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, using nonlinear large deviations results. This translates into bounds on the distance between true likelihood and mean-field likelihood, as well as bounds on the distance between approximate parameter estimates from the MLE, assuming the likelihood is not very flat. In small networks, a simple Monte Carlo exercise shows that our deterministic method provides similar estimates as the simulation-based methods with the advantage of converging in quadratic time.

Keywords: Networks, Microeconometrics, Large networks, Variational Inference, Large deviations, Mean-Field Approximations

1. INTRODUCTION

In this paper, we provide 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, 2014; Airoldi et al., 2008; Bickel et al., 2013), ERGMs allow for conditional and unconditional dependence among links (Snijders, 2002; Wasserman and Pattison, 1996). These models have recently gained attention in economics, because several works

Date: First version: June 15, 2015. This version: April 4, 2019.
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 and Demian Pouzo for comments on previous versions of this paper. The second author is partially supported by NSF Grant DMS-1613164.

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 and Jackson, 2014; Boucher and Mourifie, 2017), and in a large class of evolutionary games and social interactions models (Blume, 1993; Durlauf and 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 and Diaconis, 2013; Mele, 2017). Estimates of the MCMC-MLE converge almost surely to the MLE if the likelihoods are well-behaved (Geyer and Thompson, 1992). However, a recent literature has shown that MCMC-MLE may converge in exponential time, making estimation and approximation of the likelihood impractical or infeasible for a large class of ERGMs (Bhamidi et al., 2011; Chatterjee and Diaconis, 2013; Mele, 2017). An alternative is the Maximum Pseudo-likelihood estimator (MPLE), that 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 some special cases and when some regularity conditions are satisfied (Boucher and Mourifie, 2017; Besag, 1974). Several authors have shown that MPLE may give misleading estimates when the dependence among links is strong (Geyer and 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 and Jordan, 2008; Bishop, 2006; Chatterjee and 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 main 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 non-linear 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 and Mourifie, 2017; Chandrasekhar, 2016; dePaula, forthcoming). The result holds for dense and moderately sparse networks. Remarkably and conveniently the mean-field error converges to zero 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 the main result is a bound 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 inference in large networks. Finally, we show that our mean-field parameter estimates are close to the MLE in terms of Euclidean distance, as long as the likelihoods are well-behaved and not very flat. Because our bounds are not 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 and Mourifie, 2017; Moody, 2001; Badev, 2013) or the indian villages in Banerjee et al. (2013) include on average about 200-300 nodes. To understand the performance of our estimator, we perform simple Monte Carlo exercises in small networks and compare it to MCMC-MLE and MPLE. In terms of computational speed, our method performs as fast or faster than MCMC-MLE, because it converges in quadratic time; while MCMC-MLE may converge in exponential time; mean-field is usually slower than MPLE. Our Monte Carlo results show good performance for networks with 50 to 500 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. When we increase the network size, all three estimator improve, as expected. We conclude that our method performance is similar to available estimators in small networks.

To the best of our knowledge, this paper is one of the first works in economics to use mean-field approximations for inference in complex models. Furthermore, 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 and Diaconis, 2013; Amir et al., 2012; Mele, 2017; He and Zheng, 2013; Aristoff and 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 like in Blume (1993), Badev (2013), Durlauf and Ioannides (2010), models for bundles (Fox and Lazzati, 2017), and models of choices from menus (Kosyakova et al., 2018) 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 2 presents the theoretical model and variational approximations. Section 3 contains the main theoretical results and the error bounds. Section 4 presents the Monte Carlo results and Section 5 concludes. All the proofs and additional Monte Carlo simulations are in the Appendix.

2. NETWORK FORMATION MODEL AND VARIATIONAL METHODS

2.1. **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 and Pattison (1996), Jackson (2008), dePaula (forthcoming), Mele (2017), Moody (2001), Wimmer and 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 $i = 1, ..., n$; each node is characterized by a S-dimensional vector of observed attributes $\tau_i \in \mathcal{X} := \otimes_{j=1}^{S} \mathcal{X}_j$, $i = 1, ..., n$. The sets $\mathcal{X}_j$ can represent age, race, gender, income, etc.[1] Let α be a $n \times n$ 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

$$\pi_n(g; \alpha, \beta, \gamma) = \frac{\exp\left[Q_n(g; \alpha, \beta, \gamma)\right]}{\sum_{\omega \in \mathcal{G}_n} \exp\left[Q_n(\omega; \alpha, \beta, \gamma)\right]}, \tag{2.1}$$

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

$$Q_n(g; \alpha, \beta, \gamma) = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_{ij} g_{ij} + \frac{\beta}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} g_{ij} g_{jk} + \frac{\gamma}{6n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} g_{ij} g_{jk} g_{ki}. \tag{2.2}$$

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 likelihood (2.1) is a proper distribution. The second and third term of the potential function (2.2) are the counts of two-stars and triangles in the network, rescaled by n. We rewrite (2.1) as

$$\pi_n(g; \alpha, \beta, \gamma) = \exp\left\{n^2\left[T_n(g; \alpha, \beta, \gamma) - \psi_n(\alpha, \beta, \gamma)\right]\right\}, \tag{2.3}$$

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

$$\psi_n(\alpha,\beta,\gamma) = \frac{1}{n^2}\log\sum_{\omega\in\mathcal{G}_n}\exp\left[n^2T_n(\omega;\alpha,\beta,\gamma)\right], \tag{2.4}$$

and $\mathcal{G}_n := \{\omega = (\omega_{ij})_{1\leq i,j\leq n} : \omega_{ij} = \omega_{ji} \in \{0,1\}, \omega_{ii} = 0, 1 \leq i,j \leq n\}$ is the set of all binary matrices with n nodes. The re-scaling 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.

2.2. **Microeconomic equilibrium foundations.** ERGMs caught the attention of economists because recent works proves a behavioral and equilibrium interpretation of likelihood (2.3).[2] In fact, these likelihoods naturally arise as equilibrium of best-response dynamics in potential games (Blume (1993), Monderer and Shapley (1996), Butts (2009), Mele (2011)).

To be concrete, let's 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 players i is given by

$$u_i(g,\tau) = \sum_{j=1}^n \alpha_{ij}g_{ij} + \frac{\beta}{n}\sum_{j=1}^n\sum_{k=1}^n g_{ij}g_{jk} + \frac{\gamma}{n}\sum_{j=1}^n\sum_{k=1}^n g_{ij}g_{jk}g_{ki}, \tag{2.5}$$

Each player forms links with other nodes, maximizing utility (2.5), 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] (1) the network formation is sequential, with only two active players in each period; (2) 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; (3) before choosing whether to form or delete a link,

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

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

players receive an i.i.d. logistic shock $(\varepsilon_{ij1}, \varepsilon_{ij0})$. At time t, the link g_{ij}^t is formed if

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

Mele (2017) shows that such a model is a potential game (Monderer and Shapley, 1996) with potential function given by equation (2.2). The probability of observing network g in the long run is given by (2.3) (Theorem 1 in Mele (2017)), thus (2.3) describes the stationary behavior of the model. In the long-run we observe with high probability the pairwise stable networks, where no pair of players want to form or delete a link.[4]

2.3. **Variational Approximations.** The constant $\psi_n(\alpha, \beta, \gamma)$ in (2.4) 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 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 et al., 2009; Koskinen, 2004; Caimo and Friel, 2011; Murray et al., 2006). However, Bhamidi et al. (2011) has shown that such methods may have exponentially slow convergence for many ERGMs 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:

$$\begin{aligned} KL(q_n|\pi_n) &= \sum_{\omega \in \mathcal{G}_n} q_n(\omega) \log \left[\frac{q_n(\omega)}{\pi_n(\omega; \alpha, \beta)} \right] \\ &= \sum_{\omega \in \mathcal{G}_n} q_n(\omega) \left[\log q_n(\omega) - n^2 T_n(\omega; \alpha, \beta, \gamma) + n^2 \psi_n(\alpha, \beta, \gamma) \right] \geq 0. \end{aligned} \tag{2.6}$$

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

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

[4]In the Online Appendix D we provide more details about the microeconomic foundation of the model for the interested reader.

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

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 re-scaled 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 $\mathcal{L}(q_n)$, i.e.

$$\psi_n(\alpha,\beta,\gamma) = \sup_{q_n\in\mathcal{Q}_n} \mathcal{L}(q_n) = \sup_{q_n\in\mathcal{Q}_n}\left\{\mathbb{E}_{q_n}\left[T_n(\omega;\alpha,\beta,\gamma)\right] + \frac{1}{n^2}\mathcal{H}(q_n)\right\}, \tag{2.7}$$

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, i.e. a maximization problem.

In general, problem (2.7) has no closed-form solution, thus the literature suggests to restrict $\mathcal{Q}_n$ to be the set of all completely factorized distribution[6]

$$q_n(g) = \prod_{i,j}\mu_{ij}^{g_{ij}}(1-\mu_{ij})^{1-g_{ij}}, \tag{2.8}$$

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

$$\mathbb{E}_{q_n}\left[T_n\left(\omega;\alpha,\beta,\gamma\right)\right] = \frac{\sum_i\sum_j\alpha_{ij}\mu_{ij}}{n^2} + \beta\frac{\sum_i\sum_j\sum_k\mu_{ij}\mu_{jk}}{2n^3} + \gamma\frac{\sum_i\sum_j\sum_k\mu_{ij}\mu_{jk}\mu_{ki}}{6n^3}.$$

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 a $n\times n$ symmetric matrix $\boldsymbol{\mu}(\alpha,\beta,\gamma)$ that solves

$$\begin{aligned}
\psi_n(\alpha,\beta,\gamma) &\geq \psi_n^{MF}(\boldsymbol{\mu}(\alpha,\beta,\gamma)) \\
&= \sup_{\boldsymbol{\mu}\in[0,1]^{n^2}:\mu_{ij}=\mu_{ji},\forall i,j}\left\{\frac{1}{n^2}\sum_{i,j}\alpha_{ij}\mu_{ij} + \frac{\beta}{2n^3}\sum_{i,j,k}\mu_{ij}\mu_{jk} + \frac{\gamma}{6n^3}\sum_{i,j,k}\mu_{ij}\mu_{jk}\mu_{ki}\right. \\
&\qquad\left. - \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n\left[\mu_{ij}\log\mu_{ij} + (1-\mu_{ij})\log(1-\mu_{ij})\right]\right\}.
\end{aligned} \tag{2.9}$$

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

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]

3. Theoretical results

3.1. Convergence of the variational mean-field approximation.

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

THEOREM 3.1. *For fixed network size n, the approximation error of the variational mean-field problem is bounded as*

$$\frac{C_3(\beta,\gamma)}{n} \leq \psi_n(\alpha,\beta,\gamma) - \psi_n^{MF}(\boldsymbol{\mu}(\alpha,\beta,\gamma)) \leq C_1(\alpha,\beta,\gamma)\left(\frac{\log n}{n}\right)^{1/5} + \frac{C_2(\alpha,\beta,\gamma)}{n^{-1/2}}, \tag{3.1}$$

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:

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

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

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

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

The constants in Theorem 3.1 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, as explained in the next remark.

The estimated approximation error bounds in Theorem 3.1 allow for moderate sparsity in our model, in the sense that $|\alpha_{ij}|$, $|\beta|$ and $|\gamma|$ can have moderate growth in n instead of being bounded,

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

and the difference of ψ_n and ψ_n^{MF} goes to zero 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 zero as $n\to\infty$. On the other hand, if the graph is too sparse, e.g. $|\beta| = \Omega(n)$, $|\gamma| = \Omega(n)$, then ψ_n cannot be approximated by ψ_n^{MF}.

Our main Theorem 3.1 implies that *we can approximate the log-likelihood of the ERGM* using the mean-field approximated constant.

PROPOSITION 3.1. *Let $\ell_n(g_n,\alpha,\beta,\gamma)$ be the log-likelihood of the ERGM*

$$\ell_n(g_n,\alpha,\beta,\gamma) := n^{-2}\log\left(\pi_n(g_n,\alpha,\beta,\gamma)\right) = T_n(g_n,\alpha,\beta,\gamma) - \psi_n(\alpha,\beta,\gamma),$$

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,

$$0 \le \sup_{\alpha,\beta,\gamma\in\Theta}\left[\ell_n^{MF} - \ell_n\right] \le \sup_{\alpha,\beta,\gamma\in\Theta} C_1(\alpha,\beta,\gamma) n^{-1/5}(\log n)^{1/5} + \sup_{\alpha,\beta,\gamma\in\Theta} C_2(\alpha,\beta,\gamma) n^{-1/2}. \tag{3.2}$$

Proposition 3.1 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 and Thompson, 1992; Snijders, 2002; dePaula, forthcoming; Moller and Waagepetersen, 2004). The main difference is that our approximation is *deterministic* and does not require any simulation.

We use the bounds on the likelihoods to also derive a bound on the distance between the MLE and our mean-field estimator, when the MLE exists and it is well-behaved. Because our bounds are not sharp, this proves to be quite hard. We therefore, consider a *local* version of this convergence. We know that the ERGM likelihood is concave in parameters because it is an exponential family.

We also know that the mean-field log-constant is convex in parameters[8], therefore the approximate log-likelihood is also concave. However, to get a bound on the distance between estimates we need well-behaved objective functions, with enough curvature at least close to their maximizers. If the objective functions is too flat, the distance between the estimator may be too large in terms of our upper bounds.[9] Therefore we assume that the likelihood and its mean-field approximation have enough curvature.

