sampler or Metropolis–Hastings procedure, only variables in y_{00} are allowed to change in the simulations, that is, we treat all other ties in $y_{[1,1]}$ as fixed; and the additional condition is posed that a variable in y_{00} is allowed to change only if the resulting new value of y_{00} still yields Z_1 as the first-order zone for a snowball sample starting with seed set Z_0. Thus, for example, Gibbs sampling to simulate the ERGM for a known parameter vector proceeds as follows.

1. Sample randomly a pair $(i,\ j)$ for which $Y(i, j)$ belongs to y_{00}.
2. If replacing the value $y(i, j)$ by $1 - y(i, j)$ would imply that Z_1 is no longer the first-order zone for the one-wave snowball sample with seed set Z_0, then go to step 1.
3. Sample randomly a value for $Y(i, j)$ from its conditional distribution of $Y(i, j)$ given all other elements of $Y_{[1,1]}$, as implied by distribution (4).
4. Replace $Y(i, j)$ by this value.
5. Go to step 1.

One important consideration is the existence of conditional MLEs for any observed subgraph $y_{0+1,0+1}$. Handcock (2003) has articulated the circumstances in which MLEs may not exist in the ERGM context, and has also set out various circumstances in which MCMCMLEs may not exist even though MLEs do. In practice, when snowball samples are small and hence configurations corresponding to model statistics may be rare, conditional MCMCMLEs may be impossible or difficult to obtain because of the nonexistence of the MLE or the MCMCMLE, or outcomes close to these situations.

The approach just described makes less efficient use of the data than the method of Handcock and Gile (2010) for estimation of ERGM parameters in the presence of missing data when estimation over all of Y is feasible. In that case, it is not recommended. A difference between the two approaches is that in the conditional estimation approach we do not need to know n, the size of the node set N, whereas n is clearly required for the use of Handcock and Gile's (2010) approach. Our approach is computationally simpler in the case that the number of unobserved nodes is known to be large, because in the Handcock–Gile approach all tie variables between those nodes need to be simulated.

When a saturated snowball sample is observed – i.e., starting from some initial node set A, subsequent snowball waves are observed until no new nodes are found – a more efficient estimation method is the procedure suggested by Snijders (2010, discussion of Corollary 3). Under the assumption that the ERGM satisfies the condition of so-called component independence

(weaker than social circuit dependence), parameters for such an incompletely observed network can be estimated by an MCMCMLE where in the simulation only those changes are allowed that would yield the observed node set as the node set for a saturated snowball sample with seed set A, and the number of unobserved nodes does not need to be known.

The conditional estimation method just presented can be regarded as analogous to Besag's (1974) coding scheme approach for the Ising model. To illustrate the coding scheme approach, consider a set of binary variables associated with the points of intersection of a regular two-dimensional grid, with any pair of variables which are immediate neighbours in the grid regarded as conditionally dependent. Without loss of generality, each variable can be identified with a pair (m, h) of integers, and its immediate neighbours are $(m-1, h)$, $(m, h-1)$, $(m+1, h)$, and $(m, h+1)$. In the coding scheme approach, all variables with an even sum of integers are assigned to one class, and all variables with an odd sum of integers to the other. The partition of variables into classes is then such that all variables within one class are conditionally independent of one another given the observations in the other class, and hence a model can be estimated readily for the observations in one class conditional on the values of the variables in the other class.

While the approach just presented does not isolate a set of variables that are conditionally independent given values of variables in a second set, it does isolate a set of variables that are conditionally independent of the variables for which there are no observations in the sample. Just as for the coding scheme approach, the conditional MLE is less efficient that the MLE computed for the entire set of variables, but, as we argue below, there may be circumstances in which the latter is very difficult to obtain, and the conditional MLE is therefore a valuable substitute.

4.2. Conditional estimation based on a two-wave or higher-wave sample

In the case of a two-wave or higher-wave sample, we can proceed similarly. For the two-wave case, we obtain conditional estimates for model parameters from a model for network ties within $Z_0 \cup Z_1$ that is conditional on Y_{12} and Y_{22} as well as the composition of the node sets Z_0, Z_1 and Z_2. More generally, we obtain conditional estimates for a model for ties within $Z_0 \cup Z_1 \cup \cdots \cup Z_{k-1}$, conditional on $Y_{k-1,k}$ and Y_{kk} as well as the composition of the node sets $Z_0, Z_1, \ldots, Z_k$. Conditionality on the node sets $Z_0, Z_1, \ldots, Z_k$ is equivalent to the fact that, for each $h = 1, \ldots, k$, for all nodes i in Z_h there is at least one connecting tie to i from some node in Z_{h-1} and there are no ties to i from any nodes in Z_m for $m < h-1$. As a result, $Y_{mh} = 0$ for all $m < h-1$, and all arrays of the form $Y_{h,h+1}$ satisfy the condition that every column contains at least one unit entry. For the two-wave sample, this means that $Y_{02} = 0$, and in Y_{01} and Y_{12} every column must contain at least one unit entry.

Denote $N_{[k-1]} = Z_0 \cup Z_1 \cup \cdots \cup Z_{k-1}$, and let $Y_{[k-1,k-1]}$ be the block in the adjacency matrix corresponding to node set $N_{[k-1]}$. It may be noted that, if the seed set were $N_{[k-1]}$, the zone of order 1 would be Z_k. Therefore applying the reasoning above to $N_{[k-1]}$ and Z_k leads to the conclusion that

$$\log(\Pr(Y_{[k-1,k-1]} = y_{[k-1,k-1]} \mid Y^c_{[k-1,k-1]} = y^c_{[k-1,k-1]})) = C + \Sigma_p \theta_p z_p(y)$$

for some C which does not depend on $y_{[k-1,k-1]}$ and where ties outside $Z_{[k]}$ do not contribute to $z_p(y)$. This proves the following result.

Proposition 2. *For a $(k+1)$-wave snowball sample from an ERGM satisfying social circuit dependence,*

$$\log(\Pr(Y_{[k,k]} = y_{[k,k]} \mid Z_0, Z_1, \ldots, Z_{k+1}, Y^c_{[k,k]} = y^c_{[k,k]})) = C + \Sigma_p \theta_p z_p(y_{[k+1,k+1]}) \quad (5)$$

for a constant C independent of $y_{[k,k]}$, where $Y_{[h,h]}$ denotes the set of tie variables among nodes in $Z_0 \cup Z_1 \cup \cdots Z_h$, for outcomes $y_{[k,k]}$ that for $h = 1, \ldots, k+1$ yield Z_h as the zone of order h when starting with seed set Z_0, whereas this probability is 0 for all other values of $y_{[kk]}$.

