APPROXIMATE VARIATIONAL ESTIMATION FOR A MODEL OF NETWORK FORMATION

ANGELO MELE AND LINGJIONG ZHU

Abstract. We study an equilibrium model of sequential network formation with heterogeneous players. The payoffs depend on the number and composition of direct connections, but also the number of indirect links. We show that the network formation process is a potential game and in the long run the model converges to an exponential random graph (ERGM). Since standard simulation-based inference methods for ERGMs could have exponentially slow convergence, we propose an alternative deterministic method, based on a variational approximation of the likelihood. We compute bounds for the approximation error for a given network size and we prove that our variational method is asymptotically exact, extending results from the large deviations and graph limits literature to allow for covariates in the ERGM. A simple Monte Carlo shows that our deterministic method provides more robust estimates than standard simulation based inference.

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

1. Introduction

In this paper we study an equilibrium model of network formation with heterogeneous agents, and provide variational methods to approximate the likelihood. We contribute to the literature on the formation and evolution of social and economic networks, whose structure and composition has been shown to have important implications for economic performance, health outcomes, social mobility, the diffusion of information, criminal behavior and other socioeconomic outcomes. The spread of online social media has increased the availability of network data, however, the estimation and identification of structural models of network formation pose formidable econometric challenges (Jackson (2008), Sheng (2012), Graham (2014), Leung (2014), Menzel (2016), Chandrasekhar (2016), Mele (2017), dePaula (forthcoming)).

We develop a game theoretical equilibrium model of network formation, where individuals' make sequential decisions to form or delete ties, weighing benefits and costs of each link.[1] Players are *heterogeneous*: each individual is characterized by a set of characteristics, e.g. race, gender, income, etc. Players' payoffs depend on the number and composition of their links. In addition, players receive utility from the number of links formed by their direct

Date: First version: June 15, 2015. This version: September 20, 2017.
We are grateful to Aureo DePaula, Bryan Graham, Mert Gürbüzbalaban, Matt Jackson, Hiro Kaido, Michael Leung and Demian Pouzo for comments and suggestions. The second author is partially supported by NSF Grant DMS-1613164.

[1]For a survey of strategic models of network formation, see Jackson (2008).

neighbors. We can easily generalize this payoff function to include additional link externalities, like common friends, or cliques of four or more people: most of our results will hold without much modification in the proofs. While similar payoff functions have been used in the network economics literature (e.g. DePaula et al. (2014), Sheng (2012), Leung (2014), Chandrasekhar and Jackson (2014), Menzel (2016), Mele (2017)), in our model the network is formed sequentially: in each period two random players meet and receive a logistic matching shock. Conditional on the meeting and the value of the matching shock, they decide whether to form a link (or delete it) by maximizing the sum of their payoffs.[2] This network formation protocol is consistent with a standard pairwise stable equilibrium with transfers.[3]

We show that this network formation model is a potential game: all the incentives of the players can be summarized by an aggregate function of the state of the network, called a potential.[4] We prove that the sequence of networks generated through the random matching and the endogenous linking decisions converges to a unique stationary equilibrium. That is, in the long-run, the likelihood of observing a specific network configuration follows a discrete exponential distribution. As a consequence, our model converges to an Exponential Random Graph model (ERGM) (Wasserman and Pattison (1996), Wasserman and Faust (1994), Frank and Strauss (1986), Snijders (2002)). The latter is a random network formation model that assumes the probability of observing a specific network configuration is proportional to an exponential function of network statistics. Such model is a workhorse for empirical applications in statistics and social sciences, perhaps due to its extreme flexibility and the availability of estimation packages in open source software (Snijders (2002), Moody (2001), Goodreau et al. (2009), Wimmer and Lewis (2010)).

Our likelihood depends on an intractable normalizing constant that is impossible to compute exactly. The ERGM literature bypasses this problem and computes an approximate likelihood using Markov Chain Monte Carlo simulations.[5] However, it can be shown that for many models of interest, these simulations may converge very slowly, making estimation infeasible. Bhamidi et al. (2011) show that the standard Gibbs sampler used for simulation of ERGMs may converge exponentially slow.[6] Chatterjee and Diaconis (2013) and Mele (2017) show that this is the case even when the ERGM is asymptotically equivalent to a model with independent links.

To overcome these computational challenges, we develop alternative deterministic methods of approximations for the intractable likelihood of the ERGM. We focus on the case where the number of types (covariates) is finite, but is allowed to increase with the size of the

[2]See also Mele (2017), Badev (2013), Butts (2009), Christakis et al. (2010) for models of sequential network formation in the economics literature.

[3]See Jackson (2008) for several notions of equilibrium in the economic network literature.

