Journal of Data Science 21 (2), 295–309
April 2023
DOI: 10.6339/23-JDS1094
Computing in Data Science

Computing Pseudolikelihood Estimators for Exponential-Family Random Graph Models

Christian S. Schmid[1] and David R. Hunter[2,*]

[1]*F. Hoffmann-La Roche Ltd, Switzerland*
[2]*Pennsylvania State University, USA*

Abstract

The reputation of the maximum pseudolikelihood estimator (MPLE) for Exponential Random Graph Models (ERGM) has undergone a drastic change over the past 30 years. While first receiving broad support, mainly due to its computational feasibility and the lack of alternatives, general opinions started to change with the introduction of approximate maximum likelihood estimator (MLE) methods that became practicable due to increasing computing power and the introduction of MCMC methods. Previous comparison studies appear to yield contradicting results regarding the preference of these two point estimators; however, there is consensus that the prevailing method to obtain an MPLE's standard error by the inverse Hessian matrix generally underestimates standard errors. We propose replacing the inverse Hessian matrix by an approximation of the Godambe matrix that results in confidence intervals with appropriate coverage rates and that, in addition, enables examining for model degeneracy. Our results also provide empirical evidence for the asymptotic normality of the MPLE under certain conditions.

Keywords *Godambe matrix; maximum pseudo-likelihood; parametric bootstrap*

1 Introduction

Network data have become ubiquitous in recent years, both in the data science literature and society generally. The probabilistic modeling of networks has a long history, dating to at least Erdős and Rényi (1959) and Gilbert (1959), who introduce a model where each tie occurs independently with probability p. Holland and Leinhardt (1981) are the first to consider tie dependence within dyads, or node pairs, in their p_1 model, a model that is later generalized by the Markov random graph model of Frank and Strauss (1986). This model assumes that only dyads that share a common node can depend on each other. The exponential random graph model (ERGM), or p^* model as it is called by Wasserman and Pattison (1996), generalizes the Markov random graph model and is to this day a popular way to model complex dependency structures of networks.

The ERGM framework may be described as follows: Let a network on N nodes be represented as an $N \times N$ adjacency matrix y with $y_{ij} = 1$ if there is an edge between i and j, $i \neq j, i, j \in \mathcal{N} = \{1, \ldots, N\}$ and $y_{ij} = 0$ otherwise. For the sake of simplicity, we will confine ourselves to undirected networks, i.e., those where $y_{ij} = y_{ji}$, and disallow self-edges where $y_{ii} \neq 0$. An extension to directed networks is straightforward. The ERGM assumes that an observed network y^{obs} is a realization of matrix-like random variable Y with an underlying probability distribution

*Corresponding author. Email: dhunter@stat.psu.edu.

Received August 1, 2022; Accepted February 24, 2023

defined over all possible networks on N nodes. The ERGM takes the form

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{k(\theta)} \tag{1}$$

for $y \in \mathcal{Y}(\mathcal{N})$, where $\mathcal{Y}(\mathcal{N})$ is the sample space of allowable networks on the given set of nodes and $\theta \in \mathbb{R}^q$ is a vector of parameters. In many applications, $\mathcal{Y}(\mathcal{N})$ denotes the entire set $\{(y \in \mathbb{R}^{N\times N}, y_{ij} = y_{ji} \in \{0, 1\}, y_{ii} = 0\}$ of possible networks on N nodes, while in other applications $\mathcal{Y}(\mathcal{N})$ may be constrained to be a proper subset of this set. The sufficient statistics $g : \mathcal{Y}(\mathcal{N}) \to \mathbb{R}^q$, $y \mapsto (g_1(y), \ldots, g_q(y))$ play a central role in the model, since they enable the inclusion of traditional *exogenous* covariates like a node's observable features as well as *endogenous* statistics, i.e., statistics that allow for inference on the structure of the network. Popular endogenous statistics are a network's number of triangles or the number of ties that share one common node (two-stars). The normalizing constant

$$k(\theta) := \sum_{y^* \in \mathcal{Y}(\mathcal{N})} \exp\{\theta^\top g(y^*)\}, \tag{2}$$

a weighted sum over all possible networks on N nodes, assures that (1) defines a probability model.

2 Maximum Pseudolikelihood Estimation

The estimation of the parameter vector θ has been a major focus in ERGM literature. The challenge lies in the normalizing factor $k(\theta)$ that appears in the likelihood function and requires the calculation of a weighted sum with $2^{N(N-1)/2}$ summands for undirected networks. This number is very large for even relatively small networks, making straightforward calculation and therefore the computation of the maximum likelihood estimator (MLE) in most cases infeasible.

Frank and Strauss (1986) propose, and Strauss and Ikeda (1990) fully develop, an estimation approach based on the maximum pseudolikelihood estimator (MPLE) first introduced for lattice models by Besag (1974). The pseudolikelihood is a special form of a composite likelihood (Lindsay, 1988), an inference function where conditional or marginal densities are multiplied with one another, irrespective of whether these components are independent of each other or not in the true probability model. If independence does not hold, the inference function has the characteristics of a misspecified model's likelihood (White, 1982). For a detailed review of composite likelihood methods, we refer readers to the review by Varin et al. (2011).

Some additional notation is necessary for introducing the pseudolikelihood function. Define y^c_{ij} as the matrix y without its ij (and ji, in the undirected case) entries, often also referred to as the rest of the network; and let y^+_{ij} and y^-_{ij} be the full matrices that agree with all entries of y^c_{ij} but where y_{ij} is forced to be 1 and 0, respectively. Then based on Equation (1), the conditional probability of a tie given the rest of the network satisfies

$$\text{logit}[P_\theta(Y_{ij} = 1 | Y^c_{ij} = y^c_{ij})] = \theta^\top [g(y^+_{ij}) - g(y^-_{ij})], \tag{3}$$

where $\text{logit}(p) := \log p - \log(1 - p)$. We refer to $\Delta_{ij} := g(y^+_{ij}) - g(y^-_{ij})$ as the vector of change statistics; note its implicit dependence on y^c_{ij} even though we simplify notation by omitting the y. Equation (3) corresponds to a logistic regression model for Y_{ij}, where we assume a linear relationship between the predictor variables—here, the change statistics—and the log-odds that

