MI

Goodness of fit for stochastic actor-oriented models

Josh Lospinoso[1] and Tom AB Snijders[2,3] (iD)

Abstract
We propose a Mahalanobis distance–based Monte Carlo goodness of fit testing procedure for the family of stochastic actor-oriented models for social network evolution. A modified model distance estimator is proposed to help the researcher identify model extensions that will remediate poor fit. A limited simulation study is provided to establish baseline legitimacy for the Mahalanobis distance–based Monte Carlo test and modified model distance estimator. A forward model selection workflow is proposed, and this procedure is demonstrated on a real data set.

Keywords
Stochastic actor-oriented model, goodness of fit, social network analysis, model selection

Introduction

Social networks define an important type of data structure that has been gaining attention in recent years. Social networks are dyadic relations between social actors (e.g. individuals, organizations, or teams). Important examples are affective relationships such as trust and friendship, or task-oriented relationships like advice seeking. Networks are commonly represented by directed graphs or more complicated structures, where the nodes in the graph represent the social actors and the arcs (directed) or edges (non-directed) represent the ties between them. A classical textbook is Wasserman and Faust (1994), and recent textbooks are, for example, Kolaczyk (2009) and Borgatti et al. (2018).

For the study of network dynamics, Snijders (2001) introduced stochastic actor-oriented models (SAOMs), and proposed estimators for parameters in these models given network panel data using the method of moments (MoM). Bayesian estimators were proposed by Koskinen and Snijders (2007) and maximum likelihood estimators by Snijders et al. (2010a). The model was extended to the coevolution of networks and actor-level variables (Snijders et al., 2007; Steglich et al., 2010) and to the evolution of multivariate networks (Snijders et al., 2013). It is implemented in the R package RSiena (Ripley et al., 2019) and has been widely applied in the social sciences (e.g. Veenstra et al., 2013). An overview is given in Snijders (2017).

Compared with these developments, issues of model selection have not been elaborated in as much detail. Model selection for SAOMs usually proceeds guided by research questions, and utilizing substantive knowledge combined with forward model selection and hypothesis testing (Lospinoso et al., 2011; Schweinberger, 2012). Some guidelines were given in Snijders et al. (2010b), but not propped with formalized methods. This article proposes methods for assessing goodness of fit (GOF) of estimated models, thereby complementing the existing methodology.

In the statistical tradition of GOF (cf. Lehmann and Romano, 2005), we consider evidence as to whether the observed data is consonant with the assumption that it came from the fitted model under study. Hypothesis testing focuses on a *specific* alternative hypothesis. In contrast, the proposed GOF method has no particular alternative in mind. It evaluates how well the model fits *in general*. Because of the complex

[1]Shift5, Rosslyn, VA, USA
[2]Department of Sociology, University of Groningen, Groningen, The Netherlands
[3]Department of Statistics, Nuffield College, University of Oxford, Oxford, UK

Corresponding author:
Tom AB Snijders, Department of Sociology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands.
Email: Tom.Snijders@nuffield.ox.ac.uk

nature of network data, we do not think an omnibus GOF test is feasible. Therefore, we focus on a wide set of features for which a good fit between model and data is desirable, and study the correspondence for these features by Monte Carlo methods, much as was done by Hunter et al. (2008) for GOF studies for non-longitudinal network modeling.

We propose a data-driven methodology for assessing the fit between data and model based on a set of features that may be chosen by the user, but for which we make some suggestions, and we propose methods assisting the data analyst to find the directions, within a larger set of options, that seem most promising for improving the model fit if it is found to be lacking. These empirical considerations must be used in practice together with considerations based on substantive theories and whatever is available as knowledge of the processes determining network change; data-driven procedures can supplement, but not supplant, subject-matter knowledge.

The ideal for network modeling is that by specifying rules for network formation based on local information, the global properties of the network will also be represented in a satisfactory way. Local properties depend on neigbourhoods of the nodes, that is, other nodes directly connected (by incoming or outgoing ties) or connected through at most one intermediary. Examples of local properties are reciprocation, characteristics of tied nodes, and the frequency of various triadic configurations (Holland and Leinhardt, 1976). An example of a global property is the frequency distribution of geodesic distances, the geodesic distance between two nodes being defined as the minimum length of a path connecting them. Thus, the stochastic models will be based on local information, while the features used for checking will be based on local as well as global information.

The estimation method mainly used for SAOMs is the MoM, because it is much less time-consuming than likelihood-based approaches. Fit comparison based on likelihoods, therefore, is not available in many practical situations, and will not be considered here. Instead, fit assessment will be based on comparing features of the observed network to their expected values in the estimated distribution of networks, estimated from a large number of simulations. When this assessment leads to a conclusion of poor fit, it is desirable to remediate this by proposing model elaborations. Estimation is computationally intensive, and testing large numbers of candidate models can be time-consuming. Therefore, we also propose a computationally cheap predictor for the improvement of fit if the model were to be extended by specific additional effects. This estimator can be evaluated using only ingredients calculated already for the MoM estimation of the restricted model.

This article proceeds by first providing a brief introduction to SAOMs. We then present some possibilities for GOF features, which we call *auxiliary statistics*. Next, we propose a Monte Carlo–based GOF test based on the auxiliary statistics. Most interesting auxiliary statistics are multi-dimensional; they are therefore combined using the Mahalanobis distance (Mahalanobis distance–based Monte Carlo (MDMC)). A computationally cheap estimator for the Mahalanobis distance that would be obtained from specific model extensions, based on a first-order Taylor series, is developed to give suggestions to the researcher for choosing model extensions to remediate lack of fit, the so-called *modified model distance* (MMD) estimator. A set of simulation studies is conducted to demonstrate the effectiveness of both the GOF test and of the MMD estimator. The proposed GOF procedure is summarized in a workflow. This workflow is demonstrated in a brief forward model selection exercise. The article concludes with a discussion on some future directions.

This MDMC approach was proposed in the conference paper Lospinoso and Satchell (2011) and the unpublished DPhil dissertation Lospinoso (2012), and has already been widely applied. However, until now it was not formally described in a publication. The MMD estimator was not yet described or applied in other publications (except for the mentioned dissertation). The simulation studies are new.

Stochastic actor-oriented models

We present the model as developed in Snijders (2001); see Snijders (2017) for a treatment focusing on statistical issues and Snijders et al. (2010b) for a friendly introduction. A social network composed of n actors is modeled as a directed graph (digraph) with nodes $1,\ldots,n$, represented by an adjacency matrix $(x_{ij})_{n \times n}$, where $x_{ij} = 1$ if there is a tie from actor i to actor j, $x_{ij} = 0$ if there is no such tie, and $x_{ii} = 0$ for all i (self-ties are not permitted). For the tie variables x_{ij}, node i is called the *sender* and j the *receiver* of the tie; senders are often referred to as "ego" and receivers as "alter." The number $\sum_j x_{ij}$ of outgoing ties of actor i is called the *outdegree* and the number $\sum_j x_{ji}$ of incoming ties is the *indegree* of this actor. Well-known features of social networks are tendencies toward reciprocation of ties, and toward transitive network closure, signifying the tendency that if two actors are indirectly tied, that is, through an intermediary, they will probably also be directly tied. Network closure will lead to the existence of cohesive subgroups in the network, with internally many ties but relatively fewer ties outside; mostly, however, such subgroups are not very clearly separated.

The network is observed at discrete time points called *observations* occurring at times $t_1, t_2, \ldots, t_M \in \mathcal{T}$, for some $M \geq 2$. Intervals between successive observations are called *periods*, and the *m*th period is the interval $\{t \in \mathcal{T} : t_m \leq t \leq t_{m+1}\}$.

The model assumes that the social network evolves in continuous time over an interval $\mathcal{T} \subset \mathbb{R}$. Accordingly, the digraph $x(t)$ models the state of social relationships at time $t \in \mathcal{T}$, unobserved unless t is one of the observation times

t_m. The actors, represented by the nodes of the network, are assumed to get opportunities for changing one of their outgoing ties at discrete time points, and such opportunities are called *ministeps*.

The stochastic process $\{X(t):t \in \mathcal{T}\}$ is modeled as a Markov process on the space of all digraphs so that for any time $t^* \in \mathcal{T}$, the conditional distribution for the future $\{X(t):t > t^*\}$ given the past $\{X(t):t \leq t^*\}$ depends only on $X(t^*)$.

The SAOMs consider two principal concepts in constructing the intensity matrix: the frequency at which actors i get opportunities to update one of their outgoing tie variables X_{ij}, and their choice of which tie variable to update. The expected frequency per time unit for actor i to get an opportunity for change is called the *rate function*, and is denoted as $\lambda_i(x)$. Waiting times between opportunities for actor i to make an update to the digraph are exponentially distributed with parameter $\lambda_i(x)$, but another actor might make a change earlier, which then might change the value of x and of $\lambda_i(x)$. Therefore, the definition is that waiting times between opportunities for change by any actor are exponentially distributed with rate parameter

$$\lambda_+\left(x(t)\,|\,\alpha\right) = \sum_i \lambda_i\left(x(t)\,|\,\alpha\right) \tag{1}$$

where α is a parameter, and the conditional probability that it is actor i who has an opportunity for change, if some actor has this opportunity, is

$$\frac{\lambda_i\left(x(t)\,|\,\alpha\right)}{\lambda_+\left(x(t)\,|\,\alpha\right)} \tag{2}$$

Rate functions may depend on actor-level covariates and the network position of actors. For simplicity, we here consider rate functions depending on neither i nor x, but only on the period, that is, $\lambda_i(x) = \rho_m$ for $t_m < t \leq t_{m+1}$.

