APPROXIMATE VARIATIONAL ESTIMATION FOR A MODEL OF NETWORK FORMATION

Angelo Mele and Lingjiong Zhu*

Abstract—We develop approximate estimation methods for exponential random graph models (ERGMs), whose likelihood is proportional to an intractable normalizing constant. The usual approach approximates this constant with Monte Carlo simulations; however, convergence may be exponentially slow. We propose a deterministic method, based on a variational mean-field approximation of the ERGM's normalizing constant. We compute lower and upper bounds for the approximation error for any network size, adapting nonlinear large deviation results. This translates into bounds on the distance between true likelihood and mean-field likelihood. Monte Carlo simulations suggest that in practice, our deterministic method performs better than our conservative theoretical approximation bounds imply, for a large class of models.

I. Introduction

THIS paper studies variational mean-field methods to approximate the likelihood of exponential random graph models (ERGMs), a class of statistical network formation models that has become popular in sociology, machine learning, statistics, and, more recently, economics. While a large part of the statistical network literature is devoted to models with unconditionally or conditionally independent links (Graham, 2017; Airoldi et al., 2008; Bickel et al., 2013), ERGMs allow for conditional and unconditional dependence among links (Snijders, 2002; Wasserman & Pattison, 1996). These models have recently gained attention in economics because several works have shown that ERGMs have a microeconomic foundation. In fact, the ERGM likelihood naturally emerges as the stationary equilibrium of a potential game, where players engage in a myopic best-response dynamics of link formation (Blume, 1993; Mele, 2017; Badev, 2013; Chandrasekhar, 2016; Chandrasekhar & Jackson, 2014; Boucher & Mourifie, 2017), and in a large class of evolutionary games and social interaction models (Blume, 1993; Durlauf & Ioannides, 2010).

Estimation and inference for ERGMs are challenging because the likelihood of the observed network is proportional to an intractable, normalizing constant that cannot be computed exactly, even in small networks. Therefore, exact maximum likelihood estimation (MLE) is infeasible. The usual estimation approach, the Markov chain Monte Carlo MLE (MCMC-MLE), consists of simulating many networks using the model's conditional link probabilities and approximating the constant and the likelihood with Monte Carlo methods (Snijders, 2002; Koskinen, 2004; Chatterjee & Diaconis, 2013; Mele, 2017). Estimates of the MCMC-MLE converge almost surely to the MLE if the likelihoods are well behaved (Geyer & Thompson, 1992). However, recent literature has shown that the simulation methods used to compute the MCMC-MLE may have exponentially slow convergence, making estimation and approximation of the likelihood impractical or infeasible for a large class of ERGMs (Bhamidi, Bresler, & Sly, 2011; Chatterjee & Diaconis, 2013; Mele, 2017). An alternative is the maximum pseudo-likelihood estimator (MPLE), which finds the parameters that maximize the product of the conditional link probabilities of the model. While MPLE is simple and computationally fast, the properties of the estimator are not well understood, except in special cases, when some regularity conditions are satisfied (Boucher & Mourifie, 2017; Besag, 1974); in practice MPLE may give misleading estimates when the dependence among links is strong (Geyer & Thompson, 1992). Furthermore, since the ERGMs are exponential families, networks with the same sufficient statistics will produce the same MLE but may have different MPLE.

Our work departs from the standard methods of estimation, proposing deterministic approximations of the likelihood, based on the approximated solution of a variational problem. Our strategy is to use a mean-field algorithm to approximate the normalizing constant of the ERGM at any given parameter value (Wainwright & Jordan, 2008; Bishop, 2006; Chatterjee & Diaconis, 2013). We then maximize the resulting approximate log likelihood with respect to the parameters. To be concrete, our approximation consists of using the likelihood of a simpler model with independent links to approximate the constant of the ERGM. The mean-field approximation algorithm finds the likelihood with independent links that minimizes the Kullback-Leibler divergence from the ERGM likelihood. Using this likelihood with independent links, we compute an approximate normalizing constant. We then evaluate the log likelihood of our model, where the exact normalizing constant is replaced by its mean-field approximation.

Our contribution is the computation of exact bounds for the approximation error of the normalizing constant's mean-field estimate. Our proofs use the theoretical machinery of Chatterjee and Dembo (2016) for nonlinear large deviations in models with intractable normalizing constants. Using this powerful tool, we provide explicit lower and upper bounds to the error of approximation for the mean-field, normalizing constant. The bounds depend on the magnitude of the parameters of our model and the size of link externalities (Mele, 2017; Boucher & Mourifie, 2017; Chandrasekhar, 2016; DePaula, 2017). The result holds for dense and moderately

Received for publication September 22, 2017. Revision accepted for publication December 16, 2020. Editor: Bryan S. Graham.

*Mele: Johns Hopkins University; Zhu: Florida State University.

We are grateful to the editor and three excellent referees for their suggestions. We thank Anton Badev, Vincent Boucher, Aureo DePaula, Bryan Graham, Mert Gürbüzbalaban, Matt Jackson, Hiro Kaido, Michael Leung, Xiaodong Liu, and Demian Pouzo for comments on previous versions of this paper. L.Z. is partially supported by NSF grant DMS-1613164.

A supplemental appendix is available online at https://doi.org/10.1162/rest_a_01023.

The Review of Economics and Statistics, January 2023, 105(1): 113–124

https://doi.org/10.1162/rest_a_01023

sparse networks. Remarkably and conveniently, the mean-field error converges to 0 as the network becomes large. This guarantees that for large networks, the log-normalizing constant of an ERGM is well approximated by our mean-field log-normalizing constant.

The main implication of our main result is that we can compute bounds to the distance between the log likelihood of the ERGM and our approximate log likelihood; these also converge in sup-norm as the network grows large. As a consequence, we can use the approximated likelihood for estimation in large networks. If the likelihood is strictly concave, it is possible to show that our approximate estimator converges to the maximum likelihood estimator as long as the network grows large. Furthermore, because our bounds may not be sharp, in practice convergence could be faster than what is implied in these results.

While our method is guaranteed to perform well in large graphs, many applications involve small networks. For example, the school networks data in the National Longitudinal Study of Adoloscent Health (Add Health) (Boucher & Mourifie, 2017; Moody, 2001; Badev, 2013) or the Indian villages in Banerjee et al. (2013) include on average about 200 to 300 nodes. To understand the performance of our estimator in practice, we perform simple Monte Carlo exercises in networks with a few hundred nodes, comparing mean-field estimates to MCMC-MLE and MPLE. Our Monte Carlo results show that in practice, our estimator works better than the theoretical results suggest, for networks with 50 to 1,000 nodes. The median mean-field approximation point estimates are close to the true parameters but exhibit a small bias. Both MCMC-MLE and MPLE show a larger variability of point estimates for the two-stars and triangle parameters, measured as median absolute deviation. When we increase the network size, all three estimators improve, as expected. We conclude that our method's performance is comparable to available estimators in small networks. While our code can be made faster by exploiting efficient matrix algebra libraries and parallelization, the CPU time for estimation is comparable to the estimators implemented in the `ergm` package in R for networks with fewer than 200 nodes.

The main message of our theoretical results and Monte Carlo simulations is that the approximate mean-field approach is a valid alternative to existing methods for estimation of a large class of ERGMs. We note that our theoretical bounds may not be sharp, and in practice the mean-field algorithm may have better performance than what is implied by our conservative results, as confirmed by our Monte Carlo experiments.

To the best of our knowledge, this paper is one of the first works in economics to use mean-field approximations for approximate estimation of complex models. We show that our application of variational approximations has theoretical guarantees, and we can bound the error of approximation. While similar deterministic methods have been used to provide an approximation to the normalizing constant of the ERGM model (Chatterjee & Diaconis, 2013; Amir, Pu, & Espelage, 2012; Mele, 2017; He & Zheng, 2013; Aristoff & Zhu, 2018), we are the first to characterize the variational approximation error for a model with covariates and its computational feasibility.

Our technique can be applied to other models in economics and social sciences. For example, models of social interactions with binary decisions as in Blume (1993), Badev (2013), and Durlauf and Ioannides (2010); models for bundles (Fox & Lazzati, 2017); and models of choices from menus (Kosyakova et al., 2020) have similar likelihoods with intractable normalizing constants. Therefore, our method of approximation may allow estimation of these models for large sets of bundles or menu choices.

The rest of the paper is organized as follows. Section II presents the theoretical model and variational approximations. Section III contains the main theoretical results and the error bounds. Section IV presents the Monte Carlo results, and section V concludes. All the proofs are in the appendix. Additional results, Monte Carlo simulations, and discussions are in the online appendix.

II. Network Formation Model and Variational Methods

A. *Exponential Random Graph Models*

The class of exponential random graphs is an important generative model for networks and has been extensively used in applications in many disciplines (Wasserman & Pattison, 1996; Jackson, 2010; DePaula, 2017; Mele, 2017; Moody, 2001; Wimmer & Lewis, 2010; Amir et al., 2012). In this paper we consider a model with nodal covariates, two-stars, and triangles.