PROPOSITION 3.2. *Assume (α, β, γ) lives on a compact set Θ. Let $\hat{\theta}_n := (\hat{\alpha}_n, \hat{\beta}_n, \hat{\gamma}_n)$ and $\hat{\theta}_n^{MF} := (\hat{\alpha}_n^{MF}, \hat{\beta}_n^{MF}, \hat{\gamma}_n^{MF})$ be the maximizers of ℓ_n and ℓ_n^{MF}, respectively, in the interior of Θ. Also assume that ψ_n and ψ_n^{MF} are differentiable and μ_n- and μ_n^{MF}-strongly convex in (α, β, γ), respectively, on Θ, where $\mu_n > 0$ and $\mu_n^{MF} > 0$. Then*

$$\|\hat{\theta}_n - \hat{\theta}_n^{MF}\| \leq \frac{2}{(\mu_n + \mu_n^{MF})^{\frac{1}{2}}} \left[\sup_{\alpha,\beta,\gamma\in\Theta} C_1^{\frac{1}{2}}(\alpha, \beta, \gamma) \left(\frac{\log n}{n} \right)^{\frac{1}{10}} + \sup_{\alpha,\beta,\gamma\in\Theta} C_2^{\frac{1}{2}}(\alpha, \beta, \gamma) n^{-\frac{1}{4}} \right], \tag{3.3}$$

where C_1 and C_2 are defined in Theorem 3.1 and $\|\cdot\|$ denotes the Euclidean norm.

In Proposition 3.2, if μ_n and μ_n^{MF} goes to zero sufficiently fast as n goes zero, then the bound in (3.3) may not go to zero as n goes to zero. If for example μ_n, μ_n^{MF} are uniformly bounded from below, and both $\sup_{\alpha,\beta,\gamma\in\Theta} C_1(\alpha, \beta, \gamma)$ and $\sup_{\alpha,\beta,\gamma\in\Theta} C_2(\alpha, \beta, \gamma)$ are $O(1)$, then $\|\hat{\theta}_n - \hat{\theta}_n^{MF}\| = O(n^{-1/10}(\log n)^{1/10})$.

4. ESTIMATION EXPERIMENTS IN FINITE NETWORKS

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 and 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.

[8] ψ_n^{MF} is convex in (α, β, γ) by its definition in (2.9) since the expression inside the supremum in (2.9) is affine in (α, β, γ) and supremum over any affine function is convex.

[9] Geyer and Thompson (1992) mentions similar problems arise for the MCMC-MLE. Indeed, as mentioned above, the MLE may not exist. For example, if the number of triangles is zero in the data, it will be impossible to estimate γ and the MCMC-MLE may give an approximation with solution that tends to infinity.

4.1. **Approximation algorithm for the normalizing constant.** We implemented our variational approximation for few models in the R package `mfergm`, available in Github.[10] 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 and 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 (2.9). If we take the derivative with respect to μ_{ij} and equate to zero, we get

$$\mu_{ij}^* = \left\{1 + \exp\left[-2\alpha_{ij} - \beta n^{-1}\sum_{k=1}^{n}\left(\mu_{jk}^* + \mu_{ki}^*\right) - \gamma n^{-1}\sum_{k=1}^{n}\mu_{jk}^*\mu_{ki}^*\right]\right\}^{-1} \tag{4.1}$$

The logit equation in (4.1) characterizes a system of equations, whose fixed point is a solution of the mean-field problem. We can therefore start from a matrix $\boldsymbol{\mu}$ and iterate the updates in (4.1) until we reach a fixed point, as described in the following algorithm.

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 $t = 0, ..., T$:*
Step 1. Update the entries of matrix $\boldsymbol{\mu}^{(t)}$ for all $i, j = 1, ..., n$

$$\mu_{ij}^{(t+1)} = \left\{1 + \exp\left[-2\alpha_{ij} - \beta n^{-1}\sum_{k=1}^{n}\left(\mu_{jk}^{(t)} + \mu_{ki}^{(t)}\right) - \gamma n^{-1}\sum_{k=1}^{n}\mu_{jk}^{(t)}\mu_{ki}^{(t)}\right]\right\}^{-1}. \tag{4.2}$$

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

$$\psi_n^{MF(t)} = \frac{\sum_i\sum_j \alpha_{ij}\mu_{ij}^{(t)}}{n^2} + \beta\frac{\sum_i\sum_j\sum_k \mu_{ij}^{(t)}\mu_{jk}^{(t)}}{2n^3} + \gamma\frac{\sum_i\sum_j\sum_k \mu_{ij}^{(t)}\mu_{jk}^{(t)}\mu_{ki}^{(t)}}{6n^3}$$
$$- \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\mu_{ij}^{(t)}\log\mu_{ij}^{(t)} + (1-\mu_{ij}^{(t)})\log(1-\mu_{ij}^{(t)})\right].$$

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 (4.1) to each entry of the matrix, until the increase in the objective function is less than a tolerance

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

level . Since the problem is concave in each μ_{ij}, this iterative method is guaranteed to find a local maximum of (2.9).[11] In our simulations we use a tolerance level of $\epsilon_{tol} = 0.0001$. To improve convergence we can re-start the algorithm from different random matrices, as usually done with local optimizers.[12] 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.

4.2. **Monte Carlo design.** All the computations in this section are performed on a 2017 iMac with 4.2 GHz Intel Core i7 (8 processors) and 32GB RAM. We test our approximation using simulated networks, generated using a 10 million run of the Metropolis-Hastings sampler implemented in the `ergm` package in R. Each node i has a binary attribute x_i, i.e. $x_i \overset{iid}{\sim} Bernoulli(0.5)$. Let $z_{ij} = 1$ if $x_i = x_j$ and $z_{ij} = 0$ otherwise.

$$\begin{aligned}(4.3)\quad t_z(g) &:= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n g_{ij}z_{ij};\ t_{-z}(g) := \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n g_{ij}(1-z_{ij}),\\ t_e(g) &:= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n g_{ij};\ t_s(g) := \frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n g_{ij}g_{jk};\ t_t(g) := \frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n g_{ij}g_{jk}g_{ki},\end{aligned}$$

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

$$\ell_n(g,x;\alpha,\beta,\gamma) = \alpha_1 t_z(g) + \alpha_2 t_{-z}(g) + (\beta/2)t_s(g) + (\gamma/6)t_t(g) - \psi_n(\alpha_1,\alpha_2,\beta,\gamma). \tag{4.4}$$

For computational convenience we rewrite model (4.4) in a slightly different but equivalent way

$$\ell_n(g,x;\tilde{\alpha},\beta,\gamma) = \tilde{\alpha}_1 t_e(g) + \tilde{\alpha}_2 t_z(g) + (\beta/2)t_s(g) + (\gamma/6)t_t(g) - \psi_n(\alpha_1,\alpha_2,\beta,\gamma), \tag{4.5}$$

[11] 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.

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

where we have defined $\tilde{\alpha}_1 := \alpha_2$ and $\tilde{\alpha}_2 := \alpha_1 - \alpha_2$. This specification is usually found in applications.[13] In the rest of this section we setup the simulations and provide results for the specification in (4.5).

We generate the artificial networks as follows. We generate 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 to sample 100 networks, each separated by 10000 iterations, and after a burn-in of 10 million iterations.[14]

The MCMC-MLE estimator is solved using the Stochastic approximation method of Snijders (2002), where each simulations has a burnin of $100,000$ iterations of the Metropolis-Hastings sampler and networks are sampled every 1000 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. This clearly decreases the computational time required for convergence. All the code is available in Github for replication.

4.3. **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 4.1. We show results for $n = 50, 100, 200$ and 500; reporting median, 5th and 95th percentile point estimates for each parameter.

The median estimates of the mean-field approximation are quite stable and exhibit a small bias, which is well known in the literature (Wainwright and 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. Nonetheless the median point estimates of β and γ are slowly converging to the true parameter vector as n increases.[15] So while it is hard to claim that the mean-field approximation outperforms the other estimators, it seems to provide estimates in line with MPLE and MCMC-MLE.

[13]There are other small differences in how we have specified the model and how we have setup computations using the `statnet` package in R, that 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.

[14]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 $\gamma \neq 0$.

[15]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 4.1. Monte Carlo estimates, comparison of three methods. True parameter vector is $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta, \gamma) = (-2, 1, -1, -1)$

$n = 50$	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
median	-1.975	1.054	-1.805	-9.302	-2.016	0.995	-1.000	-1.001	-1.930	1.025	-2.847	-7.433
0.05	-2.611	0.598	-9.666	-75.431	-3.879	0.914	-1.269	-1.146	-2.548	0.763	-11.691	-45.896
0.95	-1.489	1.397	7.348	58.765	-1.926	4.276	-0.840	-0.901	-1.411	1.303	4.587	30.478
$n = 100$	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
median	-2.035	1.012	-0.765	-2.199	-2.021	0.988	-0.998	-1.001	-2.011	0.998	-0.926	-3.215
0.05	-2.312	0.730	-4.937	-52.429	-2.080	0.945	-1.031	-1.031	-2.258	0.835	-6.051	-36.920
0.95	-1.662	1.218	3.555	32.974	-1.978	1.139	-0.950	-0.939	-1.661	1.150	2.513	18.628
$n = 200$	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
median	-1.980	1.004	-1.734	-4.100	-2.029	0.988	-0.996	-0.999	-1.959	1.001	-1.897	-1.027
0.05	-2.212	0.876	-4.710	-27.070	-2.060	0.968	-1.002	-1.010	-2.156	0.920	-5.724	-18.304
0.95	-1.779	1.112	2.792	31.735	-2.005	1.028	-0.969	-0.987	-1.757	1.078	1.340	20.250
$n = 500$	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
median	-2.016	0.998	-0.858	-2.831	-2.047	1.008	-0.947	-0.999	-2.022	1.001	-0.811	-1.398
0.05	-2.154	0.946	-3.986	-18.120	-2.105	0.981	-0.998	-1.074	-2.115	0.965	-3.815	-12.214
0.95	-1.813	1.057	1.392	20.361	-2.004	1.109	-0.442	-0.964	-1.823	1.034	0.884	10.066

Results of 100 Monte Carlo estimates using the three methods. MCMC-MLE stands for the Monte Carlo Maximum Likelihood estimator of Geyer and Thompson (1992), implemented in the package `ergm` in `R`, using the stochastic approximation algorithm developed in Snijders (2002). MEAN-FIELD is our method, implemented with an iterative algorithm. MPLE is the Maximum Pseudo-Likelihood Estimate, which assumes independence of the conditional choice probabilities. Each network dataset is generated with a 10 million run of the Metropolis-Hastings sampler of the `ergm` command in R, sampling every 10000 iterations.

The second set of results is for a model with parameters $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta, \gamma) = (-2, 1, -2, 3)$, see Table 4.2. The pattern is similar to Table 4.1. Indeed our mean-field estimator seems to work relatively well in most cases, especially for the estimates of β and γ.[16] We note that both MPLE and MCMC-MLE converge to more precise estimates as n increase, which is what one would expect. For parameters $\tilde{\alpha}_1, \tilde{\alpha}_2$ our mean-field estimator (median) bias persists as n increases.

While these results are encouraging, in Appendix we report some example of nonconvergence of the mean-field algorithm.

[16]We have ran some additional experiments in which we increased the size of the network up to $n = 2000$ and found that the MPLE estimator performs well at that network size. Since the MCMC-MLE is started at the MPLE estimate, it is also well performing, but much slower in terms of speed.

TABLE 4.2. Monte Carlo estimates, comparison of three methods. True parameter vector is $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta, \gamma) = (-2, 1, -2, 3)$

$n = 50$	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
median	-2.033	0.972	-2.239	-6.134	-2.037	0.990	-2.000	3.000	-1.996	0.987	-2.793	-2.325
0.05	-2.643	0.614	-10.317	-73.906	-2.212	0.856	-2.652	2.865	-2.506	0.784	-12.485	-55.067
0.95	-1.424	1.399	6.763	68.994	-1.887	1.351	-1.875	3.314	-1.304	1.379	4.807	39.883
$n = 100$	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
median	-1.975	0.983	-2.364	3.014	-2.040	0.970	-2.000	3.000	-1.945	0.989	-2.944	2.981
0.05	-2.307	0.779	-7.526	-41.294	-2.108	0.908	-2.044	2.950	-2.226	0.863	-8.598	-22.868
0.95	-1.689	1.232	2.959	48.968	-1.995	1.048	-1.939	3.049	-1.626	1.166	0.836	28.756
$n = 200$	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
median	-2.019	1.004	-1.869	7.701	-2.049	0.976	-1.997	2.999	-2.012	1.004	-2.339	5.054
0.05	-2.267	0.890	-6.331	-34.052	-2.113	0.948	-2.020	2.970	-2.200	0.933	-6.976	-21.621
0.95	-1.738	1.116	2.277	37.341	-2.017	1.071	-1.953	3.029	-1.734	1.082	1.239	23.073
$n = 500$	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
median	-2.012	0.998	-1.799	0.669	-2.063	0.994	-1.977	2.999	-2.006	1.000	-1.796	1.069
0.05	-2.150	0.939	-3.869	-20.252	-2.121	0.967	-2.000	2.964	-2.118	0.962	-4.105	-11.200
0.95	-1.884	1.057	0.728	20.524	-2.037	1.056	-1.893	3.035	-1.872	1.035	0.352	13.874

Notes: see notes for Table 4.1.