Now suppose that an opportunity for change arises for actor i, while the current digraph is x. Define $x^{(\pm ij)} \in \mathcal{X}$ as the digraph in which, starting from digraph x, only the binary variable x_{ij} is toggled. In other words, $x_{ij}^{(\pm ij)} = 1 - x_{ij}$ and $x_{hk}^{(\pm ij)} = x_{hk}$ for all $(h,k) \neq (i,j)$, and formally define $x^{(\pm ii)} = x$ for all i. The probability that actor i selects x_{ij} as the tie variable to change, so that the next state will be $x^{(\pm ij)}$, is denoted as $p_{ij}(x)$. Actors are not required to make a change when an opportunity occurs, which is reflected by the requirement $\sum_{j \neq i} p_{ij}(x) \leq 1$, without the need for this to be equal to 1. The probabilities $p_{ij}(x)$ are dependent on the so-called *evaluation function*. This function gives an evaluation of the attraction toward each possible next state of the network, and is denoted by $f_i(x\,|\,\beta)$. This attraction is conveniently modeled as a linear combination of the relevant features of the network

$$f_i\left(x\,|\,\beta\right) = \beta^T s_i\left(x\right) \tag{3}$$

where s_i is a vector-valued function containing structural features of the digraph as seen from the point of view of actor i and covariate effects, and β is a parameter vector. The components of s_i are called *effects*.

The selection of the tie variable x_{ij} to be changed in a ministep by actor i is modeled by a conditional logit model. This leads to conditional choice probabilities $p_{ij}(x\,|\,\beta)$ given by

$$p_{ij}\left(x\,|\,\beta\right) = \frac{\exp\left\{f_i\left(x^{(\pm ij)}\,|\,\beta\right)\right\}}{\displaystyle\sum_{k=1}^{n} \exp\left\{f_i\left(x^{(\pm ik)}\,|\,\beta\right)\right\}} \tag{4}$$

In accordance with the formal definition $x^{(\pm ii)} = x$, the choice $j = i$ is interpreted as keeping the current digraph as it is, without making a change. The probabilities (equation 4) can be obtained from the assumption that actor i acts according to myopic stochastic optimization of the evaluation function for the state that will be reached by this change (Snijders, 2001). The myopic stochastic optimization can of course not be more than an "as if" assumption; it is a sufficient, but not a necessary condition for the choice probabilities given by equation (4).

A basic menu of effects s_i is given in Snijders et al. (2010b); the full set available in the R package RSiena is defined in Ripley et al. (2019). An example of some effects, used in the example, is the following. The formula for the component $s_{ki}(x)$ of $s_i(x)$ is given, with a minimal subgraph illustrating the component:

1. Outdegree effect

$$\sum_j x_{ij}$$

represents the tendency for actors to have outgoing ties. This parameter is analogous to a constant term in regression models. In practice, it is always included in the model to fit the trend in the total number of ties.

2. Reciprocity effect

$$\sum_j x_{ij} x_{ji}$$

represents the tendency for actors to reciprocate incoming ties with outgoing ties.

3. Transitive triplets

$$\sum_{j,h} x_{ij} x_{ih} x_{hj}$$

represents the tendency for an actor to create transitively closed structures, where "friends of friends are friends."

4.   Geometrically weighted shared partners ("gwesp")

$$\sum_{j=1}^{n} x_{ij} e^{\alpha} \left\{ 1 - \left(1 - e^{-\alpha} \right)^{t_{ij}} \right\}$$
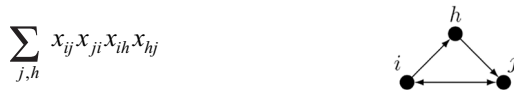
where

$$t_{ij} = \sum_{h} x_{ih} x_{hj}$$

is another representation of the tendency to create transitively closed structures, where the number of indirect connections ("two-paths") t_{ij} has a decreasing marginal effect. In this article, the parameter α is fixed to $\ln(2) \approx 0.69$. See Snijders et al. (2006) and Hunter and Handcock (2006).

5.   Three-cycles

$$\sum_{j,h} x_{ij} x_{jh} x_{hi}$$

represents the tendency for an actor to create closed cycles $i \rightarrow j \rightarrow h \rightarrow i$.

6.   Transitive reciprocated triplets

$$\sum_{j,h} x_{ij} x_{ji} x_{ih} x_{hj}$$

is an interaction between reciprocity and transitivity (see Block, 2015).

7.   Dense triads

$$\sum_{j,h} x_{ij} x_{ji} x_{ih} x_{hi} x_{hj} x_{jh}$$

represents the tendency for an actor to be a part of triads where everybody is mutually connected.

8.   In-degree popularity

$$\sum_{j} x_{ij} \left\{ \sum_{h} x_{jh} \right\}$$

represents the tendency of actors to send ties to other actors with currently high in-degrees.

9.   Out-degree popularity

$$\sum_{j} x_{ij} \left\{ \sum_{h} x_{jh} \right\}$$

represents the tendency of actors to send ties to other actors with currently high out-degrees.

10.   Out-degree activity

$$\sum_{j} x_{ij} \left\{ \sum_{h} x_{ih} \right\}$$

represents the tendency of actors with currently high out-degrees to create new ties.

11.   Reciprocated-degree activity

$$\sum_{j} x_{ij} \left\{ \sum_{h} x_{ih} x_{hi} \right\}$$

represents the tendency of actors with currently many reciprocated ties to create new ties.

12.   Same covariate

$$\sum_{j} x_{ij} \mathbf{1} \left\{ v_i = v_j \right\}$$

for a categorical actor covariate V, where $\mathbf{1}$ is the indicator function, represents the tendency of actors to create ties to other actors that have the same value of this covariate. In the picture, the covariate is binary and indicated by the node color.

13.   Similar covariate

$$\sum_{j} x_{ij} \left(1 - \frac{|v_i - v_j|}{\text{Range}(V)} \right)$$
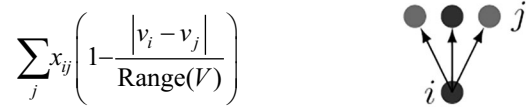
for a numerical actor covariate V represents the tendency of actors to create ties to other actors that have a similar value of this covariate.

Some of these effects can have seemingly similar consequences for the generated networks. For example, transitive triplets, gwesp, dense triads, and the same or similar covariate effects all will lead to some kind of clustering in the network. One of the purposes of the GOF studies is to determine which of these provides the best fit.

The mostly used estimation method for these models is the MoM (Snijders, 2001, 2017). Denoting the parameter by $\theta = (\rho_1, \rho_2, \ldots, \rho_{M-1}, \beta)$, and the observed sequence of

networks by $x(t_m) \, (m=1,\ldots,M)$, the MoM estimate $\hat{\theta}$ is here defined as the solution of

$$\sum_{m=1}^{M-1} E_{\hat{\theta}} \left\{ z\left(X(t_m), X(t_{m+1}) \right) \mid X(t_m) = x(t_m) \right\}$$
$$= \sum_{m=1}^{M-1} z\left(x(t_m), x(t_{m+1}) \right) \quad (5)$$

for a suitable vector of statistics z, which are sensitive to the parameters in θ. For the model specification given here, z may be chosen to contain the statistics

$$\sum_{i,j} |x_{ij}(t_m) - x_{ij}(t_{m+1})| \, (1 \le m < M) \quad (6a)$$

which are sensitive to the rate parameters ρ_m, and

$$\sum_{m=1}^{M-1} \sum_i s_i \left(x(t_{m+1}) \right) \quad (6b)$$

which is sensitive to the parameter β. The solution of equation (5) can be approximated well by stochastic approximation (Snijders, 2001, 2017). This is implemented in the RSiena package (Ripley et al., 2019).

Statistics

The "GOF" problem is one of testing the hypothesis that the model which generated the observed data is equal to the fitted model. Due to (1) the vastness of the state space of networks and (2) the idea that we have a "sample of size 1" observed over time, the approaches that have been developed for standard statistical modeling cannot be applied. Here, we follow the approach proposed by Hunter et al. (2008) and used also by Robins et al. (2009) for assessing the GOF of exponentially random graph models representing cross-sectionally observed networks. Many of the statistics proposed below are adopted from, or inspired by their work. This approach takes one *auxiliary statistic*, which is a vector of features of the data that is not directly included in the model, that is, it is not a function of the estimation statistics (equation 6). The value of this auxiliary statistic is compared between the observed data and their distribution as implied by the SAOM with the estimated parameters. In practice, usually several auxiliary statistics will be considered; the explanation of the procedure is for one vector-valued auxiliary statistic.

Before proposing the form of the GOF test, we present a number of possible auxiliary statistics. This list is by no means exhaustive, but we provide concrete examples which will be used in the simulation study and the example application later in this article:

Triad census. There are 16 possible isomorphic subgraphs of three nodes and the $3 \times 2^4 = 48$ possible edge configurations between them, as illustrated in Figure 1.
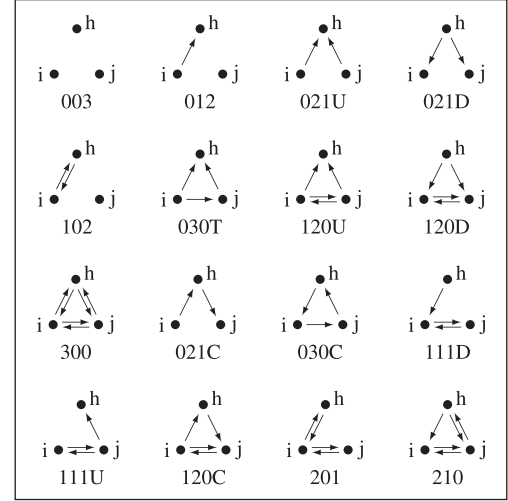

Figure 1. The 16 possible triads for transitivity in a digraph, adapted from Holland and Leinhardt (1976). The Mutual/Asymmetric/Null (MAN) notation for each triad is also given below each figure. For certain MAN classes, for example, 030C, a letter is appended to make the notation unique.

The vector of frequencies of these subgraphs in the network is called the *triad census*. Holland and Leinhardt (1976) proposed to study this as the "local network structure." Further interpretations can be found, for example, in Borgatti et al. (2018).

The triad census can be used to assess whether the nuances of local network structure, such as *network closure* (i.e. transitivity)—a fundamental feature of social networks—are accurately represented by the fitted model.