$Y_{ij} = 1$. In the network setting, the assumption of the independence of observations translates to dyadic independence, i.e., the Y_{ij} are mutually independent so that

$$P_\theta(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c) = P_\theta(Y_{ij} = 1).$$

If this is the case, the log-likelihood function is

$$\begin{aligned} p\ell(\theta) &= \log\Bigg(\prod_{ij \in \Omega(\mathcal{N})} P_\theta(Y_{ij} = 1)^{y_{ij}} P_\theta(Y_{ij} = 0)^{1-y_{ij}} \Bigg) \\ &= \sum_{ij \in \Omega(\mathcal{N})} \big[y_{ij}(\theta^\top \Delta_{ij}) - \log\big(1 + \exp\{\theta^\top \Delta_{ij}\}\big) \big], \end{aligned} \tag{4}$$

where $\Omega(\mathcal{N}) = \{ij \mid i, j \in \{1, \ldots, N\},\ i < j\}$ is defined as the set of all undirected dyads and the maximizer of (4) is equivalent to the MLE of a logistic regression fit of (3). This means that the maximizer of (4) can be obtained using standard logistic regression software. In many ERGMs, however, this independence assumption does not hold, making (4) an incorrect log-likelihood function. In this context, (4) is called the log-pseudolikelihood function and maximizing it results in what we refer to as the MPLE. Obtaining the MPLE is simple and fast, but this estimator can be imprecise, since a network's dependency structure is for the sake of simplicity deliberately ignored. In addition, even though the MPLE can be obtained using logistic regression software, the software's output should be treated with caution. We will demonstrate in the two following sections that the standard errors obtained from logistic regression software are not appropriate if the independence assumption does not hold.

Frank and Strauss (1986) and Strauss and Ikeda (1990) were the first to compare the performance of MPLE to MLE for networks. Due to the extreme difficulty of obtaining maximum likelihood estimators, these authors focus their comparisons on models with a univariate sufficient statistic and constrained the set of possible networks to those with a fixed number of edges. In these papers, the univariate sufficient statistics that were the attention of this investigation were the number of two-stars and triangles in a network, respectively. In an undirected network, a two-star is defined as any pair of edges (i, j) and (i, k) sharing a node, and a triangle is defined as any set of edges (i, j), (j, k) and (k, i) on the same three distinct nodes. Frank and Strauss (1986) call the two-star model the cluster model, while Strauss and Ikeda (1990) call the triangle model the triad model. An approximate MLE is achieved by a trial and error method that required the simulation of networks from a model defined by a given parameter value and comparing the simulated network's sufficient statistic to the sufficient statistic in the original network. An estimator is defined as the approximate MLE once the simulated networks yielded the same sufficient statistic as the observed network. The justification of this approach is rooted in the fact that in exponential family distributions, the expectation of the sufficient statistics with respect to the MLE equals the sufficient statistic in the observation (Barndorff-Nielsen, 1978). In other words, ERGMs have the appealing property that the MLE is the same as the method of moments estimator. While Frank and Strauss (1986) solely base their conclusions on the comparison of the MLE and MPLE point estimators, Strauss and Ikeda (1990) take the study one step further by also investigating the mean square error (MSE). The conclusion of this first comparison is that "the two methods appear to give estimators that are about equally good" (Strauss and Ikeda, 1990, p. 207).

Dahmström and Dahmström (1993) compare the MPLE to the MLE on the cluster and triad models for networks with 12 edges and $N = 7$ nodes, a sample space small enough to allow

computation of the exact MLE. Comparing only the point estimates, these authors conclude that the MPLE and MLE can differ significantly.

Corander et al. (1998) compare the two estimation methods for the cluster and triad models based on mean squared error for networks of 40 to 100 nodes, approximating the MLE in a fashion similar to Strauss and Ikeda (1990). The authors conclude that the MLE has a smaller MSE for networks up to size $N = 40$, but that for larger networks both methods perform nearly equivalently well. Corander et al. (1998) also mention that unlike the MLE, the MPLE is not a function of the sufficient statistics, which means that different networks with the same sufficient statistics yield the same MLE but may have different MPLEs. This in turn implies that the MPLE violates the likelihood principle (Barnard et al., 1962; Birnbaum, 1962) according to which two networks with the same sufficient statistics should yield the same estimator. Schmid and Hunter (2020) further discuss the use of MPLE as a method for aiding the search for an approximate MLE based on a stochastic algorithm, sometimes called Markov chain maximum likelihood estimation (MCMLE).

All potential sufficient statistics as well as the boundary of the corresponding convex hull can be obtained in R using `ergm.allstats()` in the *ergm* package (Handcock et al., 2022) and `gConvexHull()` in the *rgeos* package (Bivand and Rundel, 2020). The convex hull of a set of vectors, which has an important role in determining the existence of maximum likelihood estimators for exponential family models as we shall see later, is defined as the unique intersection of all convex sets containing all of the vectors or, equivalently, the smallest convex set containing all the vectors. Directly computing all possible sufficient statistic combinations can take several hours for a network on only 9 nodes, with computation increasing exponentially for each additional node. Code that implements this calculation is found in the Supplementary Material section.

The top panel of Figure 1 visualizes every possible 2-dimensional vector of number of edges and triangles of a network on $N = 9$ nodes, along with the convex hull of these vectors. Standard exponential family theory shows that the MLE for an ERGM using these sufficient statistics exists for all observable networks except those whose statistics lie on the boundary of the convex hull. This theory does not apply to the MPLE. Having 26 edges and 25 triangles, the network depicted in the bottom left panel of Figure 1 has statistics lying inside the convex hull, which guarantees the existence of an MLE, but interestingly one can show that the MPLE for this network does not exist. Konis (2007) shows that if the data may be separated, in the sense that there exists a vector β such that

$$\beta^\top [g(y_{ij}^+) - g(y_{ij}^-)] \begin{cases} < 0 & \text{when } y_{ij} = 0, \\ > 0 & \text{when } y_{ij} = 1, \end{cases}$$

then the MPLE does not exist. Konis (2007) argues that finding such β can be posed as a linear programming problem. In particular, consider the program

$$\begin{aligned} &\text{maximize} && (e^\top \bar{X})\beta \\ &\text{subject to} && \bar{X}^\top \beta \geqslant 0, \end{aligned} \tag{5}$$

where $e^\top$ is a vector of ones and $\bar{X}$ is the design matrix $g(y_{ij}^+) - g(y_{ij}^-)$ modified so that each element in a row corresponding to a dyad with no tie, i.e., for which $y_{ij} = 0$, is multiplied by -1. If (5) has a nonzero solution, then the data are separable and the MPLE does not exist.