Our model assumes that the network consists of n heterogeneous nodes, indexed by $i = 1, \ldots, n$; each node is characterized by an S-dimensional vector of observed attributes $\tau_i \in \mathcal{X} := \otimes_{j=1}^{S} \mathcal{X}_j$, $i = 1, \ldots, n$. The sets $\mathcal{X}_j$ can represent, for example, age, race, gender, and income.[1] Let α be an $n \times n$ symmetric matrix with elements $\alpha_{ij} := \nu(\tau_i, \tau_j)$, where $\nu : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric function, and let β and γ be scalars. For ease of exposition, we focus on the case in which the attributes are discrete and finite, but our results hold when this assumption is relaxed and the number of attributes is allowed to increase with the size of the network.

The likelihood $\pi_n(g, \alpha, \beta, \gamma)$ of observing the adjacency matrix g depends on the composition of links, the number of two-stars, and the number of triangles

$$\pi_n(g; \alpha, \beta, \gamma) = \frac{\exp\left[Q_n(g; \alpha, \beta, \gamma)\right]}{\sum_{\omega \in \mathcal{G}_n} \exp\left[Q_n(\omega; \alpha, \beta, \gamma)\right]}, \tag{1}$$

[1]For instance, if we consider gender and income, then $S = 2$, and we can take $\otimes_{j=1}^{2} \mathcal{X}_j = \{\text{male, female}\} \times \{\text{low, medium, high}\}$. The sets $\mathcal{X}_j$ can be both discrete and continuous. For example, if we consider gender and income, we can also take $\otimes_{j=1}^{2} \mathcal{X}_j = \{\text{male, female}\} \times [\$50{,}000, \$200{,}000]$. Below we restrict the covariates to be discrete, but we allow the number of types to grow with the size of the network.

where the function Q is called a *potential function* and takes the form

$$Q_n(g;\alpha,\beta,\gamma)=\sum_{i=1}^n\sum_{j=1}^n \alpha_{ij}g_{ij}+\frac{\beta}{2n}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n g_{ij}g_{jk} + \frac{2\gamma}{3n}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n g_{ij}g_{jk}g_{ki}, \tag{2}$$

and $c(\alpha,\beta,\gamma):=\sum_{\omega\in\mathcal{G}_n}\exp\left[Q_n(\omega;\alpha,\beta,\gamma)\right]$ is a normalizing constant that guarantees that the likelihood, equation (1), is a proper distribution. The second and third terms of the potential function, equation (2), are the counts of two-stars and triangles in the network, rescaled by n. We rewrite equation (1) as

$$\pi_n(g;\alpha,\beta,\gamma)=\exp\left\{n^2\left[T_n(g;\alpha,\beta,\gamma)-\psi_n(\alpha,\beta,\gamma)\right]\right\}, \tag{3}$$

where $T_n(g;\alpha,\beta,\gamma)=Q_n(g;\alpha,\beta,\gamma)n^{-2}$ is the potential scaled by n^2, and the log-normalizing constant (scaled by n^2) is

$$\psi_n(\alpha,\beta,\gamma)=\frac{1}{n^2}\log\sum_{\omega\in\mathcal{G}_n}\exp\left[n^2T_n(\omega;\alpha,\beta,\gamma)\right], \tag{4}$$

and $\mathcal{G}_n:=\{\omega=(\omega_{ij})_{1\le i,j\le n}:\omega_{ij}=\omega_{ji}\in\{0,1\},\omega_{ii}=0,1\le i,j\le n\}$ is the set of all binary matrices with n nodes. The rescaling of the potential and the log-normalizing constant is necessary for the asymptotic results, to avoid the explosion of the potential function as the size of the network grows large.

B. *Microeconomic Equilibrium Foundations*

ERGMs caught the attention of economists because recent work proves a behavioral and equilibrium interpretation of the likelihood, equation (3).[2] In fact, these likelihoods naturally arise as the equilibrium of best-response dynamics in potential games (Blume, 1993; Monderer & Shapley, 1996; Butts, 2009; Mele, 2011).

To be concrete, we consider the following game. Players' payoffs are a function of the composition of direct links, friends' popularity, and the number of common friends. The utility of network g for player i is given by

$$u_i(g,\tau)=\sum_{j=1}^n\alpha_{ij}g_{ij}+\frac{\beta}{n}\sum_{j=1}^n\sum_{k=1}^n g_{ij}g_{jk} + \frac{\gamma}{n}\sum_{j=1}^n\sum_{k=1}^n g_{ij}g_{jk}g_{ki}. \tag{5}$$

Each player forms links with other nodes, maximizing utility (see equation 5), but taking into account the strategies of other players. We can show that this game of network formation converges to an exponential random graph in a stationary equilibrium, under the following assumptions:[3] (a) the network formation is sequential, with only two active players in each period; (b) two players meet over time with probability $\rho_{ij}:=\rho(\tau_i,\tau_j,g_{-ij})>0$, where g_{-ij} indicate the network g but link g_{ij}, and these meetings are i.i.d. over time; and (c) before choosing whether to form or delete a link, players receive an i.i.d. logistic shock $(\varepsilon_{ij1},\varepsilon_{ij0})$. At time t, the link g_{ij}^t is formed if

$$u_i(g_{ij}^t=1,g_{-ij}^{t-1},\tau)+u_j(g_{ij}^t=1,g_{-ij}^{t-1},\tau)+\varepsilon_{ij1}^t \ge u_i(g_{ij}^t=0,g_{-ij}^{t-1},\tau)+u_j(g_{ij}^t=0,g_{-ij}^{t-1},\tau)+\varepsilon_{ij0}^t.$$

Mele (2017) shows that such a model is a potential game (Monderer & Shapley, 1996) with potential function given by equation (2). The probability of observing network g in the long run is given by equation (3) (theorem 1 in Mele, 2017); thus equation (3) describes the stationary behavior of the model. In the long run, we observe the pairwise stable networks with high probability, where no pair of players wants to form or delete a link.[4]

C. *Variational Approximations*

The constant $\psi_n(\alpha,\beta,\gamma)$ in equation (4) is intractable because it is a sum over all $2^{\binom{n}{2}}$ possible networks with n nodes; if there are $n=10$ nodes, the sum involves the computation of 2^{45} potential functions, which is infeasible.[5] In the literature on exponential family likelihoods with intractable normalizing constant, this problem is solved by approximating the normalizing constant using Markov chain Monte Carlo (Snijders, 2002; Mele, 2017; Goodreau, Kitts, & Morris, 2009; Koskinen, 2004; Caimo & Friel, 2011; Murray et al., 2006). However, Bhamidi et al. (2011) have shown that such methods may have exponentially slow convergence for many ERGM specifications.

We propose methods that avoid simulations, and we find an approximate likelihood $q_n(g)$ that minimizes the Kullback-Leibler divergence $KL(q_n|\pi_n)$ between q_n and the true likelihood π_n:

$$KL(q_n|\pi_n)=\sum_{\omega\in\mathcal{G}_n}q_n(\omega)\log\left[\frac{q_n(\omega)}{\pi_n(\omega;\alpha,\beta)}\right]$$

[2] Butts (2009), Mele (2017), Chandrasekhar and Jackson (2014), Boucher and Mourifie (2017), Badev (2013), and DePaula (2017).

[3] See Mele (2017) or Badev (2013) for more technical details and variants of these assumptions. See also Chandrasekhar (2016), DePaula (2017), Chandrasekhar and Jackson (2014), and Boucher and Mourifie (2017).

[4] In online appendix E, we provide more details about the microeconomic foundation of the model for interested readers.

[5] See Geyer and Thompson (1992), Murray, Ghahramani, and MacKay (2006), and Snijders (2002) for examples.

$$= \sum_{\omega \in \mathcal{G}_n} q_n(\omega)\big[\log q_n(\omega) - n^2 T_n(\omega; \alpha, \beta, \gamma) + n^2 \psi_n(\alpha, \beta, \gamma)\big] \geq 0. \tag{6}$$

With some algebra we obtain a lower bound for the constant $\psi_n(\alpha, \beta, \gamma)$,

$$\psi_n(\alpha, \beta, \gamma) \geq \mathbb{E}_{q_n}\left[T_n(\omega; \alpha, \beta, \gamma)\right] + \frac{1}{n^2}\mathcal{H}(q_n) := \mathcal{L}(q_n),$$

where $\mathcal{H}(q_n) = -\sum_{\omega \in \mathcal{G}_n} q_n(\omega)\log q_n(\omega)$ is the entropy of distribution q_n, and $\mathbb{E}_{q_n}\left[T_n(\omega; \alpha, \beta, \gamma)\right]$ is the expected value of the rescaled potential, computed according to the distribution q_n.