This proposition allows an MCMC procedure for simulating and estimating the ERGM from a $(k+1)$-wave snowball sample, conditional on the last wave. The algorithm has exactly the form described earlier: in the Gibbs or Metropolis–Hastings procedure to simulate the conditional distribution (5) for a given parameter value θ, the values of elements $Y(i, j)$ of $Y_{[k,k]}$ are sampled from their conditional distribution implied by (5), but proposed moves need to respect the conditioning on the structure defined by the zones $Z_0, Z_1, \ldots, Z_{k+1}$. As discussed above, this is equivalent to $Y_{mh} = 0$ for all $m < h-1$ and all columns in the arrays $Y_{h,h+1}$ containing at least one element equal to 1. It is easy to see that the collection of changes in $Y_{[k,k]}$ in which moves that violate these requirements are forbidden defines a walk on the outcome space of $Y_{[k,k]}$ conditional on $Z_0, Z_1, \ldots, Z_{k+1}$ which communicates with every other state.

4.3. Snowball sampling with a well-separated seed set or well-separated classes

In many empirical instances, the set A will be a random sample of nodes in N, as Goodman (1961) originally assumed. If the set A happens to satisfy a particular separation condition, then the estimation may be based on computations conducted in parallel, and it can potentially be carried out with considerable efficiency. We say that a partition $A = A_1 \cup A_2 \cup \cdots \cup A_s$ of the seed set A of a two-wave snowball sample into classes $\{A_1, A_2, \ldots, A_s\}$ is *well separated* for models satisfying social circuit dependence if the geodesic distance $d_{st} \geq 3$ in y for every pair of nodes $s,\ t$ from distinct classes in the partition. In this case, $\cap j\{Z_0(A_j) \cup Z_1(A_j)\} = \emptyset$, and the computation of change statistics across subarrays corresponding to $Z_0(A_j) \cup Z_1(A_j)$ may be carried out in parallel. For a k-wave snowball sample, we would consider a partition whose classes satisfy $d_{st} \geq 2k-1$ in y for every pair of nodes $s,\ t$ from distinct classes in the partition.

Indeed, one potential application of the conditional estimation approach we have described is to the development of fast approximate estimation methods in the case of very large networks; we discuss this application further below.

4.4. Assessment of the ERGM homogeneity assumption

Another potentially useful application of the method just described is to assess the plausibility of the assumption of homogeneity made in many practical applications of ERGMs. By homogeneity we mean here that parameters are constant between subsets of nodes even if these are separated by large geodesic distances. The assumption of homogeneity of ERGM effects is a strong one, particularly in large networks. A well-separated partition of some random seed set may be used to obtain multiple independent estimates of the model parameters, and these estimates may be assessed for homogeneity in the manner of Lubbers (2003).

5. Simulation studies

In this section, we present the results of three simulation studies designed to assess the effectiveness of the proposed conditional estimation strategy and to evaluate some of its properties. In each case, we began with a known model, that is, with a fixed number n of nodes and a fixed parameter vector θ. In all of the simulations reported below, the same true model is assumed, with edge, alternating star, alternating triangle, and alternating 2-path parameters equal to -4.0, 0.2, 1.0 and -0.2, respectively, with $\lambda = 2$. These parameter values are in line with estimates obtained from empirical data sets. We set the size of the network of interest to be either

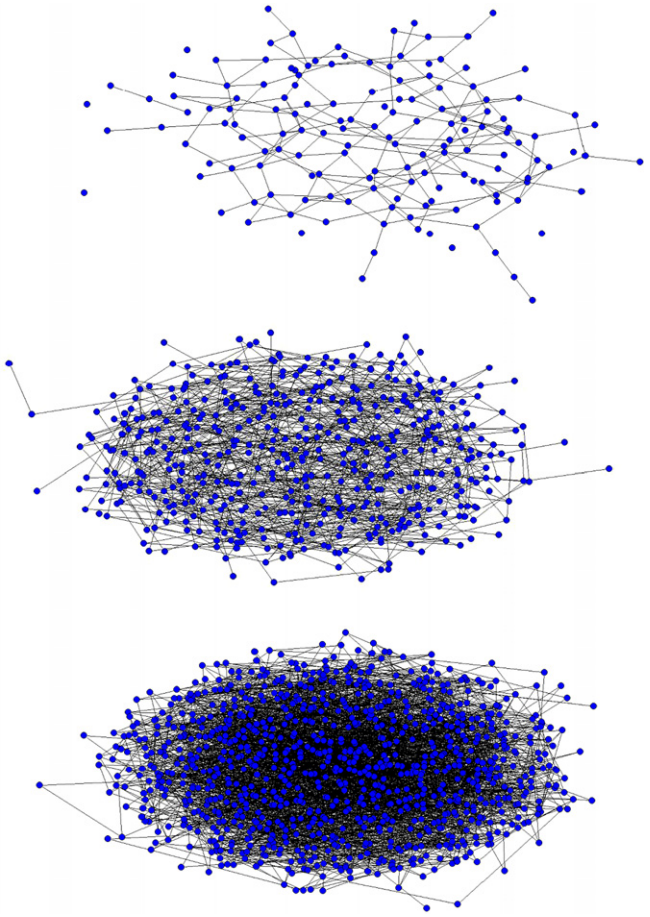

Fig. 1. Illustrative graphs drawn from exponential random graph distributions with parameters fixed at −4 (edge), 0.2 (alternating star), 1.0 (alternating triangle), and −0.2 (alternating 2-path) for networks of size 150, 500, and 1000.

150, 500, or 1000 nodes. The resulting network distributions are relatively sparse, but increasing in average degree as a function of n. The average degree is 2.8 for the 150-node distribution, 4.8 for the 500-node distribution, and 6.0 for the 1000-node distribution. The level of clustering is quite high: on average, networks in these distributions have approximately 10 times the number of triangles expected in a random graph distribution of the same density, and approximately 100 times the number of 2-triangles. A single graph sampled from each graph distribution is presented in Fig. 1.

For each of the three network distributions ($n = 150$, 500, or 1000), we simulated the random graph distribution with parameter vector $\theta = (-4.0, 0.2, 1.0, -0.2)$ and sampled a specified number of networks from each distribution. The samples were obtained by running one long MCMC simulation, sampling every 100,000th graph after a burn-in of 1,000,000 steps. A single seed set of each predetermined fixed size ($a = 10, 20, 30, 40, 50, 60$) was then selected at random from each of the sampled graphs, the corresponding snowball sample was identified, and the conditional estimation procedure described earlier was then conducted. For the cases of $n = 150$ and $n = 500$, we also obtained MCMCMLEs for the full graph for comparative purposes.

In the first simulation study, we systematically varied both n and the size $a = |A|$ of the seed set in order to assess the impact of these factors on bias and variability in the conditional estimates obtained from two-wave snowball samples.

In the second simulation study, still for a two-wave snowball sample, we focused on estimates for the case of $n = 500$ and seed sets of size 10, and evaluated variability in the conditional estimates using 100 random seed sets of size 10 from each of 100 sampled graphs. (These 100 sampled graphs were independent of the 500 samples used in the first study.) The purpose of this second study was to assess the relative contribution to variability in the conditional estimates of (a) the sampling of complete graphs from the ERGM distribution, and (b) snowball sampling within the sampled complete graph.

Table 2
Convergence rates for simulation study 1 (dashes appear for conditions not utilized in the study).