The network depicted in the bottom left of Figure 1 results in separated data, and hence does not yield an MPLE. However, this does not consequently mean that all networks with the

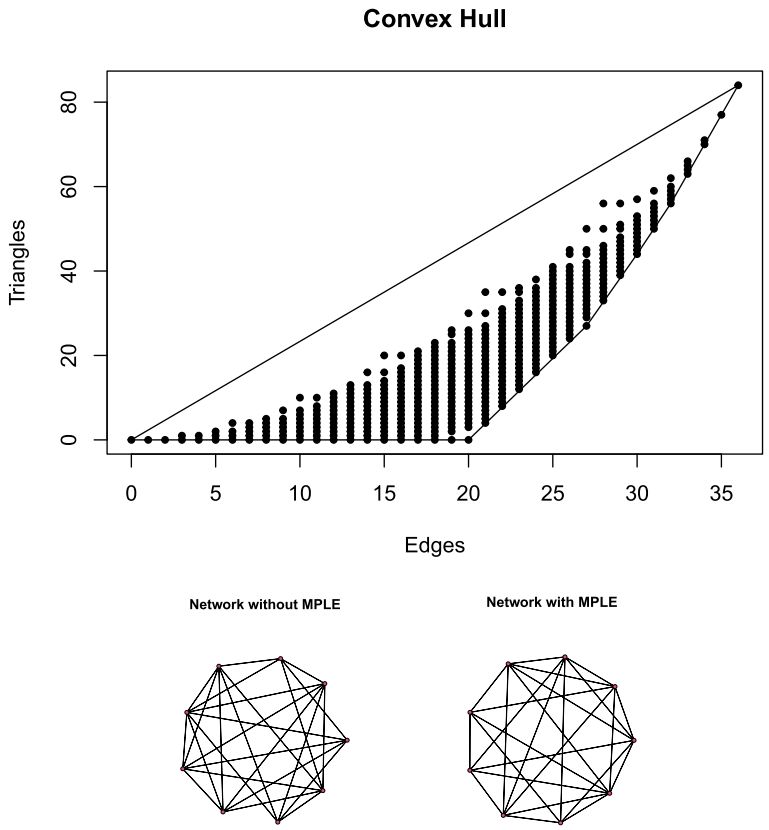

Figure 1: Top: Convex hull and potential sufficient statistics for a network of size $N = 9$ accounting for the number of edges and triangles. Bottom left: A network with 26 edges and 25 triangles that has no MPLE. Bottom right: A network with 26 edges and 25 triangles that has an MPLE.

same sufficient statistics do not have an MPLE. On the contrary, the network depicted in the bottom right of Figure 1 has 26 edges and 25 triangles just as the network to its left, with the difference of having an MPLE.

From version 4.0 onward, the *ergm* package automatically tests for the existence of the MPLE using the *rcdd* package (Geyer and Meeden, 2019). Code that creates the network depicted in the bottom left panel of Figure 1 and tests it for the existence of the MPLE is included in the Supplementary Material.

Another popular estimation approach for models with intractable normalizing constants is Markov Chain maximum likelihood estimation (MCMLE), first proposed by Geyer and Thompson (1992) and then adapted to the ERGM framework by Snijders (2002) and Hunter and Handcock (2006). This family of estimation techniques attempts to approximate the MLE by estimating the normalizing constant through networks that were sampled using MCMC methods. The idea is that for any chosen $\theta_0 \in \mathbb{R}^q$ one can approximate the log-likelihood function by

$$\ell(\theta) - \ell(\theta_0) \approx (\theta - \theta_0)^\top g(y) - \log\left(\frac{1}{L}\sum_{i=1}^{L} \exp\{(\theta - \theta_0)^\top g(y_i)\}\right),$$

where $y_1, \ldots, y_L$ are networks sampled from the distribution defined by θ_0.

The introduction of MCMLE opened further possibilities for the evaluation of the MPLE. Analyzing two data sets (Sampson, 1968; Krackhardt, 1987) and comparing MCMLE and MPLE and the resulting standard errors, Snijders (2002) conclude that the MPLE has a tendency to underestimate standard errors and should therefore not be trusted. Robins et al. (2007) and Lubbers and Snijders (2007) come to the same conclusion. These results are to be expected, especially since Strauss and Ikeda (1990) already clarify that "the quoted standard errors of the estimated parameters do not apply, because the [...] observations in the regression are certainly not independent" (p. 207). Despite the warning, Snijders (2002), Robins et al. (2007), and Lubbers and Snijders (2007) all appear to draw their conclusions from the standard output from logistic regression-based estimates of the standard errors, which are based on an incorrect model.

van Duijn et al. (2009) investigate the efficiency, bias, standard errors, and confidence interval coverage rates of MPLE and of MCMLE for an undirected network on 36 nodes (Lazega, 2001) in natural and mean value parameter space. The mean value parameter space is defined by the bijective mapping $\mu : \mathbb{R}^q \to \mathcal{C}$, $\mu(\theta) = \mathbb{E}_\theta[g(Y)]$, with $\mathcal{C}$ denoting the interior of the convex hull of the sample space of sufficient statistics. For their simulation studies, van Duijn et al. (2009) treat a model's estimated MLE as the true parameter value and simulate networks from the corresponding probability distribution. Regarding the estimators' efficiency, the authors conclude that "the MLE is substantially more efficient than the MPLE" and that the difference is even more pronounced in mean value parameter space. In their studies, the MLE has larger bias in natural parameter space, while the MPLE has larger bias in mean value parameter space. The MPLE standard errors obtained from the logistic regression output lead to MPLE-based confidence intervals with coverage rates far below the nominal confidence level, confirming the conclusion of Snijders (2002) that they are in general underestimated.