To find the best likelihood approximation, we minimize $KL(q_n|\pi_n)$ with respect to q_n, which is equivalent to finding the supremum of the lower-bound $\mathcal{L}(q_n)$, that is,

$$\psi_n(\alpha, \beta, \gamma) = \sup_{q_n \in \mathcal{Q}_n} \mathcal{L}(q_n) = \sup_{q_n \in \mathcal{Q}_n}\left\{\mathbb{E}_{q_n}\left[T_n(\omega; \alpha, \beta, \gamma)\right] + \frac{1}{n^2}\mathcal{H}(q_n)\right\}, \tag{7}$$

where $\mathcal{Q}_n$ is the set of all the probability distributions on $\mathcal{G}_n$. We have transformed the problem of computing an intractable sum into a variational problem, that is, a maximization problem.

In general, problem (7) has no closed-form solution; thus, the literature suggests restricting $\mathcal{Q}_n$ to be the set of all completely factorized distribution,[6]

$$q_n(g) = \prod_{i,j} \mu_{ij}^{g_{ij}}(1 - \mu_{ij})^{1-g_{ij}}, \tag{8}$$

where $\mu_{ij} = \mathbb{E}_{q_n}(g_{ij}) = \mathbb{P}_{q_n}(g_{ij} = 1)$. This approximation is called a *mean-field approximation* of the discrete exponential family. Straightforward algebra shows that the entropy of q_n is additive,

$$\frac{1}{n^2}\mathcal{H}(q_n) = -\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\mu_{ij}\log\mu_{ij} + (1 - \mu_{ij})\log(1 - \mu_{ij})\right],$$

and the expected potential can be computed as

$$\mathbb{E}_{q_n}\left[T_n(\omega; \alpha, \beta, \gamma)\right] = \frac{\sum_i \sum_j \alpha_{ij}\mu_{ij}}{n^2} + \beta\frac{\sum_i \sum_j \sum_k \mu_{ij}\mu_{jk}}{2n^3} + \gamma\frac{2\sum_i \sum_j \sum_k \mu_{ij}\mu_{jk}\mu_{ki}}{3n^3}.$$

The mean-field approximation leads to a lower bound of $\psi_n(\alpha, \beta, \gamma)$, because we restricted Q_n, and the simpler variational problem is to find an $n \times n$ symmetric matrix $\boldsymbol{\mu}(\alpha, \beta, \gamma)$ that solves

$$\psi_n(\alpha, \beta, \gamma) \geq \psi_n^{MF}(\boldsymbol{\mu}(\alpha, \beta, \gamma)) = \sup_{\mu \in [0,1]^{n^2}: \mu_{ij} = \mu_{ji}, \forall i,j}\left\{\frac{1}{n^2}\sum_{i,j}\alpha_{ij}\mu_{ij} + \frac{\beta}{2n^3}\sum_{i,j,k}\mu_{ij}\mu_{jk} + \frac{2\gamma}{3n^3}\sum_{i,j,k}\mu_{ij}\mu_{jk}\mu_{ki} - \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\mu_{ij}\log\mu_{ij} + (1 - \mu_{ij})\log(1 - \mu_{ij})\right]\right\}. \tag{9}$$

The mean-field problem is in general nonconvex and the maximization can be performed using any global optimization method (e.g., simulated annealing or Nelder-Mead).[7]

III. Theoretical Results

A. *Convergence of the Variational Mean-Field Approximation*

For finite n, the variational mean-field approximation contains an error of approximation. In the following theorem, we provide a lower and an upper bound to the error of approximation for our model:

Theorem 1. *For fixed network size n, the approximation error of the variational mean-field problem is bounded as*

$$\frac{C_3(\beta, \gamma)}{n} \leq \psi_n(\alpha, \beta, \gamma) - \psi_n^{MF}(\boldsymbol{\mu}(\alpha, \beta, \gamma)) \leq C_1(\alpha, \beta, \gamma)\left(\frac{\log n}{n}\right)^{1/5} + \frac{C_2(\alpha, \beta, \gamma)}{n^{1/2}}, \tag{10}$$

where $C_1(\alpha, \beta, \gamma)$, $C_2(\alpha, \beta, \gamma)$ *are constants depending on* α, β, *and* γ, *and* $C_3(\beta, \gamma)$ *is a constant depending only on* β, γ*:*

$$C_1(\alpha, \beta, \gamma) := c_1 \cdot \left(\max_{i,j}|\alpha_{ij}| + |\beta|^4 + |\gamma|^4 + 1\right),$$

$$C_2(\alpha, \beta, \gamma) := c_2 \cdot \left(\max_{i,j}|\alpha_{ij}| + |\beta| + |\gamma| + 1\right)^{1/2} \cdot (1 + |\beta|^2 + |\gamma|^2)^{1/2},$$

$$C_3(\beta, \gamma) := |\beta| + 4|\gamma|,$$

where $c_1, c_2 > 0$ *are some universal constants.*

[6] See Wainwright and Jordan (2008) and Bishop (2006).

[7] See Wainwright and Jordan (2008) and Bishop (2006) for more details.

The constants in theorem 1 are functions of the parameters α, β, and γ. The upper bound depends on the maximum payoff from direct links and the intensity of payoff from indirect connections. The lower bound only depends on the strength of indirect connections payoffs (popularity and common friends, that is, β and γ). One consequence is that our result holds when the network is dense but also when it is moderately sparse, in the sense that $|\alpha_{ij}|$, $|\beta|$, and $|\gamma|$ can have moderate growth in n instead of being bounded, and the difference of ψ_n and ψ_n^{MF} goes to 0 if $C_1(\alpha, \beta, \gamma)$ grows slower than $n^{1/5}/(\log n)^{1/5}$ and $C_2(\alpha, \beta, \gamma)$ grows slower than $n^{1/2}$ as $n \to \infty$. For example, if $\max_{i,j} |\alpha_{ij}| = O(n^{\delta_1})$, $|\beta| = O(n^{\delta_2})$, $|\gamma| = O(n^{\delta_3})$ where $\delta_1 < \frac{1}{5}$ and $\delta_2, \delta_3 < \frac{1}{20}$, then $\psi_n - \psi_n^{MF}$ goes to 0 as $n \to \infty$. But if the graph is too sparse, for example, $|\beta| = \Omega(n)$, $|\gamma| = \Omega(n)$, then ψ_n cannot be approximated by ψ_n^{MF}.

Our main theorem 1 implies that we can approximate the log likelihood of the ERGM using the mean-field approximated constant.

Proposition 1. *Let $\ell_n(g_n, \alpha, \beta, \gamma)$ be the log likelihood of the ERGM,*

$$\begin{aligned}\ell_n(g_n, \alpha, \beta, \gamma) &:= n^{-2} \log (\pi_n(g_n, \alpha, \beta, \gamma)) \\ &= T_n(g_n, \alpha, \beta, \gamma) - \psi_n(\alpha, \beta, \gamma),\end{aligned}$$

and $\ell_n^{MF}(g_n, \alpha, \beta, \gamma)$ be the mean-field log likelihood obtained by approximating ψ_n with ψ_n^{MF}:

$$\ell_n^{MF}(g_n, \alpha, \beta, \gamma) := T_n(g_n, \alpha, \beta, \gamma) - \psi_n^{MF}(\alpha, \beta, \gamma).$$

Then for any compact parameter space Θ,

$$\begin{aligned}0 \le \sup_{\alpha,\beta,\gamma\in\Theta} \left[\ell_n^{MF} - \ell_n\right] &\le \sup_{\alpha,\beta,\gamma\in\Theta} C_1(\alpha, \beta, \gamma) n^{-1/5} (\log n)^{1/5} \\ &+ \sup_{\alpha,\beta,\gamma\in\Theta} C_2(\alpha, \beta, \gamma) n^{-1/2}. \qquad (11)\end{aligned}$$

Proposition 1 shows that as the network size grows large, the mean-field approximation of the log likelihood ℓ_n^{MF} is arbitrarily close to the ERGM log likelihood ℓ_n. This approximation is similar in spirit to the MCMC-MLE method, where the log-normalizing constant is approximated via MCMC to obtain an approximated log likelihood (Geyer & Thompson, 1992; Snijders, 2002; DePaula, 2017; Moller & Waagepetersen, 2004). The main difference is that our approximation is deterministic and does not require any simulation.

Note that $\ell_n^{MF} = T_n - \psi_n^{MF}$ and $\ell_n = T_n - \psi_n$. If ℓ_n converges to ℓ_∞ uniformly on a compact parameter space Θ, then so does ℓ_n^{MF}. If ℓ_n, ℓ_n^{MF}, and ℓ_∞ are continuous and strictly concave, $\hat{\theta}_n$, $\hat{\theta}_n^{MF}$, the unique maximizers of ℓ_n and ℓ_n^{MF}, will converge to the unique maximizer of ℓ_∞ and hence $\hat{\theta}_n - \hat{\theta}_n^{MF}$ will go to 0 as $n \to \infty$. In the online appendix, we provide further results on the behavior of the mean-field approximation as $n \to \infty$, where we discuss the convergence of the log constant.[8]

The result in proposition 1 can be used to bound the distance between the mean-field estimate and the maximum likelihood estimate for any network size rather than for large n. However, such bounds require additional and stronger assumptions on the shape of the likelihood. Indeed, in appendix B, we show that a sufficient condition for computing the bound is a strongly concave likelihood. Under such assumption, we can use the bound in proposition 1 for the log likelihood to provide a bound on the distance between MLE and mean-field estimator for any network size n. However, these bounds may not be sharp, and therefore we consider them very conservative. In the next section, we show via Monte Carlo simulation that in many cases, our estimator performs better than the bounds would imply.