Convergence rate			
Seed set size	$n = 150$	$n = 500$	$n = 1000$
3	0.762	0.852	0.764
5	0.914	0.976	0.940
7	0.986	0.992	0.966
9	0.998	0.998	0.984
11	1.000	1.000	0.984
13	1.000	1.000	0.998
15	1.000	1.000	0.998
17	–	1.000	–
18	–	1.000	–
19	–	1.000	–
20	–	1.000	1.000
30	–	1.000	–
40	–	1.000	–
50	–	1.000	–

In the third simulation study, we assessed the performance of the conditional estimation approach for a one-wave snowball sample for a new set of 500 networks of size 150, 500, and 1000 sampled from the same ERGM, and for varying seed set sizes. We also compared these estimates to MCMCMLEs for the entire graph in the case of 150 nodes.

In all cases, simulations and estimations were carried out using the program *PNet* (Wang, Robins, & Pattison, 2009), which implements the MCMCMLE using the Robbins–Monro algorithm presented in Snijders (2002) but respects the snowball structure as outlined in Section 4 above. Provided that there was evidence of convergence of the estimation procedure, that is, provided that the convergence t-value for each of the statistics that correspond to parameters in the model was less than a predetermined value, set here at 0.1, we retained the estimated parameter values for the summary statistics described below. For each parameter in the model, we calculated various summary measures characterizing the distribution of estimates. Bias was estimated as the difference between the average of the converged estimates and the true parameter value; the root mean square error (RMSE) was estimated as the square root of the mean squared difference between converged estimated and true parameter value. As the sampling distribution of estimates can be long tailed, we also present the median and inter-quartile range as robust measures of central tendency and spread.

5.1. Simulation study 1: bias and variability of conditional estimates from two-wave snowball samples as a function of network size, conditioning, and seed set size

Table 2 contains the convergence rates for estimations in each condition of simulation study 1, and demonstrates that convergence rates were high; only for very small seed sets were the convergence rates not satisfactory, presumably in part because of the nonexistence of MLEs. Tables 3–6 contain summary statistics for the conditional estimates of the edge, alternating star, alternating triangle, and alternating 2-path effects, respectively, as a function of network size and seed set size.

A number of patterns are evident in Tables 3–6.

First, even allowing for differences in the scale of parameters, measures of bias and variability vary across the four effects. The

Table 3
Summary statistics for conditional estimates of the edge effect in simulation study 1 (population value −4.0).

Edge						
Network size	Seed set size	Mean	Median	IQR	Bias	RMSE
150	3	−1.394	−2.838	6.464	2.606	6.838
	5	−3.044	−3.620	4.069	0.956	4.153
	7	−3.342	−3.778	3.712	0.658	3.244
	9	−3.453	−3.699	3.143	0.547	2.509
	11	−3.603	−3.913	2.681	0.397	2.111
	13	−3.717	−3.890	2.255	0.283	1.927
	15	−3.717	−3.839	2.185	0.283	1.682
500	3	0.647	−2.214	14.703	4.647	11.593
	5	−0.667	−2.657	9.115	3.333	8.471
	7	−1.381	−3.160	8.460	2.619	7.269
	9	−2.086	−3.296	6.166	1.914	5.742
	11	−2.218	−3.291	5.375	1.782	4.951
	13	−2.241	−3.431	5.365	1.759	4.892
	15	−2.832	−3.646	4.598	1.168	4.024
	20	−2.966	−3.514	4.086	1.034	3.378
	30	−3.221	−3.637	3.041	0.779	2.594
	40	−3.551	−3.847	2.614	0.449	2.173
	50	−3.510	−3.853	2.457	0.490	1.967
1000	3	0.208	−2.348	18.231	4.208	13.625
	5	−0.529	−2.886	13.101	3.471	10.373
	7	−0.074	−2.910	12.410	3.926	10.385
	9	−0.052	−2.261	10.333	3.948	9.399
	11	−1.187	−2.918	9.514	2.813	8.149
	13	−1.394	−3.239	8.485	2.606	7.410
	15	−1.787	−3.269	7.730	2.213	6.796
	20	−2.265	−3.462	6.885	1.735	5.814

Table 4
Summary statistics for conditional estimates of alternating star effect in simulation study 1 (population value 0.2).

AS						
Network size	Seed set size	Mean	Median	IQR	Bias	RMSE
150	3	−0.963	−0.624	3.049	−1.163	2.937
	5	−0.230	−0.082	1.891	−0.430	1.649
	7	−0.001	0.087	1.610	−0.201	1.308
	9	0.026	0.097	1.430	−0.174	1.080
	11	0.089	0.146	1.145	−0.111	0.857
	13	0.123	0.209	1.076	−0.077	0.833
	15	0.111	0.161	1.002	−0.089	0.748
500	3	−1.235	−0.597	4.612	−1.435	3.630
	5	−0.807	−0.381	2.875	−1.007	2.564
	7	−0.576	−0.063	2.711	−0.776	2.189
	9	−0.375	−0.041	1.872	−0.575	1.730
	11	−0.345	−0.089	1.655	−0.545	1.527
	13	−0.315	−0.044	1.687	−0.515	1.494
	15	−0.131	0.099	1.522	−0.331	1.242
	20	−0.101	0.049	1.319	−0.301	1.041
	30	−0.031	0.047	1.008	−0.231	0.813
	40	0.076	0.165	0.822	−0.124	0.682
	50	0.058	0.138	0.755	−0.142	0.621
1000	3	−0.971	−0.524	5.197	−1.171	3.964
	5	−0.768	−0.245	3.717	−0.968	3.011
	7	−0.887	−0.186	3.472	−1.087	2.955
	9	−0.931	−0.346	3.017	−1.131	2.649
	11	−0.582	−0.125	2.716	−0.782	2.314
	13	−0.525	−0.061	2.353	−0.725	2.092
	15	−0.423	−0.018	2.169	−0.623	1.941
	20	−0.285	−0.026	1.995	−0.485	1.644

Table 5
Summary statistics for conditional estimates of alternating triangle effect in simulation study 1 (population value 1.0).

AT						
Network size	Seed set size	Mean	Median	IQR	Bias	RMSE
150	3	1.043	1.033	0.844	0.043	0.766
	5	0.991	0.972	0.515	−0.009	0.607
	7	0.972	0.965	0.401	−0.028	0.375
	9	0.970	0.980	0.353	−0.030	0.309
	11	0.989	0.992	0.313	−0.011	0.240
	13	0.990	0.981	0.243	−0.010	0.215
	15	0.986	0.974	0.220	−0.014	0.184
500	3	1.028	1.007	0.577	0.028	0.548
	5	1.003	0.988	0.366	0.003	0.304
	7	1.022	1.019	0.282	0.022	0.239
	9	0.996	0.992	0.230	−0.004	0.178
	11	1.003	1.005	0.177	0.003	0.147
	13	1.002	1.009	0.167	0.002	0.132
	15	0.995	0.994	0.153	−0.005	0.116
	20	0.999	1.001	0.129	−0.001	0.095
	30	1.000	1.000	0.095	0.000	0.075
	40	0.997	0.998	0.086	−0.003	0.061
	50	0.996	0.993	0.067	−0.004	0.053
1000	3	0.983	0.943	0.476	−0.017	0.480
	5	1.000	0.971	0.372	0.000	0.297
	7	1.004	0.988	0.255	0.004	0.203
	9	1.010	1.012	0.223	0.010	0.174
	11	0.997	1.000	0.185	−0.003	0.139
	13	1.000	1.009	0.164	0.000	0.123
	15	0.997	0.997	0.145	−0.003	0.110
	20	0.994	0.999	0.121	−0.006	0.094

Table 6
Summary statistics for conditional estimates of alternating 2-path effect in simulation study 1 (population value −0.2).

