Variational approximations for the exponential random graph model

Angelo Mele, Johns Hopkins University
Lingjiong Zhu, University of Minnesota

We study a model of sequential network formation that converges to the exponential random graph model (ERGM). The likelihood of the model is known up to an intractable normalizing constant, which is usually approximated using simulation methods. However, some of these methods are computationally very expensive and may fail to converge in reasonable time for large networks applications. In this paper, we leverage and extend recent results from the literature on large deviations for random graphs to provide an algorithm for estimation of the model that avoids simulations. Our method uses a variational mean-field approximation of the likelihood: we show that the approximation becomes exact as the number of nodes grows to infinity, providing a consistent estimate of the normalizing constant. We also provide analytical bounds for finite networks and show that we can exploit homophily to simplify the variational approximation. This method is tractable and scales to large networks.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Economics, Sociology; G.2.2 [**Graph theory**]: Network problems; G.3 [**Probability and statistics**]: Statistical Computing; I.5.1 [**Pattern Recognition**]: Statistical Models

Additional Key Words and Phrases: network formation, ERGM, mean-field approximations, estimation, graph limits, large deviations

1. INTRODUCTION

We study a model of strategic network formation with heterogeneous players, that converges to the exponential random graph model. The likelihood of observing a specific network is known up to an intractable normalizing constant, which is infeasible to compute. The standard estimation method uses a MCMC algorithm that generates samples from the ERGM to provide an estimate of the normalizing constant.[1] However, recent work by Bhamidi et al. [2011] has shown that such simulation methods may have exponentially slow convergence.

We provide an alternative tractable method of estimation for a large class of exponential random graph models. We show that a mean-field variational approximation of the likelihood provides a lower bound for the normalizing constant.[2]

The main theorem of the paper shows that this lower bound becomes exact as the network size n grows large. To obtain the latter statement we extend recent results from the large deviations literature for random graphs.[3] We also provide exact bounds

[1]Snijders [2002], Caimo and Friel [2010], Mele [2011], Geyer and Thompson [1992]
[2]Wainwright and Jordan [2008], Bishop [2006].
[3]See Chatterjee and Varadhan [2011], Chatterjee and Diaconis [2013], Chatterjee and Dembo [2014]

Author's addresses: Angelo Mele, Carey Business School, Johns Hopkins University, 100 International Drive, Baltimore, MD-21202, USA; email: angelo.mele@jhu.edu; Lingjiong Zhu, School of Mathematics, University of Minnesota, 206 Church Street S.E., Minneapolis, MN-55455, USA; email: zhul@umn.edu.

DOI:http://dx.doi.org/10.1145/0000000.0000000

for the approximation error of the variational mean-field for fixed n.[4] Lastly, we show that when the model features (extreme) homophily, we can approximate the intractable constant by solving independent univariate maximization problems. Our estimation method is tractable and scalable for large networks.

2. THEORETICAL MODEL

There is a population of n players (the nodes), characterized by an exogenous type τ_i: this vector can contain age, race, gender, income, etc. We collect all τ_i's in a 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.[5] The utility of player i is

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

where $\alpha_{ij} := \alpha(\tau_i, \tau_j)$ are symmetric. The utility of player i depends on the number of direct links, each weighted according to a function α of the types τ. Players also care about the number of links that each of their direct contacts have formed.[6]

The network formation follows a sequential best-response dynamics. In each period t, a pair of players is selected from the population with probability ρ_{ij}. Upon meeting, the pair decides whether to form a link g_{ij} by maximizing the sum of their utility. Players are myopic: when they form a new link, they do not consider the effect of that link on the future evolution of the network.
We make the following assumptions.

ASSUMPTION 2.1. *The meeting process does not depend on the state of the network, and $\rho_{ij} > 0$ for all ij and i.i.d. over time.*

The assumption means that any meeting has positive probability (however small) and the meetings are i.i.d. over time.

ASSUMPTION 2.2. *Individuals receive a logistic shock before they decide whether to form a link (i.i.d. over time and players).*

Before deciding whether to form or sever a link, the pair receives a stochastic shock ε_{ij} to the surplus generated by the relationship, which models an i.i.d. matching value. This is a standard assumption in many random utility models. We can now show that the network formation is a potential game (Monderer and Shapley [1996]).