IV. Estimation Experiments

To understand the performance of the variational approximation in smaller networks, we perform some Monte Carlo experiments. We compare the mean-field approximation with the standard simulation-based MCMC-MLE (Geyer & Thompson, 1992; Snijders, 2002) and the MPLE (Besag, 1974). Our method converges in n^2 steps, while the MCMC-MLE may converge in e^{n^2} steps. The MPLE usually converges faster.

A. Approximation Algorithm for the Normalizing Constant

We implemented our variational approximation for a few models in the R package `mfergm`, available in Github.[9] We follow the statistical machine learning literature and use an iterative algorithm that is guaranteed to converge to a local maximum of the mean-field problem (Wainwright & Jordan, 2008; Bishop, 2006). The algorithm is derived from first-order conditions of the variational mean-field problem.

Let $\boldsymbol{\mu}^*$ be the matrix that solves the variational problem (9). If we take the derivative with respect to μ_{ij} and equate to 0 we get

$$\begin{aligned}\mu_{ij}^* = \Bigg\{1 + \exp\Bigg[&- 2\alpha_{ij} - \beta n^{-1} \sum_{k=1}^{n} \left(\mu_{jk}^* + \mu_{ki}^*\right) \\ &- 4\gamma n^{-1} \sum_{k=1}^{n} \mu_{jk}^* \mu_{ki}^*\Bigg]\Bigg\}^{-1}. \qquad (12)\end{aligned}$$

The logit equation (12) characterizes a system of equations whose fixed point is a solution of the mean-field problem. We

[8] The strict concavity of the likelihood is closely related to the identification of parameters in ERGM models, for which there is a lack of general results (see Mele, 2017; Chatterjee & Diaconis, 2013; Aristoff & Zhu, 2018 for examples in special cases).

[9] See https://github.com/meleangelo/mfergm , with instructions for installation and few examples.

can therefore start from a matrix $\boldsymbol{\mu}$ and iterate the updates in equation (12) until we reach a fixed point, as described in algorithm 1.

Algorithm 1: Approximation of Log-Normalizing Constant.

Fix parameters α, β, γ, and a relatively small tolerance value ϵ_{tol}. Initialize the $n \times n$ matrix $\boldsymbol{\mu}^{(0)}$ as $\mu_{ij}^{(0)} \overset{iid}{\sim} U[0, 1]$, for all i, j. Fix the maximum number of iterations as T. Then for each $t = 0, \ldots, T$:

Step 1. Update the entries of matrix $\boldsymbol{\mu}^{(t)}$ for all $i, j = 1, \ldots, n$:

$$\mu_{ij}^{(t+1)} = \left\{1 + \exp\left[- 2\alpha_{ij} - \beta n^{-1} \sum_{k=1}^{n} \left(\mu_{jk}^{(t)} + \mu_{ki}^{(t)}\right) - 4\gamma n^{-1} \sum_{k=1}^{n} \mu_{jk}^{(t)} \mu_{ki}^{(t)} \right]\right\}^{-1}. \quad (13)$$

Step 2. Compute the value of the variational mean-field log constant $\psi_n^{MF(t)}$ as

$$\psi_n^{MF(t)} = \frac{\sum_i \sum_j \alpha_{ij} \mu_{ij}^{(t)}}{n^2} + \beta \frac{\sum_i \sum_j \sum_k \mu_{ij}^{(t)} \mu_{jk}^{(t)}}{2n^3} + \gamma \frac{2 \sum_i \sum_j \sum_k \mu_{ij}^{(t)} \mu_{jk}^{(t)} \mu_{ki}^{(t)}}{3n^3} - \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\mu_{ij}^{(t)} \log \mu_{ij}^{(t)} + (1 - \mu_{ij}^{(t)}) \log(1 - \mu_{ij}^{(t)})\right].$$

Step 3. Stop at $t^* \leq T$ if: $\psi_n^{MF(t^*)} - \psi_n^{MF(t^*-1)} \leq \epsilon_{tol}$. Otherwise go back to step 1.

The algorithm is initialized at a random uniform matrix $\boldsymbol{\mu}^{(0)}$ and iteratively applies the update, equation (1), to each entry of the matrix, until the increase in the objective function is less than a tolerance level. Since the problem is concave in each μ_{ij}, this iterative method is guaranteed to find a local maximum of equation (9).[10] In our simulations we use a tolerance level of $\epsilon_{tol} = 0.0001$. To improve convergence, we can restart the algorithm from different random matrices, as usually done with local optimizers.[11] This step is easily parallelizable, thus preserving the order n^2 convergence, while the standard MCMC-MLE is an intrinsically sequential algorithm and cannot be parallelized.

B. *Monte Carlo Design*

All the computations in this section are performed on a PC Dell T6610 with 6 Quad-core Intel i7 (48 threads) and 64 GB RAM. We test our approximation using 1,000 simulated networks. Each node i has a binary attribute x_i, $x_i \overset{iid}{\sim} Bernoulli(0.5)$. Let $z_{ij} = 1$ if $x_i = x_j$ and $z_{ij} = 0$ otherwise.

$$t_z(g) := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} z_{ij};\ t_{-z}(g) := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij}(1 - z_{ij}),$$

$$t_e(g) := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij};\ t_s(g) := \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk};\ t_t(g) := \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk} g_{ki}, \quad (14)$$

where $t_e(g)$, $t_s(g)$, and $t_t(g)$ are the fraction of links, two-stars, and triangles, respectively. And $t_z(g)$ and $t_{-z}(g)$ are the fractions of links of the same type and different type, respectively. The log likelihood of the model $\ell_n(g; \alpha, \beta, \gamma)$ is

$$\ell_n(g, x; \alpha, \beta, \gamma) = \alpha_1 t_z(g) + \alpha_2 t_{-z}(g) + (\beta/2) t_s(g) + (2\gamma/3) t_t(g) - \psi_n(\alpha_1, \alpha_2, \beta, \gamma). \quad (15)$$

For computational convenience, we rewrite model (15) in a slightly different but equivalent way,

$$\ell_n(g, x; \tilde{\alpha}, \beta, \gamma) = \tilde{\alpha}_1 t_e(g) + \tilde{\alpha}_2 t_z(g) + (\beta/2) t_s(g) + (2\gamma/3) t_t(g) - \psi_n(\alpha_1, \alpha_2, \beta, \gamma), \quad (16)$$

where we have defined $\tilde{\alpha}_1 := \alpha_2$ and $\tilde{\alpha}_2 := \alpha_1 - \alpha_2$. We use specification (16) in our simulations.[12]

To generate the artificial networks, we draw i.i.d. attributes $x_i \sim Bernoulli(0.5)$, initialize a network with n nodes as an Erdos-Renyi graph with probability $p = e^{\tilde{\alpha}_1}/(1 + e^{\tilde{\alpha}_1})$, and then run the Metropolis-Hastings network sampler using the `simulate.ergm` command in the R package `ergm` to sample 1,000 networks, each separated by 10,000 iterations, and after a burn-in of 10 million iterations.[13] The MCMC-MLE

[10] There are other alternatives to the random uniform matrix. Indeed a simple starting value could be the set of conditional probabilities of the model at parameters α, β, γ. We did not experiment with this alternative method.

[11] In the Monte Carlo exercises, we have experimented with different numbers of restarts of the iterative algorithm. However, it is not clear what would be the optimal number of restarts. A fixed number of restarts could be suboptimal. It seems reasonable to increase this number as the network grows larger.

[12] There are other small differences in how we have specified the model and how we have setup computations using the `statnet` package in R, which can affect the comparability of the simulation results, in particular the normalizations of the sufficient statistics. This is handled by our `mfergm` package to guarantee comparability of the estimates obtained with MCMC-MLE, MPLE, and mean-field approximate inference.

[13] The code is available in the Github package `mfergm`, and the function is `simulate.model#`, where # stands for the model number: 2 is the model with $\gamma = 0$, 3 is the model with $\beta = 0$, and 4 is the model with $\beta \neq 0$ and $\gamma \neq 0$.