A2P						
Network size	Seed set size	Mean	Median	IQR	Bias	RMSE
150	3	−0.052	−0.083	0.721	0.148	0.648
	5	−0.182	−0.147	0.440	0.018	0.382
	7	−0.226	−0.197	0.376	−0.026	0.308
	9	−0.212	−0.181	0.339	−0.012	0.257
	11	−0.216	−0.203	0.277	−0.016	0.205
	13	−0.217	−0.200	0.267	−0.017	0.209
	15	−0.205	−0.199	0.258	−0.005	0.182
500	3	−0.164	−0.143	0.390	0.036	0.344
	5	−0.167	−0.149	0.264	0.033	0.198
	7	−0.182	−0.171	0.212	0.018	0.166
	9	−0.178	−0.181	0.174	0.022	0.127
	11	−0.176	−0.166	0.152	0.024	0.125
	13	−0.183	−0.178	0.149	0.017	0.116
	15	−0.192	−0.189	0.141	0.008	0.102
	20	−0.190	−0.187	0.120	0.010	0.087
	30	−0.191	−0.186	0.092	0.009	0.070
	40	−0.198	−0.197	0.081	0.002	0.061
	50	−0.195	−0.189	0.074	0.005	0.054
1000	3	−0.190	−0.169	0.282	0.010	0.257
	5	−0.188	−0.180	0.236	0.012	0.175
	7	−0.181	−0.176	0.181	0.019	0.142
	9	−0.168	−0.173	0.164	0.032	0.120
	11	−0.182	−0.181	0.138	0.018	0.111
	13	−0.184	−0.179	0.123	0.016	0.095
	15	−0.186	−0.179	0.124	0.014	0.097
	20	−0.189	−0.183	0.110	0.011	0.076

bias and variability are generally very low in the case of the alternating triangle and alternating 2-path effects, and lower than those for the edge and alternating star effects. The latter exhibit more skewed distributions, and are clearly affected by some very large estimates on one side of the distribution, estimates that may be questioned as implausible in the estimation of an ERGM from a complete single observation. We discuss the skewness in estimates of the edge and alternating star effects further below.

The second general pattern evident in Tables 3–6 is that, for a given network size, the bias and variability generally decrease as a function of seed set size, and are at modest levels for seed sets of 10

Table 7
Coverage probabilities for estimates in the two-wave snowball sampling (simulation study 1).

Coverage probabilities					
Network size	Seed set size	Edge	AS	AT	A2P
150	3	0.995	0.992	0.982	0.921
	5	0.980	0.969	0.989	0.917
	7	0.963	0.951	0.963	0.929
	9	0.982	0.970	0.956	0.932
	11	0.978	0.974	0.970	0.948
	13	0.972	0.952	0.936	0.944
	15	0.974	0.962	0.956	0.964
500	3	0.988	0.993	0.974	0.958
	5	0.969	0.984	0.959	0.961
	7	0.966	0.970	0.966	0.962
	9	0.964	0.984	0.952	0.966
	11	0.960	0.974	0.954	0.948
	13	0.958	0.964	0.962	0.944
	15	0.942	0.950	0.956	0.966
	20	0.952	0.956	0.954	0.946
	30	0.974	0.968	0.940	0.952
	40	0.950	0.952	0.962	0.952
	50	0.954	0.948	0.960	0.964
1000	3	0.990	0.992	0.969	0.976
	5	0.964	0.972	0.949	0.970
	7	0.967	0.981	0.961	0.950
	9	0.982	0.984	0.945	0.967
	11	0.961	0.970	0.955	0.951
	13	0.980	0.978	0.954	0.972
	15	0.976	0.972	0.952	0.962
	20	0.952	0.958	0.948	0.962

or so. This is encouraging, and suggests that the strategy has some value in recovering parameter values from snowball samples.

A third pattern, however, is that, as the network size increases, the bias and variability appear to increase somewhat as well, particularly for the edge and alternating star effects, suggesting that larger networks are likely to need somewhat larger seed sets if estimates of these effects with low bias and variability are desired.

Fourth, and importantly, the median of the conditional estimates generally is much closer to the true value than the average estimate, confirming the impact of the shape of the distribution on the average estimates and hence on estimates of bias in Tables 3–6.

To provide some insight into the distribution of the estimates, particularly for the edge and alternating-star effects, Figs. 2 and 3 contain boxplots of estimates as a function of seed set size for networks of size 500 and 1000. It is evident from these plots that the distribution of the estimates for some effects is quite skewed. This may reflect the fact that observed graph statistics can be close to extreme values, leading to estimates that are very large in magnitude. This is an important feature of the behaviour of the conditional estimation strategy, and means that care is required in practice to ensure that observed statistics are not close to extreme values. It should also be noted that, for the edge and alternating *k*-star effects, the increase in variability of estimates with network size is also evident in Figs. 2 and 3: for a given seed set size, the variability of the estimates is greater in the case of the network of size 1000.

Table 7 contains the coverage probabilities, that is, the proportion of times the true value of the parameter was within an approximate confidence interval constructed as the mean estimate plus or minus two estimated standard deviations of the estimate. Encouragingly, the coverage probabilities are close to their nominal values of 0.95.

Overall, while these figures suggest that care needs to be exercised in interpreting results of this approach given the skewed distribution of conditional estimates in some cases, and an associated modest bias, especially for edge and alternating star effects, they also suggest that these snowball design-based methods do offer some value in obtaining approximate estimates of ERGM effects when other methods are not feasible or impractical, and they also offer the prospect of using these methods fruitfully in schemes designed to speed computation in estimating models for large networks. A strong cautionary note, however, must be made in relation to the relative size of the initial seed set compared to the size of the network: as the case of 1000 nodes illustrates, more seeds are required for networks of larger size if the impact of more variability in conditional estimates is to be avoided. This is demonstrated in Fig. 3, where it is clear that the variability of estimates remains high even for the larger seed set.

Table 8
Variation of conditional parameter estimates, based on a one-way analysis in which the graph sampled from the ERGM with edge, alternating star, alternating 2-path, and alternating triangle parameters of −4.0, 0.2, −0.2, and 1.0, respectively, is treated as a random factor. The analysis utilizes estimates for 100 random seed sets for each of the 100 sampled graphs. The seed set size is fixed at 10 and conditional estimation is based on a two-wave sample.