3 Estimating Standard Errors for MPLE

Although Strauss and Ikeda (1990) as well as van Duijn et al. (2009) acknowledge that MPLE standard errors obtained from logistic regression output are unsuitable, to the best of our knowledge, no one has yet formally introduced a correct way to specify standard errors for the MPLE in ERGMs.

Based on the log-pseudolikelihood (4), let us define $s(\theta)$ to be the vector of first derivatives,

$$s_k(\theta) = \frac{\partial}{\partial \theta_k} p\ell(\theta) = \sum_{ij \in \Omega(\mathcal{N})} \left(Y_{ij} \Delta_{ijk} - \frac{\exp\{\theta^\top \Delta_{ij}\}}{1 + \exp\{\theta^\top \Delta_{ij}\}} \Delta_{ijk} \right), \tag{6}$$

and $J(\theta)$ the negative Hessian matrix,

$$J_{kl}(\theta) = -\frac{\partial}{\partial \theta_l} s_k(\theta) = \sum_{ij \in \Omega(\mathcal{N})} \left(\frac{\exp\{\theta^\top \Delta_{ij}\}}{\left[1 + \exp\{\theta^\top \Delta_{ij}\}\right]^2} \Delta_{ijk} \Delta_{ijl} \right), \tag{7}$$

where $k, l \in \{1, \ldots, q\}$ and Δ_{ijk} denotes the kth coordinate of the vector Δ_{ij}. While Δ_{ijk} may depend on Y^c_{ij}, for now we follow standard convention in considering Δ_{ijk} to be nonrandom by defining the change statistics based on the observed network y^{obs}. This convention is justified in the case where the log-likelihood coincides with the log-pseudolikelihood, since in this case the edge indicators Y_{ij} are all mutually independent and Δ_{ijk} does not depend on Y^c_{ij}. Thus, the

Hessian matrix does not depend on the random variable Y_{ij}, as is standard in an exponential family model. We conclude that $J(\theta) = -\nabla s(\theta) = -\mathbb{E}_\theta[\nabla s(\theta)]$, i.e., the negative Hessian matrix is both the Fisher information and the observed Fisher information. Furthermore, $J(\theta)^{-1}$ is the approximate covariance matrix used by logistic regression software to estimate standard errors.

One feature of a correctly specified likelihood $\ell(\theta)$ is that the Bartlett identities,

$$\mathbb{E}_\theta[\ell'(\theta)] = 0, \tag{8}$$

$$\mathrm{Var}_\theta[\ell'(\theta)] = -\mathbb{E}_\theta[\ell''(\theta)], \tag{9}$$

hold, which justifies $J(\hat{\theta})^{-1}$ as the covariance matrix of $\hat{\theta}$ according to standard asymptotic theory for maximum likelihood estimators. However, the pseudolikelihood is a form of misspecified likelihood, where (8) and (9) do not apply anymore, which consequently makes $J(\hat{\theta})^{-1}$ an incorrect covariance matrix for an estimator $\hat{\theta}$.

A more suitable method to estimate MPLE standard errors is by the calculation of the Godambe matrix (Godambe, 1960), also known as the sandwich information matrix, as demonstrated for example by Okabayashi et al. (2011) for Potts models. The Godambe matrix is

$$G(\theta) = J(\theta)^{-1} V(\theta) J(\theta)^{-1}, \tag{10}$$

where $J(\theta)$ is referred to in this context as the sensitivity matrix and $V(\theta) = \mathrm{Var}_\theta[s(\theta)]$ is called the variability matrix. We can justify the Godambe matrix by usual Taylor approximation

$$s(\hat{\theta}) \approx s(\theta) + J(\theta)(\hat{\theta} - \theta),$$

where since $s(\hat{\theta}) = 0$ we obtain

$$\hat{\theta} - \theta \approx [-J(\theta)]^{-1} [s(\theta)]. \tag{11}$$

The usual derivation of the Godambe matrix relies on the multivariate central limit theorem. However, $s(\theta)$ is not the sum of independent and identically distributed random vectors Indeed, Shalizi and Rinaldo (2013) show that many ERGMs are not consistent under sampling. Here, "consistent" refers to the context in which the number of nodes N increases without bound, rather than the context where we observe independent networks $y_1^{obs}, y_2^{obs}, \ldots$ on the same fixed set of nodes. In fact, both MLE and MPLE are consistent and asymptotically normal in the latter context, as shown by Arnold and Strauss (1988). However, it is not the prevailing situation that multiple networks are being sampled from a common distribution; it is more common to observe a subnetwork from a hypothesized larger population network whose size we conceptualize as growing without bound.

Although standard asymptotics do not apply here, we may obtain Equation (10) as an approximation to the variance of the MPLE by simply taking the variance of Equation (11):

$$\mathrm{Var}(\hat{\theta}) \approx [-J(\theta)]^{-1} \mathrm{Var}[s(\theta)] [-J(\theta)]^{-1}.$$

The variability matrix $V(\theta)$ cannot in general be directly computed for an ERGM. For this reason, we propose to approximate $V(\theta)$ by simulating R networks $y_1, \ldots, y_R$ from the distribution defined by the MPLE and then to calculate the vector of first derivatives of the pseudolikelihood function $s^1(\theta), \ldots, s^R(\theta)$ for each of the simulated networks. Here, the superscript r

indicates the vector of first derivatives of the rth simulated network. Let $\bar{s}(\theta) = R^{-1}\sum_{r=1}^{R} s^r(\theta)$ be the sample mean vector. Then

$$\hat{V}(\theta) = \frac{1}{R-1}\sum_{r=1}^{R}\big[s^r(\theta) - \bar{s}(\theta)\big]\big[s^r(\theta) - \bar{s}(\theta)\big]^T. \tag{12}$$

If $\tilde{\theta}$ denotes the MPLE, then the Godambe matrix can be estimated as

$$\hat{G}(\tilde{\theta}) = J(\tilde{\theta})^{-1}\hat{V}(\tilde{\theta})J(\tilde{\theta})^{-1}. \tag{13}$$

The estimation based on the Godambe matrix therefore requires the simulation of networks, which may appear to be a disadvantage. However, Schmid and Desmarais (2017) point out that the simulation of networks serves the dual purpose of helping assess model degeneracy as discussed by Schweinberger (2011), among others: Simulated networks operate as a potential warning sign if the estimated probability distribution does not produce networks that appear to be sampled from the same distribution as the observed network. As we will illustrate later, the MPLE is especially prone to defining probability distributions that put most mass on networks that do not resemble the observed network.