4.4. **Computational speed.** The computational speed of the three estimators is similar for small networks. For $n = 100$, the mean-field approximation takes about $3.5s$ to estimate the model, while an MCMC-MLE with a burnin of $100,000$ and sampling every 1000 iterations takes approximately $5.5s$ and the MPLE takes about $1.7s$. For $n = 50$ the estimates take $1.6s$ for mean-field, $4s$ for MCMC-MLE and $1.2s$ for MPLE.

5. CONCLUSIONS 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 and Jordan, 2008; Bishop, 2006; Chatterjee and Diaconis, 2013; Mele, 2017). Our theoretical results use nonlinear large deviations methods (Chatterjee and Dembo, 2016) to bound the error of approximation, showing that it converges to zero 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 as the network becomes large and 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, forthcoming) in a simple Monte Carlo study for small networks. The mean-field approximation exhibits some bias, but the median estimates are similar to MCMC-MLE and MPLE. Our method converges in quadratic time, 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 seem 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 blockmodels. Second, it is not clear that the ERGM model is identified for all parameter values. Indeed some results in this literature suggest otherwise (Chatterjee and Diaconis, 2013; Mele, 2017; Boucher and Mourifie, 2017). A promising research avenue for the future is the use of the large n mean-field approximation to understand identification, similarly 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 are not sharp. This raises the question of whether there is another factorization (like in structured mean-field) that leads to better approximations and faster convergence. We hope that our work will stimulate additional research and more applications of this class of approximations.

References

Airoldi, Edoardo M., David Blei, Stephen E. Fienberg and Eric P. Xing (2008), 'Mixed membership stochastic blockmodels', *Journal of Machine Learning Research* **9**, 1981–2014.

Amir, Eyal, Wen Pu and Dorothy Espelage (2012), Approximating partition functions for exponential-family random graph models, NIPS Conference 2012.

Aristoff, David and Lingjiong Zhu (2018), 'On the phase transition curve in a directed exponential random graph model', *Advances in Applied Probability* **50**, 272–301.

Badev, Anton (2013), Discrete games in endogenous networks: Theory and policy.

Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo and Matthew O. Jackson (2013), 'The diffusion of microfinance', *Science* **341**(6144).

Besag, Julian (1974), 'Spatial interaction and the statistical analysis od lattice systems', *Journal of the Royal Statistical Society Series B (Methodological)* **36**(2), 192–236.

Bhamidi, Shankar, Guy Bresler and Allan Sly (2011), 'Mixing time of exponential random graphs', *The Annals of Applied Probability* **21**(6), 2146–2170.

Bickel, Peter, David Choi, Xiangyu Chang and Hai Zhang (2013), 'Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels', *Ann. Statist.* **41**(4), 1922–1943.

Bishop, Christopher (2006), *Pattern recognition and machine learning*, Springer, New York.

Blume, Lawrence E. (1993), 'The statistical mechanics of strategic interaction', *Games and Economic Behavior* **5**(3), 387–424.

Borgs, C., J.T. Chayes, L. Lovász, V.T. Sós and K. Vesztergombi (2008), 'Convergent sequences of dense graphs i: Subgraph frequencies, metric properties and testing', *Advances in Mathematics* **219**(6), 1801 – 1851.

Boucher, Vincent (2015), 'Structural homophily', *International Economic Review* **56**(1).

Boucher, Vincent and Ismael Mourifie (2017), 'My friends far far away: A random field approach to exponential random graph models', *Econometrics Journal* **20**(3), S14–S46.

Butts, Carter (2009), Using potential games to parameterize erg models. working paper.

Caimo, Alberto and Nial Friel (2011), 'Bayesian inference for exponential random graph models', *Social Networks* **33**(1), 41–55.

Chandrasekhar, Arun (2016), *The Oxford Handbook of the economics of Networks*, Oxford University Press, chapter Econometrics of Network Formation.

Chandrasekhar, Arun and Matthew Jackson (2014), Tractable and consistent exponential random graph models. working paper.

Chatterjee, Sourav and Amir Dembo (2016), 'Nonlinear large deviations', *Advances in Mathematics* **299**, 396–450.

Chatterjee, Sourav and Persi Diaconis (2013), 'Estimating and understanding exponential random graph models', *The Annals of Statistics* **41**(5).

Chatterjee, Sourav and S. R. S. Varadhan (2011), 'The large deviation principle for the Erdos-Rényi random graph', *European Journal of Combinatorics* **32**(7), 1000 – 1017.

dePaula, Aureo (forthcoming), 'Econometrics of network models', *Advances in Economics and Econometrics: Theory and Applications* .

DePaula, Aureo, Seth Richards-Shubik and Elie Tamer (2011), Inference approaches with network data.

Durlauf, Steven N. and Yannis M. Ioannides (2010), 'Social interactions', *Annual Review of Economics* **2**(1), 451–478.

Fox, Jeremy T. and Natalia Lazzati (2017), 'A note on identification of discrete choice models for bundles and binary games', *Quantitative Economics* **8**(3), 1021–1036.

Geyer, Charles and Elizabeth Thompson (1992), 'Constrained monte carlo maximum likelihood for depedendent data', *Journal of the Royal Statistical Society, Series B (Methodological)* **54**(3), 657–699.

Goodreau, S. M., Kitts J. A. and Morris M. (2009), 'Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks', *Demography* **46**(1), 103–125.

Graham, Bryan (2014), An empirical model of network formation: detecting homophily when agents are heterogeneous. working paper.

He, Ran and Tian Zheng (2013), Estimation of exponential random graph models for large social networks via graph limits, *in* 'Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining', ASONAM '13, ACM, New York, NY, USA, pp. 248–255.

Iijima, Ryota and Yuichiro Kamada (2014), Social distance and network structures. Working Paper.

Jackson, Matthew O. (2008), *Social and Economics Networks*, Princeton.

Koskinen, Johan (2004), Bayesian analysis of exponential random graphs - estimation of parameters and model selection, Research report 2004:2, Department of Statistics, Stockholm University.

Kosyakova, Tetyana, Thomas Otter, Sanjog Misra and Christian Neuerburg (2018), Exact mcmc for choices from menus - measuring substitution and complementarity among menu items.

Lovasz, L. (2012), *Large Networks and Graph Limits*, American Mathematical Society colloquium publications, American Mathematical Society.

Mele, Angelo (2011), Segregation in social networks: Monte carlo maximum likelihood estimation. Working Paper.

Mele, Angelo (2017), 'A structural model of dense network formation', *Econometrica* **85**, 825–850.

Moller, Jesper and Rasmus Plenge Waagepetersen (2004), *Statistical inference and simulation for spatial point processes*, Monographs on Statistics and Applied Probability 100, Chapman and Hall.

Monderer, Dov and Lloyd Shapley (1996), 'Potential games', *Games and Economic Behavior* **14**, 124–143.

Moody, James (2001), 'Race, school integration, and friendship segregation in america', *American Journal of Sociology* **103**(7), 679–716.

Murray, Iain A., Zoubin Ghahramani and David J. C. MacKay (2006), 'Mcmc for doubly-intractable distributions', *Uncertainty in Artificial Intelligence* .

Radin, Charles and Mei Yin (2013), 'Phase transitions in exponential random graphs', *The Annals of Applied Probability* **23**(6), 2458–2471.

Sheng, Shuyang (2012), Identification and estimation of network formation games. working paper.

Snijders, Tom A.B (2002), 'Markov chain Monte Carlo estimation of exponential random graph models', *Journal of Social Structure* **3**(2).

Train, Kenneth (2009), *Discrete Choice Methods with Simulation*, Cambridge University Press.

Wainwright, M.J. and M.l. Jordan (2008), 'Graphical models, exponential families, and variational inference', *Foundations and Trends@ in Machine Learning* **1**(1-2), 1–305.

Wasserman, Stanley and Philippa Pattison (1996), 'Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*', *Psychometrika* **61**(3), 401–425.

Wimmer, Andreas and Kevin Lewis (2010), 'Beyond and below racial homophily: Erg models of a friendship network documented on facebook', *American Journal of Sociology* **116**(2), 583–642.

Xing, Eric P., Michael I. Jordan and Stuart Russell (2003), A generalized mean field algorithm for variational inference in exponential families, *in* 'Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence', UAI'03, pp. 583–591.

APPENDIX

A.1. **Proof of Theorem 3.1.** In this proof we will try 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

$$f_i := \frac{\partial f}{\partial x_i}, \qquad f_{ij} := \frac{\partial^2 f}{\partial x_i \partial x_j}, \tag{A.1}$$

and let

$$a := \|f\|, \qquad b_i := \|f_i\|, \qquad c_{ij} := \|f_{ij}\|. \tag{A.2}$$

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

$$\sum_{i=1}^N (f_i(x) - d_i)^2 \leq N\epsilon^2. \tag{A.3}$$

Let us define

$$F := \log \sum_{x \in \{0,1\}^N} e^{f(x)}, \tag{A.4}$$

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

$$I(x) := \sum_{i=1}^N [x_i \log x_i + (1-x_i)\log(1-x_i)]. \tag{A.5}$$

In the proof we rely on Theorem 1.5 in Chatterjee and Dembo (2016) that we reproduce in Theorem A.1 to help the reader:

THEOREM A.1 (Chatterjee and Dembo (2016))**.** *For any* $\epsilon > 0$,

$$\sup_{x \in [0,1]^N} \{f(x) - I(x)\} - \frac{1}{2}\sum_{i=1}^N c_{ii} \leq F \leq \sup_{x \in [0,1]^N} \{f(x) - I(x)\} + \mathcal{E}_1 + \mathcal{E}_2, \tag{A.6}$$

where

$$\mathcal{E}_1 := \frac{1}{4}\left(N\sum_{i=1}^{N} b_i^2\right)^{1/2}\epsilon + 3N\epsilon + \log|\mathcal{D}(\epsilon)|, \tag{A.7}$$

and

$$\begin{aligned}\mathcal{E}_2 \quad &:= 4\left(\textstyle\sum_{i=1}^{N}(ac_{ii}+b_i^2)+\frac{1}{4}\sum_{i,j=1}^{N}(ac_{ij}^2+b_ib_jc_{ij}+4b_ic_{ij})\right)^{1/2}\\ &+\tfrac{1}{4}\left(\textstyle\sum_{i=1}^{N} b_i^2\right)^{1/2}\left(\textstyle\sum_{i=1}^{N} c_{ii}^2\right)^{1/2}+3\textstyle\sum_{i=1}^{N} c_{ii}+\log 2.\end{aligned} \tag{A.8}$$

We will use the Theorem A.1 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, i.e. $N = \binom{n}{2}$. Let

$$Z_n := \sum_{x_{ij}\in\{0,1\},x_{ij}=x_{ji},1\le i<j\le n} e^{\sum_{1\le i,j\le n}\alpha_{ij}x_{ij}+\frac{\beta}{2n}\sum_{1\le i,j,k\le n}x_{ij}x_{jk}+\frac{\gamma}{6n}\sum_{1\le i,j,k\le n}x_{ij}x_{jk}x_{ki}}, \tag{A.9}$$

be the normalizing factor and also define

$$\begin{aligned}L_n := \sup_{x_{ij}\in[0,1],x_{ij}=x_{ji},1\le i<j\le n}\Bigg\{&\frac{1}{n^2}\sum_{i,j}\alpha_{ij}x_{ij}+\frac{\beta}{2n^3}\sum_{i,j,k}x_{ij}x_{jk}+\frac{\gamma}{6n^3}\sum_{i,j,k}x_{ij}x_{jk}x_{ki}\\ &-\frac{1}{n^2}\sum_{1\le i<j\le n}[x_{ij}\log x_{ij}+(1-x_{ij})\log(1-x_{ij})]\Bigg\}.\end{aligned} \tag{A.10}$$

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

$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}x_{ij}+\frac{\beta}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}x_{ij}x_{jk}+\frac{\gamma}{6n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}x_{ij}x_{jk}x_{ki}. \tag{A.11}$$

Then, we can compute that,

$$a = \|f\| \leq \sum_{i=1}^{n}\sum_{j=1}^{n}|\alpha_{ij}| + \frac{1}{2}|\beta|n^2 + \frac{1}{6}|\gamma|n^2 \tag{A.12}$$
$$\leq n^2\left[\max_{i,j}|\alpha_{i,j}| + \frac{1}{2}|\beta| + \frac{1}{6}|\gamma|\right].$$

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

$$T(x) := \frac{1}{n^{k-2}}\sum_{q\in[n]^k}\prod_{\{\ell,\ell'\}\in E} x_{q_\ell q_{\ell'}}, \tag{A.13}$$

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

$$\left\|\frac{\partial T}{\partial x_{ij}}\right\| \leq 2|E|, \tag{A.14}$$

and

$$\left\|\frac{\partial^2 T}{\partial x_{ij}\partial x_{i'j'}}\right\| \leq \begin{cases} 4|E|(|E|-1)n^{-1} & \text{if } |\{i,j,i',j'\}| = 2 \text{ or } 3, \\ 4|E|(|E|-1)n^{-2} & \text{if } |\{i,j,i',j'\}| = 4. \end{cases} \tag{A.15}$$

Therefore, by (A.14), we can compute that

$$b_{(ij)} = \left\|\frac{\partial f}{\partial x_{ij}}\right\| \leq 2\max_{i,j}|\alpha_{ij}| + 2|\beta| + 2|\gamma|. \tag{A.16}$$