[4]Monderer and Shapley (1996) define potential games and their equilibria. In the network literature, there is a relationship between potential games and network equilibria, as described in Butts (2009), Mele (2017), Chandrasekhar and Jackson (2014), and Badev (2013).

[5]See Snijders (2002), Wasserman and Pattison (1996), Caimo and Friel (2011), Mele (2017) for examples.

[6]See also Chatterjee and Diaconis (2013) and Mele (2017) for similar results.

network. The paper provides several contributions. First, we propose a variational mean-field approximation for the discrete exponential family to approximates the likelihood.[7] Our approximation is computationally feasible, scalable to large networks, and it is implemented as a standard iterative algorithm to facilitate estimation. In addition, we compute bounds to the error of approximation for a given size of the network, extending recent results in the literature on large deviations for random graphs (Chatterjee and Varadhan (2011), Chatterjee and Diaconis (2013), Lovasz (2012), Aristoff and Zhu (2014), Radin and Yin (2013)). Specifically our approach extends the nonlinear large deviations and graph limits results of Chatterjee and Dembo (2014) and Chatterjee and Diaconis (2013) to include covariates in the ERGM.

Second, we prove that the mean-field approximation is asymptotically exact, i.e. the approximation error approaches zero as the network grows large.[8]

Third, while the general variational approximation does not have a closed form solution, we characterize the mean-field result in two important special cases, to show how the technique is implemented. The first special case is when there is *extreme homophily*, i.e. the cost of forming links across groups is very high. We show that as the network grows large, the variational approximation for this model is equivalent to the solution of independent univariate maximization problems, one for each group. The solution is the fixed point of a logit equation, studied in Chatterjee and Diaconis (2013). The second special case, *two groups of equal size*, is more complicated. We prove that the variational problem is equivalent to the maximization of a function with two variables. We study this maximization problem using techniques developed first by Radin and Yin (2013) and Aristoff and Zhu (2014), showing that when the payoff from indirect links is relatively small (or negative), there is a unique solution. The solution provides the probabilities of links within-group and across groups. We also find a phase transition: when the difference in costs of forming links within-group and across groups is relatively small with respect to the payoff from indirect links, then there are two solutions: either the network is very sparse, or it is very dense. In this phase transition identification is problematic, because the same vector of parameters may generate two completely different network datasets.

Fourth, we show that as the number of players grows large, the networks generated by our model concentrate around the solution of the mean-field approximation. In other words, the networks generated according to the approximated likelihood are arbitrarily close to the networks generate by our model in a large economy. In the special case of extreme homophily, the networks generated by our model will resemble a block-diagonal network, while in the special case of two groups with equal size the networks are similar to a stochastic block model with two groups. The general solution is more complicated to characterize and we leave that to future research.

---

[7] Wainwright and Jordan (2008) provide a self-contained introduction to variational approximations and mean-field methods for estimation and approximation. See also Bishop (2006), Jaakkola (2000), Grimmer (2011).

[8] See Lovasz (2012), Chatterjee and Varadhan (2011) and Chatterjee and Diaconis (2013) for an overview of the graph limits and their use in the random graph literature. In appendix we also provide an alternative approximation based on the graph limit of the ERGM.

We conducted simple Monte Carlo experiments in small networks, to compare our approximation with the standard simulation-based Monte Carlo Maximum Likelihood estimator (MC-MLE) (Geyer and Thompson (1992)) and the simpler Maximum Pseudo-Likelihood estimator (MPLE) (Besag (1974)) estimators for ERGMs. Our approach provides robust estimates, while exhibiting a small bias, which is a well known property of the mean-field approximation. In addition, the complexity of our algorithm is quadratic in the number of players, while the Gibbs or Metropolis-Hastings samplers used in MC-MLE can converge in a number of steps that is exponential in the number of players squared (the number of links). Therefore, while for some parameter values the simulation is practically infeasible, our mean-field approximation is guaranteed to converge to a local maximum. To improve on the approximation we restart the iterative variational algorithm several times. This step can easily be parallelized, preserving the quadratic time convergence, while the MCMC is intrinsically a sequential algorithm and cannot exploit the parallel architecture of modern computers. Finally, while our model can generate dense networks, all the results hold if we allow some sparsity. Our technique is thus applicable to empirical applications where the network is moderately sparse.

