The Annals of Applied Statistics
2010, Vol. 4, No. 2, 567–588
DOI: 10.1214/09-AOAS313

MAXIMUM LIKELIHOOD ESTIMATION FOR SOCIAL NETWORK DYNAMICS

BY TOM A. B. SNIJDERS[1], JOHAN KOSKINEN[1] AND MICHAEL SCHWEINBERGER[2]

University of Oxford, University of Groningen, University of Oxford and Penn State University

A model for network panel data is discussed, based on the assumption that the observed data are discrete observations of a continuous-time Markov process on the space of all directed graphs on a given node set, in which changes in tie variables are independent conditional on the current graph. The model for tie changes is parametric and designed for applications to social network analysis, where the network dynamics can be interpreted as being generated by choices made by the social actors represented by the nodes of the graph. An algorithm for calculating the Maximum Likelihood estimator is presented, based on data augmentation and stochastic approximation. An application to an evolving friendship network is given and a small simulation study is presented which suggests that for small data sets the Maximum Likelihood estimator is more efficient than the earlier proposed Method of Moments estimator.

1. Introduction. Relations between social actors can be studied by methods of social network analysis [e.g., Wasserman and Faust (1994); Carrington, Scott and Wasserman (2005)]. Examples are friendship between pupils in a school class or alliances between firms. A basic data structure for social networks is the directed graph or digraph, where the actors are represented by the nodes, and the arcs between the nodes indicate the social ties. Social network analysis traditionally has had a focus on rich description of network data, but the recent development of methods of statistical inference for network data [e.g., Airoldi et al. (2007); Hunter and Handcock (2006)] has the potential of moving this field toward a wider use of inferential approaches. Longitudinal studies are especially important for obtaining insight into social networks, but statistical methods for longitudinal network data that are versatile enough for realistic modeling are only just beginning to be developed.

This article considers repeated observations of a relation, or network, on a given set of actors $\mathcal{N} = \{1, \ldots, n\}$, observed according to a panel design, and repre-

Received March 2009; revised August 2009.

[1]Supported in part by the US National Institutes of Health (NIH 1R01HD052887-01A2).

[2]Supported in part by the Netherlands Organization for Scientific Research (NWO 401-01-552 and 446-06-029).

Key words and phrases. Graphs, longitudinal data, method of moments, stochastic approximation, Robbins–Monro algorithm.

sented as a sequence of digraphs $x(t_m)$ for $m = 1, \ldots, M$, where $t_1 < \cdots < t_M$ are the observation moments. The nodes represent social actors (which may be individuals, companies, etc.), and the node set $\mathcal{N}$ is the same for all observation moments. A digraph is defined here as an irreflexive relation, that is, a subset x of $\{(i, j) \in \mathcal{N}^2 \mid i \neq j\}$, and when $(i, j) \in x$ we shall say that there is an arc, or a tie, from i to j. It often is realistic to assume that the social network has been developing between the observation moments, which leads to the assumption that the observations $x(t_m)$ are realizations of stochastic digraphs $X(t_m)$ embedded in a continuous-time stochastic process $X(t)$, $t_1 \leq t \leq t_M$. Holland and Leinhardt (1977) proposed to use continuous-time Markov chains, defined on the space of all digraphs with a given node set, for modeling social network dynamics even if the observations are made at a few discrete time points and not continuously. Continuous-time Markov chains provide a natural starting point for modeling longitudinal network data. Wasserman (1979, 1980) and Leenders (1995) elaborated *dyad-independent models*, where the ties between pairs of actors (*dyads*) develop according to processes that are mutually independent between dyads. This is not realistic for social processes, because dependence between the set of ties among three or more actors can be very strong, as was found already by Davis (1970) who showed that a basic feature of many social networks is the tendency toward transitivity ("friends of my friends are my friends"). Snijders and van Duijn (1997) and Snijders (2001) proposed so-called *actor-oriented models*, explained in the next section, which do allow such higher-order dependencies.

The actor-oriented models are too complicated for the calculation of likelihoods or estimators in closed form, but they represent stochastic processes which can be easily simulated. This was exploited by the estimation method proposed in the papers mentioned, which is a Method of Moments (*MoM*) estimator implemented algorithmically by stochastic approximation [Robbins and Monro (1951); also see, e.g., Kushner and Yin (2003)]. This estimator usually performs well [some empirical applications are given in de Federico (2003), van de Bunt, van Duijn and Snijders (1999), and van Duijn et al. (2003)].

It is to be expected, however, that the statistical efficiency of the Maximum Likelihood (*ML*) estimator will be greater. ML estimation also paves the way for likelihood-based model selection, which will be a marked improvement on existing methods of model selection. This article presents an MCMC algorithm for approximating the ML estimator, combining and extending ideas in Gu and Kong (1998), Snijders (2001), and Koskinen and Snijders (2007), the latter of which proposed for this model a MCMC algorithm for Bayesian inference. Section 2 presents the model definition. The algorithm for obtaining the ML estimator is described in Section 3. Section 4 reports results of an empirical example and Section 5 presents a very small simulation study comparing the ML and MoM estimators. The paper finishes with an algorithm for approximating the likelihood ratio test in Section 6 and a discussion in Section 7.

2. Model definition. We assume that repeated observations $x(t_1), \ldots, x(t_M)$ on the network are available for some $M \geq 2$. The network, or digraph, x will be identified with its $n \times n$ adjacency matrix, of which the elements denote whether there is a tie from node i to node j ($x_{ij} = 1$) or not ($x_{ij} = 0$). Self-ties are not allowed, so that the diagonal is structurally zero. Random variables are denoted by capitals. The stochastic process $X(t)$, in which the observations $x(t_m)$ are embedded, is modeled as being right-continuous.

Various models have been proposed, most of them being Markov processes of some kind. We focus on actor-oriented models [Snijders (2001)]. Tie-oriented models [Snijders (2006)] can be treated similarly. The basic idea of actor-oriented models [Snijders (2001)] is that the nodes of the graph represent social actors, who have control, albeit under constraints, of their outgoing ties; and the graph develops as a continuous-time Markov process (even though it is observed only at M discrete time points). The constraints are that ties may change only one by one, and actors do not coordinate their changes of ties. Thus, at any given moment, one actor i may create one new tie or delete one existing tie, where the probability distribution of such changes depends on the current digraph; excluded are simultaneous changes such as swapping one tie for another, or bargaining between actors over ties. This constraint was proposed already by Holland and Leinhardt (1977), and it has the virtue of splitting up the change process in its smallest possible constituents.

The actor-oriented model is further specified as follows; further discussion and motivation is given in Snijders (2001).

1. *Opportunities for change*

Each actor i gets at stochastically determined moments the opportunity to change one of the outgoing tie variables $X_{ij}(t)$ ($j \in \mathcal{N}$, $j \neq i$). Since the process is assumed to be Markovian, waiting times between opportunities have exponential distributions. Each of the actors i has a *rate function* $\lambda_i(\alpha, x)$ which defines how quickly this actor gets an opportunity to change a tie variable, when the current value of the digraph is x, and where α is a parameter. At any time point t with $X(t) = x$, the waiting time until the next opportunity for change by any actor is exponentially distributed with parameter

$$\lambda(\alpha, x) = \sum_i \lambda_i(\alpha, x). \tag{1}$$

Given that an opportunity for change occurs, the probability that it is actor i who gets the opportunity is given by

$$\pi_i(\alpha, x) = \lambda_i(\alpha, x)/\lambda(\alpha, x). \tag{2}$$

The rate functions can be constant between observation moments, or they can depend on functions $r_{ik}(x)$ which may be covariates or positional characteristics of

the actors such as outdegrees $\sum_j x_{ij}$. A convenient assumption is to use an exponential link function,

$$\lambda_i(\alpha, x) = \exp\Bigl(\sum_k \alpha_k r_{ik}(x)\Bigr). \tag{3}$$

2. *Options for change*

When actor i gets the opportunity to make a change, this actor has a permitted set $\mathcal{A}_i(x^0)$ of values to which the digraph may be changed, where x^0 is the current value of the digraph. The assumption that the actor controls his or her outgoing ties, but can change only one tie variable at the time, is equivalent to

$$\mathcal{A}_i(x^0) \subset \{x^0\} \cup \mathcal{A}_i^{\mathrm{r}}(x^0), \tag{4a}$$

where $\mathcal{A}_i^{\mathrm{r}}(x^0)$ is the set of adjacency matrices differing from x in exactly one element,

$$\begin{aligned}\mathcal{A}_i^{\mathrm{r}}(x^0) = \{x \mid {} & x_{ij} = 1 - x_{ij}^0 \text{ for one } j \neq i, \\ & \text{and } x_{hk} = x_{hk}^0 \text{ for all other } (h,k)\}.\end{aligned} \tag{4b}$$