Edgewise shared partners (ESP). We can also consider, for a given C, the vector of statistics $A_E(x) = (A_{E1}(x), A_{E2}(x), \cdots, A_{EC}(x))^T$ containing elements

$$A_{Ec}(x) = \sum_{i,j} x_{ij} \mathbf{1} \left\{ \sum_{k \ne i, j} x_{ik} x_{jk} = c \right\} \quad (7)$$

where $\mathbf{1}\{.\}$ is the indicator function. These elements count the number of node pairs which share c outgoing partners. The ESP counts will help to capture the importance of *redundancy* (i.e. multiple indirect connections) in network closure; this feature is not represented directly in the triad census. The use of ESP counts originates from the work of Hunter and Handcock (2006) and Snijders et al. (2006).

Outdegree distribution. This is the vector of statistics $A_D(x) = (A_{D1}(x), A_{D2}(x), \cdots, A_{DC}(x))^T$ given by

$$A_{Dc}(x) = \sum_i \mathbf{1} \left\{ \sum_k x_{ik} = c \right\} \quad (8)$$

These statistics count the number of nodes with c outgoing ties. In the social networks literature, this is often interpreted as the *activity* of the nodes. The dimension C of this statistic will be chosen based on the observed network and the interests of the researcher. While average outdegree is modeled explicitly by virtually all SAOM models used in practice, the distribution of outdegrees can have many different shapes, and these will not automatically be represented well by an estimated model.

Indegree distribution. This is the vector of statistics $A_P(x) = (A_{P1}(x), A_{P2}(x), \cdots, A_{PC}(x))^T$, defined by

$$A_{Pc}(x) = \sum_j \mathbf{1}\left\{\sum_k x_{kj} = c\right\} \tag{9}$$

These elements count the number of nodes with c incoming ties, in the social networks literature often interpreted as *popularity*. Indegree and outdegree distributions are quite distinct from one another, and should be considered separately.

Geodesic distances. Define $G_{ij}(x)$, the geodesic distance, as the length of the shortest path between nodes i and j in the graph. This path length may be defined with or without taking into account directionality of edges. The geodesic distances are in the vector of statistics $A_G(x) = (A_{G1}(x), A_{G2}(x), \cdots, A_{GC}(x))^T$, containing elements

$$A_{Gc}(x) = \sum_j \mathbf{1}\left\{G_{ij}(x) = c\right\} \tag{10}$$

The distribution of geodesic distances is an *emergent* property of social networks which is important, for example, for how quickly ideas and norms can spread. Importantly, geodesic distance is among the statistics presented here the only non-local characteristic of the network.

Edgewise similarity. Covariates are denoted $V(i)$ where i is the actor, and lead to statistics such as

$$A_V = \sum_{i,j} x_{ij} \mathbf{S}\left(V(i), V(j)\right) \tag{11}$$

where $\mathbf{S}(v, v')$ is a function representing the similarity between values v and v'. This function be specified in several ways depending on the type of covariate. For binary covariates, the function

$$\mathbf{S}(v, v') = \mathbf{1}\{v = v'\}$$

is a sensible choice. For numerical covariates, a transformed mean absolute difference can be used, as in

$$\mathbf{S}(v, v') = 1 - \frac{|v - v'|}{\mathrm{Range}(V)}$$

The choice of similarity function should depend on substantive theory and field knowledge, and on which aspects of fit are deemed important.

A Monte Carlo Mahalanobis distance based GOF test

In the previous section, we outlined a number of important features of digraph and covariate data that can be used for assessing GOF. From these auxiliary statistics, we now focus on only one GOF statistic $A(x)$, but we must still consider how to make inference about the alignment of this statistic between the data and the fitted model under investigation with estimated parameter $\hat{\theta}$. Note that the statistic $A(x)$ should be quite distinct from the functions $s_i(x)$ used to model the network (see equations (3) and (6)), as we wish to assess whether the model also is adequate in the representation of features of the network that are not explicitly used to define the model.

Recall that we assume to have panel data, that is, a sequence of observed networks $x(t_1), x(t_2), \ldots, x(t_M)$ for $M \geq 2$. The approach to analyzing such a sequence advocated by Snijders (2001) is to consider each observation conditional on the previous one. The testing problem, therefore, is to compare the observed value of the auxiliary statistic $A(x(t_{m+1}))$ to the estimated conditional distribution of $A(X(t_{m+1}))$ given $x(t_m)$, for $1 \leq m \leq M-1$. This conditional distribution is not analytically tractable, and we handle this by taking recourse to Monte Carlo simulation. As we shall see, it will not be required to conduct additional simulations beyond what is already required for MoM estimation.

For simplicity, we deal with a single period. Since the estimation conditions on $x(t_1)$, this part of the observations is fixed and the only random part is $x(t_2)$, which for notational simplicity we denote by x. We will make use of expectations E_θ over the distribution implied by θ of $X(t_2)$ given $x(t_1)$. For transparency, we also drop from the notation the conditioning on $x(t_1)$ and denote $X(t_2)$ by X. The case of multiple periods is an important elaboration; however, at present we will concern ourselves with the case of one period. We return below to the issue of multiple periods.

The approach taken here is to construct a function D that, for a given vector of statistics $A(X)$, measures some notion of distance between the observed $A(x)$ and the distribution of the statistics $A(X)$ where X has the distribution implied by θ. This raises the question of how to map the multidimensional $A(X)$ onto a scalar D. A good way of doing so is provided by the Mahalanobis (1936) distance. Denoting the expected value of $A(X)$ by $\mu(\theta)$ and the covariance matrix by $\Sigma(\theta)$, we propose the Mahalanobis distance

$$D(x,\theta)=\left(A(x)-\mu(\theta)\right)^{T}\Sigma(\theta)^{-1}\left(A(x)-\mu(\theta)\right) \qquad (12)$$

as our test statistic. We cannot in general compute equation (12) analytically; therefore, we use its Monte Carlo estimate $\widehat{D}$, obtained by replacing μ and Σ by estimates. With a large number N of simulations $x_{h}^{(\mathrm{sim})}$ for $h=1,\dots,N$ from the SAOM with parameter $\hat{\theta}$, we estimate $\mu(\theta)$, $\Sigma(\theta)$, and $D(x,\theta)$ by

$$\hat{\mu}=\frac{1}{N}\sum_{h=1}^{N}A\left(x_{h}^{(\mathrm{sim})}\right)$$

$$\widehat{\Sigma}=\frac{1}{N}\sum_{h=1}^{N}\left(A\left(x_{h}^{(\mathrm{sim})}\right)-\hat{\mu}\right)\left(A\left(x_{h}^{(\mathrm{sim})}\right)-\hat{\mu}\right)^{T} \qquad (13)$$

$$\widehat{D}(x)=\left(A(x)-\hat{\mu}\right)^{T}\widehat{\Sigma}^{-1}\left(A(x)-\hat{\mu}\right)$$

This $\widehat{D}(x)$ will be briefly called the MDMC. The MDMC test is constructed by considering the p-value

$$\mathrm{P}_{\theta}\left\{\widehat{D}\left(X,\hat{\theta}\right)>\widehat{D}\left(x,\hat{\theta}\right)\right\} \qquad (14)$$

If the central limit theorem would apply, we would expect that equation (12) has approximately a chi-square distribution. However, we have no proof of this, and therefore follow a parametric bootstrap procedure. Then, plugging in the various values leads to

$$\hat{p}=\frac{1}{N}\sum_{h=1}^{N}\mathbf{1}\left\{\widehat{D}\left(x_{h}^{(\mathrm{sim})}\right)>\widehat{D}(x)\right\} \qquad (15)$$

as the estimator for the p-value (equation (14)). A very low value indicates a poor fit.

For $\hat{p}$, we should be concerned about the difference between a straight 0 and a very small positive value. Suppose that $N=1000$ simulations are carried out. If $\hat{p}=0.001$ or larger, the observed statistics are somewhere within the total multivariate cloud of simulated values of statistics $A(x_{h}^{(\mathrm{sim})})$, but if $\hat{p}=0$ precisely, the observed statistics are outside this cloud of values, and might be far away. Therefore, when interpreting the value of equation (15), the distinction between the values 0, on one hand, and $1/N$ or larger, on the other hand, should be taken very seriously.

Summarizing, this approach uses Monte Carlo simulations for parameter estimate $\hat{\theta}$ thrice. First, the simulations are obtained in the algorithm used for the usual MoM estimation (Ripley et al., 2019; Snijders, 2001) for computing standard errors and convergence checking. Next, they are used to estimate μ, Σ, and D as in equation (13). Finally, they are used to estimate the p-value (equation (14)) on which the MDMC test is based. These three computations can all use the same set of N simulations. By estimating the p-value according to

equation (15), we do not use assumptions about the distribution of $\widehat{D}$ directly, although we do use the assumption that X was generated by the SAOM and that $\hat{\theta}$ is a good estimate. It is not required that $A(X)$ have a multivariate normal distribution; although the use of the Mahalanobis distance is more justified when the multivariate distribution has a more nearly elliptical shape. This p-value addresses the hypothesis that the SAOM with parameter $\hat{\theta}$ generated the observed $A(x)$ against the unspecified composite hypothesis that some other model generated it. Due to the use of estimated parameters in equation (13), there will be some measure of overfitting; this will be the stronger for statistics highly correlated with the statistics used for estimation. This will lead to a conservative test (cf. Schweinberger, 2012). Below we will investigate the severity of this conservatism.

MMD estimator

The development of the MDMC allows the researcher to test GOF. If the fit is satisfactory, the researcher can go on with the analysis. In the event that fit is not acceptable, the researcher may have some theoretically based ideas on remediation. Even so, there may be a large number of effects that are plausible for inclusion. The richness and complexity inherent in networks provides the researcher with an enormous menu of effects to choose from, as is illustrated by the implemented effects in the RSiena package (Ripley et al., 2019). In most cases, theory and experience will not suffice to give a definite conjecture about the effect that should be added to improve the fit. Trying out many different effects will be time-consuming. This section presents an approximation to suggest which model improvements might be empirically promising, without requiring to estimate the correspondingly extended model.

The "MMD" estimator is an estimator for the Mahalanobis distance $D(x,\theta)$ for the situation where a baseline model is considered with parameter estimated as $\hat{\theta}_{0}$, the focus is on one auxiliary statistic, and a variety of effects is considered, not included in the baseline model, that might be used to extend the model. Each effect corresponds to a potential model extension, and we are interested in knowing the extent to which each of these model extensions would improve the fit as measured by $D(x,\theta)$. Thus, if for a given auxiliary function it turns out that the fit is unsatisfactory, the MMD can be calculated for a set of potential model extensions, and the extension producing the largest improvement in Mahalanobis distance can be selected as the tentative new model. Since parameter estimation for all extended models would be time-consuming, we forego the step of estimation and we are satisfied with an approximation.