By (A.15), we can also compute that

$$c_{(i,j)(i'j')} = \left\|\frac{\partial^2 f}{\partial x_{ij}\partial x_{i'j'}}\right\| \tag{A.17}$$
$$\leq \begin{cases} 4\left(\frac{1}{2}|\beta|2(2-1) + \frac{1}{6}|\gamma|3(3-1)\right)n^{-1} & \text{if } |\{i,j,i',j'\}| = 2 \text{ or } 3, \\ 4\left(\frac{1}{2}|\beta|2(2-1) + \frac{1}{6}|\gamma|3(3-1)\right)n^{-2} & \text{if } |\{i,j,i',j'\}| = 4, \end{cases}$$
$$= \begin{cases} 4(|\beta| + |\gamma|)n^{-1} & \text{if } |\{i,j,i',j'\}| = 2 \text{ or } 3, \\ 4(|\beta| + |\gamma|)n^{-2} & \text{if } |\{i,j,i',j'\}| = 4. \end{cases}$$

Next, we compute that

$$\frac{\partial f}{\partial x_{ij}} = 2\alpha_{ij} + \frac{\partial}{\partial x_{ij}}\left[\frac{\beta}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} x_{ij}x_{jk} + \frac{\gamma}{6n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} x_{ij}x_{jk}x_{ki}\right]. \tag{A.18}$$

Let T_1 and T_2 be defined as

$$T_1(x) := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} x_{ij}x_{jk}, \qquad T_2(x) := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} x_{ij}x_{jk}x_{ki}. \tag{A.19}$$

Then, we have

$$\frac{\partial f}{\partial x_{ij}} = 2\alpha_{ij} + \frac{\beta}{2}\frac{\partial T_1}{\partial x_{ij}} + \frac{\gamma}{6}\frac{\partial T_2}{\partial x_{ij}}. \tag{A.20}$$

Chatterjee and Dembo (2016) (Lemma 5.2.) showed that for the T_1 and T_2 defined above, there exist a set $\mathcal{D}_1(\epsilon)$ and $\mathcal{D}_2(\epsilon)$ satisfying the criterion (A.3) (with $f = T_1$ and $f = T_2$) so that

$$|\mathcal{D}_1(\epsilon)| \le \exp\left\{\frac{\tilde{C}_1 2^4 3^4 n}{\epsilon^4}\log\frac{\tilde{C}_2 2^4 3^4}{\epsilon^4}\right\} = \exp\left\{\frac{\tilde{C}_1 6^4 n}{\epsilon^4}\log\frac{\tilde{C}_2 6^4}{\epsilon^4}\right\}, \tag{A.21}$$

$$|\mathcal{D}_2(\epsilon)| \le \exp\left\{\frac{\tilde{C}_1 3^4 3^4 n}{\epsilon^4}\log\frac{\tilde{C}_2 3^4 3^4}{\epsilon^4}\right\} = \exp\left\{\frac{\tilde{C}_1 3^8 n}{\epsilon^4}\log\frac{\tilde{C}_2 3^8}{\epsilon^4}\right\}, \tag{A.22}$$

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

Let us define

(A.23)

$$\mathcal{D}(\epsilon) := \left\{2\alpha_{ij} + \frac{\beta}{2}d_1 + \frac{\gamma}{6}d_2 : d_1 \in \mathcal{D}_1\left(\frac{2}{\beta}\cdot\frac{\epsilon}{\sqrt{2}}\right), d_2 \in \mathcal{D}_2\left(\frac{6}{\gamma}\cdot\frac{\epsilon}{\sqrt{2}}\right), 1 \le i \le j \le n\right\}.$$

Hence, $\mathcal{D}(\epsilon)$ satisfies the criterion (A.3) and

$$\begin{aligned} |\mathcal{D}(\epsilon)| &\le \frac{1}{2}n(n+1)\left|\mathcal{D}_1\left(\sqrt{2}\epsilon/\beta\right)\right| \cdot \left|\mathcal{D}_2\left(3\sqrt{2}\epsilon/\gamma\right)\right| \\ &\le \frac{1}{2}n(n+1)\exp\left\{\frac{\tilde{C}_1 6^4\beta^4 n}{4\epsilon^4}\log\frac{\tilde{C}_2 6^4\beta^4}{4\epsilon^4}\right\}\exp\left\{\frac{\tilde{C}_1 3^8\gamma^4 n}{4\cdot 3^4\epsilon^4}\log\frac{\tilde{C}_2 3^8\gamma^4}{4\cdot 3^4\epsilon^4}\right\}. \end{aligned} \tag{A.24}$$

Therefore, by recalling $\mathcal{E}_1$ from (A.7), we get

$$\begin{aligned}
\mathcal{E}_1 &= \frac{1}{4}\left(\binom{n}{2}\sum_{1\leq i<j\leq n} b_{(ij)}^2\right)^{1/2}\epsilon + 3\binom{n}{2}\epsilon + \log|\mathcal{D}(\epsilon)| \\
&\leq \left[\frac{1}{4}\left(2\max_{i,j}|\alpha_{ij}| + 2|\beta| + 2|\gamma|\right) + 3\right]\binom{n}{2}\epsilon \\
&\qquad + \log\left(\frac{1}{2}n(n+1)\right) + \frac{\tilde{C}_1 6^4\beta^4 n}{4\epsilon^4}\log\frac{\tilde{C}_2 6^4\beta^4}{4\epsilon^4} + \frac{\tilde{C}_1 3^4\gamma^4 n}{4\epsilon^4}\log\frac{\tilde{C}_2 3^4\gamma^4}{4\epsilon^4} \\
&\leq C_1(\alpha,\beta,\gamma)n^2\epsilon + \frac{C_1(\alpha,\beta,\gamma)n}{\epsilon^4}\log\frac{C_1(\alpha,\beta,\gamma)}{\epsilon^4} \\
&= C_1(\alpha,\beta,\gamma)n^{9/5}(\log n)^{1/5},
\end{aligned} \tag{A.25}$$

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

$$C_1(\alpha,\beta,\gamma) := c_1\left(\max_{i,j}|\alpha_{ij}| + |\beta|^4 + |\gamma|^4 + 1\right), \tag{A.26}$$

where $c_1 > 0$ is some universal constant. To see why we can choose $C_1(\alpha,\beta,\gamma)$ as in (A.26) so that (A.25) holds, we first notice that it follows from (A.25) that we can choose $C_1(\alpha,\beta,\gamma)$ such that $C_1(\alpha,\beta,\gamma) \geq \max\{\tilde{c}_1\max_{ij}|\alpha_{ij}| + \tilde{c}_2|\beta| + \tilde{c}_3|\gamma| + \tilde{c}_4, \tilde{c}_5\beta^4, \tilde{c}_6\gamma^4\}$, 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 $\max\{\tilde{c}_1\max_{ij}|\alpha_{ij}| + \tilde{c}_2|\beta| + \tilde{c}_3|\gamma| + \tilde{c}_4, \tilde{c}_5\beta^4, \tilde{c}_6\gamma^4\} \leq \tilde{c}_1\max_{ij}|\alpha_{ij}| + \tilde{c}_2|\beta| + \tilde{c}_3|\gamma| + \tilde{c}_4 + \tilde{c}_5\beta^4 + \tilde{c}_6\gamma^4 \leq c_1(\max_{i,j}|\alpha_{ij}| + |\beta|^4 + |\gamma|^4 + 1)$ for some universal constant $c_1 > 0$. Thus, we can take $C_1(\alpha,\beta,\gamma)$ as in (A.26).

We can also compute from (A.8) that

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

so that

$$
\begin{aligned}
\mathcal{E}_2 \le 4\Bigg\{ & \binom{n}{2}\left(n\left(\max_{i,j}|\alpha_{ij}|+\frac{1}{2}|\beta|+\frac{1}{6}|\gamma|\right)4(|\beta|+|\gamma|)+\left(2\max_{i,j}|\alpha_{ij}|+2|\beta|+2|\gamma|\right)^2\right)\\
&+\frac{1}{4}n^2\left[\max_{i,j}|\alpha_{ij}|+\frac{1}{2}|\beta|+\frac{1}{6}|\gamma|\right]\\
&\cdot\left[\binom{n}{2}\binom{n-2}{2}4^2(|\beta|+|\gamma|)^2n^{-4}+\left(\binom{n}{2}^2-\binom{n}{2}\binom{n-2}{2}\right)4^2(|\beta|+|\gamma|)^2n^{-2}\right]\\
&+\left(2\max_{i,j}|\alpha_{ij}|+2|\beta|+2|\gamma|\right)\cdot\left(\max_{i,j}|\alpha_{ij}|+\frac{1}{2}|\beta|+\frac{1}{6}|\gamma|\right)\\
&\cdot\left[\binom{n}{2}\binom{n-2}{2}4(|\beta|+|\gamma|)n^{-2}+\left(\binom{n}{2}^2-\binom{n}{2}\binom{n-2}{2}\right)4(|\beta|+|\gamma|)n^{-1}\right]\Bigg\}^{1/2}\\
&+\frac{1}{4}\binom{n}{2}(2\max_{i,j}|\alpha_{ij}|+2|\beta|+2|\gamma|)4(|\beta|+|\gamma|)n^{-1}+3\binom{n}{2}4(|\beta|+|\gamma|)n^{-1}+\log 2\\
\le{}& C_2(\alpha,\beta,\gamma)n^{3/2},
\end{aligned}
$$

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

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

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

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

$$
C_2(\alpha,\beta,\gamma):=c_2\left(\max_{i,j}|\alpha_{ij}|+|\beta|+|\gamma|+1\right)^{1/2}(1+|\beta|^2+|\gamma|^2)^{1/2}, \tag{A.27}
$$

where $c_2>0$ is some universal constant.

Finally, to get lower bound, notice that

$$
\frac{1}{2}\sum_{1\le i<j\le n}c_{(ij)(ij)}\le\frac{1}{2}\binom{n}{2}4(|\beta|+|\gamma|)n^{-1}\le C_3(\beta,\gamma)n, \tag{A.28}
$$

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

A.2. **Proof of Proposition 3.1.** We can approximate ψ_n by ψ_n^{MF} as seen in Theorem 3.1, 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

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

This proves the result.

A.3. **Proof of Proposition 3.2.** Note that since ψ_n (resp. ψ_n^{MF}) is differentiable and μ_n-strongly convex in $\theta := (\alpha,\beta,\gamma) \in \Theta$ and

$$\ell_n = T_n - \psi_n, \qquad \ell_n^{MF} = T_n - \psi_n^{MF},$$

and T_n is linear in $\theta = (\alpha,\beta,\gamma)$, we have that ℓ_n (resp. ℓ_n^{MF}) is differentiable and μ_n-strongly concave in $\theta := (\alpha,\beta,\gamma) \in \Theta$ so that for any $x, y \in \Theta$,

$$\ell_n(y) \leq \ell_n(x) + \nabla\ell_n(x)^T(y-x) - \frac{\mu_n}{2}\|y-x\|^2, \tag{A.29}$$

and in particular,

$$\begin{aligned}\ell_n(\hat{\theta}_n^{MF}) &\leq \ell_n(\hat{\theta}_n) + \nabla \ell_n(\hat{\theta}_n)^T(\hat{\theta}_n^{MF} - \hat{\theta}_n) - \frac{\mu_n}{2}\|\hat{\theta}_n^{MF} - \hat{\theta}_n\|^2 \\ &= \ell_n(\hat{\theta}_n) - \frac{\mu_n}{2}\|\hat{\theta}_n^{MF} - \hat{\theta}_n\|^2,\end{aligned} \tag{A.30}$$

and similarly, for any $x, y \in \Theta$,

$$\ell_n^{MF}(y) \leq \ell_n^{MF}(x) + \nabla \ell_n^{MF}(x)^T(y - x) - \frac{\mu_n}{2}\|y - x\|^2, \tag{A.31}$$

and in particular,

$$\begin{aligned}\ell_n^{MF}(\hat{\theta}_n) &\leq \ell_n^{MF}(\hat{\theta}_n^{MF}) + \nabla \ell_n^{MF}(\hat{\theta}_n^{MF})^T(\hat{\theta}_n - \hat{\theta}_n^{MF}) - \frac{\mu_n^{MF}}{2}\|\hat{\theta}_n - \hat{\theta}_n^{MF}\|^2 \\ &= \ell_n^{MF}(\hat{\theta}_n^{MF}) - \frac{\mu_n^{MF}}{2}\|\hat{\theta}_n - \hat{\theta}_n^{MF}\|^2.\end{aligned} \tag{A.32}$$

Adding the inequalities (A.30) and (A.32), we get

$$\begin{aligned}\|\hat{\theta}_n - \hat{\theta}_n^{MF}\|^2 &\leq \frac{2}{\mu_n^{MF} + \mu_n}\left[\left(\ell_n^{MF}(\hat{\theta}_n^{MF}) - \ell_n(\hat{\theta}_n^{MF})\right) + \left(\ell_n(\hat{\theta}_n) - \ell_n^{MF}(\hat{\theta}_n)\right)\right] \\ &\leq \frac{4}{\mu_n^{MF} + \mu_n}\sup_{\theta \in \Theta}|\ell_n^{MF}(\theta) - \ell_n(\theta)|.\end{aligned}$$

By applying Theorem 3.1, we get

$$\begin{aligned}\|\hat{\theta}_n - \hat{\theta}_n^{MF}\| &\leq \frac{2}{(\mu_n + \mu_n^{MF})^{\frac{1}{2}}}\left[\sup_{\alpha,\beta,\gamma \in \Theta} C_1(\alpha,\beta,\gamma)n^{-\frac{1}{5}}(\log n)^{\frac{1}{5}} + \sup_{\alpha,\beta,\gamma \in \Theta} C_2(\alpha,\beta,\gamma)n^{-\frac{1}{2}}\right]^{\frac{1}{2}} \\ &\leq \frac{2}{(\mu_n + \mu_n^{MF})^{\frac{1}{2}}}\left[\sup_{\alpha,\beta,\gamma \in \Theta} C_1^{\frac{1}{2}}(\alpha,\beta,\gamma)n^{-\frac{1}{10}}(\log n)^{\frac{1}{10}} + \sup_{\alpha,\beta,\gamma \in \Theta} C_2^{\frac{1}{2}}(\alpha,\beta,\gamma)n^{-\frac{1}{4}}\right],\end{aligned}$$

where the last step is due to the inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for any $x, y \geq 0$. The proof is complete.