We contribute to the literature in several ways. First, we provide a structural equilibrium model of network formation that allows estimation using only one (large) network. Similar models have appeared in the economics literature, e.g. DePaula et al. (2014), Sheng (2012), Leung (2014), Miyauchi (2012), Chandrasekhar and Jackson (2014), Menzel (2016), Currarini et al. (2009), Christakis et al. (2010), Mele (2017). We add to this literature by showing that our network formation model extends the potential game characterization of Mele (2017) to undirected networks and in the long run behaves like an exponential random graph model (Wasserman and Pattison (1996), Frank and Strauss (1986), Snijders (2002)). As a consequence, our characterization of the equilibrium networks provides a bridge between the economics/strategic and statistical/random network literatures.[9]

Second, we provide an example of use of variational methods of inference in the economics literature. Approximate variational inference has been used in several disciplines to provide alternative approximations when Monte Carlo methods are infeasible or computationally too burdensome. For example Braun and McAuliffe (2010) provides a mean-field approximation for standard multinomial choice models used in economics and marketing, and Grimmer (2011) show uses variational inference for a Bayesian logit model. However, these methods have not been widely adopted by economists, perhaps because of the bias introduced in the estimates.[10] We prove that for our structural model, the variational approximation is asymptotically exact, and we provide Monte Carlo evidence that the estimates are comparable to standard simulation-based inference.

Third, we characterize the asymptotic properties of our model and the inference method, extending results from the recent large deviations and graph limits literature. In particular, our bounds to the approximation errors extend the approach of Chatterjee and Dembo (2014) and Chatterjee and Diaconis (2013) to include covariates in the exponential random

[9]See also Mele (2017), Chandrasekhar and Jackson (2014), Badev (2013) for models that incorporate both random and strategic network features.

[10]A notable exception is Galichon et al. (2016) for models of discrete choice.

graph. While similar ideas have been used to provide approximation to the normalizing constant of the ERGM model (e.g. Amir et al. (2012) and He and Zheng (2013)), we are the first to provide a characterization of the variational problem for a model with covariates. In addition, we also provide a complete characterization of the variational solution for special cases, and discuss identification of the parameters.

2. Theoretical Model

2.1. **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. [11] 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.[12] The utility of player i is

$$u_i(g,\tau) = \sum_{j=1}^{n} \alpha_{ij} g_{ij} + \frac{\beta}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk}, \tag{2.1}$$

where $\alpha_{ij} := \alpha(\tau_i, \tau_j)$ are symmetric functions $\alpha : \otimes_{j=1}^{S}\mathcal{X}_j \times \otimes_{j=1}^{S}\mathcal{X}_j \to \mathbb{R}$ and $\alpha(\tau_i, \tau_j) = \alpha(\tau_j, \tau_i)$ 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 α 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.[13] 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 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.[14]

Example 2.1. *(Homophily) Consider a model with $\alpha_{ij} = V - c(\tau_i, \tau_j)$, 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*

$$c(\tau_i, \tau_j) = \begin{cases} c & \text{if } \tau_i = \tau_j, \\ C & \text{if } \tau_i \neq \tau_j. \end{cases} \tag{2.2}$$

[11] 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.

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

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

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

For example, consider the parameterization $0 < c < V < C$ *and* $\beta = 0$*. 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* $\beta > 0$*, also links across groups will be formed, as long as* β *is sufficiently large.*

Example 2.2. *(Social Distance Model) Let the payoff from direct links be a function of the social distance among the individuals. Formally, let* $\alpha(\tau_i, \tau_j) := \gamma d(\tau_i, \tau_j) - c$*, where* $d(\tau_i, \tau_j)$ *is a distance function,* γ *is a parameter that determines the sensitivity to the social distance and* $c > 0$ *is the cost of forming a link.*[15] *The case with* $\gamma < 0$ *represents a world where individuals prefer linking to similar agents and* $\gamma > 0$ *represents a world where individuals prefer linking with people at larger social distance. Note that even when* $\gamma < 0$*, if we have* $\beta > 0$ *sufficiently large, individuals may still have an incentive to form links with people at larger social distance.*

2.2. **Meetings and equilibrium.** The network formation process follows a stochastic best-response dynamics:[16] 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 2.1. *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*

$$\rho_{ij} := \rho(\tau_i, \tau_j, g_{-ij}) > 0 \tag{2.3}$$

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

Assumption 2.1 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 2.2. *Individuals receive a logistic shock before they decide whether to form a link (i.i.d. over time and players).*

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

[16]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 2.1. *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*

$$Q_n(g;\alpha,\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}g_{ij} + \frac{\beta}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} g_{ij}g_{jk}. \tag{2.4}$$

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.[17] The sequential network formation follows a *Glauber* dynamics, therefore converging to a unique stationary distribution.