The method proposed here is similar to methods used in model specification of structural equation models (SEMs) (cf. Kaplan, 1990, 1991). Specifically, they are to some extent analogous to the model modification index (MMI) of Sorbøm (1989) and the expected parameter change of Saris

et al. (1987). An important difference with the use of the MMI in structural equation modeling is that in SEMs the GOF function (usually the log likelihood) is also maximized for obtaining estimates. This approach is unavailable to us in the current context. For MoM estimation, we have only an estimating function which cannot be used as a GOF function (for a MoM estimate, the estimating function is the difference between the left-hand and the right-hand sides of (5), and the estimate is defined by this being zero). Furthermore, the purpose of our GOF approach is to consider the fit specifically for statistics that were not used for the estimation.

For a given extension of the model, supposing it is the true one, denote the parameter value by θ_1. For this parameter value and the given data set x, we wish to estimate the Mahalanobis distance $D(x, \theta_1)$ without going through the full procedure of estimating θ_1 by the MoM (Snijders, 2001) or by Maximum Likelihood (Snijders et al., 2010a). Instead, we may use the approach of Schweinberger (2012) and calculate a one-step estimate $\hat{\theta}_1$ (so called because it is based on a single Newton-Raphson step for optimizing the estimation function approximated from calculations done for $\theta = \hat{\theta}_0$). The one-step estimation essentially comes for free, being based entirely on data sets simulated under parameter value $\hat{\theta}_0$; these simulations are typically conducted anyway for convergence checking and standard error calculation for $\hat{\theta}_0$ (the "third phase" of the algorithm, see Snijders, 2001).

The question now is how to approximate the value $D(x, \theta)$ as defined in equation (12) for $\theta = \hat{\theta}_1$. We assume $D(x, \theta)$ is a twice differentiable function of θ. For determining the one-step estimate $\hat{\theta}_1$, a Taylor expansion is used, and we may likewise use a Taylor expansion for approximating $D(x, \theta_1)$. As a second-order expansion is used for determining one-step estimates, it seems natural to consider a second-order expansion for $D(x, \theta)$. Such an expansion was elaborated in Lospinoso (2012), but this requires computation of the observed (Fisher) information matrix and is computationally considerably more complex. A main purpose of the present GOF test is to dovetail with the relatively expedient MoM estimation procedure. We, therefore, deal exclusively with the first-order expansion here, which contains ingredients immediately available as by-product of the MoM estimation procedure of Snijders (2001), and which proves to work satisfactorily in practice.

Letting $\nabla_\theta = \partial / \partial \theta$, the first-order expansion is

$$D(x, \theta_1) \approx D(x, \hat{\theta}_0) + (\theta_1 - \hat{\theta}_0) \nabla_\theta D(x, \theta)|_{\theta = \hat{\theta}_0} \quad (16)$$

The gradient $\nabla_\theta D(x, \theta)$ has elements

$$\frac{\partial}{\partial \theta_i} D(x, \theta) = (A(x) - \mu(\theta))^T \Xi_i(\theta)(A(x) - \mu(\theta))$$
$$- 2\mu_i'(\theta)^T \Sigma(\theta)^{-1}(A(x) - \mu(\theta)) \quad (17a)$$

where

$$\mu_i'(\theta) = E_\theta \left\{ [u_\theta(X)]_i A(X) \right\} \quad (17b)$$

$$\Xi_i(\theta) = -\Sigma(\theta)^{-1} \Gamma_i(\theta) \Sigma(\theta)^{-1}$$
$$\Gamma_i(\theta) = E_\theta \left\{ [u_\theta(X)]_i A(X) A(X)^T \right\} \quad (17c)$$

$$- \mu_i'(\theta) \mu(\theta)^T - \mu(\theta) \mu_i'(\theta)^T \quad (17d)$$

and

$$u(X) = \nabla_\theta \log p_\theta(X) \quad (17e)$$

is the score function for the probability function $p_\theta(X)$. A derivation is provided in the Appendix 1.

The derivatives are expressed in equations (17b) and (17d) using the score function $u_\theta(X)$. To keep the Monte Carlo error for estimating these quantities within reasonable bounds, it is important to use a linear control variable, as was done also in Schweinberger and Snijders (2007). Therefore, we rewrite equations (17b) and (17d) as

$$\mu_i'(\theta) = E_\theta \left\{ [u_\theta(X)]_i (A(X) - A(x)) \right\}$$
$$\Gamma_i(\theta) = E_\theta \left\{ \begin{array}{l} [u_\theta(X)]_i (A(X) - A(x)) \\ (A(X) - A(x))^T \end{array} \right\} \quad (18)$$

using the property that

$$E_\theta \left\{ u_\theta(X) \right\} = 0$$

This leads to the Monte Carlo estimators

$$\widehat{\mu_i'}(\theta) = \frac{1}{N} \sum_{h=1}^{N} \left\{ [u_\theta(x_h^{(sim)})]_i (A(x_h^{(sim)}) - A(x)) \right\}$$
$$\hat{\Gamma}_i(\theta) = \frac{1}{N} \sum_{h=1}^{N} \left\{ [u_\theta(x_h^{(sim)})]_i \times \right.$$
$$\left. (A(x_h^{(sim)}) - A(x))(A(x_h^{(sim)}) - A(x))^T \right\} \quad (19)$$

The MMD estimator for $D(x, \theta_1)$ uses a Monte Carlo simulation set-up as above with simulations for the estimated value $\hat{\theta}_0$. The estimation proceeds as follows: estimate θ_1 by the one-step estimator $\hat{\theta}_1$ from Schweinberger and Snijders (2007); use equation (13) to estimate $\mu(\hat{\theta}_0)$ and $\Sigma(\hat{\theta}_0)$; estimate $\mu_i'(\check{\theta}_0)$ and $\Gamma_i(\check{\theta}_0)$ by (17c), with $\theta = \hat{\theta}_0$; estimate $\Xi_i(\check{\theta}_0)$ using (17a), and finally, plug these results into equations (17a) and (16).

Concluding, this approximation can be computed using simulations $x_h^{(\text{sim})}$ under parameter estimate $\hat{\theta}_0$, where in addition to the auxiliary statistics $A(x_h^{(\text{sim})})$ we also need to compute the score functions $u_\theta(x_h^{(\text{sim})})$.

Two warnings should be given here. First, the MDMC for an extended model does not need to be smaller than for the baseline model. This is because the MoM estimator is not oriented toward minimizing the function $D(x,\theta)$. Second, the value of the MMD is not guaranteed to be positive. This is because it is an approximation only. Therefore, a researcher should not be surprised in sometimes finding that a model extension seems to lead to a worse fit for some auxiliary statistic, nor in sometimes finding that the MMD predicts a worse fit, or a negative Mahalanobis distance. However, these are exceptions.

For a given auxiliary statistic, when considering a set of potential model improvements, the improvement yielding the largest decrease in the Mahalanobis distance may be considered to be the best choice. Let us call this the MMD-1 model modification. We are just considering one-dimensional modifications, so there is no issue of different a priori advantages because of involving different degrees of freedom. A more subtle possibility is available when a set of several auxiliary statistics and a set of model improvements is under consideration. Then one may use the following procedure, to be called MMD-2, for model modification:

1. If all MDMC p-values are larger than some threshold, for example, the conventional $\alpha = 0.05$, then continue with the baseline model;
2. If for some auxiliary statistics the MDMC p-value is less than α, use for the following step the auxiliary statistic having the smallest MDMC p-value;
3. Choose the effect that gives for this auxiliary statistic the best improvement as predicted by the MMD, and add this to the model.

Evidently, this procedure can be iterated.

Simulation study

In this section, we provide a small simulation study of (1) the validity and power of the proposed MDMC test, and (2) the effectiveness of the one-step Mahalanobis distance estimators in guiding model selection. We use a subset of the Teenage Friends and Lifestyle Study (TFLS) as the basis for our study. The TFLS data set was collected by West and Sweeting (1996) and utilized in many publications including Michell and Amos (1997), Pearson and Michell (2000), and Steglich et al. (2010). The panel data were recorded over a 3-year period starting in 1995, when the pupils were aged 13, and ending in 1997. A total of 160 pupils took part in the study, 129 of whom were present at all three measurement points. We utilize a subset of 50 girls from the study, called the TFLS-50, who were present at all three measurement

points, chosen only for demonstration purposes, and distributed with the RSiena package (Ripley et al., 2019). Friendship networks were formed by allowing the pupils to name up to six best friends.

We simulate data in the following way: using the first observation of the TFLS-50 as the time-1 measurement, perform independent Monte Carlo simulations, according to three different SAOM specifications, each yielding 250 networks for the time-2 measurement. A binary actor covariate V is constructed, with values 0 for the first 25 and 1 for the last 25 girls. Since the order of the individuals in the data set is arbitrary, this is like a random covariate. We simulate with a constant rate $\rho = 4$ and the following evaluation function:

$$f_{ij}(x \mid \beta) = \beta_1 \sum_j x_{ij} \quad \text{(outdegree effect)}$$

$$+ \beta_2 \sum_j x_{ij} x_{ji} \quad \text{(reciprocity effect)}$$

$$+ \beta_3 \sum_{j,k \neq j} x_{ij} x_{ik} x_{jk} \quad \text{(transitive triplets effect)}$$

$$+ \beta_4 \sum_j x_{ij} \sum_k x_{kj} \quad \text{(indegree popularity effect)}$$

$$+ \beta_5 \sum_j x_{ij} \mathbf{1}\{v_i = v_j\} \quad \text{(same } V \text{ effect)}$$

These 750 simulations are then used as the time-2 observation.