The *ergm* package can also handle an offset in the model, i.e., a model of the form

$$\text{logit}[P_\theta(Y_{ij} = 1 | Y^c_{ij} = y^c_{ij})] = \theta^\top[g(y^+_{ij}) - g(y^-_{ij})] + \beta[t(y^+_{ij}) - t(y^-_{ij})],$$

where $\beta : \mathbb{R}^d \to \mathbb{R}$ is a function of $\delta_{ij} = t(y^+_{ij}) - t(y^-_{ij})$, the change statistic corresponding to a d-dimensional statistic vector $t(y)$ defined for the offset. Common forms of the β function are $\log(\delta_{ij})$ in the case $d = 1$ and $\gamma^\top\delta_{ij}$ for a known $\gamma \in \mathbb{R}^d$. If an offset is present, then the calculation of the inverse Hessian matrix and the estimation of the Godambe matrix have to be adjusted. In particular, the score function (6) changes to

$$s_k(\theta) = \frac{\partial}{\partial\theta_k} p\ell(\theta) = \sum_{ij\in\Omega(\mathcal{N})}\left(y_{ij}\Delta_{ijk} - \frac{\exp\{\theta^\top\Delta_{ij} + \beta(\delta_{ij})\}}{1 + \exp\{\theta^\top\Delta_{ij} + \beta(\delta_{ij})\}}\Delta_{ijk}\right),$$

and the negative Hessian matrix (7) becomes

$$J_{kl}(\theta) = -\frac{\partial}{\partial\theta_l}s_k(\theta) = \sum_{ij\in\Omega(\mathcal{N})}\left(\frac{\exp\{\theta^\top\Delta_{ij} + \beta(\delta_{ij})\}}{\big[1 + \exp\{\theta^\top\Delta_{ij} + \beta(\delta_{ij})\}\big]^2}\Delta_{ijk}\Delta_{ijl}\right).$$

In addition to offset models, the *ergm* package can estimate models with a wide range of sample space constraints. One particular sample space constraint, which applies for example to citation networks, is that a document that has been published before another document can not cite the latter; see Schmid et al. (2021) for a citation-network application. Let $\Psi(\mathcal{N}) \subset \Omega(\mathcal{N})$ be the constrained sample space of a network on N nodes. Then the pseudolikelihood simplifies to

$$p\ell_\Psi(\theta) = \sum_{ij\in\Psi(\mathcal{N})}\big[y_{ij}(\theta^\top\Delta_{ij}) - \log\big(1 + \exp\{\theta^\top\Delta_{ij}\}\big)\big], \tag{14}$$

and the score function (6) and the negative Hessian matrix (7) are similarly modified.

4 Simulation Studies

We calculate MPLE confidence intervals based on the Fisher information and Godambe matrices and study the coverage rates of both methods. In addition, we compare coverage rates of MCMLE confidence intervals (Hunter and Handcock, 2006) and 95% parametric bootstrap confidence intervals of the MPLE as introduced by Schmid and Desmarais (2017). The parametric bootstrap confidence intervals are obtained by simulating 500 new networks from the distribution defined by the MPLE, and then estimating the MPLEs for each of the 500 simulated networks. The standard error obtained from the 500 MPLEs is taken as the standard error that results in the bootstrap confidence interval.

We compare the coverage rates of all four methods in a simulation study and on a real-life network. For the simulation study, we follow Desmarais and Cranmer (2012) and consider the undirected (ρ, σ, τ)-model as introduced by Frank and Strauss (1986), where $\theta = (\rho, \sigma, \tau)$ represent the parameters for a network's numbers of edges, two-stars, and triangles, respectively. We set $\rho = -0.25$, $\sigma = -0.2$, and $\tau = 0.5$ and simulate 500 undirected networks for four different sizes: $N = 50$, 100, 200, and 300. ERGMs with triangle statistics and/or two-star statistics are well-known to be subject to problems of degeneracy (see, for instance, Robins et al., 2007; Handcock et al., 2008). While a full discussion of degeneracy is beyond the scope of this article, we reiterate that, as noted in Section 3, the fact that our studies require multiple simulated networks from this model using these parameter settings—which reflect settings used in previous work on MPLE—provides strong evidence that degeneracy is not a practical concern in these tests.

For each simulated network, we obtain the MCMLE with its corresponding 95% confidence interval as well as the MPLE with 95% confidence intervals estimated by the Fisher matrix, the Godambe matrix, and parametric bootstrapping. Code used to implement these tests is included in the Supplementary Material. The results are summarized in Table (1).

As expected, MCMLE confidence intervals yield coverage rates close to the anticipated 95% regardless of network size. On the other hand, the Fisher intervals are nowhere close to the

Table 1: Coverage rates of 95% confidence intervals for the (ρ, σ, τ)-model for four different network sizes ($N = 50, 100, 200, 300$).