Including x^0 in $\mathcal{A}_i(x^0)$ can be important for expressing the property that actors who are satisfied with the current network will prefer to keep it unchanged. Therefore, the usual model is $\mathcal{A}_i(x^0) = \{x^0\} \cup \mathcal{A}_i^{\mathrm{r}}(x^0)$. This does not lead to identifiability problems because the ratio between not making a change and making a change is not a free parameter, but fixed by assumption (5) for the conditional probabilities of the new state of the network.

Some alternatives are models with structurally impossible ties, where the impossible digraphs are left out of $\mathcal{A}_i(x^0)$, and models where the actor is required to make a change whenever there is the opportunity, obtained by leaving the current element x^0 out of $\mathcal{A}_i(x^0)$.

It is assumed that the network dynamics is driven by a so-called *objective function* $f_i(\beta, x^0, x)$ that can be interpreted as the relative attractiveness for actor i of moving from the network represented by x^0 to the network x, and where β is a parameter. Under the condition that the current digraph is x^0 and actor i gets the opportunity to make a change, the conditional probability that the next digraph is x is modeled as

$$p_i(\beta, x^0, x) = \begin{cases} \exp(f_i(\beta, x^0, x)) / \displaystyle\sum_{\tilde{x}} \exp(f_i(\beta, x^0, \tilde{x})), & x \in \mathcal{A}_i(x^0), \\ 0, & x \notin \mathcal{A}_i(x^0), \end{cases} \tag{5}$$

where the summation extends over $\tilde{x} \in \mathcal{A}_i(x^0)$. This formula can be motivated by a random utility argument as used in econometrics [see, e.g., Maddala (1983)], where it is assumed that the actor maximizes $f_i(\beta, x^0, x)$ plus a random disturbance having a standard Gumbel distribution. Assumption (4) implies that instead

of $p_i(\beta, x^0, x)$ we can also write $p_{ij}(\beta, x^0)$ where the correspondence between x and j is defined as follows: if $x \neq x^0$, j is the unique element of $\mathcal{N}$ for which $x_{ij} \neq x^0_{ij}$; if $x = x^0$, $j = i$. This less redundant notation will be used in the sequel. Thus, for $j \neq i$, $p_{ij}(\beta, x^0)$ is the probability that, under the condition that actor i has the opportunity to make a change and the current digraph is x^0, the change will be to $x_{ij} = 1 - x^0_{ij}$ with the rest unchanged, while $p_{ii}(\beta, x^0)$ is the probability that, under the same condition, the digraph will not be changed.

The most usual models are based on objective functions that depend on x only. This has the interpretation that actors wish to maximize a function $f_i(\beta, x)$ independently of "where they come from." The greater generality of (5), where the objective function can depend also on the previous state x^0, makes it possible to model path-dependencies, or hysteresis, where the loss suffered from withdrawing a given tie differs from the gain from creating this tie, even if the rest of the network has remained unchanged.

Various ingredients for specifying the objective function were proposed in Snijders (2001). A linear form is convenient,

$$f_i(\beta, x^0, x) = \sum_{k=1}^{L} \beta_k s_{ik}(x^0, x), \tag{6}$$

where the functions $s_{ik}(x^0, x)$ are determined by subject-matter knowledge and available social scientific theory. These functions can represent essential aspects of the network structure, assessed from the point of view of actor i, such as

$$s_{ik}(x^0, x) = \sum_j x_{ij} \qquad \text{(outdegree)} \tag{7}$$

$$\sum_j x_{ij} x_{ji} \qquad \text{(reciprocated ties)} \tag{8}$$

$$\sum_{j,k} x_{ij} x_{jk} x_{ik} \qquad \text{(transitive triplets)} \tag{9}$$

$$\sum_j (1 - x_{ij}) \max_k x_{ik} x_{kj} \qquad \text{(indirect ties)} \tag{10}$$

$$\sum_j x^0_{ij} x^0_{ji} x_{ij} x_{ji} \qquad \text{(persistent reciprocity);} \tag{11}$$

they can also depend on covariates—such as resources and preferences of the actors, or costs of exchange between pairs of actors—or combinations of network structure and covariates. For example, de Federico (2003) found in a study of friendship between foreign exchange students that friendship formation tends to be reciprocal, with a negative parameter for forming indirect ties (thus leading to relatively closed networks), and that friendships are more likely to be formed

between individuals from the same region (a covariate effect), but that reciprocation adds less to the tendency to form ties between persons from the same region than between arbitrary individuals (negative interaction between covariate and reciprocity).

3. *Intensity matrix; time-homogeneity*

The model description given above defines $X(t)$ as a continuous-time Markov process with Q-matrix or intensity matrix [e.g., Norris (1997)] for $x \neq x^0$ given by

$$q(x^0, x) = \begin{cases} \lambda_i(\alpha, x^0) p_i(\beta, x^0, x), & \text{if } x \in \mathcal{A}_i(x^0), i \in \mathcal{N}, \\ 0, & \text{if } x \notin \bigcup_i \mathcal{A}_i(x^0). \end{cases} \tag{12}$$

The assumptions do not imply that the distribution of $X(t)$ is stationary. The intensity matrix is time-homogeneous, however, except for time dependence reflected by time-varying components in the functions $s_{ik}(x^0, x)$.

For the data-collection designs to which this paper is devoted, where observations are done at discrete time points $t_1, \ldots, t_M$, these time points can be used for marking time-heterogeneity of the transition distribution (cf. the example in Section 6). For example, covariates may be included with values allowed to change at the observation moments. A special role is played here by the time durations $t_m - t_{m-1}$ between successive observations. Standard theory for continuous-time Markov chains [e.g., Norris (1997)] shows that the matrix of transition probabilities from $X(t_m)$ to $X(t_{m-1})$ is $e^{(t_m - t_{m-1})Q}$, where Q is the matrix with elements (12). Thus, changing the duration $t_m - t_{m-1}$ can be compensated by multiplication of the rate function $\lambda_i(\alpha, x)$ by a constant. Since the connection between an externally defined real-valued time variable t_m and the rapidity of network change is tenuous at best, it usually is advisable [e.g., see Snijders (2001)] to include a multiplicative parameter in the rate function which is constant between observation moments t_m but differs freely between periods (t_{m-1}, t_m). With the inclusion of such a parameter, the numerical values t_m become unimportant because changes in their values can be compensated by the multiplicative rate parameters.

2.1. *Comparison with discrete-time Markov chain models.* Popular models for analyzing cross-sectional network data are exponential families of graph distributions such as the Markov model of Frank and Strauss (1986), generalized to the p^* Exponential Random Graph (ERG) model by Frank (1991) and Wasserman and Pattison (1996), and elaborated and further specified by Hunter and Handcock (2006) and Snijders et al. (2006). Discrete-time Markov chain models, as longitudinal models of this kind, were proposed by Robins and Pattison (2001), Krackhardt and Handcock (2007), and Hanneke and Xing (2007). There are a number of essential differences between these models and the actor-oriented model treated here, with respect to interpretation as well as statistical procedures.