APPENDIX B. ADDITIONAL SIMULATION RESULTS

B.1. **No covariates, edges and two-stars model.** We have estimated a model with no covariates. This corresponds to a model in which $\tilde{\alpha}_2 = 0$ or $\alpha_1 = \alpha_2 = \alpha$. The results of our simulations for small networks are in Table B.1. Our method performs relatively well in this simpler case. Indeed in this case there are results that would allow us to solve the variational problem in closed form for large n (Chatterjee and Diaconis, 2013; Mele, 2017; Aristoff and Zhu, 2018; Radin and Yin, 2013). The MPLE and MCMC-MLE median estimate seems to converge to the true value as we increase n, but our approximation seems to perform slightly better here.

TABLE B.1. Monte Carlo estimates, comparison of three methods. True parameter vector is $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta) = (-2, 0, 1)$

$n = 50$	MCMC-MLE			MEAN-FIELD			MPLE		
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β
median	-2.063	0.016	-0.324	-2.021	0.007	0.999	-1.983	0.018	-1.006
0.05	-2.692	-0.614	-23.828	-2.412	-0.372	0.975	-2.439	-0.368	-34.177
0.95	-1.363	0.657	22.738	-1.783	0.413	1.015	-1.449	0.401	14.465
$n = 100$	MCMC-MLE			MEAN-FIELD			MPLE		
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β
median	-1.970	-0.042	0.221	-1.981	-0.017	1.000	-1.949	-0.023	-1.231
0.05	-2.241	-0.333	-13.226	-2.101	-0.194	0.993	-2.168	-0.196	-14.402
0.95	-1.602	0.249	16.316	-1.874	0.134	1.012	-1.643	0.142	9.328
$n = 200$	MCMC-MLE			MEAN-FIELD			MPLE		
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β
median	-2.012	-0.005	1.483	-1.998	0.002	1.000	-2.003	-0.001	1.225
0.05	-2.214	-0.184	-9.515	-2.067	-0.093	0.997	-2.160	-0.095	-9.682
0.95	-1.796	0.161	12.179	-1.935	0.091	1.003	-1.790	0.095	8.784

Notes. See notes for Table 4.1.

B.2. **Model with 2-stars.** In this subsection we report estimates of a model where the triangle term is excluded from the specification ($\gamma = 0$ in log-likelihood (4.5)). In Table B.2 we report results for 100 simulations of a model with $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta) = (-2, 1, 2)$. We run simulations for networks of size $n = 50, 100, 200$, to show how our method compares to MCMC-MLE and MPLE when the size of the network grows. In general, we expect more precise results as n grows large.

The results are encouraging and the mean-field approximation seems to behave as expected. Indeed, the median estimate is very close to the true parameters that generate the data. As the size of the network grows from $n = 50$ to $n = 200$, both MCMC-MLE and MPLE also improve in precision. The fastest method in terms of computational time is the MPLE. This is because the MPLE's speed depends on the number of parameters. Our mean-field approximation is as fast as the MCMC-MLE.

TABLE B.2. Monte Carlo estimates, comparison of three methods. True parameter vector is $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta) = (-2, 1, 2)$

$n = 50$	MCMC-MLE			MEAN-FIELD			MPLE		
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β
median	-2.015	0.999	2.303	-1.993	1.000	2.004	-1.996	0.998	1.820
0.05	-2.433	0.641	-1.085	-2.060	0.885	1.916	-2.325	0.780	-2.556
0.95	-1.666	1.337	6.118	-1.905	1.090	2.087	-1.573	1.273	4.783
$n = 100$	MCMC-MLE			MEAN-FIELD			MPLE		
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β
median	-1.995	1.012	1.932	-1.980	1.011	2.011	-1.980	1.010	1.783
0.05	-2.189	0.861	0.701	-2.032	0.969	1.992	-2.175	0.901	0.329
0.95	-1.833	1.157	3.314	-1.944	1.044	2.088	-1.816	1.141	2.867
$n = 200$	MCMC-MLE			MEAN-FIELD			MPLE		
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β
median	-2.000	1.009	1.938	-1.986	1.005	2.016	-1.997	1.007	1.930
0.05	-2.182	0.925	0.843	-2.004	0.932	1.999	-2.176	0.950	0.592
0.95	-1.882	1.087	4.119	-1.935	1.028	2.214	-1.847	1.069	3.541

Notes. See notes for Table 4.1.

The second set of Monte Carlo experiments is reported in Table B.3, where the data are generated by parameter vector $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta) = (-2, 1, 3)$. The pattern is similar to the previous table, but the mean field estimates exhibit a little more bias.

TABLE B.3. Monte Carlo estimates, comparison of three methods. True parameter vector is $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta) = (-2, 1, 3)$

$n = 50$	MCMC-MLE			MEAN-FIELD			MPLE		
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β
median	-1.978	1.010	2.742	-1.958	1.026	3.025	-1.921	1.016	2.357
0.05	-2.308	0.745	1.342	-2.045	0.878	2.938	-2.201	0.823	-0.742
0.95	-1.689	1.229	4.466	-1.811	1.141	3.468	-1.547	1.202	4.288
$n = 100$	MCMC-MLE			MEAN-FIELD			MPLE		
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β
median	-2.005	1.002	3.022	-1.851	1.091	3.166	-1.997	1.001	3.009
0.05	-2.116	0.892	2.665	-2.274	0.866	2.998	-2.098	0.924	2.514
0.95	-1.902	1.110	3.414	-1.670	1.861	4.092	-1.895	1.096	3.425
$n = 200$	MCMC-MLE			MEAN-FIELD			MPLE		
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	β
median	-2.003	1.000	2.959	-1.923	1.030	3.107	-1.984	1.000	2.847
0.05	-2.151	0.934	2.314	-2.059	0.922	3.000	-2.104	0.951	2.096
0.95	-1.902	1.064	3.944	-1.836	1.164	4.222	-1.861	1.039	3.666

Notes. See notes for Table 4.1.

B.3. **Model with triangles.** The second set of simulations involves a model with no two-stars, that is $\beta = 0$, in Table B.4. In this specification our mean-field approximation seems to do better than the other estimators, at least for this small networks.

TABLE B.4. Monte Carlo estimates, comparison of three methods. True parameter vector is $(\tilde{\alpha}_1, \tilde{\alpha}_2, \gamma) = (-2, 1, -2)$

$n = 50$	MCMC-MLE			MEAN-FIELD			MPLE		
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	γ	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	γ	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	γ
median	-2.024	1.026	-13.959	-2.000	1.005	-2.000	-2.031	1.012	-9.804
0.05	-2.384	0.622	-60.419	-2.321	0.168	-6.425	-2.398	0.758	-45.881
0.95	-1.689	1.457	49.585	-0.739	2.246	-1.777	-1.809	1.394	21.696

$n = 100$	MCMC-MLE			MEAN-FIELD			MPLE		
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	γ	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	γ	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	γ
median	-2.006	1.019	-6.053	-1.967	1.035	-2.007	-2.002	1.015	-4.980
0.05	-2.164	0.832	-35.171	-3.472	0.951	-7.368	-2.124	0.876	-23.937
0.95	-1.824	1.183	27.361	-1.388	3.763	-1.910	-1.890	1.153	13.519

$n = 200$	MCMC-MLE			MEAN-FIELD			MPLE		
	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	γ	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	γ	$\tilde{\alpha}_1$	$\tilde{\alpha}_2$	γ
median	-2.007	1.001	-1.002	-1.972	1.031	-2.006	-2.003	1.000	-1.913
0.05	-2.083	0.901	-23.049	-2.014	1.008	-2.115	-2.061	0.929	-15.721
0.95	-1.931	1.095	16.760	-1.473	1.636	-1.983	-1.952	1.072	9.153

Notes. See notes for Table 4.1.

B.4. **Some examples of nonconvergence.** In Table B.5 we show an example in which the mean-field approximation performs worse. While the median point estimate is in line with the other estimators, the range of values is quite large. Table B.6 also shows a case in which parameter β is estimated poorly.

There are several possible explanations for this poor convergence. First, it may be that we are not finding the maximizer of the approximation variational problem (2.9), given the local nature of updates (4.1). In these simulations we do not start the matrix $\boldsymbol{\mu}^{(0)}$ at different initial values, therefore we converge to a local maximum that may not be global. Our code allows the researcher to initialize $\boldsymbol{\mu}^{(0)}$ at different random starting points. This can improve convergence. In principle we should increase the number of re-starts as n grows, as it is known that these models may have multiple modes. Ideally, one can use a Nelder-Mead or Simulated Annealing algorithm to find the maximizer of the variational problem, but this is more time-consuming. All these ideas lead to simple parallelizations of our package functions that are beyond the scope of the present work. Second, the tolerance level that we use $\epsilon_{tol} = 0.0001$ may be too large. Third, the likelihood may exhibit a phase transition and thus a small difference in parameters may cause a large change in the behavior of the model. We conjecture that some of these issues are related to identification and we plan to explore this in future work.

TABLE B.5. Monte Carlo estimates, comparison of three methods. True parameter vector is $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta, \gamma) = (-2, 1, -2, 1)$

$n = 50$	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
median	-2.053	1.000	-0.719	-9.252	-2.031	0.979	-1.996	1.001	-2.025	0.994	-1.111	-7.920
0.05	-2.463	0.425	-10.661	-116.527	-2.164	0.872	-2.104	0.921	-2.405	0.698	-12.867	-66.600
0.95	-1.525	1.404	7.641	54.050	-1.934	1.123	-1.880	1.123	-1.536	1.359	2.807	28.835
$n = 100$	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
median	-1.975	0.998	-1.982	4.475	-2.047	0.972	-1.996	1.005	-1.971	1.005	-2.366	5.529
0.05	-2.251	0.788	-6.647	-64.548	-2.103	0.927	-2.051	0.878	-2.211	0.866	-7.472	-37.758
0.95	-1.734	1.234	2.649	61.304	-1.999	1.043	-1.940	1.048	-1.715	1.176	1.158	27.884
$n = 200$	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
median	-1.998	1.010	-2.095	-2.905	-2.049	0.971	-1.997	1.001	-1.977	1.006	-2.428	0.042
0.05	-2.244	0.885	-5.865	-32.144	-2.115	0.952	-2.015	0.977	-2.171	0.927	-6.820	-20.504
0.95	-1.810	1.128	2.239	37.402	-2.025	1.046	-1.951	1.026	-1.775	1.087	0.914	22.035
$n = 500$	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
median	-1.997	1.008	-2.190	-1.908	-2.071	0.997	-1.986	0.999	-1.993	1.006	-2.229	-0.901
0.05	-2.107	0.958	-4.097	-19.067	-5.907	0.966	-2.634	0.930	-2.094	0.974	-4.127	-12.291
0.95	-1.889	1.060	0.304	16.596	-2.043	5.974	-1.861	1.168	-1.883	1.034	-0.335	10.046

Notes: see notes for Table 4.1.

TABLE B.6. Monte Carlo estimates, comparison of three methods. True parameter vector is $(\tilde{\alpha}_1, \tilde{\alpha}_2, \beta, \gamma) = (-2, 1, 2, -1)$

$n = 50$	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
median	-1.999	0.991	1.982	-2.664	-1.887	0.934	5.845	-1.000	-1.956	0.996	1.547	-0.469
0.05	-2.476	0.626	-2.400	-34.797	-2.289	0.675	3.688	-1.516	-2.453	0.765	-4.344	-26.769
0.95	-1.516	1.279	7.879	27.501	-1.673	1.344	9.026	-0.888	-1.346	1.250	6.045	15.593
$n = 100$	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
median	-1.989	0.980	2.025	0.280	-1.850	0.940	4.793	-1.001	-1.987	0.987	1.823	-0.149
0.05	-2.245	0.788	-0.044	-17.684	-2.074	0.729	4.128	-1.103	-2.239	0.844	-0.802	-15.650
0.95	-1.762	1.163	4.261	15.150	-1.693	1.157	6.897	-0.885	-1.711	1.150	4.115	11.113
$n = 200$	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
median	-1.976	0.986	1.578	0.483	-1.844	0.916	5.223	-1.000	-1.970	0.994	1.551	-0.960
0.05	-2.191	0.894	-0.231	-13.421	-1.978	0.812	4.321	-1.084	-2.173	0.924	-0.822	-11.144
0.95	-1.805	1.089	3.920	14.010	-1.783	1.033	6.900	-0.920	-1.736	1.076	3.857	11.229

Notes. See notes for Table 4.1.