PROPOSITION 2.3. *If assumptions 2.1 and 2.2 hold, then 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*

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

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

---

[4] He and Zheng [2013] use a similar approximation, but they do not provide approximation errors.
[5] Extensions to directed networks are straightforward (see Mele [2011])
[6] The normalization of β by n is necessary for the asymptotic analysis.

the potential function.[7] The sequential network formation follows a *Glauber* dynamics, therefore converging to a unique stationary distribution.

THEOREM 2.4. *In the long run, the model converges to the stationary distribution π_n, defined as*

$$\pi_n(g;\alpha,\beta) = \frac{\exp\left[Q_n(g;\alpha,\beta)\right]}{\sum_{\omega\in\mathcal{G}}\exp\left[Q_n(\omega;\alpha,\beta)\right]} = \exp\left\{n^2\left[T_n(g;\alpha,\beta) - \psi_n(\alpha,\beta)\right]\right\}, \tag{3}$$

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

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

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

The statements in Proposition 2.3 and Theorem 2.4 are straightforward extensions of Mele [2011]) and Chandrasekhar and Jackson [2014]. Notice that the likelihood (3) corresponds to an ERGM model with two-stars.

3. VARIATIONAL APPROXIMATIONS

The constant $\psi_n(\alpha,\beta)$ in (4) is intractable because it involves a sum over all $2^{\binom{n}{2}}$ possible networks with n players. The usual strategy consists of approximating the constant using an MCMC sampler (Snijders [2002]). At each iteration, a random link g_{ij} is selected and it is proposed to swap its value to $1 - g_{ij}$; the swap is accepted according to a Metropolis-Hastings ratio. However, recent work by Bhamidi et al. [2011] has shown that such a local sampler may have exponentially slow convergence for many non-trivial parameter vectors.

We propose an alternative estimation method that does not rely on simulations. Our method consists of finding 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:

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

With some algebra we obtain

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

where $\mathcal{H}(q_n) = -\sum_{\omega\in\mathcal{G}} q_n(\omega)\log q_n(\omega)$ is the entropy of distribution q_n.

We want to find the best likelihood approximation, so we minimize $KL(q_n|\pi_n)$ with respect to q_n, which is equivalent to

$$\psi_n(\alpha,\beta) = \sup_{q_n\in\mathcal{Q}_n}\mathcal{L}(q_n) = \sup_{q_n\in\mathcal{Q}_n}\left\{\mathbb{E}_q\left[T_n(\omega;\alpha,\beta)\right] + \frac{1}{n^2}\mathcal{H}(q_n)\right\}. \tag{5}$$

In most cases this variational problem has no closed-form solution. The machine learning literature suggests to restrict the set $\mathcal{Q}_n$ to find a tractable approximation.[8] A pop-

[7] A network g is pairwise stable with transfers if: (1) $g_{ij} = 1 \Rightarrow u_i(g,\tau) + u_j(g,\tau) \ge u_i(g-ij,\tau) + u_j(g-ij,\tau)$ and (2) $g_{ij} = 0 \Rightarrow u_i(g,\tau) + u_j(g,\tau) \ge u_i(g+ij,\tau) + u_j(g+ij,\tau)$; 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.

[8] See Wainwright and Jordan [2008], Bishop [2006]

ular choice for the set $\mathcal{Q}_n$ is the set of all completely factorized distribution

$$q_n(g) = \prod_{i,j} \mu_{ij}^{g_{ij}} (1-\mu_{ij})^{1-g_{ij}}, \tag{6}$$

where $\mu_{ij} = \mathbb{E}_q(g_{ij}) = \mathbb{P}_q(g_{ij} = 1)$. Straightforward algebra shows that the entropy of q_n is additive in each link's entropy

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

and the expected potential is computed as

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

The mean-field approximation leads to a lower bound. The maximization problem is now to find a matrix $\boldsymbol{\mu}(\alpha,\beta)$

$$\begin{aligned}\psi_n(\alpha,\beta) \;&\geq\; \psi_n^{MF}(\boldsymbol{\mu}(\alpha,\beta)) \\ &= \sup_{\boldsymbol{\mu}\in[0,1]^{n^2}}\left\{\frac{\sum_i\sum_j \alpha_{ij}\mu_{ij}}{n^2} + \beta\frac{\sum_i\sum_j\sum_k \mu_{ij}\mu_{jk}}{2n^3}\right.\\ &\left.-\;\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\mu_{ij}\log\mu_{ij} + (1-\mu_{ij})\log(1-\mu_{ij})\right]\right\}.\end{aligned}$$