When applying the actor-oriented model to a sequence of two or more repeated observations, these observations are embedded in a continuous-time model. This has clear face validity in applications where changes in the network take place at arbitrary moments between observations. Singer (2008) gives an overview of the use of this principle for continuously distributed data. This approach yields a major advantage over the cited longitudinal ERG models: our probability model is defined independently of the observational design. Data from irregularly spaced observation intervals can be analyzed without the need to make any adaptations. Another advantage is that the dynamic process is defined parsimoniously as a function of its elementary constituents—in this case, the conditional probability of a single tie change. The random utility interpretation of (5), discussed above, gives an interpretation in terms of myopic optimization of an objective function to which a random disturbance is added, and which can be used to represent a "social mechanism" that could have given rise to the observed dynamics.

The discrete-time ERG models constitute exponential families, which has the advantage that many standard techniques can be directly applied. The distribution of the continuous-time process $\{X(t), t_1 \leq t \leq t_M\}$ for the actor-oriented model constitutes an exponential family, so for discrete observation moments our model can be regarded as an incompletely observed exponential family. For both types of model, inference is computer-intensive and time-consuming because of the need to implement elaborate MCMC procedures. The definition of the model given above can be used directly to simulate data from the probability distribution, conditional on an initial state of the network. This contrasts with models in the ERG family, which can be simulated only indirectly, by regarding them as the stationary distribution of an auxiliary Markov chain, and applying a Gibbs or Metropolis–Hastings algorithm. The near degeneracy problem [Snijders et al. (2006)] which plagues some specifications of ERG models is, in practice, not a problem for the actor-oriented model.

3. ML estimation. This section presents an algorithm for MCMC approximation of the maximum likelihood estimate. Estimation is done conditional on the first observation $x(t_1)$. This has the advantage that no model assumptions need to be invoked concerning the probability distribution that may have led to the first observed network $x(t_1)$, and the estimated parameters refer exclusively to the dynamics of the network.

The algorithm proposed below can be sketched as follows. For each m $(m = 2, \ldots, M)$ the observed data are augmented with random draws from the sequence of intermediate digraphs $x(t)$ that could have led from one observation, $x(t_{m-1})$, to the next, $x(t_m)$ (Section 3.1). These draws can be simulated using a Metropolis–Hastings algorithm (Section 3.3). They are then used in the updating steps of a Robbins–Monro algorithm, following ideas of Gu and Kong (1998), to find the solution of the likelihood equation (Section 3.2). These elements are put together in Section 3.4.

3.1. *Augmented data.* The likelihood for the observed data $x(t_2), \ldots, x(t_M)$ conditional on $x(t_1)$ cannot generally be expressed in a computable form. Therefore, the observed data will be augmented with data such that an easily computable likelihood is obtained, employing the general data augmentation principle proposed by Tanner and Wong (1987). The data augmentation can be done for each period (t_{m-1}, t_m) separately and, therefore, this section considers only the observations $x(t_1)$ and $x(t_2)$.

Denote the time points of an opportunity for change by T_r and their total number between t_1 and t_2 by R, the time points being ordered increasingly so that $t_1 = T_0 < T_1 < T_2 < \cdots < T_R < t_2$. The model assumptions imply that at each time T_r, there is one actor, denoted I_r, who gets an opportunity for change at this time moment. Define J_r as the actor toward whom the tie variable is changed, and define formally $J_r = I_r$ if there is no change [i.e., if $x(T_r) = x(T_{r-1})$]:

$$(I_r, J_r) \text{ is the only } (i, j) \text{ for which } x_{ij}(T_{r-1}) \neq x_{ij}(T_r) \text{ if there is such an } (i, j); \text{ else } J_r = I_r.$$

Given $x(t_1)$, the outcome of the stochastic process (T_r, I_r, J_r), $r = 1, \ldots, R$ completely determines $x(t)$, $t_1 < t \leq t_2$.

The augmenting data consist of R and (I_r, J_r), $r = 1, \ldots, R$, without the time points T_r. It may be noted that this differs from the definition of a sample path in Koskinen and Snijders (2007) in that the times in between opportunities for change are integrated out; the reason is to obtain a simpler MCMC algorithm. The possible outcomes of the augmenting data are determined by the condition

$$\sharp\{r \mid 1 \leq r \leq R, (i_r, j_r) = (i, j)\} = \begin{cases} \text{even}, & \text{if } x_{ij}(t_2) = x_{ij}(t_1), \\ \text{odd}, & \text{if } x_{ij}(t_2) \neq x_{ij}(t_1), \end{cases} \tag{13}$$

for all (i, j) with $i \neq j$. The stochastic process $V = ((I_r, J_r), r = 1, \ldots, R)$ will be referred to as the sample path; the elements for which $I_r = J_r$, although redundant to calculate $x(t_2)$ from $x(t_1)$, are retained because they facilitate the computation of the likelihood. Define $x^{(k)} = x(T_k)$; the digraphs $x^{(k)}$ and $x^{(k-1)}$ differ in element (I_k, J_k) provided $I_k \neq J_k$, and in no other elements.

The probability function of the sample path, conditional on $x(t_1)$, is given by

$$\begin{aligned} &p_{\mathrm{sp}}\{V = ((i_1, j_1), \ldots, (i_R, j_R)); \alpha, \beta\} \\ &\quad = P_{\alpha,\beta}\{T_R \leq t_2 < T_{R+1} | x^{(0)}, (i_1, j_1), \ldots, (i_R, j_R)\} \\ &\qquad \times \prod_{r=1}^{R} \pi_{i_r}(\alpha, x^{(r-1)}) p_{i_r, j_r}(\beta, x^{(r-1)}), \end{aligned} \tag{14}$$

where π_i is defined in (2) and p_{ij} in and just after (5). Denote the first component of (14) by

$$\begin{aligned} &\kappa(\alpha, x^{(0)}, (i_1, j_1), \ldots, (i_R, j_R)) \\ &\quad = P_{\alpha,\beta}\{T_R \leq t_2 < T_{R+1} | x^{(0)}, (i_1, j_1), \ldots, (i_R, j_R)\}. \end{aligned} \tag{15}$$

Conditioning on $x^{(0)}, (i_1, j_1), (i_2, j_2), \ldots,$ [and not on $x(t_2)$!], the differences $T_{r+1} - T_r$ are independently exponentially distributed with parameters $\lambda(\alpha, x^{(r)})$. Hence, under this conditioning the distribution of $T_R - t_1$ is the convolution of exponential distributions with parameters $\lambda(\alpha, x^{(r)})$ for $r = 0, \ldots, R-1$. In the special case that the actor-level rates of change $\lambda_i(\alpha, x)$ are constant, denoted by α_1, R has a Poisson distribution with parameter $n\alpha_1(t_2 - t_1)$; (15) then is given by

$$\kappa\big(\alpha, x^{(0)}, (i_1, j_1), \ldots, (i_R, j_R)\big) = \exp\big(-n\alpha_1(t_2 - t_1)\big)\frac{(n\alpha_1(t_2 - t_1))^R}{R!}. \tag{16}$$

In the general case where the change rates are nonconstant, the probability (15) can be approximated as follows. Denote by $p_{T_R}(t)$ the density function of T_R, conditional on $x^{(0)}, (i_1, j_1), (i_2, j_2), \ldots,$ or, equivalently, on the embedded process $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$. The distribution of $T_R - t_1$ is a convolution of exponential distributions and, therefore, the central limit theorem implies that the density function $p_{T_R}(t)$ is approximately the normal density with expected value

$$\mu_\alpha = \sum_{r=0}^{R-1} \big\{\lambda\big(\alpha, x^{(r)}\big)\big\}^{-1} \tag{17}$$

and variance

$$\sigma_\alpha^2 = \sum_{r=0}^{R-1} \big\{\lambda\big(\alpha, x^{(r)}\big)\big\}^{-2}. \tag{18}$$

Hence,

$$p_{T_R}(t) \approx \frac{1}{\sqrt{(2\pi\sigma_\alpha^2)}} \exp\left(\frac{-(t - t_1 - \mu_\alpha)^2}{2\sigma_\alpha^2}\right). \tag{19}$$