ONLINE APPENDIX - NOT FOR PUBLICATION

APPENDIX C. ASYMPTOTIC RESULTS

In this section we consider the model as $n \to \infty$. We have seen previously that the log normalizing constant $\psi_n(\alpha, \beta, \gamma)$ can be approximated by $\psi_n^{MF}(\boldsymbol{\mu}(\alpha, \beta, \gamma))$ by the mean-field approximation, where $\boldsymbol{\mu}(\alpha, \beta, \gamma)$ solves the optimization problem in (2.9) and $\psi_n^{MF}(\boldsymbol{\mu}(\alpha, \beta, \gamma))$ is its optimal value, where we recall that

$$\psi_n^{MF}(\boldsymbol{\mu}(\alpha,\beta,\gamma)) = \sup_{\boldsymbol{\mu}\in[0,1]^{n^2}:\mu_{ij}=\mu_{ji},\forall i,j} \left\{ \frac{1}{n^2}\sum_{i,j}\alpha_{ij}\mu_{ij} + \frac{\beta}{2n^3}\sum_{i,j,k}\mu_{ij}\mu_{jk} + \frac{\gamma}{6n^3}\sum_{i,j,k}\mu_{ij}\mu_{jk}\mu_{ki} \right.$$
$$\left. - \frac{1}{2n^2}\sum_{i,j}[\mu_{ij}\log\mu_{ij} + (1-\mu_{ij})\log(1-\mu_{ij})] \right\},$$

We will study the limit as $n \to \infty$. Before we proceed, we need a representation of the vector α in the infinite network. The following assumption guarantee that we can switch from the discrete to the continuum.

ASSUMPTION C.1. *Assume that*

$$\alpha_{ij} = \alpha\left(i/n, j/n\right),$$

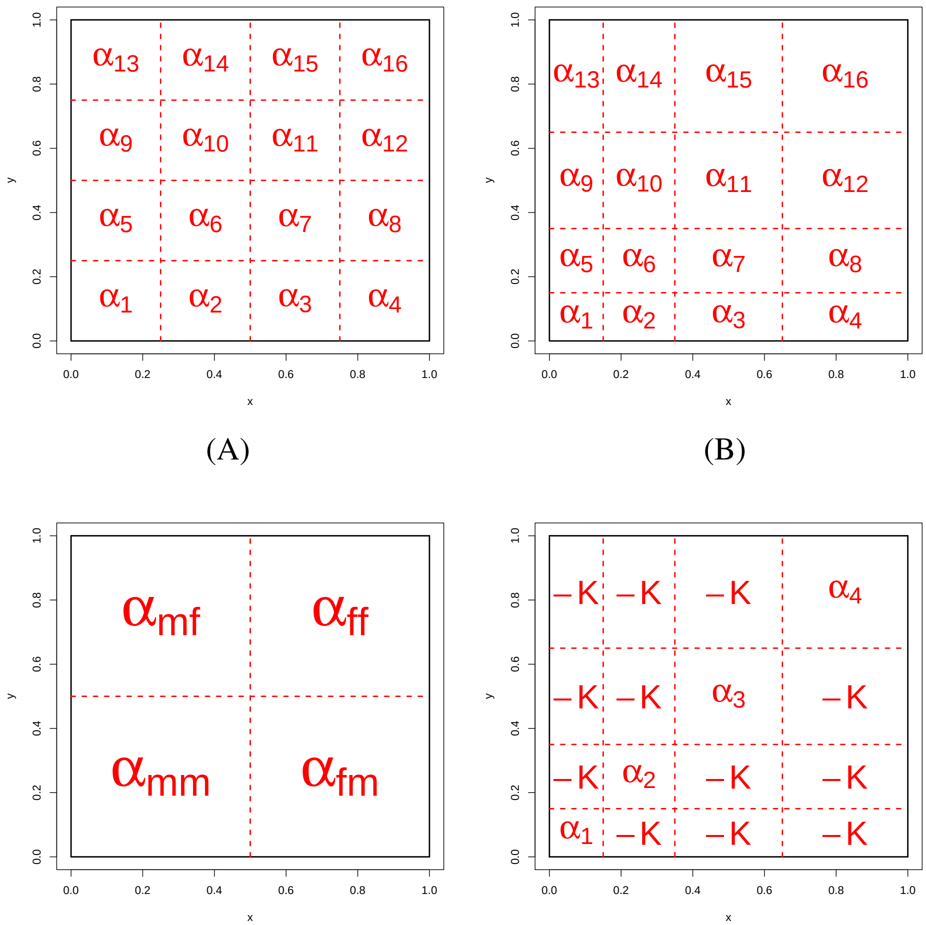
where $\alpha(x, y) : [0, 1]^2 \to \mathbb{R}$ *is a deterministic exogenous function that is symmetric, i.e.,* $\alpha(x, y) = \alpha(y, x)$. [17]

Since we have n players, the number of types for the players must be finite, although it may grow as n grows. α_{ij} are symmetric, and can take at most $\frac{n(n+1)}{2}$ values. As $n \to \infty$, the number of types can become infinite and $\alpha(x, y)$ may take infinitely many values. On the other hand, in terms of practical applications, finitely many values often suffice [18].

[17] To ease the notations, we project $\otimes_{j=1}^{S}\mathcal{X}_j$ onto $[0, 1]$ and the function $\alpha(\tau_i, \tau_j)$ defined previously is now re-defined from $[0, 1]^2$ to $\mathbb{R}$.

[18] If an entry of the vector τ_i is continuous, we can always transform the variable in a discrete vector using thresholds. For example, if $\mathcal{X}_j$ = [\$50,000,\$200,000], we can bucket the incomes into three levels, low: [\$50,000,\$100,000), medium [\$100,000,\$150,000) and high: [\$150,000, \$200,000].

FIGURE C.1. Examples of function $\alpha(x, y)$.



(A) (B)

(C) (D)

The figure provides several examples of possible partitions of the net benefit function $\alpha(x, y)$ with finite covariates. The asymptotic version of this function is defined over the unit square.

ASSUMPTION C.2. *We assume that $\alpha(x, y)$ is unifomly bounded in x and y:*

$$\sup_{(x,y)\in[0,1]^2} |\alpha(x, y)| < \infty. \tag{C.1}$$

As a simple example, let us consider gender: the population consists of males and female agents. For example, half of the nodes (population) are males, say $i = 1, 2, \ldots, \frac{n}{2}$ and the other half are females, $i = \frac{n}{2} + 1, \frac{n}{2} + 2, \ldots, n$.[19] That means, $\alpha(x, y)$ takes three values according to the three

[19]Here, we assume without loss of generality that n is an even number.

regions:

$$\left\{(x,y) : 0 < x, y < \tfrac{1}{2}\right\},$$

$$\left\{(x,y) : \tfrac{1}{2} < x, y < 1\right\},$$

$$\left\{(x,y) : 0 < x < \tfrac{1}{2} < y < 1\right\} \bigcup \left\{(x,y) : 0 < y < \tfrac{1}{2} < x < 1\right\},$$

and these three regions correspond precisely to pairs: male-male, female-female, and male-female. This example is represented in Figure C.1(C).

The work of Chatterjee and Diaconis (2013) show that the variational problem in (2.7) translates into an analogous variational problem for the graph limit.[20] In the special case $\alpha(x,y) \equiv \alpha$, it is shown in Chatterjee and Diaconis (2013) that as $n \to \infty$ the log-constant of the ERGM converges to the solution of the variational problem (C.3), that is

$$\psi_n(\alpha, \beta, \gamma) \to \psi(\alpha, \beta, \gamma), \tag{C.2}$$

where

(C.3)

$$\psi(\alpha,\beta,\gamma) = \sup_{h \in \mathcal{W}} \left\{ \alpha \int_0^1 \int_0^1 h(x,y)dxdy + \frac{\beta}{2} \int_0^1 \int_0^1 \int_0^1 h(x,y)h(y,z)dxdydz \right. \\ \left. + \frac{\gamma}{6} \int_0^1 \int_0^1 \int_0^1 h(x,y)h(y,z)h(z,x)dxdydz - \frac{1}{2} \int_0^1 \int_0^1 I(h(x,y))dxdy \right\},$$

where

$$\mathcal{W} := \left\{ h : [0,1]^2 \to [0,1], h(x,y) = h(y,x), 0 \le x, y \le 1 \right\}, \tag{C.4}$$

and we define the entropy function:

$$I(x) := x \log x + (1-x)\log(1-x), \qquad 0 \le x \le 1,$$

with $I(0) = I(1) = 0$.

[20]See also Mele (2017) for similar results in a directed network.

In essence the first three terms in (C.3) correspond to the expected potential function in the continuum, while the last term in (C.3) corresponds to the entropy of the graph limit.

We will show that (C.2) holds with

(C.5)
$$\psi(\alpha,\beta,\gamma)=\sup_{h\in\mathcal{W}}\left\{\int_0^1\int_0^1\alpha(x,y)h(x,y)dxdy+\frac{\beta}{2}\int_0^1\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz\right.$$
$$\left.+\frac{\gamma}{6}\int_0^1\int_0^1\int_0^1 h(x,y)h(y,z)h(z,x)dxdydz-\frac{1}{2}\int_0^1\int_0^1 I(h(x,y))dxdy\right\},$$

The function h in the expressions above is known as the *graphon* from the graph limits literature [21], large deviations literature for random graphs[22] and analysis of the resulting variational problem.[23] and it is a representation of an infinite network, where h is a simple symmetric function $h:[0,1]^2\rightarrow[0,1]$, and $h(x,y)=h(y,x)$. Note that our goal is to approximate ψ_n^{MF} and hence ψ_n by ψ, whose definition involves the function h, and we call such a function a graphon in the rest of the paper, to be consistent with the literature, while we are not attempting here to establish a theory of graph limits to allow nodal covariates. That is an interesting research direction worth investigating in the future, but is out of the scope of the current paper.

The following proposition shows that for a model with finitely many types the variational approximation is asymptotically exact.

PROPOSITION C.1. *Under Assumptions C.1 and C.2, as $n\rightarrow\infty$*

$$\psi_n(\alpha,\beta,\gamma)\rightarrow\psi(\alpha,\beta,\gamma),$$

where $\psi(\alpha,\beta,\gamma)$ is defined in (C.5).

Proof. It follows directly from Theorem 3.1 and $\psi_n^{MF}(\boldsymbol{\mu}(\alpha,\beta,\gamma))\rightarrow\psi(\alpha,\beta,\gamma)$, as $n\rightarrow\infty$. □

The proposition states that as n becomes large, we can approximate the exponential random graph using a model with independent links (conditional on finitely many types). This is a very

[21]See Lovasz (2012), Borgs et al. (2008)
[22]See Chatterjee and Varadhan (2011), Chatterjee and Diaconis (2013)
[23]See Aristoff and Zhu (2018), Radin and Yin (2013) among others.

useful result because the latter approximation is simple and tractable, while the exponential random graph model contains complex dependence patterns that make estimation computationally expensive.

C.1. **Approximation of the limit log normalizing constant.** We can analyze and provide an approximation of the log-constant in the large network limit. The variational formula for $\psi(\alpha, \beta, \gamma)$ is an infinite-dimensional problem which is intractable in most cases. Nevertheless, we can always bound the infinite dimensional problem with finite dimensional ones (both lower and upper bounds), at least in the absence of transitivity. For details, see Proposition E.2 in the Online Appendix. The lower-bound in Proposition E.2 coincides with the structured mean-field approach of Xing et al. (2003). In a model with homogeneous players, the lower-bound corresponds to the computational approximation of graph limits implemented in He and Zheng (2013).

In the case of extreme homophily, we can also obtain finite-dimensional approximation, see Proposition E.1 in the Online Appendix.

C.2. **Characterization of the variational problem.** We recall that the log normalizing constant in the $n \to \infty$ limit is given by the variational problem:

$$\begin{aligned}\psi(\alpha, \beta, \gamma) = \sup_{h \in \mathcal{W}} \Bigg\{ &\int_0^1 \int_0^1 \alpha(x,y) h(x,y) dxdy + \frac{\beta}{2} \int_0^1 \int_0^1 \int_0^1 h(x,y)h(y,z) dxdydz \\ &+ \frac{\gamma}{6} \int_0^1 \int_0^1 \int_0^1 h(x,y)h(y,z)h(z,x) dxdydz \\ &- \frac{1}{2} \int_0^1 \int_0^1 \left[h(x,y) \log h(x,y) + (1-h(x,y)) \log(1-h(x,y)) \right] dxdy \Bigg\}.\end{aligned} \tag{C.6}$$

PROPOSITION C.2. *The optimal graphon h that solves the variational problem* (C.6) *satisfies the Euler-Lagrange equation:*

$$2\alpha(x,y) + \beta \int_0^1 h(x,y) dx + \beta \int_0^1 h(x,y) dy + \gamma \int_0^1 h(x,z)h(y,z) dz = \log \left(\frac{h(x,y)}{1-h(x,y)} \right). \tag{C.7}$$

Proof. The proof follows from the same argument as in Theorem 6.1. in Chatterjee and Diaconis (2013). □

COROLLARY 1. *If $\alpha(x,y)$ is not a constant function, then the optimal graphon h that solves the variational problem (C.6) is not a constant function.*

Proof. If the optimal graphon h is a constant function, then (C.7) implies that α is a constant function. Contradiction. □

In general, if a graphon satisfies the Euler-Lagrange equation, that only indicates that the graphon is a stationary point, and it is not clear if the graphon is the local maximizer, local minimizer or neither. In the next result, we will show that when β is negative, any graphon satisfying the Euler-Lagrange equation in our model is indeed a local maximizer.