Three combinations of parameter values are used; in these combinations, one of the values $\beta_3, \beta_4,$ and β_5 is non-zero, the other two are zero. The value of β_1 in all three specifications is tuned so that the average degree for the simulated data is between 4 and 5, a value that is reasonable in practice. The reciprocity parameter is set at 2.0, also a value close to what is often found for friendship networks. To achieve some further similarity between the specifications, the non-zero parameter among $\beta_3, \beta_4,$ and β_5 is determined (based on trial and error) so that the value of $\mathrm{E}(\hat{\beta}_k)/\mathrm{SE}(\hat{\beta}_k)$ for each of $k = 3,4,5$ is about 3, corresponding to a power of very roughly 0.8 for detecting this effect by a one-parameter test. (This reasoning is based on an approximation where $\hat{\beta}_k/\mathrm{SE}(\hat{\beta}_k)$ is assumed to have a normal distribution with mean equal to 0 if $\beta_k = 0$ and continuously increasing as a function of β_k, and variance equal to 1.) This leads to the parameter combinations A, B, and C in Table 1.

In this demonstration, we use for the GOF study the following auxiliary statistics:

*   The triad census (see Figure 1);
*   The geodesic distance distribution for $C = 5$, where C is the number of statistics (equation (10)), that is, the dimension of the auxiliary statistic;
*   The indegree distribution for $C = 5$;
*   The outdegree distribution for $C = 5$.

Table 1. Specifications of three models for simulation.

Model	A	B	C
β_1 outdegree	−1.30	−1.65	−1.35
β_2 reciprocity	2.00	2.00	2.00
β_3 transitive triplets	0.27	0.0	0.0
β_4 indegree popularity	0.0	0.12	0.0
β_5 same V	0.0	0.0	0.45

Table 2. Frequency intervals for $\hat{\beta}^{(0)}$ under misspecification.

	A	B	C
Outdegree $\hat{\beta}_1^{(0)}$ (L)	−1.35	−1.34	−1.32
Outdegree $\hat{\beta}_1^{(0)}$ (U)	−0.62	−0.75	−0.77
Reciprocity $\hat{\beta}_2^{(0)}$ (L)	1.63	1.46	1.59
Reciprocity $\hat{\beta}_2^{(0)}$ (U)	2.59	2.34	2.47

Lower (L) and upper (U) end points for 95% relative frequency intervals obtained via simulations.
Columns refer to Models A, B, and C in Table 1.

The number of Monte Carlo simulations to compute the $\hat{p}$-values (equation (15)) is $N = 1000$. The procedure is comprised of the following steps:

1. Estimate the parameters of an *improperly specified* base model with an incorrect evaluation function

$$f_{ij}^{(0)}(x \mid \beta) = \beta_1^{(0)} \sum_j x_{ij} \ (\text{outdegree effect})$$
$$+ \beta_2^{(0)} \sum_j x_{ij} x_{ji} \ (\text{reciprocity effect}) \tag{20}$$

This may be considered to be a minimal model.

2. The proposed GOF test is evaluated at $\hat{\beta}^{(0)}$, for each of the four auxiliary statistics.
3. Three model elaborations are considered. Each elaboration entails adding one of the following terms to the evaluation function

$$\beta_3^{(TT)} \sum_{j,k \neq j} x_{ij} x_{jk} x_{ij} \ (\text{transitive triplets}) \tag{21}$$

$$\beta_4^{(IP)} \sum_i x_{ij} \sum_k x_{kj} \ (\text{indegree popularity}) \tag{22}$$

$$\beta_5^{(SV)} \sum_j x_{ij} \mathbf{1}(v_i = v_j) \ (\text{same } V) \tag{23}$$

These effects were explained above. For each elaboration, the MMD is calculated. Note that for each simulated data set, one of these three is the properly specified model. The frequencies of selecting the correct model are reported. Using the MMD values, model selection procedure MMD-2 is applied.

4. As a sanity check, for the correct model we perform MoM estimation, and assess whether the estimates according to the properly specified model are close enough to the true parameter values.
5. For a new set of parameter values, we perform Steps 1–3 and evaluate the MDMC values $\widehat{D}(x)$ of equation (13) at $\hat{\beta}^{(TT)}, \hat{\beta}^{(SC)}, \hat{\beta}^{(IP)}$. The MMD values calculated in Step 3 are compared with these MDMC values, to see whether they are good enough approximations.

The results of these steps are as follows.

Step 1: estimation improperly specified model

Table 2 provides observed 95% intervals for $\hat{\beta}^{(0)}$ estimated using the misspecified model (equation 20). The number of cases for each column here is 250, the number of generated time-2 networks for each specification. Across all three models, the reciprocity parameter $\beta_2^{(0)}$ is contained in the corresponding frequency interval, while outdegree $\beta_1^{(0)}$ is in this interval only for model A. Although the simulation study is quite limited in scope, it is encouraging that the estimates for the reciprocity parameter seem reasonably robust to misspecification. For the outdegree parameter, differences are to be expected because the omitted effect leads to a different reference point (this is the point where the other effects all are 0).

Step 2: perform the MDMC test

The proposed MDMC test is executed at the improperly specified model. Assembling the resulting tests into receiver operating characteristic (ROC) curves will allow us to investigate how well the test detects the misspecification. These curves tell us, for a specified false positive rate, the probability of rejecting the model. See Fawcett (2006) for more information on ROC curves. This is estimated here by calculating, for any given $\alpha \in (0,1)$, the proportion among the simulation results for a given model specification and a given auxiliary statistic that the $\hat{p}$-value (equation 15) is less than α. The test has power if the ROC curve is above the diagonal, and power is better when the curve is higher. The results in Figure 2 show that power varies considerably depending on the effect/auxiliary statistic combination. For same covariate, none of the four auxiliary statistics were effective at detecting the misspecification. In fact, some of the tests appear a bit conservative (indicated by the concavity of the ROC curve). For indegree popularity, the results are mixed; indegree distribution and triad census both have good power to detect the misspecification (indicated by the convexity of their ROC curves) whereas outdegree distribution and geodesic distribution do not. The transitive triplets effect is detected only by the triad census. These results reflect the
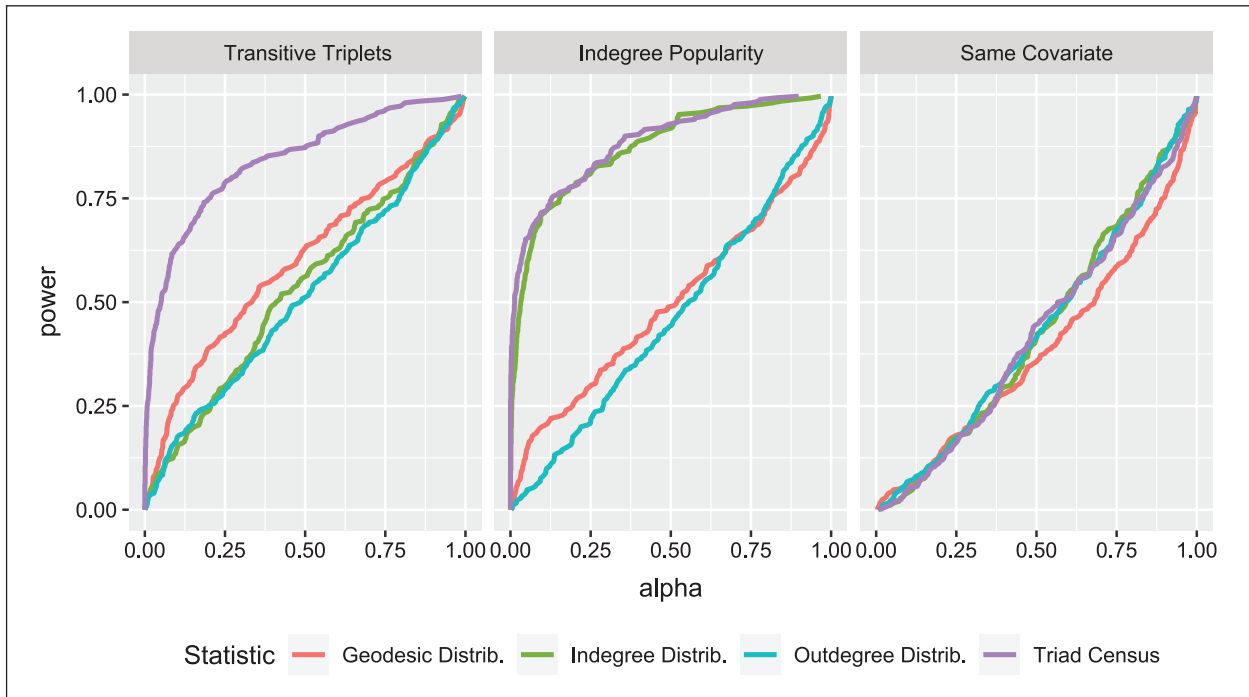
Figure 2. Receiver operating characteristic curve: the ROC curve gives indication of the power versus false positive ("alpha") trade-off of MDMC using various auxiliary statistic specifications.

differential connections between the effects and the auxiliary statistics: for example, the indegree distribution has an association with the indegree popularity effect, while the triad census has a connection with the transitive triplets effect; for some other combinations, the connection is weak. This potentially loose connection should entreat researchers to consider several auxiliary statistics for GOF testing rather than relying on one or two.

To interpret the conservative nature of the results of all auxiliary statistics for the same covariate effect, we point out two issues. These results suggest that misspecification with respect to covariates is not easily detected by auxiliary variables reflecting network structure; more research is needed to investigate whether this is generalizable. Other auxiliary statistics, considering specifically the occurrence of ties depending on the covariate values for sender and receiver, may be expected to have a better sensitivity for this type of misspecification. Second, insensitivity leads to conservativeness because the Mahalanobis $\hat{p}$-value is not corrected for the use of estimated parameter values; this is analogous to what would happen if, for example, the chi-square test for a contingency table were used without decreasing the degrees of freedom to account for the estimation of the marginal distributions.

Step 3: MMD for candidate model elaborations

The most practically important characteristic of an MMD estimator is how often it would lead the researcher to select the appropriate model elaboration, given a misspecification.

Table 3 gives the distribution of the rankings of MMD evaluated for each model elaboration. Again, the number of cases is 250, the number of simulated data sets.