Probability (15) now can be expressed as

$$\begin{aligned}
&\kappa\big(\alpha, x^{(0)}, (i_1, j_1), \ldots, (i_R, j_R)\big) \\
&\quad = \int_{t_1}^{t_2} p_{T_R}(s) P\{T_{R+1} - T_R > t_2 - T_R \mid T_R = s\}\, ds \\
&\quad = \int_{t_1}^{t_2} p_{T_R}(s) \exp\big(-\lambda\big(\alpha, x^{(R)}\big)(t_2 - s)\big)\, ds \qquad (20) \\
&\quad \approx p_{T_R}(t_2) \int_{t_1}^{t_2} \exp\big(-\lambda\big(\alpha, x^{(R)}\big)(t_2 - s)\big)\, ds \\
&\quad \approx \frac{p_{T_R}(t_2)}{\lambda(\alpha, x^{(R)})}.
\end{aligned}$$

The approximations are valid for large R and $t_2 - t_1$, under boundedness conditions on the rate functions λ_i. The first approximation in (20) is based on splitting the

integration interval into two intervals $(t_1, t_2 - L)$ and $(t_2 - L, t_2)$ for a bounded but large L; the first interval then gives an asymptotically negligible contribution and on the second interval $p_{T_R}(s)$ is approximately constant since $\text{var}(T_R) = \mathcal{O}(R)$. The second approximation uses that $t_2 - t_1$ is large. Combining the preceding equations yields

$$\begin{aligned}&\kappa(\alpha, x^{(0)}, (i_1, j_1), \ldots, (i_R, j_R))\\ &\quad\approx \frac{1}{\lambda(\alpha, x^{(R)})\sqrt{(2\pi\sigma_\alpha^2)}} \exp\left(\frac{-(t_2 - t_1 - \mu_\alpha)^2}{2\sigma_\alpha^2}\right).\end{aligned} \tag{21}$$

This shows that, for observed data $(x(t_1), x(t_2))$ augmented by the sample path, the likelihood conditional on $x(t_1)$ can be expressed directly, either exactly or in a good approximation.

3.2. *Missing data principle and stochastic approximation.* An MCMC algorithm will be used that finds the ML estimator based on augmented data. Several methods for MCMC maximum likelihood estimation have been proposed in the literature. We shall use the Markov Chain Stochastic Approximation (MCSA) algorithm proposed by Gu and Kong (1998) and used (in a slightly different specification) also by Gu and Zhu (2001). This algorithm is based on the missing information principle of Orchard and Woodbury (1972) and Louis (1982)—going back to Fisher (1925); cf. Efron (1977). The principle can be summarized as follows. Suppose that x is observed, with probability distribution parameterized by θ and having probability density $p_X(x; \theta)$ w.r.t. some σ-finite measure. To facilitate estimation, the observed data is augmented by extra data v (regarded as missing data) such that the joint density is $p_{XV}(x, v; \theta)$. Denote the observed data score function $\partial \log(p_X(x; \theta))/\partial\theta$ by $S_X(\theta; x)$ and the total data score function $\partial \log(p_{XV}(x, v; \theta))/\partial\theta$ by $S_{XV}(\theta; x, v)$. It is not hard to prove (see the cited literature) that

$$E_\theta\{S_{XV}(\theta; x, V) \mid X = x\} = S_X(\theta; x). \tag{22}$$

This is the first part of the missing information principle. This equation implies that the likelihood equation can be expressed as

$$E_\theta\{S_{XV}(\theta; x, V) \mid X = x\} = 0, \tag{23}$$

and, therefore, ML estimates can be determined, under regularity conditions, as the solution of (23).

This is applied in situations where the observed data score function $S_X(\theta; x)$ is too difficult to calculate, while the total data score function $S_{XV}(\theta; x, v)$ is computable. In our case, we condition on $X(t_1)$, so this is treated as being fixed; we observe $X = (X(t_2), \ldots, X(t_M))$; and between each pair of consecutive observations $X(t_{m-1})$ and $X(t_m)$ the data are augmented, as discussed in the previous

section, by the sample path that could have brought the network from $X(t_{m-1})$ to $X(t_m)$. These sample paths combined for $m = 2$ to M constitute V. The following two subsections supply the additional elements for how the augmenting data are used.

In the MCSA algorithm of Gu and Kong (1998), the solution of (23) is obtained by stochastic approximation [Robbins and Monro (1951); Kushner and Yin (2003)] which is defined by the updating step

$$\hat{\theta}^{(N+1)} = \hat{\theta}^{(N)} + a_N D^{-1} S_{XV}\big(\hat{\theta}^{(N)}; x, V^{(N)}\big), \tag{24}$$

where $V^{(N)}$ is generated according to the conditional distribution of V, given $X = x$, with parameter value $\hat{\theta}^{(N)}$. The sequence a_N, called the gain sequence, consists of positive numbers, tending to 0. The matrix D is a suitable matrix. It is efficient [see Kushner and Yin (2003)] to use a gain sequence tending to zero as $a_N \sim N^{-c}$ for a $c < 1$, and to estimate θ not by the last element $\hat{\theta}^{(N)}$ produced by the algorithm, but by a tail average $(N - n_0 + 1)^{-1} \sum_{n=n_0}^{N} \hat{\theta}^{(n)}$. For fixed n_0 and $N \to \infty$, this average converges to the solution of (23) for a wide range of positive definite matrices D.

The second part of the missing information principle [Orchard and Woodbury (1972)] is that the observed Fisher information matrix for the observed data can be expressed as

$$\begin{aligned} D_X(\theta) &= -\partial S_X(\theta; x)/\partial\theta \\ &= E_\theta\{D_{XV}(\theta) \mid X = x\} - \mathrm{Cov}_\theta\{S_{XV}(\theta; x, V) \mid X = x\}, \end{aligned} \tag{25}$$

where $D_{XV}(\theta)$ is the complete data observed Fisher information matrix,

$$D_{XV}(\theta) = -\partial S_{XV}(\theta; x, V)/\partial\theta. \tag{26}$$

Expression (25) can be interpreted loosely as "information is total (but partially unobserved) information minus missing information." This formula allows us to calculate standard errors.

3.3. *Simulating the sample path.* The MCSA method relies on Monte Carlo simulations of the missing data V. The Markov property allows us to treat all periods (t_{m-1}, t_m) separately and, therefore, we simplify the treatment and notation here by treating only one of those periods, temporarily assuming that $M = 2$. The missing data then is the sample path $((I_1, J_1), \ldots, (I_R, J_R))$ which specifies the sequence of tie changes that brings the network from $X(t_1)$ to $X(t_2)$. The advantage of having integrated out the time steps T_r—with the use of the expressions (16) and (20)—is that less random noise is introduced, and the simulated variable V is discrete rather than including a Euclidean vector with a varying dimension R, which would require a more complicated MCMC procedure.

The set of all sample paths connecting $x(t_1)$ and $x(t_2)$ is the set of all finite sequences of pairs (i, j), $i, j \in \mathcal{N}$, where for $i \neq j$ the parity of the number of occurrences of (i, j) is given by (13). Such sequences will be denoted

by $\underline{v} = ((i_1, j_1), \ldots, (i_R, j_R))$, and the set of all these sequences is denoted $\mathcal{V}$. The probability of the sample path conditional on $X(t_1) = x(t_1)$, $X(t_2) = x(t_2)$ is proportional to (14), rewritten here as

$$\kappa\bigl(\alpha, x^{(0)}, (i_1, j_1), \ldots, (i_R, j_R)\bigr) \times \prod_{r=1}^{R} \pi_{I_r}\bigl(\alpha, x^{(r-1)}\bigr) p_{I_r, J_r}\bigl(\beta, x^{(r-1)}\bigr), \tag{27}$$

where κ is given by (16) or approximated by (21), depending on the specification of λ_i. Draws from this distribution can be generated by the Metropolis–Hastings algorithm, provided that a proposal distribution is used which connects any two elements of $\mathcal{V}$.