PROPOSITION C.3. *Assume that $\beta < 0$ and $\gamma = 0$. If h is a graphon that satisfies the Euler-Lagrange equation (C.7), then h is a local maximizer of the variational problem (C.6).*

Proof. Let us define

$$\begin{aligned}\Lambda[h] := &\int_0^1\int_0^1 \alpha(x,y)h(x,y)dxdy + \frac{\beta}{2}\int_0^1\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz \\ &-\frac{1}{2}\int_0^1\int_0^1 \left[h(x,y)\log h(x,y) + (1-h(x,y))\log(1-h(x,y))\right]dxdy.\end{aligned} \tag{C.8}$$

Let h satisfy (C.7) and for any symmetric function g and $\epsilon > 0$ sufficiently small, we have

$$\begin{aligned}&\Lambda[h+\epsilon g] - \Lambda[h] \\ &= \epsilon^2\left[\frac{\beta}{2}\int_0^1\left(\int_0^1 g(x,y)dy\right)^2 dx - \frac{1}{4}\int_0^1\int_0^1 I''(h(x,y))g^2(x,y)dxdy\right] + O(\epsilon^3) \\ &= \epsilon^2\left[\frac{\beta}{2}\int_0^1\left(\int_0^1 g(x,y)dy\right)^2 dx - \frac{1}{4}\int_0^1\int_0^1 \frac{g^2(x,y)}{h(x,y)(1-h(x,y))}dxdy\right] + O(\epsilon^3),\end{aligned} \tag{C.9}$$

and since $\beta < 0$, we conclude that h is a local maximizer in (C.6). □

Remark C.1. *In general, the variational problem for the graphons and the corresponding Euler-Lagrange equation (C.7) does not yield a closed form solution. In the special case $\beta = \gamma = 0$,*

$$\psi(\alpha, 0, 0) = \sup_{h \in \mathcal{W}} \left\{ \iint_{[0,1]^2} \alpha(x,y)h(x,y)dxdy - \frac{1}{2} \iint_{[0,1]^2} I(h(x,y))dxdy \right\}, \tag{C.10}$$

where $I(x) := x \log x + (1-x)\log(1-x)$ and it is easy to see that the optimal graphon $h(x,y)$ is given by $h(x,y) = \frac{e^{2\alpha(x,y)}}{e^{2\alpha(x,y)}+1}$, and therefore, $\psi(\alpha,0,0) = \frac{1}{2}\iint_{[0,1]^2} \log(1+e^{2\alpha(x,y)})dxdy$.

Appendix D. Details of Equilibrium Economic Foundations

D.1. **Setup and preferences.** Consider a population of n heterogeneous players (the nodes), each characterized by an exogenous type $\tau_i \in \otimes_{j=1}^{S} \mathcal{X}_j$, $i = 1, ..., n$. The attribute τ_i is an S-dimensional vector and the sets $\mathcal{X}_j$ can represent age, race, gender, income, etc. [24] We collect all τ_i's in an $n \times S$ matrix τ. The network's adjacency matrix g has entries $g_{ij} = 1$ if i and j are linked; and $g_{ij} = 0$ otherwise. The network is undirected, i.e. $g_{ij} = g_{ji}$, and $g_{ii} = 0$, for all i's.[25] The utility of player i is

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

where $\alpha_{ij} := \nu(\tau_i, \tau_j)$ are symmetric functions $\nu : \otimes_{j=1}^{S} \mathcal{X}_j \times \otimes_{j=1}^{S} \mathcal{X}_j \to \mathbb{R}$ and $\nu(\tau_i, \tau_j) = \nu(\tau_j, \tau_i)$ for all i, j; and β is a scalar. The utility of player i depends on the number of direct links, each weighted according to a function ν of the types τ. This payoff structure implies that the net benefit of forming a direct connection depends on the characteristics of the two individuals involved in the link.

Players also care about the number of links that each of their direct contacts have formed.[26] For example, when $\beta > 0$, there is an incentive to form links to people that have many friends, e.g. popular kids in school. On the other hand, when $\beta < 0$ the incentive is reversed. For example, one

[24]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.

[25]Extensions to directed networks are straightforward (see Mele (2017)).

[26]The normalization of β by n is necessary for the asymptotic analysis.

can think that forming links to a person with many connections could decrease our visibility and decrease the effectiveness of interactions. Similar utility functions have been used extensively in the empirical network formation literature.[27]

The preferences in (D.1) include only direct links and friends' populatity. However, we can also include other types of link externalities. For example, in many applications the researcher is interested in estimating preferences for common neighbors. This is an important network statistics to measure transitity and clustering in networks. In our model we can easily add an utility component to capture these effects.

$$u_i(g,\tau) = \sum_{j=1}^{n} \alpha_{ij} g_{ij} + \frac{\beta}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk} + \frac{\gamma}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk} g_{ki}, \tag{D.2}$$

These preferences include an additional parameter γ that measures the effect of common neighbors. The potential function for this model is

$$Q_n(g;\alpha,\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij} g_{ij} + \frac{\beta}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk} + \frac{\gamma}{6n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk} g_{ki}. \tag{D.3}$$

In general, all the results that we show below extend to more general utility functions that include payoffs for link externalities similar to (2.5).

The probability that i and j meet can depend on their networks: it could be a function of their common neighbors, or a function of their degrees and centralities, for example. In Assumption D.1, we assume that the existence of a link between i and j does not affect their probability of meeting. This is because we prove the existence and functional form of the stationary distribution (2.3) using the detailed balance condition, which is not satisfied if we allow the meeting probabilities to depend on the link between i and j.

The model can easily be extended to directed networks and the results on equilibria and long-run stationary distribution will hold. The results about the approximations of the likelihood shown below will also hold for directed networks, with minimal modifications of the proofs.

[27]See Mele (2017), Sheng (2012), DePaula et al. (2011), Chandrasekhar and Jackson (2014), Badev (2013), Butts (2009).

Finally, while our model generates dense graphs, the approximations using variational methods and nonlinear large deviations that we develop in the rest of the paper also work in moderately sparse graphs. More precisely, the utility of player i is given by

$$u_i(g,\tau) = \sum_{j=1}^{n} \alpha_{ij}^{(n)} g_{ij} + \frac{\beta^{(n)}}{n} \sum_{j=1}^{n}\sum_{k=1}^{n} g_{ij}g_{jk} + \frac{\gamma^{(n)}}{n} \sum_{j=1}^{n}\sum_{k=1}^{n} g_{ij}g_{jk}g_{ki}, \tag{D.4}$$

where $|\alpha_{ij}^{(n)}|$, $|\beta^{(n)}|$ and $|\gamma^{(n)}|$ can have moderate growth in n instead of being bounded. We will give more details later in our paper. [28]

Example D.1. *(Homophily) Consider a model with* $\nu(\tau_i,\tau_j) = V - c(\tau_i,\tau_j)$*, where* $V > 0$ *is the benefit of a link and* $c(\tau_i,\tau_j)$ $(= c(\tau_j,\tau_i))$ *is the cost of the link between* i *and* j*. To model homophily in this framework let the cost function be*

$$c(\tau_i,\tau_j) = \begin{cases} c & \text{if } \tau_i = \tau_j, \\ C & \text{if } \tau_i \neq \tau_j. \end{cases} \tag{D.5}$$

For example, consider the parameterization $0 < c < V < C$ *and* $\beta = 0$, $\gamma = 0$*. In this case the players have no incentive to form links with agents of other groups. On the other hand, if we have* $0 < c < V < C$ *and* $\beta, \gamma > 0$*, also links across groups will be formed, as long as* β, γ *are sufficiently large.*

Example D.2. *(Social Distance Model) Let the payoff from direct links be a function of the social distance among the individuals. Formally, let* $\nu(\tau_i,\tau_j) := \eta d(\tau_i,\tau_j) - c$*, where* $d(\tau_i,\tau_j)$ *is a distance function,* η *is a parameter that determines the sensitivity to the social distance and* $c > 0$ *is the cost of forming a link.*[29] *The case with* $\eta < 0$ *represents a world where individuals prefer linking to similar agents and* $\eta > 0$ *represents a world where individuals prefer linking with people at larger social distance. Note that even when* $\eta < 0$*, if we have* $\beta, \gamma > 0$ *sufficiently large, individuals may still have an incentive to form links with people at larger social distance.*

[28] See Chatterjee and Dembo (2016) for additional applications of nonlinear large deviations.

[29] See Iijima and Kamada (2014) for a more general example of such model.

D.2. **Meetings and equilibrium.** The network formation process follows a stochastic best-response dynamics:[30] in each period t, two random players meet with probability ρ_{ij}; upon meeting they have the opportunity to form a link (or delete it, if already in place). Players are myopic: when they form a new link, they do not consider how the new link will affect the incentives of the other player in the future evolution of the network.

ASSUMPTION D.1. *The meeting process is a function of types and the network. Let g_{-ij} indicate the network g without considering the link g_{ij}. Then the probability that i and j meet is*

$$\rho_{ij} := \rho(\tau_i, \tau_j, g_{-ij}) > 0 \tag{D.6}$$

for all pairs i and j, and i.i.d. over time.

Assumption D.1 implies that the meeting process can depend on covariates and the state of the network. For example, if two players have many friends in common they may meet with high probability; or people that share some demographics may meet more often. Crucially, every pair of players has a strictly positive probability of meeting. This guarantees that each link of the network has the opportunity of being revised.

Upon meetings, players decide whether to form or delete a link by maximizing the sum of their current utilities, i.e. the total surplus generated by the relationship. We are implicitly assuming that individuals can transfer utilities. When deciding whether to form a new link or deleting an existing link, players receive a random matching shock ε_{ij} that shifts their preferences.

At time t, the links g_{ij} is formed if

$$u_i(g_{ij}=1, g_{-ij}, \tau) + u_j(g_{ij}=1, g_{-ij}, \tau) + \varepsilon_{ij}(1) \geq u_i(g_{ij}=0, g_{-ij}, \tau) + u_j(g_{ij}=0, g_{-ij}, \tau) + \varepsilon_{ij}(0)\,.$$

We make the following assumptions on the matching value.

ASSUMPTION D.2. *Individuals receive a logistic shock before they decide whether to form a link (i.i.d. over time and players).*

[30]See Blume (1993), Mele (2017), Badev (2013).

The logistic assumption is standard in many discrete choice models in economics and statistics (Train (2009)).

We can now characterize the equilibria of the model, following Mele (2017) and Chandrasekhar and Jackson (2014). In particular, we can show that the network formation is a potential game (Monderer and Shapley (1996)).

PROPOSITION D.1. *The network formation is a potential game, and there exists a potential function $Q_n(g;\alpha,\beta)$ that characterizes the incentives of all the players in any state of the network*

$$Q_n(g;\alpha,\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}g_{ij} + \frac{\beta}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} g_{ij}g_{jk} + \frac{\gamma}{6n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} g_{ij}g_{jk}g_{ki}. \tag{D.7}$$

Proof. The proposition follows the same lines as Proposition 1 in Mele (2017) and it is omitted for brevity. □

The potential function $Q_n(g;\alpha,\beta)$ is such that, for any g_{ij}

$$Q_n(g;\alpha,\beta) - Q_n(g-ij;\alpha,\beta) = u_i(g) + u_j(g) - \left[u_i(g-ij) + u_j(g-ij)\right].$$

Thus we can keep track of all players' incentives using the scalar $Q_n(g;\alpha,\beta)$. It is easy to show that all the pairwise stable (with transfers) networks are the local maxima of the potential function.[31] The sequential network formation follows a *Glauber* dynamics, therefore converging to a unique stationary distribution.

THEOREM D.1. *In the long run, the model converges to the stationary distribution π_n, defined as*

$$\pi_n(g;\alpha,\beta) = \frac{\exp\left[Q_n(g;\alpha,\beta)\right]}{\sum_{\omega\in\mathcal{G}}\exp\left[Q_n(\omega;\alpha,\beta)\right]} = \exp\left\{n^2\left[T_n(g;\alpha,\beta) - \psi_n(\alpha,\beta)\right]\right\}, \tag{D.8}$$

where $T_n(g;\alpha,\beta) = n^{-2}Q_n(g;\alpha,\beta)$,

$$\psi_n(\alpha,\beta) = \frac{1}{n^2}\log\sum_{\omega\in\mathcal{G}}\exp\left[n^2 T_n(\omega;\alpha,\beta)\right], \tag{D.9}$$

[31] A network g is pairwise stable with transfers if: (1) $g_{ij} = 1 \Rightarrow u_i(g,\tau) + u_j(g,\tau) \geq u_i(g-ij,\tau) + u_j(g-ij,\tau)$ and (2) $g_{ij} = 0 \Rightarrow u_i(g,\tau) + u_j(g,\tau) \geq u_i(g+ij,\tau) + u_j(g+ij,\tau)$; where $g+ij$ represents network g with the addition of link g_{ij} and network $g-ij$ represents network g without link g_{ij}. See Jackson (2008) for more details.

and $\mathcal{G} := \{\omega = (\omega_{ij})_{1\leq i,j\leq n} : \omega_{ij} = \omega_{ji} \in \{0,1\}, \omega_{ii} = 0, 1 \leq i,j \leq n\}$.

Proof. The proof is an extension of Theorem 1 in Mele (2017). See also Chandrasekhar and Jackson (2014) and Butts (2009). □

Notice that the likelihood (2.3) corresponds to an ERGM model with heterogeneous nodes and two-stars. As a consequence our model inherits all the estimation and identification challenges of the ERGM model.