N=50	Edges	Two-stars	Triangles	N=100	Edges	Two-stars	Triangles
MCMLE	0.952	0.956	0.964	MCMLE	0.940	0.948	0.940
Fisher	0.744	0.742	0.770	Fisher	0.676	0.702	0.750
Godambe	0.952	0.948	0.964	Godambe	0.954	0.952	0.936
Bootstrap	0.902	0.902	0.962	Bootstrap	0.918	0.926	0.942

N=200	Edges	Two-stars	Triangles	N=300	Edges	Two-stars	Triangles
MCMLE	0.960	0.948	0.946	MCMLE	0.950	0.946	0.956
Fisher	0.610	0.620	0.746	Fisher	0.608	0.620	0.762
Godambe	0.952	0.954	0.946	Godambe	0.950	0.950	0.942
Bootstrap	0.920	0.924	0.936	Bootstrap	0.912	0.922	0.938

desired coverage rate, with results ranging between 60% and 76%. These results, however, are not surprising, since an improper method to obtain standard errors was applied. Interestingly, the coverage rate for this method appears to worsen as the network size increases. Calculating the MPLE standard errors based on the Godambe matrix, however, yields confidence intervals that perform just as well as MCMLE confidence intervals. The fourth method, confidence intervals by parametric bootstrap, clearly outperforms the logistic regression results, but also appears to not quite reach the anticipated coverage rates.

Next we apply the four methods on the same data used by van Duijn et al. (2009), a collaboration network between 36 partners within a New England law firm (Lazega, 2001). We treat this network as an undirected network and only consider an edge between two partners if both sides indicate to have collaborated with each other. The model we use is specified by Hunter and Handcock (2006) and only slightly modified by van Duijn et al. (2009). The endogenous statistics consist of the number of edges, which is equivalent to the intercept in a logistic model, and the geometrically weighted edgewise shared partners statistic (GWESP), a statistic used to model the tendency towards triangles and clustering (Hunter and Handcock, 2006). The decay parameter for the GWESP statistic has been fixed at its MLE of 0.7781. Among the exogenous statistics, we include `seniority`, each partner's seniority rank divided by 36, and `practice` (corporate or litigation) as nodal attributes, as well as `practice`, `gender`, and `office` as dyadic homophily attributes.

As usual for real data sets, the true parameter θ is unknown. Therefore, we obtain a MCMLE and treat this estimate as the truth. At this point we conduct another simulation study, where we simulate 1000 networks from the distribution defined by the MCMLE and calculate 95% confidence intervals for all four methods. Finally, we test whether the true value falls within a computed interval. Coverage rates for the four methods are reported in Table 2.

The coverage rates for the MCMLE as well as for the MPLE using the Fisher matrix align with the results of van Duijn et al. (2009). While the MCMLE yields coverage rates close to 95%, MPLE coverage rates appear to overestimate standard errors with the exception of the `GWESP` statistic, whose coverage rate is clearly too small. MPLE coverage rates that were obtained using an approximated Godambe matrix provide similarly satisfying results as the MCMLE rates for structural and nodal variables. However, standard errors of variables accounting for homophily appear to be too small. The fourth method, parametric bootstrap intervals for the MPLE, provides adequate results.

The third simulation study investigates the performance of the inverse Hessian and Godambe matrix as estimates of the covariance matrix changes as the network size increases. We

Table 2: Coverage Rates for the Lazega Law Firm Collaboration Network.

	Structural		Nodal		Homophily		
	Edges	GWESP	Seniority	Practice	Practice	Gender	Office
MCMLE	0.944	0.920	0.937	0.940	0.958	0.954	0.957
Fisher	0.981	0.763	0.980	0.976	0.978	0.982	0.982
Godambe	0.942	0.922	0.944	0.933	0.939	0.917	0.922
Bootstrap	0.930	0.982	0.966	0.960	0.962	0.943	0.928

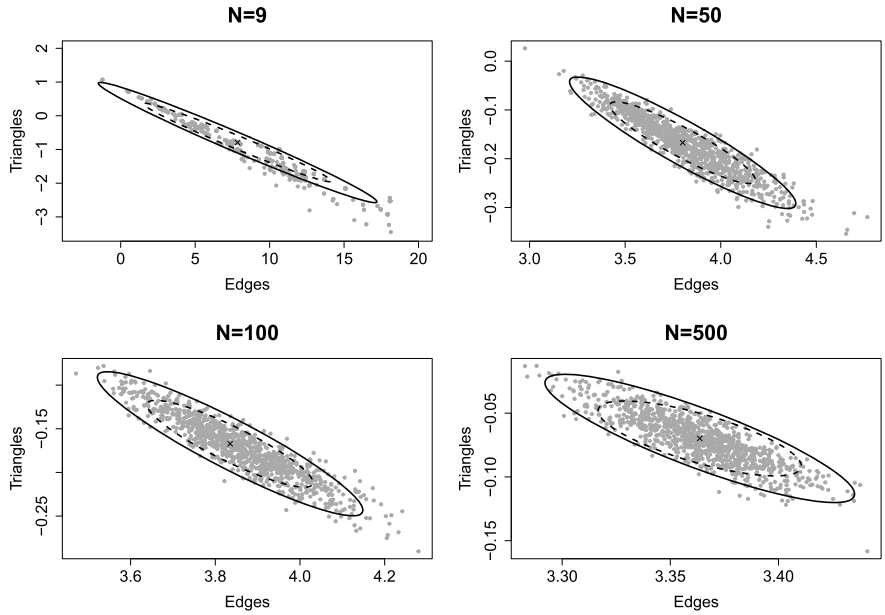

Figure 2: 95% confidence ellipses of the edges-triangle model with n nodes calculated using the inverse Hessian matrix (dashed) and Godambe matrix (solid). The 'x' indicates the parameters of the true model distribution, and every gray dot represents the MPLE of a network that was sampled from that distribution.

follow Krivitsky et al. (2011) by adding an offset to the model to adjust for network size ensuring that a node's mean degree remains the same as the network size increases. Since the number of edges is proportional to the density of a network, which must tend to zero as $N \to \infty$ if the mean degree of a node remains fixed, the parameter for the edges statistic must depend on N. For a given network size N, we simulate an initial network from the ERGM consisting of the number of edges and the number of triangles with parameters set to 4 and -0.2, respectively. The offset is set to $\log(1/N)$ as in Krivitsky et al. (2011). Next, we obtain the MPLE of the simulated network, treat the MPLE as the truth, and simulate 1000 networks. Since the model defined by the MPLE represents the true underlying model, we can calculate the actual inverse Hessian matrix $J(\theta)$ using equation (7). For the estimation of the variability matrix $V(\theta)$, we calculate the MPLE for each of the 1000 simulated networks and apply Equation (12).