Such a proposal distribution will now be given, denoting the proposal probabilities by $u(\tilde{\underline{v}}|\underline{v})$ and the target probabilities, which are proportional to (27), by $p(\underline{v})$. Then the acceptance probabilities in the Metropolis–Hastings algorithm, for a current state $\underline{v}$ and a proposed state $\tilde{\underline{v}}$, are

$$\min\left\{1, \frac{p(\tilde{\underline{v}})u(\underline{v}|\tilde{\underline{v}})}{p(\underline{v})u(\tilde{\underline{v}}|\underline{v})}\right\}. \tag{28}$$

A proposal distribution is used that consists of small changes in $\underline{v}$. The construction of the proposal distribution was based on the considerations that the proposal distribution should mix well in the set of all sample paths, and the Metropolis–Hastings ratios in (28) should be computable relatively easily. This led to proposal distributions consisting of the following types of small changes in $\underline{v}$:

1. *Paired deletions*. Of all pairs of indices r_1, r_2 with $(i_{r_1}, j_{r_1}) = (i_{r_2}, j_{r_2})$, $i_{r_1} \neq j_{r_1}$, one pair is randomly selected, and (i_{r_1}, j_{r_1}) and (i_{r_2}, j_{r_2}) are deleted from $\underline{v}$.
2. *Paired insertions*. A pair $(i, j) \in \mathcal{N}^2$ with $i \neq j$ is randomly chosen, two indices r_1, r_2 are randomly chosen, and the element (i, j) is inserted immediately before r_1 and before r_2.
3. *Single insertions*. At a random place in the path (allowing beginning and end), the element (i, i) is inserted for a random $i \in \mathcal{N}$.
4. *Single deletions*. Of all elements (i_r, j_r) satisfying $i_r = j_r$, a randomly chosen one is deleted.
5. *Permutations*. For randomly chosen $r_1 < r_2$, where $r_2 - r_1$ is bounded from above by some convenient number to avoid too lengthy calculations, the segment of consecutive elements (i_r, j_r), $r = r_1, \ldots, r_2$ is randomly permuted.

It is evident that these five operations yield new sequences within the permitted set $\mathcal{V}$ [cf. (13)]. Paired insertions and paired deletions are each others' inverse operations, and the same holds for single insertions and single deletions. These four types of operation together are sufficient for all elements of $\mathcal{V}$ to communicate. Permutations are added to achieve better mixing properties.

The detailed specification of how these elements are combined in the proposal distribution can be obtained from the authors. Two considerations guided this specification. In the first place, transparency of the algorithm. Paired insertions and paired deletions are not always unique inverses of each other. Nonuniqueness of the inverse operation would lead to complicated counting procedures to determine proposal probabilities. Therefore, the restriction is made that pairs of elements (i, j) can only be inserted at, or deleted from a pair of positions in the sequence, if there are no occurrences of the same (i, j) between these positions; and only if at least one other (i', j') (with $i' \neq i$ or $j' \neq j$) occurs in between these positions. In this restricted set of operations, paired insertions and paired deletions are each other's unique inverses, which simplifies proposal probabilities. Due to the presence of permutations, all elements of the space $\mathcal{V}$ still are reachable from any element.

The second consideration is computational efficiency. When proposing that a pair of elements (i, j) is inserted or deleted at certain positions, then also proposing to permute a segment between those positions entails no increase in computational load for calculating the Metropolis–Hastings ratios, and this permutation will lead to larger changes in $\underline{v}$, and thereby, hopefully, better mixing for a given number of computations.

3.4. *Putting it together.* This subsection combines the elements presented in the preceding subsections to define an algorithm for MCMC approximation of the ML estimate. It is now assumed that an arbitrary number $M \geq 2$ of repeated observations has been made: $x(t_1), x(t_2), \ldots, x(t_M)$. As argued before, we condition on the first observation $x(t_1)$. The further data is denoted by $x = (x(t_2), \ldots, x(t_M))$. The parameter (α, β) is denoted by θ.

The Markov property entails that the observed data score function, conditional on $X(t_1) = x(t_1)$, can be decomposed as

$$\begin{aligned} &S_{X|X(t_1)}(\theta; x(t_2), \ldots, x(t_M)|x(t_1)) \\ &\quad = \sum_{m=2}^{M} S_{X(t_m)|X(t_{m-1})}(\theta; x(t_m)|x(t_{m-1})), \end{aligned} \tag{29}$$

where $S_{X(t_m)|X(t_{m-1})}$ is the score function based on the conditional distribution of $X(t_m)$, given $X(t_{m-1})$.

For the period from t_{m-1} to t_m, the data set is augmented by V_m, which defines the order in which ties are changed between these time points, as described in Section 3.1; this can be denoted by

$$V_m = ((I_{m1}, J_{m1}), \ldots, (I_{mR_m}, J_{mR_m}))$$

with outcome v_m. The augmenting variable as used in Section 3.1 is $V = (V_2, \ldots, V_M)$. Denote the probability (14) for the period from t_{m-1} to t_m by

$p_m(v_m; \theta | x(t_{m-1}))$, and the corresponding total data score function by

$$S_m(\theta; x(t_{m-1}), v_m) = \frac{\partial \log p_m(v_m; \theta | x(t_{m-1}))}{\partial \theta}. \tag{30}$$

From (22) and (29), and using the Markov property, it can be concluded that the observed data score function now can be written as

$$\begin{aligned} &S_{X|X(t_1)}(\theta; x | x(t_1)) \\ &\quad = \sum_{m=2}^{M} E_\theta \{ S_m(\theta; x(t_{m-1}), V_m) | X(t_{m-1}) = x(t_{m-1}), X(t_m) = x(t_m) \}. \end{aligned} \tag{31}$$

The ML estimate is obtained as the value of θ for which (31) equals 0 [cf. (23)].

The algorithm is iterative, and the Nth updating step now can be represented as follows:

1. For each $m = 2, \ldots, M$, make a large number of the Metropolis–Hastings steps as described in Section 3.3, yielding $v^{(N)} = (v_2^{(N)}, \ldots, v_M^{(N)})$.
2. Compute
$$S_{XV}\bigl(\hat{\theta}^{(N)}; x, v^{(N)}\bigr) = \sum_{m=2}^{M} S_m\bigl(\hat{\theta}^{(N)}; x(t_{m-1}), v_m^{(N)}\bigr),$$
using (30) with $p_m(v_m; \theta | x(t_{m-1}))$ defined by (14) for the period from t_{m-1} to t_m, and using (15) and (16) or (21).
3. Update the provisional parameter estimate by
$$\hat{\theta}^{(N+1)} = \hat{\theta}^{(N)} + a_N D^{-1} S_{XV}\bigl(\hat{\theta}^{(N)}; x, v^{(N)}\bigr).$$

As mentioned in Section 3.2, the estimate $\hat{\theta}_{\mathrm{ML}}$ is calculated as a tail average of the values $\hat{\theta}^{(N)}$ generated by this algorithm. The covariance matrix of the ML estimator is estimated using (25), where the expected values in the right-hand side are approximated by Monte Carlo simulation of the conditional distribution of V given $X = x$, for $\theta = \hat{\theta}_{\mathrm{ML}}$. This involves the matrix of partial derivatives (26), which can be estimated by a score-function method as elaborated in Schweinberger and Snijders (2006).

The main implementation details are the following:

a. The Method of Moments (MoM) estimator [Snijders (2001)], in practice, yields a good initial value $\hat{\theta}^{(1)}$.
b. To avoid long burn-in times for step (1.), the Metropolis–Hastings algorithm for generating $V^{(N)}$ can be started from the previous value, $V^{(N-1)}$, rather than independently. This leads to some autocorrelation in the updates defined in step (3.), and the number of Metropolis–Hastings steps must be large enough that this autocorrelation is not too high. It may be noted that the Robbins Monro algorithm is robust to moderate dependence between successive updates [Kushner and Yin (2003)]. We have found good results when the number of steps is tuned so that the autocorrelations between the elements of $\hat{\theta}^{(N)}$ are less than 0.3.

c. For D in step (3.) we use a Monte Carlo estimate of the complete data observed Fisher information matrix (26), estimated for $\theta = \hat{\theta}^{(1)}$ before making the iteration steps.
d. For the other numerical parameters of the algorithm we use the same values as described, for the stochastic approximation algorithm to compute the MoM estimator, in Snijders (2001).