Estimate	Mean square (between sampled graphs)	Mean square (within sampled graphs)	Intraclass correlation
Edge	146.53	26.20	0.821
Alternating star	15.45	2.41	0.844
Alternating 2-path	0.155	0.015	0.900
Alternating triangle	0.198	0.025	0.875

5.2. Simulation study 2: the effects of graph sampling and seed set sampling on the variability of conditional estimates

In this second simulation study, 100 graphs were sampled from the same fixed exponential random graph distribution with $n = 500$ nodes (that is, the parameter values were the same as those for the first study). For each of these 100 graphs, 100 random seed sets of size 10 were used to obtain conditional parameter estimates based on a two-wave sampling design. The number of nodes included in the two-wave samples ranged from 120 to 274, with a mean of 200.9 and standard deviation of 20.0. The purpose was to assess, at this seed set size, the relative contribution to variability in the conditional estimates of the sampling of complete graphs from the ERGM distribution and the sampling of seed sets from those sampled complete graphs. The contributions were assessed by performing a one-way analysis of variance in which sampled complete graphs constituted a random factor. The intraclass correlation coefficient is also reported; it indicates the proportion of variation in the conditional estimates that is attributable to sampling a complete network from the ERGM.

The results are presented in Table 8. Notably, in all cases, the major amount of variation in the conditional estimates is associated with the sampling of the initial complete network from the ERGM. The intraclass correlations ranged from 0.82 to 0.90, and they show that a substantial portion of the variability among conditional estimates can be attributed to the sampling of networks from the ERGM rather than to the sampling of seed sets within those sampled graphs. The effect is most marked in the case of the alternating 2-path and alternating triangle effects, but is nonetheless still strong for the edge and alternating star effects. Of course, the size of the initial seed set (here, 10) is a factor in determining consistency among conditional estimates based on different seed sets of that size: smaller seed sets, and larger node sets, would almost certainly yield less consistency across seed sets.

5.3. Simulation study 3: estimation using a one-wave snowball sample

Table 9 provides summary statistics including bias and RMSE for the case of conditional estimation using a one-wave sample.

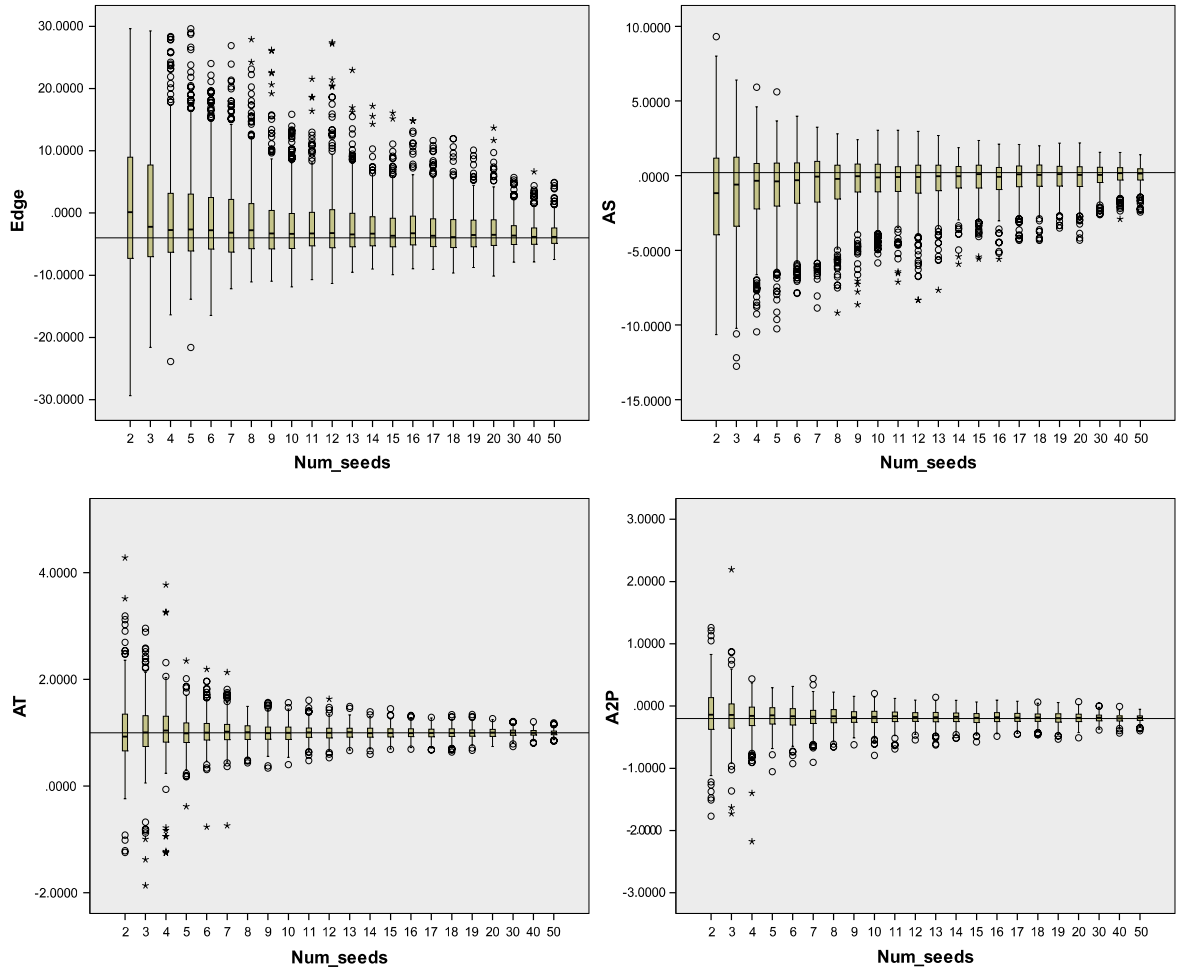

Fig. 2. Boxplots for conditional estimates from a two-wave snowball sample as a function of seed set size for the case $n = 500$.

Table 9
Summary statistics for estimates from the one-wave snowball sampling simulation as a function of network size and seed set size.

Network size	Seed set size	Effect	Bias	RMSE	Effect	Bias	RMSE
150	30	Edge	−0.460	2.922	Alt-star	0.267	1.367
	50		0.068	1.286		0.023	0.679
	100		0.101	0.681		−0.011	0.380
	150		0.081	0.307		−0.007	0.532
	30	Alt-triangle	−0.035	0.503	Alt-2-path	−0.121	0.436
	50		−0.011	0.204		−0.039	0.223
	100		−0.006	0.111		−0.015	0.123
	150		−0.004	0.095		−0.011	0.098
500	30	Edge	−1.521	11.139	Alt-star	0.597	3.649
	50		−0.915	7.198		0.332	2.314
	100		0.126	2.975		−0.016	0.990
	200		0.264	1.476		−0.076	0.503
	500		0.073	0.730		−0.007	0.261
	30	Alt-triangle	−0.259	1.292	Alt-2-path	−0.129	0.470
	50		−0.046	0.609		−0.052	0.240
	100		−0.004	0.118		−0.011	0.114
	200		−0.001	0.062		0.061	0.061
	500		−0.002	0.035		−0.005	0.035
1000	30	Edge	0.444	15.901	Alt-star	0.062	4.814
	50		−0.151	13.047		0.107	3.849
	100		0.146	8.131		−0.016	2.388
	200		0.562	3.670		−0.153	1.088
	30	Alt-triangle	−0.619	1.971	Alt-2-path	−0.098	0.460
	50		−0.200	0.958		−0.041	0.265
	100		−0.011	0.154		−0.011	0.136
	200		−0.003	0.072		0.001	0.067