Table 1.—Monte Carlo Estimates: Comparison of Three Methods
True Parameter Vector: $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta, \gamma) = (-2, 1, 1, 1)$

	MCMC-MLE				Mean Field				MPLE			
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	γ	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	γ	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	γ
$n = 50$												
median	−2.002	1.024	0.716	−2.042	−2.000	0.998	1.000	0.999	−1.957	1.016	0.118	−0.584
mad	0.295	0.238	3.412	26.132	0.044	0.040	0.012	0.012	0.268	0.179	3.261	16.540
$n = 100$												
median	−1.991	0.991	0.886	1.183	−2.002	0.995	1.001	0.999	−1.974	0.991	0.713	1.020
mad	0.197	0.117	2.324	16.150	0.020	0.017	0.005	0.005	0.178	0.085	2.237	10.478
$n = 200$												
median	−2.000	1.000	1.043	0.438	−2.003	0.995	1.001	0.999	−1.990	1.000	0.853	0.657
mad	0.127	0.064	1.686	10.627	0.009	0.009	0.002	0.002	0.125	0.046	1.613	7.950
$n = 500$												
median	−2.000	1.001	1.000	0.706	−2.002	0.994	1.016	0.992	−1.994	1.001	0.912	0.762
mad	0.084	0.033	1.090	6.962	0.007	0.008	0.023	0.011	0.074	0.023	0.945	4.691

Results of 1,000 Monte Carlo estimates using three methods. MCMC-MLE is the Monte Carlo maximum likelihood estimator of Geyer and Thompson (1992), as implemented in `ergm` in `R`, with a stochastic approximation algorithm (Snijders, 2002). Mean Field is our method. MPLE is the maximum pseudo-likelihood estimate. Each network is generated with a 10 million run of the Metropolis-Hastings sampler of the `ergm` command in R, sampling every 10,000 iterations. mad is the median absolute deviation.

Table 2.—Monte Carlo Estimates: Comparison of Three Methods
True Parameter Vector: $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta, \gamma) = (-3, 2, 1, 3)$

	MCMC-MLE				Mean Field				MPLE			
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	γ	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	γ	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	γ
$n = 50$												
median	−3.041	2.064	0.743	−0.512	−3.007	1.993	1.000	3.000	−3.026	2.083	0.215	1.764
mad	0.476	0.424	3.811	25.109	0.026	0.026	0.013	0.014	0.514	0.401	3.593	16.538
$n = 100$												
median	−3.006	2.015	0.932	0.587	−3.011	1.989	1.000	2.999	−2.991	2.018	0.682	1.773
mad	0.261	0.206	2.538	17.905	0.016	0.016	0.008	0.008	0.259	0.194	2.364	12.123
$n = 200$												
median	−3.012	2.007	1.069	2.807	−3.011	1.988	1.000	2.999	−3.005	2.011	0.932	2.988
mad	0.158	0.117	1.822	11.360	0.008	0.008	0.004	0.004	0.156	0.109	1.714	8.144
$n = 500$												
median	−2.998	2.000	0.951	3.047	−3.011	1.988	1.002	2.999	−2.998	2.001	0.921	3.117
mad	0.096	0.061	1.276	7.191	0.003	0.003	0.002	0.002	0.083	0.049	1.077	5.378

See the notes for table 1.

estimator is solved using the stochastic approximation method of Snijders (2002), where each simulation has a burn-in of 100,000 iterations of the Metropolis-Hastings sampler and networks are sampled every 1,000 iterations. The other convergence parameters are kept at default of the `ergm` package. The MPLE estimate is obtained using the default parameters in `ergm`. To be sure that our results do not depend on the initialization of the parameters, we start each estimator at the true parameter value, thus decreasing the computational time required for convergence. All the code is available in Github for replication.

C. *Results*

The first model has true parameter vector $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta, \gamma) = (-2, 1, 1, 1)$, and the summaries of point estimates are reported in table 1. We show results for $n = 50$, 100, 200, and 500, reporting median and median absolute deviation (mad) of point estimates for each parameter.

The median estimates of the mean-field approximation are quite stable and exhibit a small bias, as is well known in the literature (Wainwright & Jordan, 2008; Bishop, 2006). The median results for MCMC-MLE and MPLE are quite precise for $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ but vary a lot for β and γ, as shown by the large median absolute deviation. Nonetheless the median point estimates of β and γ are slowly converging to the true parameter vector as n increases.[14] Therefore, the mean-field approximation provides estimates in line with MPLE and MCMC-MLE, with more reliability for β and γ in these small sample estimation exercises.

The second set of results is for a model with parameters $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta, \gamma) = (-3, 2, 1, 3$; see table 2). The pattern is similar to table 1. Indeed the mean-field estimator seems to work relatively well in most cases, especially for the estimates of β and γ. For parameters $\tilde{\alpha}_1, \tilde{\alpha}_2$ our mean-field estimator (median) bias persists as n increases. Finally, we also report a simulation with a larger network with $n = 500$, 1,000 in table 3. The results are the same as the other tables, and the mean-field approximation is robustly close to the true parameter values in most simulations.

Monte Carlo experiments suggest that our approximation method performs well in practice. We conclude that in most

[14] Some of the bias in the mean-field approximation may be due to the fact that we only initialize $\boldsymbol{\mu}$ once in these simulations.

Table 3.—Monte Carlo Estimates: Comparison of Three Methods
True Parameter Vector: $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta, \gamma) = (-3, 1, 2, 1)$

	MCMC-MLE				Mean Field				MPLE			
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	γ	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	γ	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	γ
$n = 500$												
median	−3.001	0.998	2.028	−19.034	−3.000	1.000	2.000	1.000	−2.996	1.000	1.488	−7.923
mad	0.086	0.065	7.205	165.600	0.011	0.011	0.0001	0.0001	0.078	0.044	6.345	84.681
$n = 1{,}000$												
median	−2.999	1.004	1.809	−0.716	−3.000	1.000	2.000	1.000	−2.999	1.002	1.757	0.540
mad	0.057	0.037	4.891	125.293	0.005	0.005	0.0001	0.0001	0.049	0.022	4.113	61.328

See the notes for table 1. The case with $n = 1{,}000$ contains only 500 Monte Carlo replications.

cases the mean-field approximation algorithm works better than our conservative theoretical results suggest.[15]

V. Conclusion and Future Work

We have shown that for a large class of exponential random graph models (ERGM), we can approximate the normalizing constant of the likelihood using a mean-field variational approximation algorithm (Wainwright & Jordan, 2008; Bishop, 2006; Chatterjee & Diaconis, 2013; Mele, 2017). Our theoretical results use nonlinear large deviations methods (Chatterjee & Dembo, 2016) to bound the error of approximation, showing that it converges to 0 as the network grows.

Our estimation method consists of replacing the log-normalizing constant in the log likelihood of the ERGM with the value approximated by the mean-field algorithm; we then find the parameters that maximize such approximate log likelihood. Since our approximated constant converges to the true constant in large networks, the approximate log likelihood converges to the correct log likelihood in sup-norm, as the network becomes large. If the likelihoods are well behaved and not too flat around the maximizers, we can also show that our estimate converges to MLE.

Using an iterative procedure to find the approximate mean-field constant, we compare our method to MCMC-MLE and MPLE (Snijders, 2002; Boucher, 2015; Besag, 1974; DePaula, 2017) in a simple Monte Carlo study for small networks. The mean-field approximation exhibits a small bias, but the median estimates are similar to MCMC-MLE and MPLE. Theoretically, our method converges in a number of steps proportional to the number of potential links of a network, while MCMC-MLE could be exponentially slow.

While these results are encouraging, there are several open problems and possible research directions. First, it is not clear that the mean-field estimates are consistent. Our small Monte Carlo seems to indicate that there is a persistent bias term, but there is no general proof in this setting along the lines of Bickel et al. (2013) for stochastic block models. Second, it is not clear that the ERGM model is identified for all parameter values. Indeed some results in this literature suggest otherwise (Chatterjee & Diaconis, 2013; Mele, 2017; Boucher & Mourifie, 2017). A promising research avenue for the future is the use of the large n mean-field approximation to understand identification, similar to what has been done with graph limits in Chatterjee and Diaconis (2013). Third, while the mean-field approximation is simple and we are able to compute the approximation errors, our lower and upper bounds may not be sharp. This raises the question of whether there is another factorization (as in structured mean field) that leads to better approximations and faster convergence (Xing, Jordan, & Russell, 2003). We hope that our work will stimulate additional research and more applications of this class of approximations.

[15] While these results are encouraging, in the appendix, we report some examples of nonconvergence of the mean-field algorithm, mostly due to our iterative algorithm getting trapped in a local maximum in some simulations.