This algorithm is implemented in Siena version 3.3 [Snijders et al. (2009)]. The executable program as well as the code can be found at http://www.stats.ox.ac.uk/siena/.

4. Empirical example. As an illustrative example, the data set is used that was also analyzed in van de Bunt, van Duijn and Snijders (1999) and in Snijders (2001). This is a friendship network between 32 freshman students in a given discipline at a Dutch university. Initially most of the students were unknown to each other. There were six waves denoted $t_1 - t_6$ of data collection, with 3 weeks between waves for the start of the academic year, and 6 weeks in between later. The relation studied is "being friends or close friends." The data set is obtainable from website http://www.stats.ox.ac.uk/siena/.

The transitions between observations t_2 to t_3, and t_3 to t_4, will here be studied separately. To identify the rate function, we assume (arbitrarily but conveniently) that the duration of the time periods is unity, $t_3 - t_2 = t_4 - t_3 = 1$. To keep the model specification relatively simple, the rate function is supposed to be constant across actors, given by α_1 from t_2 to t_3 and by α_2 from t_3 to t_4; and the objective function (6) is chosen independent of the previous state x^0, containing contributions of the effects of outdegree, number of reciprocated ties, number of transitive triplets, number of 3-cycles, gender of the sender of the tie ("ego"), gender of the receiver of the tie ("alter"), and gender similarity:

$$\begin{aligned} f_i(\beta, x) = {} & \beta_1 \sum_j x_{ij} + \beta_2 \sum_j x_{ij} x_{ji} \\ & + \beta_3 \sum_{j,k} x_{ij} x_{jk} x_{ik} + \beta_4 \sum_{j,k} x_{ij} x_{jk} x_{ki} + \beta_5 \sum_j x_{ij} (z_j - \bar{z}) \\ & + \beta_6 \sum_j x_{ij} (z_i - \bar{z}) + \beta_7 \sum_j x_{ij} \{1 - |z_i - z_j|\}, \end{aligned}$$

where variable z_i indicates the gender of actor i ($F = 0$, $M = 1$) and $\bar{z}$ its average over the 32 individuals. This model illustrates two types of triadic dependence (transitive triplets and 3-cycles) and the use of covariates (gender).

For this model the parameters are estimated for the two transitions $t_2 - t_3$ and $t_3 - t_4$ separately both by the Method of Moments (MoM) and by Maximum Likelihood. For models where the functions s_{ik} used in (6) depend only on x and not on x^0, so that they can be expressed as $s_{ik}(x)$, the MoM estimator as defined by

Snijders (2001) is based on the vector with components $\sum_{i,j} |X_{ij}(t_m) - X_{ij}(t_{m-1})|$ for $m = 2, \ldots, M$ and $\sum_{m=2}^{M} \sum_i s_{ik}(X(t_m))$ for $k = 1, \ldots, L$. The MoM estimator is defined as the parameter vector for which the observed and expected values of this statistic are equal, and can be determined by stochastic approximation.

Both the moment equation and the likelihood equation can be represented as $E_\theta S = 0$, where S is the difference between observed and estimated moments or the complete-data score function, respectively. For both estimators, after running the stochastic approximation algorithm, convergence was checked by simulating the model for the obtained parameters for 2000 runs and calculating, for each component S_h of S, the ratio of the average simulated S_h to its standard deviation. In all cases, this ratio was less than 0.1, indicating adequate convergence.

Parameter estimates and standard errors are reported in Table 1. Assuming that the estimators are approximately normally distributed (which is supported by the simulations reported in the next section, although we have no proof), the significance can be tested by referring the ratios of estimate to standard error to a standard normal distribution. The Method of Moments and Maximum Likelihood estimates are different but lead to the same substantive conclusions. The parameters reflecting network structure, β_1 to β_4, give the same picture for both transitions. The negative $\hat{\beta}_1$ indicates that an outgoing tie which is not reciprocated and not transitively embedded is not considered attractive; the positive $\hat{\beta}_2$ and $\hat{\beta}_3$ indicate that there is evidence for tendencies toward reciprocation and transitivity of friendship choices, when controlling for all other effects in the model. For interpreting the 3-cycles effect, note that closed 3-cycles are structures denying a hierarchically ordered relation. The negative $\hat{\beta}_4$ indicates a tendency away from closed 3-cycles, which—in

TABLE 1
Parameter estimates (*Method of Moments and Maximum Likelihood*) *for data set of van de Bunt, van Duijn and Snijders* (1999)

	Effect	**(t_2, t_3)**				**(t_3, t_4)**			
		MoM		**ML**		**MoM**		**ML**	
		Est.	**S.E.**	**Est.**	**S.E.**	**Est.**	**S.E.**	**Est.**	**S.E.**
Rate function									
α	Rate parameter	3.95	0.64	2.61	0.36	3.43	0.59	5.67	0.75
Objective function									
β_1	Outdegree	−1.66	0.26	−1.02	0.26	−2.19	0.30	−2.00	0.22
β_2	Reciprocated ties	2.06	0.47	1.94	0.39	2.26	0.55	1.70	0.35
β_3	Transitive triplets	0.30	0.06	0.18	0.06	0.36	0.07	0.26	0.05
β_4	3-cycles	−0.59	0.23	−0.42	0.17	−0.59	0.27	−0.31	0.14
β_5	Gender alter	0.28	0.30	0.14	0.36	0.70	0.39	0.61	0.33
β_6	Gender ego	−0.34	0.32	−0.41	0.41	−0.04	0.38	−0.10	0.33
β_7	Gender similarity	0.33	0.30	0.34	0.36	0.48	0.37	0.44	0.32

conjunction with the positive transitive triplets parameter—could be interpreted as a tendency toward a local (i.e., triadic) popularity hierarchy in friendship. For gender, there is only a close to significant effect of gender alter for the t_3–t_4 transition, suggesting that male students tend to be more popular as friends, when controlling for the other effects.

From this single data example, of course no conclusions can be drawn concerning the relative value of these two estimation methods.

5. Simulation examples. A comparison between the finite sample behavior of the MoM and the ML estimators can be based on simulations. Here we present a small simulation study as a very limited exploration of the relative efficiency of the two estimators, which is expected to be in favor of the ML estimator. The limited nature of this simulation study does not allow generalization, but the study design is meant to be typical for applications to friendship networks in rather small groups, and replicates approximately the empirical study of the previous section.

The model is identical to that of the previous section: there are three repeated observations, 32 actors, one binary covariate called gender, and the first observed network as well as the distribution of the covariate are identical to the van de Bunt data set at observation t_2. Therefore, the observation moments are again referred to as t_2, t_3, t_4. The parameter values are rounded figures close to the estimates obtained in the preceding section. Networks for times t_3 and t_4 are generated and parameters estimated under the assumption that parameters β_k are the same in periods (t_2, t_3) and (t_3, t_4). The simulation model has parameters $\alpha_1 = 2.5$ and $\alpha_2 = 3.5$, $\beta_1 = -2$, $\beta_2 = 1$, $\beta_3 = 0.2$, $\beta_4 = 0$, $\beta_5 = 0.5$, $\beta_6 = -0.25$, and $\beta_7 = 0.5$. A total of 1000 data sets were generated and the estimates calculated by both methods. Table 2 reports the average estimates, the root mean squared errors, the rejection rates for testing the data-generating value of the parameter as the null hypothesis (estimating type-I error rates), and the rejection rates for testing that the parameters equal 0 (estimating power), where the tests were two-sided tests based on the t-ratio for the corresponding parameter estimate, assuming a standard normal reference distribution, at a nominal significance level of 5%.