Figure 2 visualizes the 95% confidence ellipses calculated from covariance matrices for $N =$ 9, 50, 100, and 500. The small 'x' at the center represents the true MPLE, while every gray dot represents the MPLE of one of the 1000 simulated networks. The dashed lines indicate the 95% confidence ellipses calculated from the inverse Hessian covariance estimate, while the solid lines indicate the 95% confidence ellipses obtained from the estimated Godambe matrix. The code for creating the $N = 50$ panel in Figure 2 is included in the Supplementary Material.

The multiplier to create confidence ellipses are obtained from a $\chi^2(2)$-distribution, based on standard asymptotic theory, and the corresponding cdf is depicted as solid line in Figure 3. In addition, Figure 3 depicts the empirical coverage rates of the confidence ellipses that were obtained using the Godambe matrix (dashed line) and inverse Hessian matrix (dotted line). Based on this figure, we can compare the expected coverage rates of a correctly specified confidence ellipse with the empirical coverage rates.

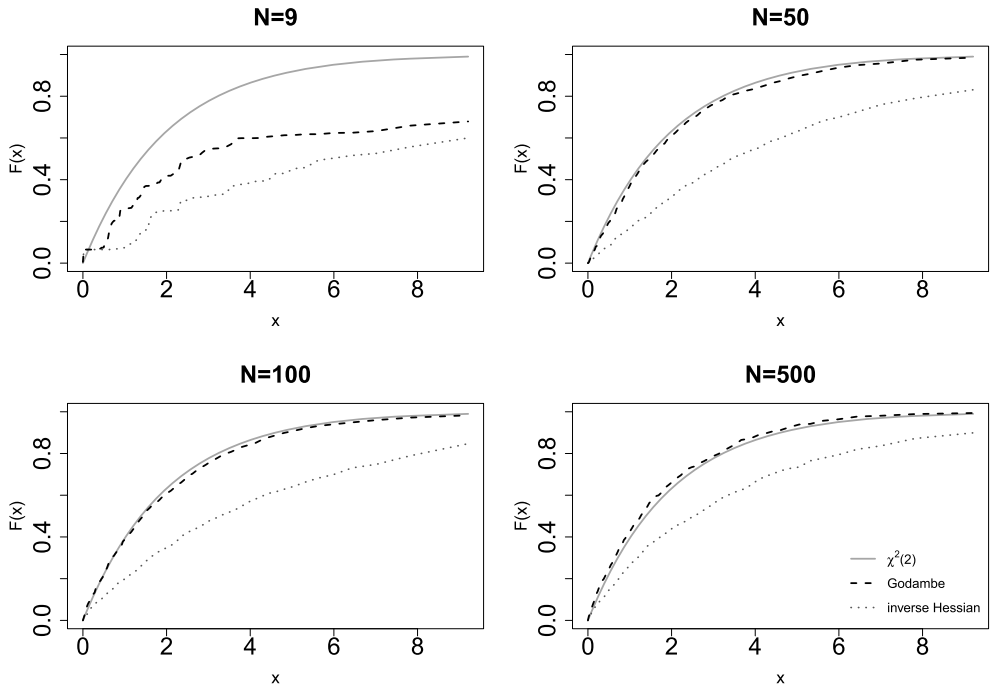

Figure 3: Empirical coverage rates of confidence ellipses obtained by covariance matrix estimates based on the Godambe matrix, depicted by the dashed line (- - -), and the inverse Hessian matrix, depicted by the dotted line (···), as a function of the usual χ^2_2-based multiplier governing the size of the ellipses. The solid gray line shows the true distribution function of a χ^2_2 distribution.

The results underline the conclusions made in the previous simulation studies that confidence intervals that are based on the inverse Hessian matrix are not reliable: Coverage rates of the inverse Hessian confidence ellipses are not close to the intended coverage rates, and the situation does not appear to improve as sample size increases. In contrast, coverage rates of Godambe matrix ellipses are similar to the intended coverage rates for networks except for the small value $N = 9$. Furthermore, it appears that the MPLEs are approximately normally distributed, which opens the question as to whether this framework, with its offset term for correcting for the overall density of the network, could yield a provably asymptotically normal distribution.

5 Discussion

This paper proposes to estimate MPLE standard errors for ERGMs using an estimated Godambe matrix. Even though the exact calculation of the Godambe matrix is infeasible for most ERGMs, the approximated Godambe matrix performed exceptionally well for moderately-sized networks in our simulation studies. Since the approximation requires the simulation of networks from the ERGM defined by the parameter values specified by the MPLE, this method potentially provides a simple check of model degeneracy—a trait all-too-easily ignored when using MPLE due to the ease of obtaining a point estimate.

It seems there is still a lot to learn about the behavior of MPLEs in the context of ERGMs for networks. The possibility that the MPLE in the model used in the `edges + triangles` model of Section 4 might be provably asymptotically normal is particularly interesting. This is because

Shalizi and Rinaldo (2013) demonstrate that the model in question, because its `triangle` term destroys the probabilistic independence of the Y_{ij} edge indicators, cannot enjoy a property called projectivity; that is, given an ERGM for a large network, it cannot be the case that the induced model for any sub-network must be the same. This means among other things that standard asymptotic results cannot hold. However, our results suggest that when parameters are allowed to depend on N, as they do when we use an offset term in Section 4 to control for the overall density of the network, asymptotic results may yet be provable since in this framework, the main result of Shalizi and Rinaldo (2013) does not apply. More study is clearly warranted. Furthermore, if the MPLE may be shown to be asymptotically normal for, say, an `edges + triangles` model when an offset term is employed, perhaps the MLE too enjoys asymptotic normality in this context.