Appendix: Proof of Theorem 1

In this proof we endeavor to follow closely the notation in Chatterjee and Dembo (2016). Suppose that $f : [0, 1]^N \to \mathbb{R}$ is twice continuously differentiable in $(0, 1)^N$, so that f and all its first- and second-order derivatives extend continuously to the boundary. Let $\|f\|$ denote the supremum norm of $f : [0, 1]^N \to \mathbb{R}$. For each i and j, denote

$$f_i := \frac{\partial f}{\partial x_i}, \quad f_{ij} := \frac{\partial^2 f}{\partial x_i \partial x_j}, \tag{A1}$$

and let

$$a := \|f\|, \quad b_i := \|f_i\|, \quad c_{ij} := \|f_{ij}\|. \tag{A2}$$

Given $\epsilon > 0$, $\mathcal{D}(\epsilon)$ is the finite subset of $\mathbb{R}^N$ so that for any $x \in \{0, 1\}^N$, there exists $d = (d_1, \ldots, d_N) \in \mathcal{D}(\epsilon)$ such that

$$\sum_{i=1}^{N} (f_i(x) - d_i)^2 \leq N\epsilon^2. \tag{A3}$$

Let us define

$$F := \log \sum_{x \in \{0,1\}^N} e^{f(x)}, \tag{A4}$$

and for any $x = (x_1, \ldots, x_N) \in [0, 1]^N$,

$$I(x) := \sum_{i=1}^{N} [x_i \log x_i + (1 - x_i) \log(1 - x_i)]. \tag{A5}$$

In the proof, we rely on theorem 1.5 in Chatterjee and Dembo (2016) that we reproduce in theorem 2

Theorem 2 (Chatterjee & Dembo, 2016). *For any* $\epsilon > 0$,

$$\sup_{x\in[0,1]^N}\{f(x)-I(x)\}-\frac{1}{2}\sum_{i=1}^{N}c_{ii}\leq F \leq \sup_{x\in[0,1]^N}\{f(x)-I(x)\}+\mathcal{E}_1+\mathcal{E}_2, \quad \text{(A6)}$$

where

$$\mathcal{E}_1 := \frac{1}{4}\left(N\sum_{i=1}^{N}b_i^2\right)^{1/2}\epsilon+3N\epsilon+\log|\mathcal{D}(\epsilon)|, \quad \text{(A7)}$$

and

$$\mathcal{E}_2 := 4\left(\sum_{i=1}^{N}(ac_{ii}+b_i^2)+\frac{1}{4}\sum_{i,j=1}^{N}(ac_{ij}^2+b_ib_jc_{ij}+4b_ic_{ij})\right)^{1/2} + \frac{1}{4}\left(\sum_{i=1}^{N}b_i^2\right)^{1/2}\left(\sum_{i=1}^{N}c_{ii}^2\right)^{1/2}+3\sum_{i=1}^{N}c_{ii}+\log 2. \quad \text{(A8)}$$

We will use the theorem 2 to derive the lower and upper bound of the mean-field approximation problem. Our results extend theorem 1.7. in Chatterjee and Dembo (2016) from the ERGM with two-stars and triangles to the model that allows nodal covariates. Notice that in our case, the N of the theorem is the number of links: $N = \binom{n}{2}$. Let

$$Z_n := \sum_{x_{ij}\in\{0,1\},x_{ij}=x_{ji},1\leq i<j\leq n} \times e^{\sum_{1\leq i,j\leq n}\alpha_{ij}x_{ij}+\frac{\beta}{2n}\sum_{1\leq i,j,k\leq n}x_{ij}x_{jk}+\frac{2\gamma}{3n}\sum_{1\leq i,j,k\leq n}x_{ij}x_{jk}x_{ki}} \quad \text{(A9)}$$

be the normalizing factor and also define

$$L_n := \sup_{x_{ij}\in[0,1],x_{ij}=x_{ji},1\leq i<j\leq n}\left\{\frac{1}{n^2}\sum_{i,j}\alpha_{ij}x_{ij}+\frac{\beta}{2n^3}\sum_{i,j,k}x_{ij}x_{jk} +\frac{2\gamma}{3n^3}\sum_{i,j,k}x_{ij}x_{jk}x_{ki}-\frac{1}{n^2}\sum_{1\leq i<j\leq n}[x_{ij}\log x_{ij} + (1-x_{ij})\log(1-x_{ij})]\right\}. \quad \text{(A10)}$$

Notice that $n^{-2}Z_n = \psi_n$ and $L_n = \psi_n^{MF}$.

For our model, the function $f : [0,1]^{\binom{n}{2}} \to \mathbb{R}$ is defined as

$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}x_{ij}+\frac{\beta}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}x_{ij}x_{jk} + \frac{2\gamma}{3n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}x_{ij}x_{jk}x_{ki}. \quad \text{(A11)}$$

Then we can compute that

$$a = \|f\| \leq \sum_{i=1}^{n}\sum_{j=1}^{n}|\alpha_{ij}|+\frac{1}{2}|\beta|n^2+\frac{2}{3}|\gamma|n^2 \leq n^2\left[\max_{i,j}|\alpha_{i,j}|+\frac{1}{2}|\beta|+\frac{2}{3}|\gamma|\right]. \quad \text{(A12)}$$

Let $k \in \mathbb{N}$, and H be a finite simple graph on the vertex set $[k] := \{1,\ldots,k\}$. Let E be the set of edges of H and $|E|$ be its cardinality. For a function $T : [0,1]^{\binom{n}{2}} \to \mathbb{R}$,

$$T(x) := \frac{1}{n^{k-2}}\sum_{q\in[n]^k}\prod_{\{\ell,\ell'\}\in E}x_{q_\ell q_{\ell'}}, \quad \text{(A13)}$$

Chatterjee and Dembo (2016, lemma 5.1) showed that for any $i<j$, $i'<j'$,

$$\left\|\frac{\partial T}{\partial x_{ij}}\right\| \leq 2|E|, \quad \text{(A14)}$$

and

$$\left\|\frac{\partial^2 T}{\partial x_{ij}\partial x_{i'j'}}\right\| \leq \begin{cases} 4|E|(|E|-1)n^{-1} & \text{if } |\{i,j,i',j'\}| = 2 \text{ or } 3, \\ 4|E|(|E|-1)n^{-2} & \text{if } |\{i,j,i',j'\}| = 4. \end{cases} \quad \text{(A15)}$$

Therefore, by equation (A14), we can compute that

$$b_{(ij)} = \left\|\frac{\partial f}{\partial x_{ij}}\right\| \leq 2\max_{i,j}|\alpha_{ij}|+2|\beta|+8|\gamma|. \quad \text{(A16)}$$

By equation (A15), we can also compute that

$$c_{(i,j)(i'j')} = \left\|\frac{\partial^2 f}{\partial x_{ij}\partial x_{i'j'}}\right\| \leq \begin{cases} 4\left(\frac{1}{2}|\beta|2(2-1) + \frac{2}{3}|\gamma|3(3-1)\right)n^{-1} & \text{if } |\{i,j,i',j'\}| = 2 \text{ or } 3, \\ 4\left(\frac{1}{2}|\beta|2(2-1) + \frac{2}{3}|\gamma|3(3-1)\right)n^{-2} & \text{if } |\{i,j,i',j'\}| = 4, \end{cases} = \begin{cases} 4\,(|\beta|+4|\gamma|)\,n^{-1} & \text{if } |\{i,j,i',j'\}| = 2 \text{ or } 3, \\ 4\,(|\beta|+4|\gamma|)\,n^{-2} & \text{if } |\{i,j,i',j'\}| = 4. \end{cases} \quad \text{(A17)}$$

Next, we compute that

$$\frac{\partial f}{\partial x_{ij}} = 2\alpha_{ij}+\frac{\partial}{\partial x_{ij}}\left[\frac{\beta}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}x_{ij}x_{jk} + \frac{2\gamma}{3n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}x_{ij}x_{jk}x_{ki}\right]. \quad \text{(A18)}$$

Let T_1 and T_2 be defined as

$$T_1(x) := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} x_{ij}x_{jk},$$

$$T_2(x) := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} x_{ij}x_{jk}x_{ki}. \quad \text{(A19)}$$

Then we have

$$\frac{\partial f}{\partial x_{ij}} = 2\alpha_{ij} + \frac{\beta}{2}\frac{\partial T_1}{\partial x_{ij}} + \frac{2\gamma}{3}\frac{\partial T_2}{\partial x_{ij}}. \quad \text{(A20)}$$

Chatterjee and Dembo (2016, lemma 5.2) showed that for the T_1 and T_2 defined above, there exists a set $\mathcal{D}_1(\epsilon)$ and $\mathcal{D}_2(\epsilon)$ satisfying the criterion (A3) (with $f = T_1$ and $f = T_2$) so that

$$|\mathcal{D}_1(\epsilon)| \le \exp\left\{\frac{\tilde{C}_1 2^4 3^4 n}{\epsilon^4}\log\frac{\tilde{C}_2 2^4 3^4}{\epsilon^4}\right\} = \exp\left\{\frac{\tilde{C}_1 6^4 n}{\epsilon^4}\log\frac{\tilde{C}_2 6^4}{\epsilon^4}\right\}, \quad \text{(A21)}$$

$$|\mathcal{D}_2(\epsilon)| \le \exp\left\{\frac{\tilde{C}_1 3^4 3^4 n}{\epsilon^4}\log\frac{\tilde{C}_2 3^4 3^4}{\epsilon^4}\right\} = \exp\left\{\frac{\tilde{C}_1 3^8 n}{\epsilon^4}\log\frac{\tilde{C}_2 3^8}{\epsilon^4}\right\}, \quad \text{(A22)}$$

where $\tilde{C}_1$ and $\tilde{C}_2$ are universal constants. We define

$$\mathcal{D}(\epsilon) := \left\{2\alpha_{ij} + \frac{\beta}{2}d_1 + \frac{2\gamma}{3}d_2 : d_1 \in \mathcal{D}_1\left(\frac{2}{\beta}\cdot\frac{\epsilon}{\sqrt{2}}\right), d_2 \in \mathcal{D}_2\left(\frac{3}{2\gamma}\cdot\frac{\epsilon}{\sqrt{2}}\right), 1 \le i \le j \le n\right\}. \quad \text{(A23)}$$