When the model misspecification is A, the omission of the transitive triplets effect (equation 21), the results in the first panel of Table 3 show that the MMD estimators based on geodesic distance, triad census, and outdegree distribution have the highest probability of selecting this model extension indeed. Only the indegree distribution selects more frequently the indegree popularity effect.

For the indegree popularity model B defined by equation (22), the triad census and the indegree distribution-based MMD estimators select the correct model more than 80% of the time. The results are much weaker for the geodesic and outdegree distributions, although for the latter the correct model still has the highest estimated probability.

The MMD estimator orderings for the same covariate model C, given by equation (23), do not perform well. None of the auxiliary statistics results in more than a 41% selection rate of the correct elaboration. This corresponds to the conservative nature for all auxiliary statistics for model C of the MDMC tests in Step 2.

The results of model selection procedure MMD-2 are given in Table 4. These are in line with the earlier results: the detectability of the transitive triplets effect and especially the indegree popularity effect are good, whereas the same covariate effect is not detectable by these auxiliary statistics. We remind the reader that the parameter values in the three cases were chosen so that the Wald test directed at this specific

Table 3. Probability distribution of MMD rankings at Step 3.

	A	B	C
Transitive triad census	0.97	0.03	0.00
Outdegree distribution	0.58	0.10	0.32
Geodesic distance distribution	0.88	0.02	0.10
Indegree distribution	0.28	0.42	0.30
	A	**B**	C
Transitive triad census	0.03	0.87	0.10
Outdegree distribution	0.33	0.39	0.28
Geodesic distance distribution	0.24	0.36	0.40
Indegree distribution	0.00	0.82	0.18
	A	B	**C**
Transitive triad census	0.52	0.38	0.10
Outdegree distribution	0.38	0.35	0.26
Geodesic distance distribution	0.53	0.24	0.23
Indegree distribution	0.09	0.50	0.41

MMD: modified model distance.
The three panels correspond to the three true models, indicated by the bold face column header.
Auxiliary statistics are in the rows. For each corresponding MMD estimator, the row gives the estimated probabilities of selecting the candidate elaboration of the column, that is, rows sum to 1.

effect has a power of approximately 0.80. In this sense the effect sizes in the three cases are equally strong, and the difference in detectability depends on the correspondence between the auxiliary statistics and the misspecification, not on differences in effect size.

Step 4: MoM estimation of candidate models

During this step, we do a complete MoM estimation of each candidate model. All 95% coverage frequency intervals (not given for lack of space) include the true model parameters, a result in line with simulation studies of, for example, Snijders (2001) and Lospinoso et al. (2011).

Step 5: MDMC tests of candidate models

To assess the performance of the MMD estimator, we conduct a set of simulations varying the true extent to which one of the three model extensions is called for.

Define β_k as the columns of Table 1 for $k = $ A, B, C, and $\beta_0 = (-1, 2, 0, 0, 0)$. The value of β_0 is similar to the parameters in Table 1 in the sense that it also leads to time-2 networks having an average degree between 4 and 5. The simulations are done for parameter values $\lambda \beta_k + (1 - \lambda)\beta_0$, and also for $\lambda(\beta_h + \beta_m) + (1 - \lambda)\beta_0$, where k is one of A, B, C, and h and m are two different out of A, B, C. For each of these six combinations of A, B, C, a total of 100 simulations is done with random values of λ, drawn from the uniform distribution on $(0.5, 1.5)$. This gives a set of 600 data sets

Table 4. Relative frequencies of models selected by procedure MMD-2 in Step 3.

	Selected extension			
	None	A	B	C
True model extension				
Transitive triplets	0.42	0.51	0.06	0.01
Indegree popularity	0.22	0.02	0.76	0.00
Same covariate	0.92	0.04	0.03	0.01

For each true model extension, the proportions of models selected by procedure MMD-2 are given; "none" means all goodness of fit $\hat{p}$-values for the baseline model were above 0.05.

generated by parameter values where for each of the parameters $\beta_3^{(TT)}$, $\beta_4^{(IP)}$, and $\beta_5^{(SV)}$, a total of 300 have the value 0 and another 300 have a positive value.

For each data set, parameters are estimated under the base model (equation 20), and for each of the four auxiliary statistics, the MMD estimators for the three candidate models are calculated. Negative MMD values are truncated to 0. Then for each of the three candidate models, the full MoM estimation is carried through and the four Mahalanobis GOF MDMC statistics are computed. Thus, for each of the 600 data sets, there are 12 pairs of a MMD estimator and an MDMC statistic. To each of these pairs corresponds an MDMC value for the base model. The question now is, whether the MMD yields an adequate prediction of the improvement, that is, decrease, in the Mahalanobis distance when comparing the base model to the estimated candidate model. In other words, how good is the MMD as an approximation of the MDMC of the candidate model, where the MDMC of the base model can be used as a reference value.

Figure 3 presents the improvement of the Mahalanobis distance with respect to the base model, as realized (vertical axis) and as predicted by the MMD (horizontal axis). In the axis labels, MDMC(0) refers to the MDMC for the base model. The MDMC refers to the realized MDMC for the candidate model, and MMD refers to the MMD truncated to nonnegative values. In view of the skewness, all Mahalanobis distances were transformed by the square root.

The figure shows a strong agreement. The realized improvements tend, however, to be smaller than the predicted improvements, as most points are below the equality line. The correlation between the differences MDMD—MDMC(0) and MMD—MDMC(0) is equal to 0.94, which is a confirmation of the value of this approximation.

Workflow

In this section, we suggest a possible GOF assessment approach for SAOM analysis that incorporates the MDMC testing procedure into the model fitting context. This can be carried out using the R package RSiena (Ripley et al., 2019), in which the MDMC test and the MMD estimator are implemented in the function sienaGOF.
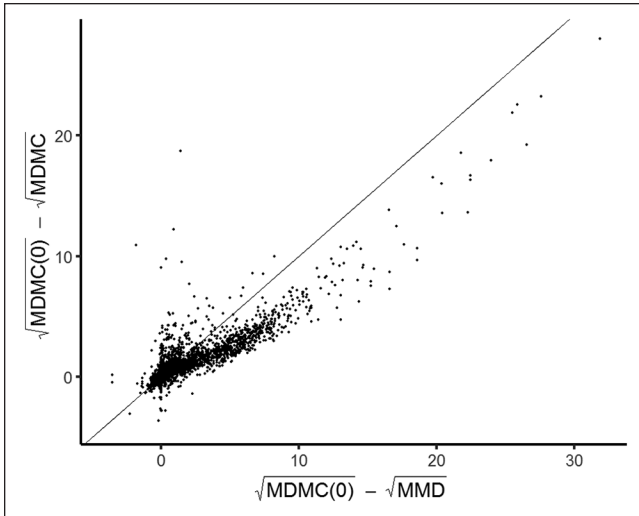
Figure 3. The decrease in Mahalanobis distance compared with the base model, as predicted by the MMD (horizontal axis) and as realized (vertical axis). The straight line indicates equality. All distances square root transformed; MMD left truncated at 0.

First, we should point attention to time heterogeneity, which is a special kind of misspecification that is possible for data with $M \geq 3$ waves. There is time heterogeneity if parameters β in equation (3) differ between periods. This can be tested in RSiena by the function sienaTimeTest, implementing a procedure of Lospinoso et al. (2011). This means that in practice, a researcher fitting SAOMs has two approaches available to assess whether the model fits the data, the one comparing data and model expectations for auxiliary statistics, and the other comparing parameters between periods. It is impossible to establish a rule that a researcher should first pay attention to one and subsequently to the other. In the case of lack of fit, the researcher may have to switch in unforeseen ways between these two approaches. Some further considerations and examples are given in Lospinoso and Satchell (2011).

A proposed workflow is the following. The preceding remarks imply that the order of Steps 5, on the one hand, and 6–7 combined, on the other hand, is not fixed.

1. Select a provisional SAOM specification: this model should be parsimonious, based on the research question and existing knowledge about the processes driving the evolution of the network under study.
2. Reflect about time heterogeneity: if the data have three or more waves, and the average degree per wave has important jumps up as well as down, this is a sign of time heterogeneity for which it might be necessary to include time as a covariate. Whether this should be a linear time effect or some transformation of time depends on the pattern shown by the average degrees over the waves. One possibility is to include dummy variables for the waves.

In the case of three or more waves, if there is a suspicion of strong time heterogeneity, or if there is persisting lack of fit or lack of convergence, it is advisable to estimate by period, that is, for each pair of subsequent waves separately. This is possible only if the data set is large enough to make estimation by period feasible.

3. Estimate the parameters for the provisional model.
4. Check the convergence of the estimation: for MoM estimation, good convergence means that simulations drawn from the fitted model yield simulated values of the estimating statistics (6) that are very close to the observed values. The manual (Ripley et al., 2019) gives criteria. If convergence is poor, try re-estimating the model; the manual offers advice for how to proceed. If poor convergence is systemic, this may be evidence of poor agreement between the provisional model and the data. Experience has shown that systemic poor convergence may reflect, for example, that covariates should be included reflecting meeting opportunities (e.g. "same classroom" if the network is situated in a school context); or that additional degree-related effects should be included (e.g. reflecting isolated nodes); or that outdegree effects on the rate of change should be included; or that there is major time heterogeneity. In this case, the researcher should go back to Steps 1 or 2.
5. Check for time heterogeneity, if appropriate: if there are three or more waves, time heterogeneity can be tested using the function sienaTimeTest, as mentioned above, with remedial possibilities mentioned in Step 2.
6. Assess GOF with the proposed MDMC GOF test: the GOF test can be carried out for several auxiliary statistics. We recommend in any case to use the two degree distributions and the triad census, with the geodesic distances distribution as a valuable addition.
   The GOF test provides a $\hat{p}$-value (equation 15). The conventional threshold of $\alpha = 0.05$ is here even more arbitrary than in other cases—it may be used as a benchmark but should not considered as a precise and important borderline value.
   If the observed $\hat{p}$-value is considered too low, we recommend visualizing the simulated auxiliary statistics versus the observed auxiliary statistics to get an intuitive feel for what is not fitting well. A plot for this purpose is also available in the RSiena package. Examples are given in Figures 4 and 5. This plot consists of a sequence of violin plots (Hintze and Nelson, 1998) for each dimension of the simulated auxiliary statistics, with confidence bands, and an overlay of the observed statistics. The plot helps us to see where the fit is poor.
7. Extend the model, if the fit is inadequate: if the GOF test points out that the fit is not satisfactory,

the model will have to be extended. The plot, further knowledge about the data, and theoretical considerations may point us to a list of candidate effects that could be added to the model. These candidate effects can be added directly to the model, or they may be evaluated approximately for their promise to improve fit by the MMD test described above. If there is a choice between several effects indistinguishable by theory, the one that yields the lowest MMD for the modified model may be chosen. The results of the simulation study above demonstrate that this is not a fail-safe procedure but it nevertheless can give meaningful guidance.