Out of the 1000 generated data sets, 5 were excluded because they did not produce well converging results in the default specification of the algorithm for one or both estimators. Table 2 shows that the results for the two estimators are very similar, and the type-I error rates are close to the nominal value, except for the inflated type-I rates of the ML estimators for the two rate parameters. The latter is related to the skewed distribution of the rate parameter estimators. The correlations between the estimators are more than 0.93 for all coordinates; note that correlations are attenuated due to the stochastic nature of the algorithms. It can be concluded that for this type of model, characterized by 32 actors and 3 waves with rates of change 2.5 and 3.5, with 7 parameters in the objective function, MoM and ML estimation yield quite similar results.

TABLE 2

Simulation results, 3 *waves for* 32 *actors*: *average parameter estimates* (*"Ave"*), *root mean squared errors* (*"RMSE"*), *estimated type-I error rates* (*"α"*), *estimated power* (*"β"*)

	Effect	MoM estimator				ML estimator			
		Ave	**RMSE**	α	β	**Ave**	**RMSE**	α	β
Rate function									
$\alpha_1 = 2.5$	Rate $t_2 - t_3$	2.46	0.44	0.063	–	2.37	0.44	0.114	–
$\alpha_2 = 3.5$	Rate $t_3 - t_4$	3.47	0.55	0.045	–	3.39	0.54	0.099	–
Objective function									
$\beta_1 = -2.0$	Outdegree	−2.01	0.15	0.053	1.00	−1.96	0.14	0.047	1.00
$\beta_2 = 1.0$	Reciprocation	1.03	0.26	0.044	0.96	0.97	0.26	0.042	0.95
$\beta_3 = 0.2$	Transitivity	0.186	0.052	0.043	0.94	0.180	0.054	0.064	0.93
$\beta_4 = 0.0$	3-cycles	0.02	0.14	0.042	–	0.03	0.14	0.060	–
$\beta_5 = 0.5$	Gender alter	0.53	0.24	0.043	0.65	0.51	0.25	0.046	0.57
$\beta_6 = -0.25$	Gender ego	−0.28	0.26	0.062	0.21	−0.28	0.27	0.059	0.20
$\beta_7 = 0.5$	Gender sim.	0.52	0.24	0.031	0.64	0.52	0.25	0.052	0.60

To explore data sets with less information, a similar simulation study was conducted with 20 actors, where the first network was induced by the t_2 network of 20 of the actors in the van de Bunt data set, and the rest of the simulation design differed from the previous study by including an interaction effect of reciprocity by gender similarity, represented by the term

$$\beta_8 \sum_j x_{ij} x_{ji} \{1 - |z_i - z_j|\},$$

with parameter $\beta_8 = 0.5$. This effect is included to achieve a higher correlation of the parameter estimators, which together with the smaller number of actors should lead to greater difficulties in estimation.

The results are shown in Table 3. Of the 1000 generated data sets, 20 were excluded from the results because of questionable convergence of the algorithm. The table shows that the ML estimator here performs clearly better than the MoM estimator. The estimated relative efficiency of MoM compared to ML expressed as ratio of mean squared errors ranges for these 10 parameters from 0.50 to 0.99. The correlation between the estimators ranges from 0.71 (gender ego) to 0.98 (rate period 1). For all parameters except β_2, the estimated power of the test based on the ML estimator is higher than that of the MoM-based test, which can be traced back to the combination of a smaller mean squared error and a less conservative test. The surprisingly low power of the MoM-based test for the gender-ego effect reflects high standard errors that tend to be obtained for estimating β_6.

6. Likelihood ratio tests. One of the important advantages of a likelihood approach is the possibility of elaborating model selection procedures. Here we only explain how to conduct a likelihood ratio test.

TABLE 3

Simulation results, 3 waves for 20 actors: average parameter estimates ("Ave"), root mean squared errors ("RMSE"), estimated type-I error rates ("α"), estimated power ("β")

	Effect	MoM estimator				ML estimator			
		Ave	**RMSE**	**α**	**β**	**Ave**	**RMSE**	**α**	**β**
Rate function									
$\alpha_1 = 2.5$	Rate $t_2 - t_3$	2.49	0.50	0.059	–	2.37	0.48	0.093	–
$\alpha_2 = 3.5$	Rate $t_3 - t_4$	3.51	0.70	0.044	–	3.36	0.67	0.094	–
Objective function									
$\beta_1 = -2.0$	Outdegree	−2.09	0.29	0.009	0.98	−2.02	0.21	0.025	1.00
$\beta_2 = 1.0$	Reciprocation	1.08	0.35	0.043	0.86	0.97	0.33	0.052	0.82
$\beta_3 = 0.2$	Transitivity	0.170	0.099	0.022	0.51	0.158	0.096	0.059	0.52
$\beta_4 = 0.0$	3-cycles	0.01	0.23	0.034	–	0.05	0.23	0.057	–
$\beta_5 = 0.5$	Gender alter	0.59	0.44	0.020	0.15	0.60	0.38	0.048	0.43
$\beta_6 = -0.25$	Gender ego	−0.35	0.50	0.024	0.00	−0.34	0.36	0.034	0.11
$\beta_7 = 0.5$	Gender sim.	0.61	0.50	0.013	0.12	0.63	0.43	0.051	0.35
$\beta_8 = 0.5$	G. sim. × rec.	0.47	0.63	0.039	0.10	0.50	0.60	0.043	0.14

A convenient way to estimate the likelihood ratio is by a simple implementation of the idea of the path sampling method described by Gelman and Meng (1998), who took this method from the statistical physics literature where it is known as thermodynamical integration. For arbitrary θ_0 and θ_1, defining the function $\theta(t) = t\theta_1 + (1-t)\theta_0$, with $\dot{\theta} = \partial\theta/\partial t = \theta_1 - \theta_0$, this method is based on the equation

$$\log\left(\frac{p_X(x;\theta_1)}{p_X(x;\theta_0)}\right) = \int_0^1 \dot{\theta}\, S_X(x;\theta(t))\, dt. \tag{32}$$

This integral can be approximated by replacing the integral by a finite sum and using (22). Applying a simulation procedure which for each $\theta(h/H)$, $h = 0, \ldots, H$, generates $L \geq 1$ draws $V_{h\ell}$ from the conditional distribution of V given $X = x$ under parameter $\theta(h/H)$, the log likelihood ratio (32) can be approximated by

$$\frac{(\theta_1 - \theta_0)}{L(H+1)} \sum_{\ell=1}^{L} \sum_{h=0}^{H} S_{XV}\big(\theta(h/H); x, V_{h\ell}\big). \tag{33}$$

Burn-in time can be considerably reduced when starting the MCMC algorithm for generating V_{h1} by the end result of the algorithm used to generate $V_{(h-1),L}$.

This was applied to the van de Bunt data set used in Section 4 to test the null hypothesis that all parameters $\beta_1, \ldots, \beta_7$ are for period (t_3, t_4) the same as for (t_2, t_3) against the alternative that they are allowed to be different. The rate parameters α were allowed to be different under the null as well as the alternative hypothesis. For calculating (33) we used $L = 10$, $H = 1000$. The estimated likelihood ratio was 18.7, which for a χ^2_7 distribution yields $p < 0.01$, leading to rejecting the null hypothesis at conventional levels of significance.

7. Discussion. This article has presented a model for longitudinal network data collected in a panel design and an algorithm for calculating the ML estimator. The model can represent triadic and other complicated dependencies that are characteristic for social networks. It is designed for applications in social network analysis, but related models may be applied for modeling networks in other sciences, such as biology. Earlier, a method of moments (MoM) estimator for this model was proposed by Snijders (2001) and Bayesian inference methods by Koskinen and Snijders (2007). The algorithm was constructed using stochastic approximation according to the approach proposed by Gu and Kong (1998), and employs Monte Carlo simulations of the unobserved changes between the panel waves, conditional on the observed data. The simulation design used is more efficient than that used in Koskinen and Snijders (2007) because here the waiting times between unobserved changes are integrated out.

No proof is yet available for the consistency and asymptotic normality of the ML estimator, which are intuitively plausible for the situation where n tends to infinity, t_1 and t_2 are fixed, and the parameters are such that the average degree $n^{-1}\sum_{i,j} x_{ij}$ tends to a positive finite limit. Limited simulation results do support the expectation that the estimators are asymptotically normal. The nonstandard assumptions (lack of independence) imply, however, that a proof may be expected to be rather complicated.