The maximization can be performed using any global optimization method. The machine learning literature proposes to use an iterative method that is guaranteed to converge to a local maximum.[9] If we take the first order conditions of the mean-field problem with respect to each μ_{ij} we obtain

$$\mu_{ij} = \frac{\exp\left[2\alpha_{ij} + \frac{\beta}{n}\sum_{k=1}^{n}\left(\mu_{jk}+\mu_{ki}\right)\right]}{1+\exp\left[2\alpha_{ij} + \frac{\beta}{n}\sum_{k=1}^{n}\left(\mu_{jk}+\mu_{ki}\right)\right]}, \quad i,j = 1,...,n. \tag{7}$$

We initialize the matrix $\boldsymbol{\mu}$ and iteratively solve (7) for each entry of the matrix. We can restart the algorithm several times to get a better approximation. Notice that this is easily parallelizable.

4. ASYMPTOTIC RESULTS

4.1. Convergence of the variational mean-field approximation

In this section we consider the model as $n \to \infty$. We use and extend results from the graph limits literature,[10] large deviations literature for random graphs[11] and analysis of the resulting variational problem.[12] Let h be a simple symmetric function $h : [0,1]^2 \to [0,1]$, and $h(x,y) = h(y,x)$. This function is called a graphon and it is a representation of an infinite network. We also need a representation of the vector α in the infinite network. The following assumptions restrict our model to discrete types.

[9]See Wainwright and Jordan [2008]
[10]See Lovasz [2012], Borgs et al. [2008]
[11]See Chatterjee and Varadhan [2011], Chatterjee and Diaconis [2013]
[12]See Aristoff and Zhu [2014], Radin and Yin [2013] among others.

ASSUMPTION 4.1. *Assume that*

$$\alpha_{ij} = \alpha\left(i/n, j/n\right), \tag{8}$$

where $\alpha(x,y) : [0,1]^2 \to [0,1]$*, are deterministic exogenous functions that are symmetric, i.e.,* $\alpha(x,y) = \alpha(y,x)$*,*

We allow finitely many types for the players and therefore $\alpha(x,y)$ is a multipodal function, that is, they take only finitely many values.[13]

ASSUMPTION 4.2. *Assume that* α_{ij} *take finitely many values. More precisely,* α_{ij} *takes values* $\alpha_1, \ldots, \alpha_p$ *and*

$$\alpha_{ij} = \alpha_\ell \quad \textit{if } (i,j) \in \mathcal{A}_\ell, 1 \le \ell \le p, \tag{9}$$

where $\{\mathcal{A}_\ell\}_{1\le\ell\le p}$ *is a partition of* $\{(i,j) : 1 \le i \ne j \le n\}$.

As a simple example, let us consider only gender: males and females. For example, half of the nodes (population) are males, say $i = 1, 2, \ldots, \frac{n}{2}$ and the other half are females, $i = \frac{n}{2}+1, \frac{n}{2}+2, \ldots, n$.[14] That means, $\alpha(x,y)$ takes three values according to the three regions:

$$\left\{(x,y) : 0 < x, y < \tfrac{1}{2}\right\}, \tag{10}$$

$$\left\{(x,y) : \tfrac{1}{2} < x, y < 1\right\}, \tag{11}$$

$$\left\{(x,y) : 0 < x < \tfrac{1}{2} < y < 1\right\} \bigcup \left\{(x,y) : 0 < y < \tfrac{1}{2} < x < 1\right\}, \tag{12}$$

and these three regions correspond precisely to pairs: male-male, female-female, and male-female.
The work of Chatterjee and Diaconis [2013] show that there is a variational problem analogous to the one shown above for the graph limit. Our extension has a similar flavor, and the variational problem for the graphon is

$$\begin{aligned}\psi(\alpha,\beta) \;=\; & \sup_{h\in\mathcal{W}} \left\{ \int_0^1\int_0^1 \alpha(x,y)h(x,y)dxdy + \frac{\beta}{2}\int_0^1\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz \right. \\ & \left. -\; \frac{1}{2}\int_0^1\int_0^1 \left[h(x,y)\log h(x,y) + (1-h(x,y))\log(1-h(x,y))\right]dxdy \right\},\end{aligned} \tag{13}$$

where $\mathcal{W} := \{h : [0,1]^2 \to [0,1], h(x,y) = h(y,x), 0 \le x, y \le 1\}$.
For finite n, the variational mean-field approximation contains an error of approximation. The next result quantifies and provides lower and upper bound to the error of approximation.