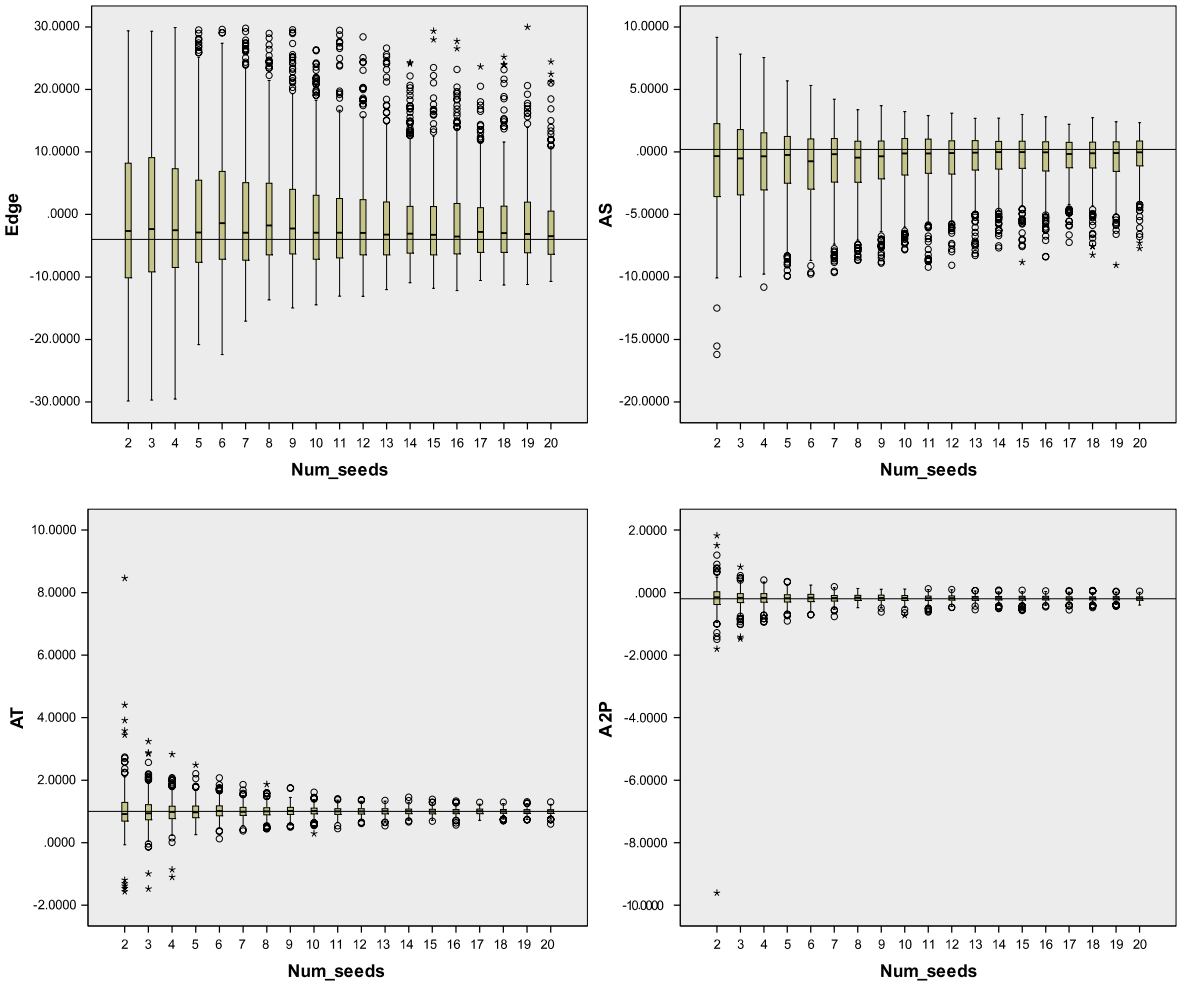

Fig. 3. Boxplots for conditional estimates from a two-wave snowball sample as a function of seed set size for the case $n = 1000$.

The statistics are reported for varying seed set sizes and for networks of size 150, 500, and 1000.

These figures exhibit a number of systematic trends similar to those for the two-wave case. First, for small seed sets, the bias and RMSE are higher. However, the situation is greatly improved for larger seed sets and, for each network size, the bias and RMSE are generally decreasing functions of seed set size, and are reasonably small for all but the lower samples sizes. There are some cases where the estimated bias does not go down with increasing seed set size, but these deviations from monotonicity are not significant. As before, for a given seed set size, the bias and RMSE appear to increase as a function of network size, particularly for the edge and alternating star effects, again suggesting the worth of using larger samples from larger networks. Nonetheless, the triangle and alternating 2-path effects appear to be well recovered at more modest seed set sizes.

For the networks of size 150 and 500, we can compare the conditional estimates with the MCMCMLEs. The MCMCMLEs are, in fact, the values corresponding to a seed set size of 150 or 500, respectively: in this case, all nodes are in Z_0, Z_1 is empty, and the conditional estimates are equivalent to the unconditional MCMCMLEs. The RMSEs are then empirical estimates of the variability of the estimates.

A comparison of the RMSEs for smaller seed sets with the RMSE of estimates based on the complete network also provides a quantitative guide to the loss of efficiency in using the partial snowball data rather than a complete network.

6. Approximate estimates for large networks

We observed earlier that this conditional estimation approach could be used to obtain approximate estimates for large complete networks, whether to obtain a good starting point for MCMCMLE or as a substitute where MCMCMLE is not feasible. The estimation in this case can be improved by taking a random sample of seed sets, and averaging the conditional estimates obtained. The average over a good number of random seed sets will have a variance that approximates the variance of a single conditional estimate, multiplied by the intraclass correlation studied in simulation study 2. We illustrate this potential using results from the simulation studies described in the preceding section.

For the same ERGM treated in the earlier section, we selected a single complete network of $n = 500$ nodes from the distribution. For this network, we obtained the MCMCMLE for each effect in the model. The estimated effects (standard errors) for edge, alternating star, alternating 2-path, and alternating triangle were -3.277 (0.911), 0.030 (0.315), -0.210 (0.037), and 1.000 (0.038), respectively. From this network, we also selected from 1 to 100 random seed sets of size 10, and used these seed sets to construct two-wave snowball samples from which we estimated conditional MCMCLEs for each effect.

Fig. 4 compares the MCMCMLE for each effect in the model (the horizontal line) with the distribution of conditional estimates for the effect (displayed in boxplot form) as we increase the number of random seeds to obtain the sample informing the average effect.