We conclude by pointing out that our work does not directly compare maximum pseudolikelihood and maximum likelihood as methods of estimation in cases when the two methods do not coincide. Certainly each method has its advantages: MPLE is much simpler to compute, whereas MLE satisfies the likelihood principle mentioned in Section 2 among many other theoretical properties of maximum likelihood estimators that apply to all exponential family models. Because of the theoretical advantages enjoyed by MLE, we feel that if computation were irrelevant and if it were possible to easily maximize the likelihood function in all situations, then estimation based on pseudolikelihood would probably be superfluous. Indeed, MPLE has often been used merely as a starting point for stochastic algorithms designed to approximate the MLE, a subject explored in much greater detail elsewhere (e.g., Schmid and Hunter, 2020). On the other hand, this article illustrates that MPLE, when its covariance is properly estimated, holds promise as an estimation method in its own right.

Supplementary Material

The R code file and the ergm package that implements the new methods.

References

Arnold BC, Strauss D (1988). Pseudolikelhood estimation, *Technical Report 164*, University of California, Riverside, Department of Statistics.

Barnard GA, Jenkins GM, Winsten CB (1962). Likelihood inference and time series. *Journal of the Royal Statistical Society. Series A. General*, 125(3): 321–372. https://doi.org/10.2307/2982406

Barndorff-Nielsen O (1978). *Information and Exponential Families in Statistical Theory*. Wiley.

Besag J (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B, Methodological*, 36(2): 192–236. https://doi.org/10.1111/j.2517-6161.1974.tb00999.x

Birnbaum A (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298): 269–306. https://doi.org/10.1080/01621459.1962.10480660

Bivand R, Rundel C (2020). rgeos: Interface to Geometry Engine – Open Source ('GEOS'). R package version 0.5-3.

Corander J, Dahmström K, Dahmström P (1998). Maximum likelihood estimation for Markov graphs, *Research Report*, Department of Statistics, University of Stockholm, 8.

Dahmström K, Dahmström P (1993). ML-estimation of the clustering parameter in a Markov graph model, *Research Report*, Department of Statistics, University of Stockholm.

Desmarais BA, Cranmer SJ (2012). Statistical mechanics of networks: Estimation and uncertainty. *Physica. A*, 391(4): 1865–1876. https://doi.org/10.1016/j.physa.2011.10.018

Erdős P, Rényi A (1959). On random graphs. *Publicationes Mathematicae Debrecen*, 6: 290–297. https://doi.org/10.5486/PMD.1959.6.3-4.12

Frank O, Strauss D (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395): 832–842. https://doi.org/10.1080/01621459.1986.10478342

Geyer CJ, Meeden GD (2019). *rcdd: Computational Geometry*. R package version 1.2-2.

Geyer CJ, Thompson EA (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B, Methodological*, 54(3): 657–699. https://doi.org/10.1111/j.2517-6161.1992.tb01443.x

Gilbert EN (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4): 1141–1144. https://doi.org/10.1214/aoms/1177706098

Godambe VP (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4): 1208–1211. https://doi.org/10.1214/aoms/1177705693

Handcock MS, Hunter DR, Butts CT, Goodreau SM, Krivitsky PN, Morris M (2022). ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks. The Statnet Project (https://statnet.org). R package version 4.2.2.

Handcock MS, Hunter DR, Butts CT, Goodreau SM, Morris M (2008). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1): 1–11. https://doi.org/10.18637/jss.v024.i01

Holland P, Leinhardt S (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373): 33–50. https://doi.org/10.1080/01621459.1981.10477598

Hunter DR, Handcock MS (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3): 565–583. https://doi.org/10.1198/106186006X133069

Konis K (2007). Linear programming algorithms for detecting separated data in binary logistic regression models (Ph.D, thesis, Worcester College, Oxford University.

Krackhardt D (1987). Cognitive social structure. *Social Networks*, 9: 109–134. https://doi.org/10.1016/0378-8733(87)90009-8

Krivitsky PN, Handcock MS, Morris M (2011). Adjusting for network size and composition effects in exponential-family random graph models. *Statistical Methodology*, 8(4): 319–339. https://doi.org/10.1016/j.stamet.2011.01.005

Lazega E (2001). *The Collegial Phenomenon: The Social Mechanism and Cooperation Among Peers in a Corporate Law Partnership*. Oxford University Press.

Lindsay BG (1988). Composite likelihood methods. *Contemporary Mathematics*, 80(1): 221–239.

Lubbers MJ, Snijders TAB (2007). A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes. *Social Networks*, 29(2): 489–507. https://doi.org/10.1016/j.socnet.2007.03.002

Okabayashi S, Johnson L, Geyer C (2011). Extending pseudo-likelihood for Potts models. *Statistica Sinica*, 21(1): 331–347.

Robins G, Snijders T, Wang P, Handcock M, Pattison P (2007). Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29(2): 192–215. https://doi.org/10.1016/j.socnet.2006.08.003

Sampson SF (1968). A novitiate in a period of change: An experimental and case study of relationships, Ph.D. dissertation, Department of Sociology, Cornell University.

Schmid CS, Chen THY, Desmarais BA (2021). *Generative dynamics of Supreme Court citations: Analysis with a new statistical network model. Political Analysis.*

Schmid CS, Desmarais BA (2017). *Exponential random graph models with big networks: Maximum pseudolikelihood estimation and the parametric bootstrap.* In: 2017 IEEE International Conference on Big Data (Big Data). 116–121.

Schmid CS, Hunter DR (2020). Improving ERGM starting values using simulated annealing.

Schweinberger M (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496): 1361–1370. https://doi.org/10.1198/jasa.2011.tm10747

Shalizi CR, Rinaldo A (2013). Consistency under sampling of exponential random graph models. *The Annals of Statistics*, 41(2): 508–535.

Snijders TA (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2): 1–40.

Strauss D, Ikeda M (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409): 202–212. https://doi.org/10.1080/01621459.1990.10475327

van Duijn MA, Gile KJ, Handcock MS (2009). A framework for the comparison of maxmimum pseudo-likelikood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, 31(1): 52–62. https://doi.org/10.1016/j.socnet.2008.10.003

Varin C, Reid N, Firth D (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21: 5–42.

Wasserman S, Pattison PE (1996). Logit models and logistic regression for social networks: I. an introduction to markov graphs and p^*. *Electronic Journal of Statistics*, 61(3): 401–425.

White H (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1): 1–25. https://doi.org/10.2307/1912526