THEOREM 4.3. *Under Assumption 4.2 and for fixed network size* n*, the approximation error of the variational mean-field problem is bounded as*

$$C_3(\beta)n^{-1} \le \psi_n(\alpha,\beta) - \psi_n^{MF}(\alpha,\beta) \le C_1(\alpha,\beta)n^{-1/5}(\log n)^{1/5} + C_2(\alpha,\beta)n^{-1/2}, \tag{14}$$

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

From Chatterjee and Diaconis [2013] we know that as $n \to \infty$ we have $\psi_n(\alpha,\beta) \to \psi(\alpha,\beta)$. The following proposition shows that for a model with finitely many types the variational approximation is asymptotically exact.

[13]If an entry of the vector τ_i is continuous, we can always transform the variable in a discrete vector using thresholds.
[14]Here, we assume without loss of generality that n is an even number.

PROPOSITION 4.4. *Under Assumption 4.1 and 4.2, the mean-field approximation becomes exact as $n \to \infty$*

$$\psi_n^{MF}(\boldsymbol{\mu}(\alpha,\beta)) \to \psi(\alpha,\beta). \tag{15}$$

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

4.2. A model with homophily

We can also 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. Let $I(x) := x \log x + (1-x)\log(1-x)$. Then we have

PROPOSITION 4.5. *Let $0 = a_0 < a_1 < \cdots < a_M = 1$ be a given sequence. Assume that*

$$\alpha(x,y) = \alpha_i, \qquad \text{if } a_{i-1} < x, y < a_i, \qquad i = 1, 2, \ldots, M. \tag{16}$$

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

$$\psi(\alpha,\beta;-\infty) = \sum_{i=1}^{M}(a_i - a_{i-1})^2 \sup_{0\leq x\leq 1}\left\{\alpha_i x + \frac{\beta}{2}x^2 - \frac{1}{2}I(x)\right\}. \tag{17}$$

Essentially this result means that with extreme homophily, the model can be approximated using a stochastic block model. An equivalent formulation imposes (almost) perfect segregation through the meeting process by assuming that $\rho_{ij} = 0$ if $\tau_i \neq \tau_j$.

5. CONCLUSIONS

We provided a simple model of sequential network formation that evolves according to a Glauber dynamics and converges to an exponential random graph in the limit. The usual estimation strategy for ERGM models consists of using state-of-the-art MCMC methods to approximate the normalizing constant of the likelihood. However, these simulations are computationally expensive and may suffer convergence problems.

We propose an alternative method that avoids simulations, and delivers an asymptotically exact estimate of the normalizing constant. Extending recent results from the large deviations literature for random graphs, we show convergence for large n and we bound the approximation error for fixed n. We also show that we can exploit homophily to simplify the approximation problem to solving several univariate maximization problems.

The method proposed here does not suffer the problem of MCMC simulations. In addition, it is parallelizable and scales well to large networks.

APPENDIX

Remark A.1. In general, the variational problem for the graphons does not yield a closed form solution. In the special case $\beta = 0$,

$$\psi(\alpha, 0) = \sup_{h \in \mathcal{W}} \left\{ \iint_{[0,1]^2} \alpha(x,y) h(x,y) dxdy - \frac{1}{2} \iint_{[0,1]^2} I(h(x,y)) dxdy \right\}, \tag{18}$$

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

$$h(x,y) = \frac{e^{2\alpha(x,y)}}{e^{2\alpha(x,y)} + 1}, \tag{19}$$

and therefore,

$$\psi(\alpha, 0) = \frac{1}{2} \iint_{[0,1]^2} \log(1 + e^{2\alpha(x,y)}) dxdy. \tag{20}$$