Our experience in the reported simulations and in working with empirical data sets is that the algorithm converges well unless data sets are small given the number of estimated parameters, but the algorithm is time-consuming (e.g., each ML estimation for one data set in Table 2 took about 35 minutes on a regular personal computer, while each MoM estimation took about 2 minutes). Further improvements in the algorithm are desirable for the practical use of ML estimators in these models. This is the subject of future work.

Judging from our very limited simulations, the advantages of ML estimation over MoM estimation in terms of root mean squared error and power of the associated tests seem strong for small data sets and small for medium to large data sets. Further simulation studies are necessary. However, even if the statistical efficiency is similar, likelihood-based estimation can have additional advantages, for example, the possibility of extensions to more complicated models and of elaborating model selection procedures.

Software implementing the procedures in this paper is available at http://www.stats.ox.ac.uk/siena/.

Acknowledgments. Part of this work was done while M. Schweinberger was working at the University of Groningen and the University of Washington. Part of this work was done while J. Koskinen was working at the University of Stockholm and the University of Melbourne.

REFERENCES

AIROLDI, E., BLEI, D. M., FIENBERG, S. E., GOLDENBERG, A., XING, E. P. and ZHENG, A. X. (2007). *Statistical Network Analysis: Models, Issues and New Directions (ICML 2006)*. *Lecture Notes in Computer Science* **4503**. Springer, Berlin.

CARRINGTON, P. J., SCOTT, J. and WASSERMAN, S., eds. (2005). *Models and Methods in Social Network Analysis*. Cambridge Univ. Press.

DAVIS, J. A. (1970). Clustering and hierarchy in interpersonal relations: Testing two theoretical models on 742 sociograms. *American Sociological Review* **35** 843–852.

DE FEDERICO DE LA RÚA, A. (2003). La dinámica de las redes de amistad. La elección de amigos en el programa Erasmus. *REDES* **4.3** 1–44.

EFRON, B. (1977). Discussion of Dempster, Laird and Rubin (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 29. MR0501537

FISHER, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* **22** 700–725.

FRANK, O. (1991). Statistical analysis of change in networks. *Statist. Neerlandica* **45** 283–293. MR1142085

FRANK, O. and STRAUSS, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.* **81** 832–842. MR0860518

GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. MR1647507

GU, M. G. and KONG, F. H. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo Method for incomplete data estimation problems. *Proc. Natl. Acad. Sci. USA* **95** 7270–7274. MR1630899

GU, M. G. and ZHU, H.-T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *J. Roy. Statist. Soc. Ser. B* **63** 339–355. MR1841419

HANNEKE, S. and XING, E. P. (2007). Discrete temporal models of social networks. In *Statistical Network Analysis: Models, Issues and New Directions (ICML 2006)* (E. Airoldi, D. M. Blei, S. E. Fienberg, A. Goldenberg, E. P. Xing and A. X. Zheng, eds.). *Lecture Notes in Computer Science* **4503** 115–125. Springer, Berlin.

HOLLAND, P. and LEINHARDT, S. (1977). A dynamic model for social networks. *Journal of Mathematical Sociology* **5** 5–20. MR0446596

HUNTER, D. R. and HANDCOCK, M. S. (2006). Inference in curved exponential family models for networks. *J. Graph. Comput. Statist.* **15** 565–583. MR2291264

KRACKHARDT, D. and HANDCOCK, M. S. (2007). Heider vs Simmel: Emergent features in dynamic structures. In *Statistical Network Analysis: Models, Issues and New Directions (ICML 2006)*. (E. Airoldi, D. M. Blei, S. E. Fienberg, A. Goldenberg, E. P. Xing and A. X. Zheng, eds.) *Lecture Notes in Computer Science* **4503** 14–27. Springer, Berlin.

KOSKINEN, J. H. and SNIJDERS, T. A. B. (2007). Bayesian inference for dynamic network data. *J. Statist. Plan. Inference* **137** 3930–3938. MR2368538

KUSHNER, H. J. and YIN, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. Springer, New York. MR1993642

LEENDERS, R. T. A. J. (1995). Models for network dynamics: A Markovian framework. *Journal of Mathematical Sociology* **20** 1–21.

LOUIS, T. A. (1982). Finding observed information when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44** 226–233. MR0676213

MADDALA, G. S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge Univ. Press. MR0799154

NORRIS, J. R. (1997). *Markov Chains*. Cambridge Univ. Press. MR1600720

ORCHARD, T. and WOODBURY, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings Sixth Berkeley Sympos. Math. Statist. Probab.* **1** 697–715. Univ. California Press, Berkeley. MR0400516

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22** 400–407. MR0042668

ROBINS, G. and PATTISON, P. (2001). Random graph models for temporal processes in social networks. *Journal of Mathematical Sociology* **25** 5–41.

SCHWEINBERGER, M. and SNIJDERS, T. A. B. (2006). Markov models for digraph panel data: Monte Carlo-based derivative estimation. *Comput. Statist. Data Anal.* **51** 4465–4483. MR2364459

SINGER, H. (2008). Nonlinear continuous time modeling approaches in panel research. *Statist. Neerlandica* **62** 29–57. MR2387645

SNIJDERS, T. A. B. (2001). The statistical evaluation of social network dynamics. In *Sociological Methodology—2001* (M. E. Sobel and M. P. Becker, eds.) 361–395. Blackwell, London.

SNIJDERS, T. A. B. (2006). Statistical methods for network dynamics. In *Proceedings of the XLIII Scientific Meeting, Italian Statistical Society*, (S. R. Luchini et al. eds.) 281–296. CLEUP, Padova.

SNIJDERS, T. A. B. and VAN DUIJN, M. A. J. (1997). Simulation for statistical inference in dynamic network models. In *Simulating Social Phenomena* (R. Conte, R. Hegselmann and P. Terna, eds.) 493–512. Springer, Berlin.

SNIJDERS, T. A. B., PATTISON, P. E., ROBINS, G. L. and HANDCOCK, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology* **36** 99–153.

SNIJDERS, T. A. B., STEGLICH, C. E. G., SCHWEINBERGER, M., HUISMAN, M. (2009). Manual for SIENA version 3.2. Dept. Statistics, Univ. Oxford. Univ. Groningen, ICS. Available at http://www.stats.ox.ac.uk/siena/.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

VAN DE BUNT, G. G., VAN DUIJN, M. A. J. and SNIJDERS, T. A. B. (1999). Friendship networks through time: An actor-oriented statistical network model. *Computational and Mathematical Organization Theory* **5** 167–192.

VAN DUIJN, M. A. J., ZEGGELINK, E. P. H., HUISMAN, M., STOKMAN, F. M. and WASSEUR, F. W. (2003). Evolution of sociology freshmen into a friendship network. *Journal of Mathematical Sociology* **27** 153–191.

WASSERMAN, S. (1979). A stochastic model for directed graphs with transition rates determined by reciprocity. *Sociological Methodology—1980* (K. F. Schuessler, ed.) 392–412. Jossey-Bass, San Francisco.

WASSERMAN, S. (1980). Analyzing social networks as stochastic processes. *J. Amer. Statist. Assoc.* **75** 280–294.

WASSERMAN, S. and FAUST, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge Univ. Press.

WASSERMAN, S. and PATTISON, P. (1996). Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika* **61** 401–425. MR1424909

T. A. B. SNIJDERS
J. KOSKINEN
NUFFIELD COLLEGE
NEW ROAD, OXFORD OX1 1NF
UNITED KINGDOM
E-MAIL: tom.snijders@nuffield.ox.ac.uk
E-MAIL: johan.koskinen@nuffield.ox.ac.uk

M. SCHWEINBERGER
DEPARTMENT OF STATISTICS
PENN STATE UNIVERSITY
326 THOMAS BUILDING
UNIVERSITY PARK, PENNSYLVANIA 16802
USA
E-MAIL: michael.schweinberger@stat.psu.edu