Hence, $\mathcal{D}(\epsilon)$ satisfies criterion equation (A3) and

$$|\mathcal{D}(\epsilon)| \le \frac{1}{2}n(n+1)\left|\mathcal{D}_1\left(\sqrt{2}\epsilon/\beta\right)\right| \cdot \left|\mathcal{D}_2\left(3\epsilon/2\sqrt{2}\gamma\right)\right| \le \frac{1}{2}n(n+1)\exp\left\{\frac{\tilde{C}_1 6^4\beta^4 n}{4\epsilon^4}\log\frac{\tilde{C}_2 6^4\beta^4}{4\epsilon^4}\right\} \times \exp\left\{\frac{\tilde{C}_1 3^8 2^6\gamma^4 n}{3^4\epsilon^4}\log\frac{\tilde{C}_2 3^8 2^6\gamma^4}{3^4\epsilon^4}\right\}. \quad \text{(A24)}$$

Therefore, by recalling $\mathcal{E}_1$ from equation (A7), we get

$$\begin{aligned}\mathcal{E}_1 &= \frac{1}{4}\left(\binom{n}{2}\sum_{1\le i<j\le n} b_{(ij)}^2\right)^{1/2}\epsilon + 3\binom{n}{2}\epsilon + \log|\mathcal{D}(\epsilon)| \\ &\le \left[\frac{1}{4}\left(2\max_{i,j}|\alpha_{ij}| + 2|\beta| + 8|\gamma|\right) + 3\right]\binom{n}{2}\epsilon \\ &\quad + \log\left(\frac{1}{2}n(n+1)\right) + \frac{\tilde{C}_1 6^4\beta^4 n}{4\epsilon^4}\log\frac{\tilde{C}_2 6^4\beta^4}{4\epsilon^4} \\ &\quad + \frac{\tilde{C}_1 3^4 2^6\gamma^4 n}{\epsilon^4}\log\frac{\tilde{C}_2 3^4 2^6\gamma^4}{\epsilon^4} \\ &\le C_1(\alpha,\beta,\gamma)n^2\epsilon + \frac{C_1(\alpha,\beta,\gamma)n}{\epsilon^4}\log\frac{C_1(\alpha,\beta,\gamma)}{\epsilon^4} \\ &= C_1(\alpha,\beta,\gamma)n^{9/5}(\log n)^{1/5},\end{aligned} \quad \text{(A25)}$$

by choosing $\epsilon = (\frac{\log n}{n})^{1/5}$, where $C_1(\alpha, \beta, \gamma)$ is a constant depending only on α, β, γ:

$$C_1(\alpha,\beta,\gamma) := c_1\left(\max_{i,j}|\alpha_{ij}| + |\beta|^4 + |\gamma|^4 + 1\right), \quad \text{(A26)}$$

where $c_1 > 0$ is some universal constant. To see why we can choose $C_1(\alpha, \beta, \gamma)$ as in equation (A26) so that equation (A25) holds, we first notice that it follows from equation (A25) that we can choose $C_1(\alpha, \beta, \gamma)$ such that $C_1(\alpha, \beta, \gamma) \ge \max\{\tilde{c}_1 \max_{ij}|\alpha_{ij}| + \tilde{c}_2|\beta| + \tilde{c}_3|\gamma| + \tilde{c}_4, \tilde{c}_5\beta^4, \tilde{c}_6\gamma^4\}$, where $\tilde{c}_1, \tilde{c}_2, \tilde{c}_3, \tilde{c}_4, \tilde{c}_5, \tilde{c}_6 > 0$ are some universal constants. Note that $\max\{\tilde{c}_1 \max_{ij}|\alpha_{ij}| + \tilde{c}_2|\beta| + \tilde{c}_3|\gamma| + \tilde{c}_4, \tilde{c}_5\beta^4, \tilde{c}_6\gamma^4\} \le \tilde{c}_1 \max_{ij}|\alpha_{ij}| + \tilde{c}_2|\beta| + \tilde{c}_3|\gamma| + \tilde{c}_4 + \tilde{c}_5\beta^4 + \tilde{c}_6\gamma^4 \le c_1(\max_{i,j}|\alpha_{ij}| + |\beta|^4 + |\gamma|^4 + 1)$ for some universal constant $c_1 > 0$. Thus, we can take $C_1(\alpha, \beta, \gamma)$ as in equation (A26).

We can also compute from equation (A8) that

$$\begin{aligned}\mathcal{E}_2 = 4\Bigg(&\sum_{1\le i<j\le n}(ac_{(ij)(ij)} + b_{(ij)}^2) \\ &+ \frac{1}{4}\sum_{1\le i<j\le n, 1\le i'<j'\le n}\Big(ac_{(ij)(i'j')}^2 + b_{(ij)}b_{(i'j')}c_{(ij)(i'j')} \\ &+ 4b_{(ij)}c_{(ij)(i'j')}\Big)\Bigg)^{1/2} + \frac{1}{4}\left(\sum_{1\le i<j\le n} b_{(ij)}^2\right)^{1/2} \\ &\times\left(\sum_{1\le i<j\le n} c_{(ij)(ij)}^2\right)^{1/2} + 3\sum_{1\le i<j\le n} c_{(ij)(ij)} + \log 2,\end{aligned}$$

so that

$$\begin{aligned}\mathcal{E}_2 \le 4\Bigg\{&\binom{n}{2}\Bigg(n\left(\max_{i,j}|\alpha_{ij}| + \frac{1}{2}|\beta| + \frac{2}{3}|\gamma|\right)4(|\beta| + 4|\gamma|) \\ &+ \left(2\max_{i,j}|\alpha_{ij}| + 2|\beta| + 8|\gamma|\right)^2\Bigg) \\ &+ \frac{1}{4}n^2\left[\max_{i,j}|\alpha_{ij}| + \frac{1}{2}|\beta| + \frac{2}{3}|\gamma|\right] \\ &\cdot\left[\binom{n}{2}\binom{n-2}{2}4^2(|\beta| + 4|\gamma|)^2 n^{-4}\right.\end{aligned}$$

$$+\left(\binom{n}{2}^2-\binom{n}{2}\binom{n-2}{2}\right)4^2(|\beta|+4|\gamma|)^2 n^{-2}\Bigg]$$
$$+\left(2\max_{i,j}|\alpha_{ij}|+2|\beta|+8|\gamma|\right)$$
$$\cdot\left(\max_{i,j}|\alpha_{ij}|+\frac{1}{2}|\beta|+\frac{2}{3}|\gamma|\right)$$
$$\cdot\Bigg[\binom{n}{2}\binom{n-2}{2}4(|\beta|+4|\gamma|)n^{-2}$$
$$+\left(\binom{n}{2}^2-\binom{n}{2}\binom{n-2}{2}\right)4(|\beta|+4|\gamma|)n^{-1}\Bigg]\Bigg\}^{1/2}$$
$$+\frac{1}{4}\binom{n}{2}\left(2\max_{i,j}|\alpha_{ij}|+2|\beta|+8|\gamma|\right)4(|\beta|+4|\gamma|)n^{-1}$$
$$+3\binom{n}{2}4(|\beta|+4|\gamma|)n^{-1}+\log 2$$
$$\leq C_2(\alpha,\beta,\gamma)n^{3/2},$$

where we used the formulas for a, $b_{(ij)}$ and $c_{(ij)(i'j')}$ that we derived earlier and the combinatorics identities:

$$\sum_{1\leq i<j\leq n,1\leq i'<j'\leq n,|\{i,j,i',j'\}|=4} 1=\sum_{1\leq i<j\leq n}\sum_{1\leq i'<j'\leq n,|\{i,j,i',j'\}|=4} 1=\binom{n}{2}\binom{n-2}{2},$$

$$\sum_{1\leq i<j\leq n,1\leq i'<j'\leq n,|\{i,j,i',j'\}|=2\text{ or }3} 1=\binom{n}{2}^2-\binom{n}{2}\binom{n-2}{2},$$

and $C_2(\alpha,\beta,\gamma)$ is a constant depending only on α,β,γ that can be chosen as

$$C_2(\alpha,\beta,\gamma):=c_2\left(\max_{i,j}|\alpha_{ij}|+|\beta|+|\gamma|+1\right)^{1/2}\times(1+|\beta|^2+|\gamma|^2)^{1/2}, \tag{A27}$$

where $c_2>0$ is some universal constant.

Finally, to get lower bound, notice that

$$\frac{1}{2}\sum_{1\leq i<j\leq n}c_{(ij)(ij)}\leq\frac{1}{2}\binom{n}{2}4(|\beta|+4|\gamma|)n^{-1}\leq C_3(\beta,\gamma)n, \tag{A28}$$

where $C_3(\beta,\gamma)$ is a constant depending only on β,γ, and we can simply take $C_3(\beta,\gamma)=|\beta|+4|\gamma|$.