A.1. Proof of Theorem 4.3

In this proof we will try to follow closely the notation in Chatterjee and Dembo [2014]. 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

$$f_i := \frac{\partial f}{\partial x_i}, \qquad f_{ij} := \frac{\partial^2 f}{\partial x_i \partial x_j}, \tag{21}$$

and let

$$a := \|f\|, \qquad b_i := \|f_i\|, \qquad c_{ij} := \|f_{ij}\|. \tag{22}$$

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

$$\sum_{i=1}^{N} (f_i(x) - d_i)^2 \leq N\epsilon^2. \tag{23}$$

Let us define

$$F := \log \sum_{x \in \{0,1\}^N} e^{f(x)}, \tag{24}$$

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

$$I(x) := \sum_{i=1}^{N} [x_i \log x_i + (1 - x_i)\log(1 - x_i)]. \tag{25}$$

Theorem 1.5. in Chatterjee and Dembo [2014] says the following:

THEOREM A.2 (CHATTERJEE AND DEMBO [2014]). *For any* $\epsilon > 0$*,*

$$\sup_{x \in [0,1]^N} \{f(x) - I(x)\} - \frac{1}{2} \sum_{i=1}^{N} c_{ii} \leq F \leq \sup_{x \in [0,1]^N} \{f(x) - I(x)\} + \mathcal{E}_1 + \mathcal{E}_2, \tag{26}$$

where

$$\mathcal{E}_1 := \frac{1}{4} \left(N \sum_{i=1}^{N} b_i^2 \right)^{1/2} \epsilon + 3N\epsilon + \log |\mathcal{D}(\epsilon)|, \tag{27}$$

and

$$\mathcal{E}_2 := 4\left(\sum_{i=1}^N (ac_{ii}+b_i^2)+\frac{1}{4}\sum_{i,j=1}^N (ac_{ij}^2+b_ib_jc_{ij}+4b_ic_{ij})\right)^{1/2} + \frac{1}{4}\left(\sum_{i=1}^N b_i^2\right)^{1/2}\left(\sum_{i=1}^N c_{ii}^2\right)^{1/2} + 3\sum_{i=1}^N c_{ii} + \log 2. \tag{28}$$

We will use the previous theorem to derive the lower and upper bound of the approximation of the mean-field approximation problem. Notice that in our case the N of the theorem is the number of links, i.e. $N = \binom{n}{2}$. Let

$$Z_n := \sum_{x_{ij}\in\{0,1\}, x_{ij}=x_{ji}, 1\le i<j\le n} e^{\sum_{1\le i,j\le n}\alpha_{ij}x_{ij} + \frac{\beta}{2n}\sum_{1\le i,j,k\le n} x_{ij}x_{jk}}, \tag{29}$$

be the normalizing factor and also define

$$L_n := \sup_{x_{ij}\in[0,1], x_{ij}=x_{ji}, 1\le i<j\le n}\left\{\frac{1}{n^2}\sum_{i,j}\alpha_{ij}x_{ij} + \frac{\beta}{2n^3}\sum_{i,j,k}x_{ij}x_{jk} - \frac{1}{n^2}\sum_{1\le i<j\le n}[x_{ij}\log x_{ij} + (1-x_{ij})\log(1-x_{ij})]\right\}. \tag{30}$$

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

$$f(x) = \sum_{i=1}^n\sum_{j=1}^n \alpha_{ij}x_{ij} + \frac{\beta}{2n}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n x_{ij}x_{jk}. \tag{31}$$

Then, we can compute that, for sufficiently large n,

$$a = \|f\| \le \sum_{i=1}^n\sum_{j=1}^n |\alpha_{ij}| + \sum_{i=1}^n\sum_{k=1}^n \frac{1}{2}|\beta| \le n^2\left[\int_{[0,1]^2}|\alpha(x,y)|dxdy + \frac{1}{2}|\beta| + 1\right]. \tag{32}$$

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 $T : [0,1]^{\binom{n}{2}} \to \mathbb{R}$

$$T(x) := \frac{1}{n^{k-2}}\sum_{q\in[n]^k}\prod_{\{\ell,\ell'\}\in E} x_{q_\ell q_{\ell'}}, \tag{33}$$