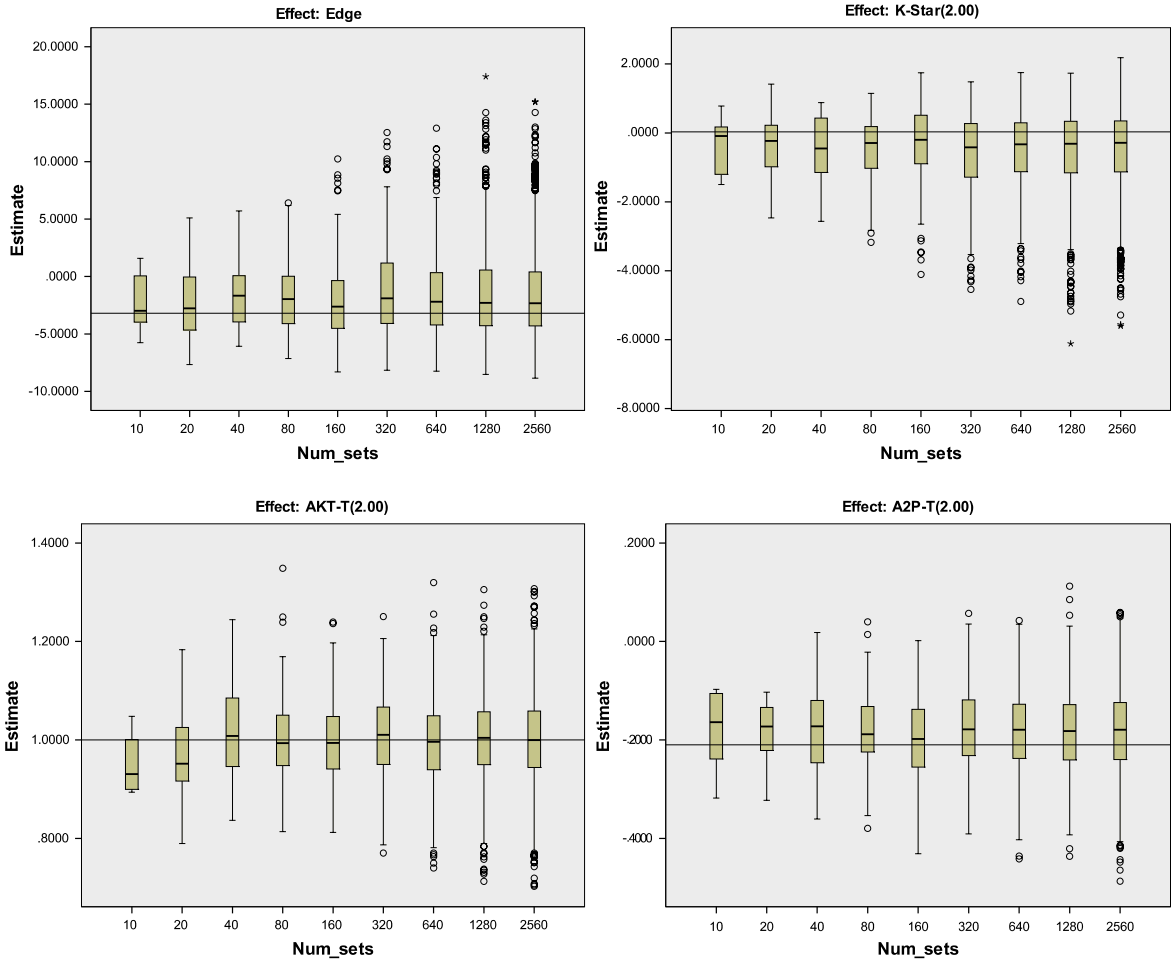

Fig. 4. Boxplots of conditional estimates for effects from seed sets of size 10, as a function of the number of random seed sets used. The MLEs for the complete network are shown as horizontal lines and are −3.277, −0.030, 1.000, and −0.210 for edge, alternating star, alternating triangle, and alternating 2-path effects, respectively.

It can be seen that, once we are combining the estimates from 40 or more random seed sets, approximation to the MCMCMLEs is reasonable.

This is illustrative only, but it suggests that the approach of combining conditional estimates from multiple random seed sets for a single complete network is worthy of further exploration, particularly as the computations for multiple random seed sets may be conducted in parallel. In the present case, the time for MCMCMLE for the full estimate (approximately 30 min) can be compared with the time for a single conditional MCMCMLE using a seed set of size 10 (of the order of 60 s).[1]

7. Discussion and extensions

The approach we have described offers the potential to extend the circumstances in which ERGMs might be employed to analyse snowball sampled network data and also to facilitate estimation and analysis for very large networks by obtaining fast approximate estimates as well as an understanding of the variability of estimates. If the sample size of the network is known and the network is not too large, the approach of Handcock and Gile (2010) is clearly optimal and should be used. If, however, the sample size is not known, or if the network is too large to consider imputation over the full set of nonobserved variables, then the approach we have outlined offers a useful method.

The conditional estimation procedure, applied as suggested to a very large network using a snowball sample from a random seed set, may also be used to study heterogeneity between different parts of the large network. For very large networks, it is implausible that an ERGM with constant parameters would apply throughout the network, as local conditions – exogenous as well as endogenous – may lead to variations in parameter values. With a snowball sample from a small seed set one will obtain a small part of the large network; variation in parameter estimates between the snowball samples may give information about possible deviations from constancy of the ERGM parameters across the network. It might be possible to combine this with ideas from Snijders (2010) to estimate parameters only from connected parts of snowball samples, thereby extracting parameter estimates from different small and relatively cohesive parts of the network.

There are also several further potential extensions of what we have proposed.

The conditional estimation method just described relies on being able to identify a subset of observed ties variables that (a) can be modelled in terms of the same parameters as the model for the full network, (b) are conditionally independent of what has not been observed, and (c) can be estimated conditional on the values

[1] Performance may vary depending on hardware and estimation configurations (as in Snijders' (2002) estimation algorithm). Here, the time for a single estimation run was approximated based on a PC with 2.4 GHz CPU. The estimation requires 26 Mb of RAM for the full network, and the estimation configurations are the same for both the full network and the conditional estimations.

of other observed variables. This is clearly a model-dependent strategy, and the approach may need to be modified for other model specifications. For example, the Markov and social circuit models described earlier can be estimated, in principle, from a one-wave snowball sample, just as we demonstrated earlier. More complex dependence assumptions may require additional data. For example, a 3-*path* model (Pattison & Robins, 2002) assumes that two ties $Y(i, j)$ and $Y(k, l)$ are conditionally dependent in the event that there is at least one tie linking one or both of $\{i, j\}$ to one or both of $\{k, l\}$. If nodes i and j are both seed nodes and therefore members of Z_0, the tie $Y(i, j)$ may then be dependent on a tie that links a node in Z_1 to a node in Z_2, but is necessarily conditionally independent of ties that link nodes within Z_2 or ties that extend beyond Z_2. This means that, in order to use a conditional estimation strategy of the form we have described, we would need to at least observe ties within Y_{12}, but we would not need necessarily to observe ties within Y_{22} or beyond.