APPENDIX E. SPECIAL CASE: THE EDGE-STAR MODEL

The general solution of the variational problem (C.3) is complicated. However, there are some special cases where we can characterize the solution with extreme detail. These examples show how we can solve the variational approximation in stylized settings, and we use them to explain how the method works in practice. In this section, we consider the special case in the absence of transitivity, i.e. $\gamma = 0$ and we get further results for the edge-star model.

E.1. **Extreme homophily.** We can exploit homophily to obtain a tractable approximation. Suppose that there are M types in the population. The cost of forming links among individuals of the same group is finite, but there is a large cost of forming links among people of different groups (potentially infinite). We show that in this case the normalizing constant can be approximated by solving M independent univariate maximization problems. In the special case of extreme homophily, our model converges to a block-diagonal model.

PROPOSITION E.1. *Let* $0 = a_0 < a_1 < \cdots < a_M = 1$ *be a given sequence. Assume that*

$$\alpha(x,y) = \alpha_{mm}, \qquad \text{if } a_{m-1} < x, y < a_m, \qquad m = 1,2,\dots,M. \tag{E.1}$$

and $\alpha(x,y) \leq -K$ *otherwise is a given function. Let* $\psi(\alpha,\beta,0;-K)$ *be the variational problem for the graphons and* $\psi(\alpha,\beta,0;-\infty) = \lim_{K\to\infty}\psi(\alpha,\beta,0;-K)$. *Then, we have*

$$\psi(\alpha,\beta,0;-\infty) = \sum_{m=1}^{M}(a_m - a_{m-1})^2 \sup_{0\leq x\leq 1}\left\{\alpha_{mm}x + \frac{\beta}{2}x^2 - \frac{1}{2}I(x)\right\}. \tag{E.2}$$

Proof. First, observe that

$$\begin{aligned}
&\psi(\alpha,\beta,0;-\infty) \\
&= \sup_{h\in\mathcal{W}^-}\left\{\sum_{i=1}^{M}\alpha_i\iint_{[a_{i-1},a_i]^2}h(x,y)dxdy+\frac{\beta}{2}\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz\right.\\
&\qquad\left.-\frac{1}{2}\sum_{i=1}^{M}\iint_{[a_{i-1},a_i]^2}I(h(x,y))dxdy\right\}\\
&= \sup_{h\in\mathcal{W}^-}\left\{\sum_{i=1}^{M}\alpha_i\iint_{[a_{i-1},a_i]^2}h(x,y)dxdy+\frac{\beta}{2}\sum_{i=1}^{M}\int_{a_{i-1}}^{a_i}\left(\int_{a_{i-1}}^{a_i}h(x,y)dy\right)^2dx\right.\\
&\qquad\left.-\frac{1}{2}\sum_{i=1}^{M}\iint_{[a_{i-1},a_i]^2}I(h(x,y))dxdy\right\}\\
&= \sum_{i=1}^{M}\sup_{\substack{h:[a_{i-1},a_i]^2\to[0,1]\\ h(x,y)=h(y,x)}}\left\{\alpha_i\iint_{[a_{i-1},a_i]^2}h(x,y)dxdy+\frac{\beta}{2}\int_{a_{i-1}}^{a_i}\left(\int_{a_{i-1}}^{a_i}h(x,y)dy\right)^2dx\right.\\
&\qquad\left.-\frac{1}{2}\iint_{[a_{i-1},a_i]^2}I(h(x,y))dxdy\right\},
\end{aligned}\tag{E.3}$$

where

$$\mathcal{W}^- := \left\{h\in\mathcal{W} : h(x,y)=0 \text{ for any } (x,y)\notin\bigcup_{i=1}^{M}[a_{i-1},a_i]^2\right\}.\tag{E.4}$$

By taking h to be a constant on $[a_{i-1},a_i]^2$, it is clear that

$$\psi(\alpha,\beta,0;-\infty)\geq\sum_{i=1}^{M}(a_i-a_{i-1})^2\sup_{0\leq x\leq 1}\left\{\alpha_i x+\frac{\beta}{2}x^2-\frac{1}{2}I(x)\right\}.\tag{E.5}$$

By Jensen's inequality

$$\begin{aligned}\text{(E.6)} \quad \psi(\alpha,\beta,0;-\infty) \leq \sum_{i=1}^{M} \sup_{\substack{h:[a_{i-1},a_i]^2\to[0,1] \\ h(x,y)=h(y,x)}} \Bigg\{ & \alpha_i \int_{a_{i-1}}^{a_i} \left(\int_{a_{i-1}}^{a_i} h(x,y)dy \right) dx \\ & + \frac{\beta}{2} \int_{a_{i-1}}^{a_i} \left(\int_{a_{i-1}}^{a_i} h(x,y)dy \right)^2 dx \\ & - \frac{1}{2}(a_i - a_{i-1}) \int_{a_{i-1}}^{a_i} I\left(\frac{1}{a_i - a_{i-1}} \int_{a_{i-1}}^{a_i} h(x,y)dy \right) dx \Bigg\} \\ \leq \sum_{i=1}^{M} (a_i - a_{i-1})^2 \sup_{0\leq x\leq 1} & \left\{ \alpha_i x + \frac{\beta}{2}x^2 - \frac{1}{2}I(x) \right\}.\end{aligned}$$

□

The net benefit function $\alpha(x,y)$ assumed in the Proposition is shown in Figure C.1(D). Essentially this result means that with extreme homophily, we can approximate the model, assuming perfect segregation: thus we can independently solve the variational problem of each type. This approach is computationally very simple, since each variational problem becomes a univariate maximization problem.

The solution of such univariate problem has been studied and characterized in previous work by Chatterjee and Diaconis (2013), Radin and Yin (2013), Aristoff and Zhu (2018) and Mele (2017). It can be shown that the solutions μ_m^*, where $m = 1, .., M$, are the fixed point of equations

$$\mu_m = \frac{\exp\left[\alpha_{mm} + \beta\mu_m\right]}{1 + \exp\left[\alpha_{mm} + \beta\mu_m\right]}, \tag{E.7}$$

for each group m, and $\beta\mu_m^*(1-\mu_m^*) < 1$. The global maximizer μ_m^* is unique except on a phase transition curve $\{(\alpha_{mm},\beta) : \alpha_{mm} + \beta = 0, \alpha_{mm} < -1\}$, see e.g. Radin and Yin (2013); Aristoff and Zhu (2018). It is shown in Chatterjee and Diaconis (2013) that the network of each group corresponds to an Erdős-Rényi graph with probability of a link equal to μ_m^*.

E.2. **Analytically Tractable Bounds.** In this section, for the edge-star model, we provide analytically tractable bounds for $\psi(\alpha,\beta,\gamma)$ when $\gamma = 0$.

PROPOSITION E.2. *Let $\gamma = 0$ and $0 = a_0 < a_1 < \cdots < a_{M-1} < a_M = 1$ be a given sequence. Let us assume that*

$$\alpha(x,y) = \alpha_{ml}, \qquad \text{if } a_{m-1} < x < a_m \text{ and } a_{l-1} < y < a_l, \text{ where } 1 \leq m,l \leq M.$$

Then, we have

$$\begin{aligned}
&\sup_{\substack{0 \leq u_{ml} \leq 1 \\ u_{ml} = u_{lm}, 1 \leq m,l \leq M}} \sum_{m=1}^{M} (a_m - a_{m-1}) \Bigg\{ \sum_{l=1}^{M} (a_l - a_{l-1}) \alpha_{ml} u_{ml} \\
&\qquad + \frac{\beta}{2} \left(\sum_{l=1}^{M} (a_l - a_{l-1}) u_{ml} \right)^2 - \frac{1}{2} \sum_{l=1}^{M} (a_l - a_{l-1}) I(u_{ml}) \Bigg\} \\
&\leq \psi(\alpha, \beta, 0) \leq \sum_{m=1}^{M} (a_m - a_{m-1}) \sup_{\substack{0 \leq u_{ml} \leq 1 \\ 1 \leq l \leq M}} \Bigg\{ \sum_{l=1}^{M} (a_l - a_{l-1}) \alpha_{ml} u_{ml} + \frac{\beta}{2} \left(\sum_{l=1}^{M} (a_l - a_{l-1}) u_{ml} \right)^2 \\
&\qquad\qquad - \frac{1}{2} \sum_{l=1}^{M} (a_l - a_{l-1}) I(u_{ml}) \Bigg\}.
\end{aligned}$$

Proof. To compute the lower and upper bounds, let us define

$$u_{ij}(x) = \frac{1}{a_j - a_{j-1}} \int_{a_{j-1}}^{a_j} h(x,y) dy, \qquad \text{for any } a_{i-1} < x < a_i. \tag{E.8}$$

We can compute that

$$\iint_{[0,1]^2} \alpha(x,y) h(x,y) dx dy = \sum_{i=1}^{M} \sum_{j=1}^{M} (a_j - a_{j-1}) \int_{a_{i-1}}^{a_i} \alpha_{ij} u_{ij}(x) dx. \tag{E.9}$$

Moreover,

$$\begin{aligned}
\frac{\beta}{2} \int_0^1 \int_0^1 \int_0^1 h(x,y) h(y,z) dx dy dz &= \frac{\beta}{2} \int_0^1 \left(\int_0^1 h(x,y) dy \right)^2 dx \\
&= \frac{\beta}{2} \sum_{i=1}^{M} \int_{a_{i-1}}^{a_i} \left(\sum_{j=1}^{M} (a_j - a_{j-1}) u_{ij}(x) \right)^2 dx.
\end{aligned} \tag{E.10}$$

By Jensen's inequality, we can also compute that

(E.11)
$$\begin{aligned}
\frac{1}{2}\int_0^1\int_0^1 I(h(x,y))dxdy &= \frac{1}{2}\sum_{i=1}^M \int_{a_{i-1}}^{a_i}\left[\sum_{j=1}^M \int_{a_{j-1}}^{a_j} I(h(x,y))dy\right]dx \\
&= \frac{1}{2}\sum_{i=1}^M \int_{a_{i-1}}^{a_i}\left[\sum_{j=1}^M (a_j-a_{j-1})\frac{1}{a_j-a_{j-1}}\int_{a_{j-1}}^{a_j} I(h(x,y))dy\right]dx \\
&\geq \frac{1}{2}\sum_{i=1}^M \int_{a_{i-1}}^{a_i}\left[\sum_{j=1}^M (a_j-a_{j-1})I\left(\frac{1}{a_j-a_{j-1}}\int_{a_{j-1}}^{a_j} h(x,y)dy\right)\right]dx \\
&= \frac{1}{2}\sum_{i=1}^M \int_{a_{i-1}}^{a_i}\sum_{j=1}^M (a_j-a_{j-1})I(u_{ij}(x))dx
\end{aligned}$$

Hence, by (E.9), (E.10), (E.11), we get

$$\begin{aligned}
\psi(\alpha,\beta,0) &\leq \sum_{i=1}^M\sum_{j=1}^M (a_j-a_{j-1})\int_{a_{i-1}}^{a_i}\alpha_{ij}u_{ij}(x)dx + \frac{\beta}{2}\sum_{i=1}^M\int_{a_{i-1}}^{a_i}\left(\sum_{j=1}^M (a_j-a_{j-1})u_{ij}(x)\right)^2 dx \\
&\qquad - \frac{1}{2}\sum_{i=1}^M\int_{a_{i-1}}^{a_i}\sum_{j=1}^M (a_j-a_{j-1})I(u_{ij}(x))dx \\
&\leq \sum_{i=1}^M (a_i-a_{i-1})\sup_{\substack{0\leq u_{ij}\leq 1\\ 1\leq j\leq M}}\left\{\sum_{j=1}^M (a_j-a_{j-1})\alpha_{ij}u_{ij} + \frac{\beta}{2}\left(\sum_{j=1}^M (a_j-a_{j-1})u_{ij}\right)^2\right. \\
&\qquad \left. - \frac{1}{2}\sum_{j=1}^M (a_j-a_{j-1})I(u_{ij})\right\}
\end{aligned}$$

On the other hand, by restricting the supremum over the graphons $h(x,y)$

(E.12) $$h(x,y)=u_{ij}, \qquad \text{if } a_{i-1}<x<a_i \text{ and } a_{j-1}<y<a_j, \text{ where } 1\leq i,j\leq M,$$

where $(u_{ij})_{1\leq i,j\leq M}$ is a symmetric matrix of the constants, and optimize over all the possible values $0 \leq u_{ij} \leq 1$, we get the lower bound:

$$\begin{aligned}\psi(\alpha,\beta,0) \geq \sup_{\substack{0\leq u_{ij}\leq 1\\ u_{ij}=u_{ji},1\leq i,j\leq M}} \sum_{i=1}^{M}(a_i - a_{i-1})\Bigg\{ \sum_{j=1}^{M}(a_j - a_{j-1})\alpha_{ij}u_{ij} \\ + \frac{\beta}{2}\left(\sum_{j=1}^{M}(a_j - a_{j-1})u_{ij}\right)^2 - \frac{1}{2}\sum_{j=1}^{M}(a_j - a_{j-1})I(u_{ij})\Bigg\}.\end{aligned} \tag{E.13}$$

□

CAREY BUSINESS SCHOOL, JOHNS HOPKINS UNIVERSITY, 100 INTERNATIONAL DR, BALTIMORE, MD 21202

DEPARTMENT OF MATHEMATICS, FLORIDA STATE UNIVERSITY, 208 LOVE BUILDING, 1017 ACADEMIC WAY, TALLAHASSEE, FL 32306