8. Iterate: depending on the results, some of the steps may have to be repeated.

There are potential theoretical pitfalls here, however. (1) As with any model selection procedure, if it is applied without theory in mind, it will be much harder to defend the validity of the final model. (2) There may well be different possibilities for improving fit when it is found to be inadequate, so the result of this procedure may depend on random circumstances. (3) It may be desirable also to drop effects from the model at certain moments during the procedure—depending on theory and insight into the data; this is illustrated by the example below. Considerations about the research question and theoretical insights will always need to play a fundamental role.

Example

We provide a small example of our model fitting procedure applied to the subset of the TFLS data introduced above. The data set includes three observations of friendship networks and alcohol consumption habits (on a scale with values 1–5). We demonstrate a forward stepping model selection exercise, assuming—for the sake of this example only—that the researcher wishes to specify a priori only the reciprocity effect, has a list of candidate effects, and wishes to inductively select a set of effects that lead to an acceptable fit with respect to the indegree and outdegree distributions, the triad census, and the distribution of geodesic distances. "Acceptable" fit is defined by the usual 5% level for the MDMC $\hat{p}$-value (equation 15). Note that we regard this as an extreme example because it is totally inductive and devoid of theory; we do not favor this kind of theory-blind forward model selection in practice. But we hope it is a useful demonstration.

The list of candidate effects consists of transitive triplets, "gwesp," three-cycles, transitive reciprocated triplets, dense triads, indegree popularity, outdegree popularity, outdegree activity, reciprocated degree activity, and covariate similarity (for alcohol consumption). These effects were defined above. All these effects are used in practice by researchers
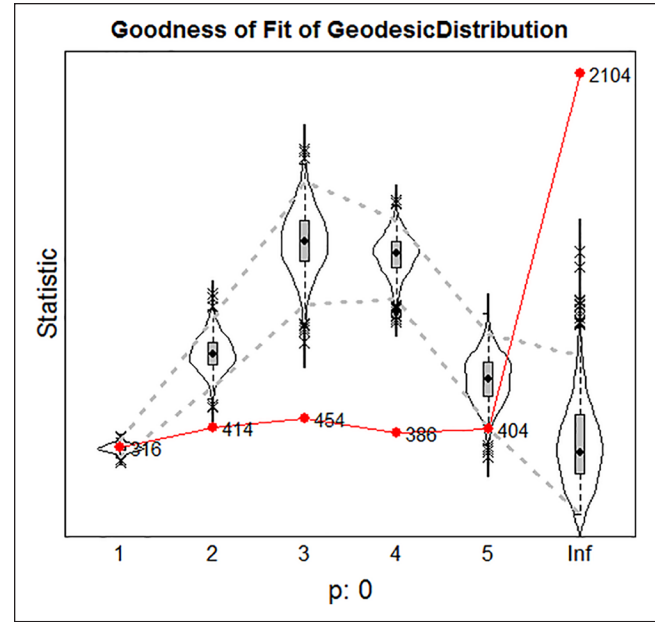

Figure 4. Goodness of fit diagnostic plot for the first model, with as auxiliary statistic the geodesic distance distribution. Observed values are indicated by numbers connected by a line. The simulated statistics are represented by the violin plots. Dotted lines give 95th percentile bands. The $\hat{p}$-value for the Mahalanobis distance-combination is given at the bottom.

applying SAOMs, although the dense triads effects is not used frequently.

The procedure was planned to have the following steps:

1. An initial model was estimated including only the outdegree and reciprocity effects.
2. Repeat:
   ▸ Add the candidate effect that is predicted by the MMD to have the greatest decrease for the first auxiliary statistic that does not have an acceptable ($\hat{p} \geq 0.05$) GOF, and estimate this model. The word "first" is used here in the order (1) indegree distribution, (2) outdegree distribution, (3) triad census, and (4) geodesic distances distribution. The first in this list to have a $\hat{p}$-value less than 0.05 is the "critical" statistic, determining the selection of the next included effect.
3. Stop when all four auxiliary statistics have an acceptable ($\hat{p} > 0.05$) GOF.

This led to a sequence of six models. They are presented in Tables 5 and 6. Each column presents the estimated model and the four GOF $\hat{p}$-values. Some candidate effects are not mentioned at all, because they were never selected. From the first model on, the indegree and outdegree distributions had an acceptable fit, so they never played an explicit role.

In view of the length of this article, we present only few plots. Figure 4 is a plot for the GOF results for the first model, containing only the outdegree and reciprocity effects, for the geodesic distance distribution of the auxiliary statistic. The fit is clearly very poor, with $\hat{p}=0$; the simulated distributions for the number of pairs with geodesic distance from 2 to 5 are much higher than the observed frequencies, and for the number of pairs at infinite distance, that is, not being in the same connected component, the reverse is true. In other words, the simulated networks have too few components, and within the components the distances are too large.

In the first two steps the critical statistic was the triad census, and the effects added were transitive triplets and reciprocated degree-activity. In the third step the critical statistic was the geodesic distance distribution, and the gwesp effect was added. This led to an unforeseen result: the transitive triplets effect, included earlier, became insignificant. Also, the triad census GoF $\hat{p}$-value dropped below 0.05. It was decided to leave out the transitive triplets effect; this increased the triad census GoF $\hat{p}$-value somewhat, but it remained less than 0.05. The triad census still was the critical statistic in this step, and the MMD now led to including the dense triads effect. This was the sixth model, and it satisfied all GOF requirements.

For the final model, a plot for the GOF results is in Figure 5. For the visual comparison of Figures 4 and 5, keep in mind that the observed values are the same, and the simulated distributions now are much closer to the observations.

Discussion

This article proposed a GOF testing procedure for SAOMs which relies on a battery of auxiliary statistics, selected by the researcher, as GOF criteria. These statistics are used to construct a Monte Carlo Mahalanobis distance based test. Because remediating poor fit on these statistics can be a complex and time-consuming undertaking, we proposed the *MMD estimator* for the Mahalanobis distance, evaluated at some provisional model, to assess which model among a set of candidates can be expected to improve fit best.
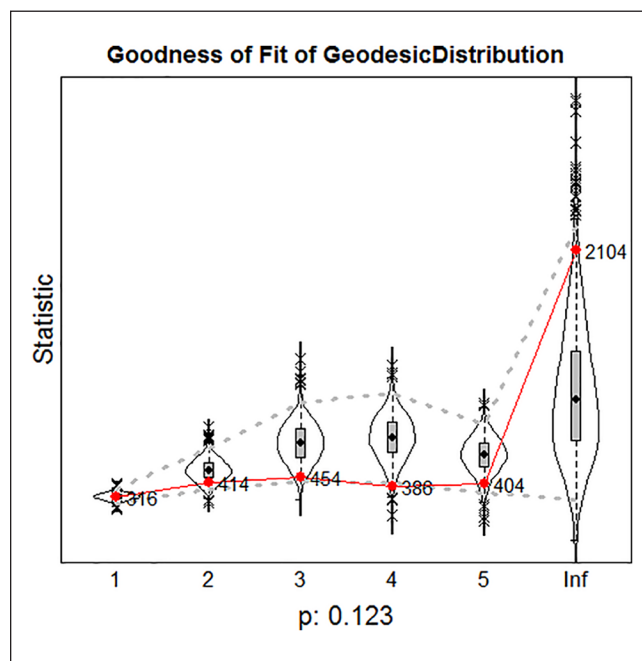

Figure 5. Goodness of fit diagnostic plot for the final model with as auxiliary statistic the geodesic distance distribution. Observed values are indicated by numbers connected by a line. The simulated statistics are represented by the violin plots. Dotted lines give 95th percentile bands. The $\hat{p}$-value for the Mahalanobis distance-combination is given at the bottom.

Table 5. MoM results first three models: parameter estimates (par.), standard errors (SEs), and goodness of fit (GOF) $\hat{p}$-values.

Effect	Par. (SE)	Par. (SE)	Par. (SE)
Rate 1	5.79 (0.92)	6.50 (1.08)	7.44 (1.41)
Rate 2	4.49 (0.67)	5.22 (0.89)	5.69 (1.05)
Outdegree	−2.38*** (0.10)	−2.69*** (0.12)	−2.32*** (0.13)
Reciprocity	2.86*** (0.19)	2.46*** (0.20)	3.61*** (0.35)
Transitive triplets	–	0.62*** (0.07)	0.80*** (0.08)
Reciprocated degree-activity	–	–	−0.39*** (0.09)

Statistic	GoF $\hat{p}$-value	GoF $\hat{p}$-value	GoF $\hat{p}$-value
Indegree distribution	0.52	0.38	0.75
Outdegree distribution	0.50	0.22	0.62
Triad census	0	0.003	0.11
Geodesic distance distribution	0	0	0

***$p < 0.001$; overall maximum convergence ratios $\leqslant 0.1$.

Table 6. MoM results last three models: parameter estimates (par.), standard errors (SEs), and goodness of fit (GOF) $\hat{p}$ -values.