In the case of directed networks, a large variety of potential ERGMs can be described, but a directed version of the social circuit model (Robins, Pattison, & Wang, 2009; Snijders et al., 2006) has been demonstrated to be useful in a number of modelling situations. The strategy outlined for the nondirected case may be adapted for directed networks, though the simplest adaptation is associated with an assumption that both parties to a potential tie agree on its presence or absence, and can therefore report on it with complete accuracy. In this case, respondents should be asked about both outgoing and incoming ties in order to accommodate the full range of dependencies within the version of the social circuit model described by Robins et al. However, since this is an unrealistic assumption to make in many contexts, the directed case is likely to need either (a) models in which dependencies are described in terms of outgoing ties only; or (b) some further developments in the handling of measurement error in networks. Both of these are likely fruitful directions for further work. One case, for example, where the first of these strategies may be illuminating is in the domain where snowball sampling first arose, namely, in the assessment of directed influence networks.

In the case of bipartite networks (Wang, Pattison, & Robins, 2013; Wang, Robins et al., 2009; Wang, Sharpe, Robins, & Pattison, 2009), the adaptation of the method proposed here is arguably more straightforward in principle because ties are nondirected; however, the context of application will determine how readily the ties of the two different types of node may be elicited in a snowball scheme.

Finally, we have described the conditional estimation strategy for snowball sampling, but the approach can be described in a more abstract form. In each application we have proposed a partitioning of tie variables in a network to be modelled into three classes: a subset of observed tie variables to be modelled conditionally (possibly subject to constraints); a subset of observed tie variables on which the ties to be modelled are assumed conditionally dependent; and a subset of (unobserved) tie variables that are conditionally independent of the ties to be modelled. Other sampling designs may give rise to partitions on the tie set of this general structure.

References

Baddeley, A., & Möller, J. (1989). Nearest-neighbour Markov point processes and random sets. *International Statistical Review*, *57*, 89–121.

Barton, A. H. (2001). Paul Lazarsfeld as institutional inventor. *International Journal of Public Opinion Research*, *13*, 245–269.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B*, *36*, 96–127 (with discussion).

Bieleman, B., Diaz, A., Merlo, G., & Kaplan, C. (1993). *Lines across Europe: nature and extent of cocaine use in Barcelona, Rotterdam and Turin*. Amsterdam: Swets & Zeitlinger.

Bott, E. (1957). *Family and social network*. London: Tavistock Publications.

Coleman, J. S. (1958). Relational analysis: the study of social organizations with survey methods. *Human Organization*, *16*, 28–36.

Dávid, B., & Snijders, T. A. B. (2002). Estimating the size of the homeless population in Budapest, Hungary. *Quality and Quantity*, *36*, 291–303.

Erdös, P., & Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, *6*, 290–297.

Frank, O. (2005). Network sampling and model fitting. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 31–56). New York: Cambridge University Press.

Frank, O., & Snijders, T. A. B. (1994). Estimating hidden populations using snowball sampling. *Journal of Official Statistics*, *10*, 53–67.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, *81*, 832–842.

Goodman, L. A. (1961). Snowball sampling. *The Annals of Mathematical Statistics*, *32*, 148–170.

Goodreau, S. (2007). Advances in exponential random graph (p^*) models applied to a large social network. *Social Networks*, *29*, 231–248.

Handcock, M. S. (2003). Statistical models for social networks. In R. L. Breiger, K. M. Carley, & P. E. Pattison (Eds.), *Dynamic social network modeling and analysis*. Washington, DC: National Academies Press.

Handcock, M.S. (2004). Assessing degeneracy in statistical models for social networks. Center for statistics in the social sciences working paper no. 39, University of Washington.

Handcock, M.S., & Gile, K.J. (2007). Modelling social networks with sampled or missing data. CSSS Working paper no. 75, University of Washington.

Handcock, M. S., & Gile, K. J. (2010). Modelling networks from sampled data. *Annals of Applied Statistics*, *4*, 5–25.

Handcock, M. S., Hunter, D., Butts, C., Goodreau, S., & Morris, M. (2008). Statnet: software tools for representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, *24*(1). URL: http://www.jtatsoft.org/v24/i01.

Heckathorn, D. D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems*, *44*, 174–199.

Holland, P., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, *76*, 33–50.

Hunter, D. R. (2007). Curved exponential family models for social networks. *Social Networks*, *29*, 216–230.

Hunter, D. R., & Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, *15*, 565–583.

Katz, E., & Lazarsfeld, P. (1955). *Personal influence*. Glencoe, IL: The Free press.

Lazarsfeld, P., Berelson, B., & Gaudet, H. (1944). *The people's choice: how the voter makes up his mind in a presidential election*. New York: Duell, Sloan & Pearce.

Lubbers, M. J. (2003). Group composition and network structure in school classes: a multilevel application of the p^* model. *Social Networks*, *25*, 309–332.

Merton, R. K. (1957). *Social theory and social structure* (rev. ed.). Glencoe, IL: The Free Press.

Milgram, S. (1967). The small world problem. *Psychology Today*, *1*, 60–67.

Pattison, P. E., & Robins, G. L. (2002). Neighbourhood-based models for social networks. *Sociological Methodology*, *32*, 301–337.

Pattison, P., & Snijders, T. A. B. (2013). Modelling social networks: next steps. In D. Lusher, J. Koskinen, & G. Robins (Eds.), *Exponential Random Graph Models for Social Networks: Theory, Method and Applications* (pp. 287–302). Cambridge University Press.

Robins, G. L., Pattison, P. E., & Wang, P. (2009). Closure, connectivity and degree distributions: exponential random graph (p^*) models for directed social networks. *Social Networks*, *31*, 105–117.

Robins, G. L., Pattison, P. E., & Woolcock, J. (2005). Small and other worlds: global network structures from local processes. *American Journal of Sociology*, *110*, 894–936.

Robins, G. L., Snijders, T. A. B., Wang, P., Handcock, M. S., & Pattison, P. E. (2007). Recent developments in exponential random graph (p^*) models. *Social Networks*, *29*, 216–230.

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. (2010). Conditional marginalization for exponential random graph models. *Journal of Mathematical Sociology*, *34*, 239–252.

Snijders, T. A. B., Pattison, P., Robins, G. L., & Handcock, M. (2006). New specifications for exponential random graph models. *Sociological Methodology*, *36*, 99–153.

Thompson, S. K., & Collins, L. M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence*, *68*, S57–S67.

Thompson, S. K., & Frank, O. (2000). Model-based estimation with link-tracing designs. *Survey Methodology*, *26*, 87–98.

Wang, P., Pattison, P. E., & Robins, G. L. (2013). Exponential random graph model specifications for bipartite networks—a dependence hierarchy. *Social Networks*, *35*(2), 211–222.

Wang, P., Robins, G., & Pattison, P. (2009). PNet: program for the estimation and simulation of p^* exponential random graph models, user manual. Department of Psychology, University of Melbourne.

Wang, P., Sharpe, K., Robins, G., & Pattison, P. (2009). Exponential random graph (p^*) models for affiliation networks. *Social Networks*, *31*, 12–23.

Wasserman, S., & Pattison, P. E. (1996). Logit models and logistic regressions for social networks, I. An introduction to Markov graphs and p^*. *Psychometrika*, *61*, 401–425.