Chatterjee and Dembo [2014] (Lemma 5.1.) showed that, for any $i<j$, $i'<j'$,

$$\left\|\frac{\partial T}{\partial x_{ij}}\right\| \le 2|E|, \tag{34}$$

and

$$\left\|\frac{\partial^2 T}{\partial x_{ij}\partial x_{i'j'}}\right\| \le \begin{cases} 4|E|(|E|-1)n^{-1} & \text{if } |\{i,j,i',j'\}| = 2 \text{ or } 3, \\ 4|E|(|E|-1)n^{-2} & \text{if } |\{i,j,i',j'\}| = 4. \end{cases} \tag{35}$$

Therefore, by (34), we can compute that

$$b_{(ij)} = \left\|\frac{\partial f}{\partial x_{ij}}\right\| \le 2\sup_{0\le x,y\le 1}|\alpha(x,y)| + 2|\beta|. \tag{36}$$

By (35), we can also compute that

$$c_{(i,j)(i'j')} = \left\| \frac{\partial^2 f}{\partial x_{ij} \partial x_{i'j'}} \right\| \le \begin{cases} 4|\beta| n^{-1} & \text{if } |\{i,j,i',j'\}| = 2 \text{ or } 3, \\ 4|\beta| n^{-2} & \text{if } |\{i,j,i',j'\}| = 4. \end{cases} \tag{37}$$

Next, we compute that

$$\frac{\partial f}{\partial x_{ij}} = 2\alpha_{ij} + \frac{\partial}{\partial x_{ij}} \frac{\beta}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} x_{ij} x_{jk}. \tag{38}$$

Let T be defined as

$$T(x) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} x_{ij} x_{jk}. \tag{39}$$

Then, we have

$$\frac{\partial f}{\partial x_{ij}} = 2\alpha_{ij} + \frac{\beta}{2} \frac{\partial T}{\partial x_{ij}}. \tag{40}$$

Chatterjee and Dembo [2014] (Lemma 5.2.) showed that for the T defined above, there exists a set $\tilde{\mathcal{D}}(\epsilon)$ satisfying the criterion (23) (with $f = T$) so that

$$|\tilde{\mathcal{D}}(\epsilon)| \le \exp\left\{ \frac{\tilde{C}_1 2^4 3^4 n}{\epsilon^4} \log \frac{\tilde{C}_2 2^4 3^4}{\epsilon^4} \right\} = \exp\left\{ \frac{C_1 n}{\epsilon^4} \log \frac{C_2}{\epsilon^4} \right\}, \tag{41}$$

where $C_i = 2^4 3^4 \tilde{C}_i$, $i = 1, 2$, are universal constants. Let us define

$$\mathcal{D}(\epsilon) := \left\{ 2\alpha_\ell + \frac{\beta}{2} d : d \in \tilde{\mathcal{D}}(2\epsilon/\beta), \ell = 1, \dots, p \right\}. \tag{42}$$

Hence, $\mathcal{D}(\epsilon)$ satisfies the criterion (23) and

$$|\mathcal{D}(\epsilon)| \le p |\tilde{\mathcal{D}}(2\epsilon/\beta)| \le p \exp\left\{ \frac{C_1 \beta^4 n}{2^4 \epsilon^4} \log \frac{C_2 \beta^4}{2^4 \epsilon^4} \right\}. \tag{43}$$

Therefore,

$$\begin{aligned} \mathcal{E}_1 &= \frac{1}{4} \left(\binom{n}{2} \sum_{1 \le i < j \le n} b_{(ij)}^2 \right)^{1/2} \epsilon + 3 \binom{n}{2} \epsilon + \log |\mathcal{D}(\epsilon)| \\ &\le \left[\frac{1}{4} (2\|\alpha\|_\infty + 2|\beta|) + 3 \right] \binom{n}{2} \epsilon + \log p + \frac{C_1 \beta^4 n}{2^4 \epsilon^4} \log \frac{C_2 \beta^4}{2^4 \epsilon^4} \\ &\le C_1(\alpha, \beta) n^2 \epsilon + \frac{C_1(\alpha, \beta) n}{\epsilon^4} \log \frac{C_1(\alpha, \beta)}{\epsilon^4} \\ &= C_1(\alpha, \beta) n^{9/5} (\log n)^{1/5}, \end{aligned} \tag{44}$$