Effect	Par. (SE)	Par. (SE)	Par. (SE)
Rate 1	7.55 (1.42)	7.88 (1.50)	7.38 (1.33)
Rate 2	5.68 (1.03)	5.91 (1.11)	5.67 (0.99)
Outdegree	−2.78*** (0.23)	−2.62*** (0.14)	−2.89*** (0.20)
Reciprocity	4.08*** (0.49)	3.84*** (0.34)	4.19*** (0.40)
Transitive triplets	−0.39 (0.38)	–	–
Reciprocated degree-activity	−0.35*** (0.07)	−0.36*** (0.07)	−0.33*** (0.07)
gwesp	2.98** (0.98)	2.05*** (0.17)	2.35*** (0.21)
Dense triads	–	–	−0.42** (0.16)

Statistic	GoF $\hat{p}$ -value	GoF $\hat{p}$ -value	GoF $\hat{p}$ -value
Indegree distribution	0.42	0.54	0.49
Outdegree distribution	0.68	0.73	0.79
Triad census	0.003	0.017	0.07
Geodesic distance distribution	0.012	0.014	0.12

gwesp: geometrically weighted shared partners.
**$p < 0.01$; ***$p < 0.001$; overall maximum convergence ratios $\leqslant 0.1$.

The techniques proposed in this article can be directly extended to more elaborate SAOMs, for example, to studies of networks and behavior (Steglich et al., 2010).

The GOF test proposed and applied in this article, and the example data set, are freely available in the R package RSiena through the sienaGOF function. For more information on RSiena, the reader is referred to its homepage at http://www.stats.ox.ac.uk/~snijders/siena/. Scripts using this function, complete with annotations, are available from the homepage.

ORCID iD

Tom AB Snijders https://orcid.org/0000-0003-3157-4157

References

Block P (2015) Reciprocity, transitivity, and the mysterious three-cycle. *Social Networks* 40: 163–173.

Borgatti SP, Everett MG and Johnson JC (2018) *Analyzing Social Networks* (2nd edn). London; Thousand Oaks, CA; New Delhi, India; Singapore: SAGE.

Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognition Letters* 27: 861–874.

Harville DA (1997) *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Hintze JL and Nelson R (1998) Violin plots: A box plot-density trace synergism. *American Statistician* 52: 181–184.

Holland PW and Leinhardt S (1976) Local structure in social networks. *Sociological Methodology* 7: 1–45.

Hunter DR and Handcock MS (2006) Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics* 15: 565–583.

Hunter DR, Goodreau SM and Handcock MS (2008) Goodness of fit of social network models. *Journal of the American Statistical Association* 103: 248–258.

Kaplan D (1990) Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research* 25: 137–155.

Kaplan D (1991) On the modification and predictive validity of covariance structure models. *Quality and Quantity* 25: 307–314.

Kolaczyk ED (2009) *Statistical Analysis of Network Data: Methods and Models*. New York: Springer.

Koskinen JH and Snijders TAB (2007) Bayesian inference for dynamic social network data. *Journal of Statistical Planning and Inference* 137: 3930–3938.

Lehmann EL and Romano JP (2005) *Testing Statistical Hypotheses* (3rd edn). New York: Springer.

Lospinoso J (2012) *Statistical models for social network dynamics*. PhD Thesis, University of Oxford, Oxford.

Lospinoso J and Satchell D (2011) Smoking behavior and friendship formation: The importance of time heterogeneity in studying social network dynamics. In: *Proceedings of 2011 44th Hawaii international conference on system sciences*, Kauai, HI, 4–7 January.

Lospinoso J, Schweinberger M, Snijders TAB, et al. (2011) Assessing and accounting for time heterogeneity in stochastic actor oriented models. *Advances in Data Analysis and Classification* 5: 147–176.

Mahalanobis P (1936) On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India* 2: 49–55.

Michell L and Amos A (1997) Girls, pecking order and smoking. *Social Science and Medicine* 44: 1861–1869.

Pearson M and Michell L (2000) Smoke rings: Social network analysis of friendship groups, smoking, and drug-taking. *Drugs: Education, Prevention, and Policy* 7: 21–37.

Ripley RM, Snijders TAB, Bóda Z, et al. (2019) Manual for Siena version 4.0. Technical report. Oxford: Department of Statistics, Nuffield College, University of Oxford. Available at: http://www.stats.ox.ac.uk/~snijders/siena/

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

Saris WE, Satorra A and Sorbøm D (1987) The detection and correction of specification errors in structural equation models. *Sociological Methodology* 17: 105–129.

Schweinberger M (2012) Statistical modeling of network panel data: Goodness-of-fit. *British Journal of Statistical and Mathematical Psychology* 65: 263–281.

Schweinberger M and Snijders TAB (2007) Markov models for digraph panel data: Monte Carlo-based derivative estimation. *Computational Statistics and Data Analysis* 51(9): 4465–4483.

Snijders TAB (2001) The statistical evaluation of social network dynamics. *Sociological Methodology* 31: 361–395.

Snijders TAB (2017) Stochastic actor-oriented models for network dynamics. *Annual Review of Statistics and Its Application* 4: 343–363.

Snijders TAB, Koskinen J and Schweinberger M (2010a) Maximum likelihood estimation for social network dynamics. *The Annals of Applied Statistics* 4: 567–588.

Snijders TAB, Lomi A and Torlò V (2013) A model for the multiplex dynamics of two-mode and one-mode networks, with an application to employment preference, friendship, and advice. *Social Networks* 35: 265–276.

Snijders TAB, Pattison P, Robins G, et al. (2006) New specifications for exponential random graph models. *Sociological Methodology* 36: 99–153.

Snijders TAB, Steglich CEG and van de Bunt GG (2010b) Introduction to actor-based models for network dynamics. *Social Networks* 32: 44–60.

Snijders TAB, Steglich CEG and Schweinberger M (2007) Modeling the co-evolution of networks and behavior. In: van Montfort K, Oud H and Satorra A (eds) *Longitudinal Models in the Behavioral and Related Sciences*. Mahwah, NJ: Lawrence Erlbaum, pp. 41–71.

Sorbøm D (1989) Model modification. *Psychometrika* 54: 371–384.

Steglich CEG, Snijders TAB and Pearson M (2010) Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology* 40: 329–393.

Veenstra R, Dijkstra JK, Steglich CEG, et al. (2013) Network–behavior dynamics. *Journal of Research on Adolescence* 23: 399–412.

Wasserman S and Faust K (1994) *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

West P and Sweeting H (1996) *Background, rationale and design of the West of Scotland 11 to 16 study*. MRC Medical Sociology Unit Working Paper No. 52. Glasgow: MRC Medical Sociology Unit.

Author biographies

Josh Lospinoso is Chief Research Officer of Shift 5 Inc., a cybersecurity company based in the United States. He is also the author of C++ Crash Course (No Starch Press, 2019). Josh was a Rhodes scholar at the University of Oxford and completed a DPhil in Statistics in 2012.

Tom A.B. Snijders is professor of Methodology and Statistics in the Department of Sociology, University of Groningen; and emeritus fellow of Nuffield College, University of Oxford. He specializes in statistical methods for network analysis and in multilevel analysis.

Appendix 1

For the derivation of (17a), we begin with the i th coordinate of the derivative

$$\frac{\partial}{\partial \theta_i} D(x, \theta) = \frac{\partial}{\partial \theta_i} \left[\left(A(x) - \mu(\theta) \right)^T \Sigma(\theta)^{-1} \left(A(x) - \mu(\theta) \right) \right]$$

Use the chain rule to obtain

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} D(x, \theta) = & \left[\frac{\partial}{\partial \theta_i} \left(A(x) - \mu(\theta) \right)^T \right] \\
& \left[\Sigma(\theta)^{-1} \left(A(x) - \mu(\theta) \right) \right] \\
& + \left(A(x) - \mu(\theta) \right)^T \\
& \left[\frac{\partial}{\partial \theta_i} \Sigma(\theta)^{-1} \right] \left(A(x) - \mu(\theta) \right) \\
& + \left[\left(A(x) - \mu(\theta) \right)^T \left[\Sigma(\theta)^{-1} \right] \right] \\
& \left[\frac{\partial}{\partial \theta_i} \left(A(x) - \mu(\theta) \right) \right]
\end{aligned} \tag{24}$$

Noting that $A(x)$ does not depend on θ

$$\nabla_\theta \left(A(x) - \mu(\theta) \right) = -\mu'(\theta) = -E_\theta \left(u_\theta(X) A(X) \right) \tag{25}$$

Denote

$$\mu_i'(\theta) = E_\theta \left(\left[u_\theta(X) \right]_i A(X) \right) \tag{26}$$

The middle term requires some more work. Since $\Sigma(\theta)^{-1}$ is symmetric, we have (cf. Harville, 1997)

$$\begin{aligned}
\Xi_i(\theta) = \frac{\partial}{\partial \theta_i} \Sigma(\theta)^{-1} &= -\Sigma(\theta)^{-1} \left[\frac{\partial}{\partial \theta_i} \Sigma(\theta) \right] \Sigma(\theta)^{-1} \\
&= -\Sigma(\theta)^{-1} \Gamma_i(\theta) \Sigma(\theta)^{-1}
\end{aligned} \tag{27}$$

where we have denoted

$$\Gamma_i(\theta) = \frac{\partial}{\partial \theta_i} \Sigma(\theta)$$

$$= \frac{\partial}{\partial \theta_i} \begin{bmatrix} E_\theta\left\{A(X)A(X)^T\right\} \\ -E_\theta\left\{A(X)\right\} E_\theta\left\{\left(A(X)^T\right)\right\} \end{bmatrix}$$

$$= \frac{\partial}{\partial \theta_i} E_\theta\left\{A(X)A(X)^T\right\}$$

$$- \left[\frac{\partial}{\partial \theta_i} E_\theta\left\{A(X)\right\}\right] E_\theta\left\{A(X)^T\right\}$$

$$- E_\theta\left\{A(X)\right\} \left[\frac{\partial}{\partial \theta_i} E_\theta\left\{A(X)^T\right\}\right]$$

$$= E_\theta\left\{[u_\theta(X)]_i A(X)A(X)^T\right\}$$

$$- \mu_i'(\theta)\mu(\theta)^T - \mu(\theta)\mu_i'(\theta)^T$$

(28)

Putting these together, we obtain

$$\frac{\partial}{\partial \theta_i} D(x,\theta) = (A(x) - \mu(\theta))^T \Xi_i(\theta)(A(x) - \mu(\theta))$$

$$- 2\mu_i'(\theta)^T \Sigma(\theta)^{-1} (A(x) - \mu(\theta))$$

(29)