Proof of Proposition 1. We can approximate ψ_n by ψ_n^{MF} as seen in theorem 1, and as a result, we can approximate the log likelihood as follows:

$$\ell_n(g,\alpha,\beta,\gamma):=\frac{1}{n^2}\log(\pi_n(g,\alpha,\beta,\gamma))=T_n(g,\alpha,\beta,\gamma)-\psi_n(\alpha,\beta,\gamma),$$

by the mean-field log likelihood:

$$\ell_n^{MF}(g,\alpha,\beta,\gamma):=T_n(g,\alpha,\beta,\gamma)-\psi_n^{MF}(\alpha,\beta,\gamma).$$

Then the difference between the mean-field likelihood and the ERGM likelihood is bounded uniformly over $g\in\mathcal{G}$, for any α,β,γ:

$$0\leq\ell_n^{MF}(g,\alpha,\beta,\gamma)-\ell_n(g,\alpha,\beta,\gamma)\leq C_1(\alpha,\beta,\gamma)n^{-1/5}(\log n)^{1/5}+C_2(\alpha,\beta,\gamma)n^{-1/2}.$$

Therefore, for any compact Θ, we have

$$0\leq\sup_{\alpha,\beta,\gamma\in\Theta}\left[\ell_n^{MF}(g,\alpha,\beta,\gamma)-\ell_n(g,\alpha,\beta,\gamma)\right]$$
$$\leq\sup_{\alpha,\beta,\gamma\in\Theta}\left[C_1(\alpha,\beta,\gamma)n^{-1/5}(\log n)^{1/5}+C_2(\alpha,\beta,\gamma)n^{-1/2}\right]$$
$$\leq\sup_{\alpha,\beta,\gamma\in\Theta}C_1(\alpha,\beta,\gamma)n^{-1/5}(\log n)^{1/5}+\sup_{\alpha,\beta,\gamma\in\Theta}C_2(\alpha,\beta,\gamma)n^{-1/2}.$$

This proves the result. □

REFERENCES

Airoldi, Edoardo M., David Blei, Stephen E. Fienberg, and Eric P. Xing, "Mixed Membership Stochastic Blockmodels," *Journal of Machine Learning Research* 9 (2008), 1981–2014.

Amir, Eyal, Wen Pu, and Dorothy Espelage, "Approximating Partition Functions for Exponential-Family Random Graph Models," in F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems*, 25 (Red Hook, NY: Curran, 2012).

Aristoff, David, and Lingjiong Zhu, "On the Phase Transition Curve in a Directed Exponential Random Graph Model," *Advances in Applied Probability* 50 (2018), 272–301. 10.1017/apr.2018.13

Badev, Anton, "Discrete Games in Endogenous Networks: Theory and Policy," PhD diss., University of Pennsylvania (2013).

Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson, "The Diffusion of Microfinance," *Science* 341:6144 (2013). 10.1126/science.1236498

Besag, Julian, "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society Series B (Methodological)* 36:2 (1974), 192–236. 10.1111/j.2517-6161.1974.tb00999.x

Bhamidi, Shankar, Guy Bresler, and Allan Sly, "Mixing Time of Exponential Random Graphs," *Annals of Applied Probability* 21:6 (2011), 2146–2170. 10.1214/10-AAP740

Bickel, Peter, David Choi, Xiangyu Chang, and Hai Zhang, "Asymptotic Normality of Maximum Likelihood and Its Variational Approximation for Stochastic Blockmodels," *Ann. Statist.* 41:4 (2013), 1922–1943. 10.1214/13-AOS1124

Bishop, Christopher, *Pattern Recognition and Machine Learning* (New York: Springer, 2006).

Blume, Lawrence E., "The Statistical Mechanics of Strategic Interaction," *Games and Economic Behavior* 5:3 (1993), 387–424. 10.1006/game.1993.1023

Boucher, Vincent, "Structural Homophily," *International Economic Review* 56:1 (2015), 235–264. 10.1111/iere.12101

Boucher, Vincent, and Ismael Mourifie, "My Friends Far Far Away: A Random Field Approach to Exponential Random Graph Models," *Econometrics Journal* 20:3 (2017), S14–S46. 10.1111/ectj.12096

Butts, Carter, "Using Potential Games to Parameterize ERG Models," University of California, Irvine working paper (2009).

Caimo, Alberto, and Nial Friel, "Bayesian Inference for Exponential Random Graph Models," *Social Networks* 33:1 (2011), 41–55. 10.1016/j.socnet.2010.09.004

Chandrasekhar, Arun, "Econometrics of Network Formation," in Y. Bramoulle, A. Galeotti, and B. W. Rogers, eds., *The Oxford Handbook of the Economics of Networks* (Oxford: Oxford University Press, 2016).

Chandrasekhar, Arun, and Matthew Jackson, "Tractable and Consistent Exponential Random Graph Models," working paper (2014).

Chatterjee, Sourav, and Amir Dembo, "Nonlinear Large Deviations," *Advances in Mathematics* 299 (2016), 396–450. 10.1016/j.aim.2016.05.017

Chatterjee, Sourav, and Persi Diaconis, "Estimating and Understanding Exponential Random Graph Models," *Annals of Statistics* 41:5 (2013).

DePaula, Aureo, "Econometrics of Network Models," in B. Honore, A. Pakes, M. Piazzesi, and L. Samuelson, eds., *Advances in Economics and Econometrics: Eleventh World Congress* (Cambridge: Cambridge University Press, 2017).

Durlauf, Steven N., and Yannis M. Ioannides, "Social Interactions," *Annual Review of Economics* 2:1 (2010), 451–478. 10.1146/annurev.economics.050708.143312

Fox, Jeremy T., and Natalia Lazzati, "A Note on Identification of Discrete Choice Models for Bundles and Binary Games," *Quantitative Economics* 8:3 (2017), 1021–1036. 10.3982/QE489

Geyer, Charles, and Elizabeth Thompson, "Constrained Monte Carlo Maximum Likelihood for Dependent Data," *Journal of the Royal Statistical Society, Series B (Methodological)* 54:3 (1992), 657–699. 10.1111/j.2517-6161.1992.tb01443.x

Goodreau, S. M., J. A. Kitts, and M. Morris, "Birds of a Feather, or Friend of a Friend? Using Exponential Random Graph Models to Investigate Adolescent Social Networks," *Demography* 46:1 (2009), 103–125. 10.1353/dem.0.0045

Graham, Bryan, "An Empirical Model of Network Formation: With Degree Heterogeneity," *Econometrica* 85:4 (2017), 1033–1063. 10.3982/ECTA12679

He, Ran, and Tian Zheng, "Estimation of Exponential Random Graph Models for Large Social Networks via Graph Limits" (pp. 248–255), in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (New York: ACM, 2013).

Jackson, Matthew O., *Social and Economics Networks* (Princeton, NJ: Princeton University Press, 2010).

Koskinen, Johan, "Bayesian Analysis of Exponential Random Graphs: Estimation of Parameters and Model Selection," Department of Statistics, Stockholm University research report 2004:2 (2004).

Kosyakova, Tetyana, Thomas Otter, Sanjog Misra, and Christian Neuerburg, "Exact MCMC for Choices from Menus: Measuring Substitution and Complementarity among Menu Items," *Marketing Science* 39:2 (2020). 10.1287/mksc.2019.1191

Mele, Angelo, "Segregation in Social Networks: Monte Carlo Maximum Likelihood Estimation," working paper (2011), https://www.proquest.com/docview/1009684546?parentSessionId=yQfN9YCHTFSYg%2BkW5DbH6FWNRN8IMcTEw4NnW3SVCSc%3D

——— "A Structural Model of Dense Network Formation," *Econometrica* 85 (2017), 825–850. 10.3982/ECTA10400

Moller, Jesper, and Rasmus Plenge Waagepetersen, *Statistical Inference and Simulation for Spatial Point Processes* (London: Chapman and Hall, 2004).

Monderer, Dov, and Lloyd Shapley, "Potential Games," *Games and Economic Behavior* 14 (1996), 124–143. 10.1006/game.1996.0044

Moody, James, "Race, School Integration, and Friendship Segregation in America," *American Journal of Sociology* 103:7 (2001), 679–716. 10.1086/338954

Murray, Iain A., Zoubin Ghahramanim, and David J. C. MacKay, "MCMC for Doubly-Intractable Distributions" (pp. 359–366), in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (AUAI Press, 2006).

Snijders, Tom A. B., "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models," *Journal of Social Structure* 3:2 (2002).

Wainwright, M. J., and M. I. Jordan, "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends in Machine Learning* 1:1–2 (2008), 1–305. 10.1561/2200000001

Wasserman, Stanley, and Philippa Pattison, "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p*," *Psychometrika* 61:3 (1996), 401–425. 10.1007/BF02294547

Wimmer, Andreas, and Kevin Lewis, "Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook," *American Journal of Sociology* 116:2 (2010), 583–642. 10.1086/653658

Xing, Eric P., Michael I. Jordan, and Stuart Russell, "A Generalized Mean Field Algorithm for Variational Inference in Exponential Families" (pp. 583–591), in *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence* (San Mateo, CA: Morgan Kaufmann, 2003).