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

We can also compute that

$$
\begin{aligned}
\mathcal{E}_2 = 4\Bigg(& \sum_{1\le i<j\le n} (ac_{(ij)(ij)} + b^2_{(ij)}) \\
& + \frac{1}{4} \sum_{1\le i<j\le n, 1\le i'<j'\le n} \left(ac^2_{(ij)(i'j')} + b_{ij} b_{i'j'} c_{(ij)(i'j')} + 4b_{(ij)} c_{(ij)(i'j')} \right) \Bigg)^{1/2} \\
& + \frac{1}{4} \left(\sum_{1\le i<j\le n} b^2_{(ij)} \right)^{1/2} \left(\sum_{1\le i<j\le n} c^2_{(ij)(ij)} \right)^{1/2} + 3 \sum_{1\le i<j\le n} c_{(ij)(ij)} + \log 2 \\
\le 4\Bigg(& \binom{n}{2} \left(n \left(\|\alpha\|_1 + \frac{1}{2}|\beta| + 1 \right) 4|\beta| + (2\|\alpha\|_\infty + 2|\beta|)^2 \right) \\
& + \frac{1}{4} n^2 \left[\|\alpha\|_1 + \frac{1}{2}|\beta| + 1 \right] \left[\binom{n}{4} 4^2 |\beta|^2 n^{-4} + \left(\binom{n}{2}^2 - \binom{n}{4} \right) 4^2 |\beta|^2 n^{-2} \right] \\
& + (2\|\alpha\|_\infty + 2|\beta|) \left(\frac{\|\alpha\|_\infty}{2} + \frac{1}{2}|\beta| + 1 \right) \\
& \quad \cdot \left[\binom{n}{4} 4|\beta| n^{-2} + \left(\binom{n}{2}^2 - \binom{N}{4} \right) 4|\beta| n^{-1} \right] \Bigg)^{1/2} \\
& + \frac{1}{4} \binom{n}{2} (2\|\alpha\|_\infty + 2|\beta|) 4|\beta| n^{-1} + 3 \binom{n}{2} 4|\beta| n^{-1} + \log 2 \\
\le C_2 & (\alpha, \beta) n^{3/2},
\end{aligned} \tag{45}
$$

where $C_2(\alpha, \beta)$ is a constant depending only on α, β.

Finally, to get lower bound, notice that

$$
\frac{1}{2} \sum_{1\le i<j\le n} c_{(ij)(ij)} = \frac{1}{2} \binom{n}{2} 4|\beta| n^{-1} \le C_3(\beta) n, \tag{46}
$$

where $C_3(\beta)$ is a constant depending only on β.

A.2. Proof of Proposition 4.5

First, observe that

$$
\begin{aligned}
&\psi(\alpha,\beta;-\infty) \qquad (47)\\
&= \sup_{h\in\mathcal{W}^-}\left\{\sum_{i=1}^M \alpha_i \iint_{[a_{i-1},a_i]^2} h(x,y)dxdy + \frac{\beta}{2}\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz \right.\\
&\qquad\qquad \left. - \frac{1}{2}\sum_{i=1}^M \iint_{[a_{i-1},a_i]^2} I(h(x,y))dxdy\right\}\\
&= \sup_{h\in\mathcal{W}^-}\left\{\sum_{i=1}^M \alpha_i \iint_{[a_{i-1},a_i]^2} h(x,y)dxdy + \frac{\beta}{2}\sum_{i=1}^M\int_{a_{i-1}}^{a_i}\left(\int_{a_{i-1}}^{a_i} h(x,y)dy\right)^2 dx \right.\\
&\qquad\qquad \left. - \frac{1}{2}\sum_{i=1}^M \iint_{[a_{i-1},a_i]^2} I(h(x,y))dxdy\right\}\\
&= \sum_{i=1}^M \sup_{\substack{h:[a_{i-1},a_i]^2\to[0,1]\\ h(x,y)=h(y,x)}}\left\{\alpha_i \iint_{[a_{i-1},a_i]^2} h(x,y)dxdy + \frac{\beta}{2}\int_{a_{i-1}}^{a_i}\left(\int_{a_{i-1}}^{a_i} h(x,y)dy\right)^2 dx \right.\\
&\qquad\qquad \left. - \frac{1}{2}\iint_{[a_{i-1},a_i]^2} I(h(x,y))dxdy\right\},
\end{aligned}
$$

where

$$
\mathcal{W}^- := \left\{ h\in\mathcal{W} : h(x,y)=0 \text{ for any } (x,y)\notin \bigcup_{i=1}^M [a_{i-1},a_i]^2 \right\}. \qquad (48)
$$

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

$$
\psi(\alpha,\beta;-\infty) \geq \sum_{i=1}^M (a_i-a_{i-1})^2 \sup_{0\leq x\leq 1}\left\{\alpha_i x + \frac{\beta}{2}x^2 - \frac{1}{2}I(x)\right\}. \qquad (49)
$$

By Jensen's inequality

$$
\begin{aligned}
\psi(\alpha,\beta;-\infty) &\leq \sum_{i=1}^M \sup_{\substack{h:[a_{i-1},a_i]^2\to[0,1]\\ h(x,y)=h(y,x)}}\left\{\alpha_i\int_{a_{i-1}}^{a_i}\left(\int_{a_{i-1}}^{a_i} h(x,y)dy\right)dx \right. \qquad (50)\\
&\qquad + \frac{\beta}{2}\int_{a_{i-1}}^{a_i}\left(\int_{a_{i-1}}^{a_i} h(x,y)dy\right)^2 dx\\
&\qquad\qquad \left. - \frac{1}{2}(a_i-a_{i-1})\int_{a_{i-1}}^{a_i} I\left(\frac{1}{a_i-a_{i-1}}\int_{a_{i-1}}^{a_i} h(x,y)dy\right)dx\right\}\\
&\leq \sum_{i=1}^M (a_i-a_{i-1})^2 \sup_{0\leq x\leq 1}\left\{\alpha_i x + \frac{\beta}{2}x^2 - \frac{1}{2}I(x)\right\}.
\end{aligned}
$$

ACKNOWLEDGMENTS

Angelo Mele thanks Roger Koenker for introducing him to variational approximations.

REFERENCES

ARISTOFF, D. AND ZHU, L. 2014. On the phase transition curve in a directed exponential random graph model. Working Paper.

BHAMIDI, S., BRESLER, G., AND SLY, A. 2011. Mixing time of exponential random graphs. *The Annals of Applied Probability 21,* 6, 2146–2170.

BISHOP, C. 2006. *Pattern recognition and machine learning*. Springer, New York.

BORGS, C., CHAYES, J., LOVÁSZ, L., SÓS, V., AND VESZTERGOMBI, K. 2008. Convergent sequences of dense graphs i: Subgraph frequencies, metric properties and testing. *Advances in Mathematics 219,* 6, 1801 – 1851.

CAIMO, A. AND FRIEL, N. 2010. Bayesian inference for exponential random graph models. *Social Networks forthcoming*.

CHANDRASEKHAR, A. AND JACKSON, M. 2014. Tractable and consistent exponential random graph models.

CHATTERJEE, S. AND DEMBO, A. 2014. Nonlinear large deviations. Working Paper.

CHATTERJEE, S. AND DIACONIS, P. 2013. Estimating and understanding exponential random graph models. *The Annals of Statistics 41,* 5.

CHATTERJEE, S. AND VARADHAN, S. R. S. 2011. The large deviation principle for the Erdos-Rényi random graph. *European Journal of Combinatorics 32,* 7, 1000 – 1017. Homomorphisms and Limits.

GEYER, C. AND THOMPSON, E. 1992. Constrained monte carlo maximum likelihood for depedendent data. *Journal of the Royal Statistical Society, Series B (Methodological) 54,* 3, 657–699.

HE, R. AND ZHENG, T. 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, 248–255.

JACKSON, M. O. 2008. *Social and Economics Networks*. Princeton.

LOVASZ, L. 2012. *Large Networks and Graph Limits*. American Mathematical Society colloquium publications. American Mathematical Society.

MELE, A. 2011. A structural model of segregation in social networks. Working Paper.

MONDERER, D. AND SHAPLEY, L. 1996. Potential games. *Games and Economic Behavior 14*, 124–143.

RADIN, C. AND YIN, M. 2013. Phase transitions in exponential random graphs. *The Annals of Applied Probability 23,* 6, 2458–2471.

SNIJDERS, T. A. 2002. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure 3,* 2.

WAINWRIGHT, M. AND JORDAN, M. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends@ in Machine Learning 1,* 1-2, 1–305.