THEOREM 2.1. *In the long run, the model converges to the stationary distribution π_n, defined as*

$$\pi_n(g;\alpha,\beta) = \frac{\exp\left[Q_n(g;\alpha,\beta)\right]}{\sum_{\omega\in\mathcal{G}}\exp\left[Q_n(\omega;\alpha,\beta)\right]} = \exp\left\{n^2\left[T_n(g;\alpha,\beta) - \psi_n(\alpha,\beta)\right]\right\}, \tag{2.5}$$

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

$$\psi_n(\alpha,\beta) = \frac{1}{n^2}\log\sum_{\omega\in\mathcal{G}}\exp\left[n^2 T_n(\omega;\alpha,\beta)\right], \tag{2.6}$$

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

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 (2.5) 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.

[17] 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.

2.3. **Extensions and discussion.** The preferences 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.

$$u_i(g,\tau) = \sum_{j=1}^{n} \alpha_{ij} g_{ij} + \frac{\beta}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk} + \frac{\gamma}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk} g_{ki}, \tag{2.7}$$

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

$$Q_n(g;\alpha,\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij} g_{ij} + \frac{\beta}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk} + \frac{\gamma}{6n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk} g_{ki}. \tag{2.8}$$

In general, all the results that we show below extend to more general utility functions that include payoffs for link externalities similar to (2.7).

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 2.1, 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 (2.5) 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.

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

$$u_i(g,\tau) = \sum_{j=1}^{n} \alpha_{ij}^{(n)} g_{ij} + \frac{\beta^{(n)}}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk}, \tag{2.9}$$

where $|\alpha_{ij}^{(n)}|$ and $|\beta^{(n)}|$ can have moderate growth in n instead of being bounded. We will give more details later in our paper. [18]

3. Variational Approximations

The constant $\psi_n(\alpha,\beta)$ in (2.6) is intractable because it involves a sum over all $2^{\binom{n}{2}}$ possible networks with n players. To be concrete, if the network has $n = 10$ nodes, the computation of $\psi_n(\alpha,\beta)$ involves the computation of 2^{45} potential functions, which is clearly an infeasible strategy.[19]

[18] See Chatterjee and Dembo (2014) for additional applications of nonlinear large deviations.

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

The usual strategy in the literature consists of approximating the constant using an MCMC algorithm (Snijders (2002)). At each iteration of the sampler, 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.[20]

We propose alternative methods that do not rely on simulations. Our first 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:

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

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

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

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

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.

$$\psi_n(\alpha,\beta) = \sup_{q_n\in\mathcal{Q}_n}\mathcal{L}(q_n) = \sup_{q_n\in\mathcal{Q}_n}\left\{\mathbb{E}_{q_n}\left[T_n(\omega;\alpha,\beta)\right] + \frac{1}{n^2}\mathcal{H}(q_n)\right\}, \tag{3.1}$$

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

Unfortunately, in most cases the variational problem (3.1) has no closed-form solution. The machine learning literature suggests to restrict the set $\mathcal{Q}_n$ to find a tractable approximation.[21] A popular choice for the set $\mathcal{Q}_n$ is the set of all completely factorized distribution

$$q_n(g) = \prod_{i,j}\mu_{ij}^{g_{ij}}(1-\mu_{ij})^{1-g_{ij}}, \tag{3.2}$$

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 in each link's entropy

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

[20] A sampler is defined *local* if it proposes to modify $o(n)$ links per iteration.

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

and the expected potential can be 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 of $\psi_n(\alpha,\beta)$, because we have restricted the set of probability distributions of the maximization. The variational problem involves finding a matrix $\boldsymbol{\mu}(\alpha,\beta)$

$$\begin{aligned}\psi_n(\alpha,\beta) &\geq \psi_n^{MF}(\boldsymbol{\mu}(\alpha,\beta)) \\ &= \sup_{\boldsymbol{\mu}\in[0,1]^{n^2}}\left\{\frac{\sum_i\sum_j \alpha_{ij}\mu_{ij}}{n^2} + \beta\frac{\sum_i\sum_j\sum_k \mu_{ij}\mu_{jk}}{2n^3}\right. \\ &\qquad \left. - \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n\left[\mu_{ij}\log\mu_{ij} + (1-\mu_{ij})\log(1-\mu_{ij})\right]\right\}. \end{aligned} \tag{3.3}$$

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.[22]

4. Asymptotic Results, Large Deviations and Graph Limits

In this section we consider the model as $n\to\infty$. We use and extend results from the graph limits literature,[23] large deviations literature for random graphs[24] and analysis of the resulting variational problem.[25] 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 *graph**on* and it is a representation of an infinite network.

We also 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 4.1. *Assume that*

$$\alpha_{ij} = \alpha\left(i/n, j/n\right), \tag{4.1}$$

where $\alpha(x,y):[0,1]^2\to\mathbb{R}$, are deterministic exogenous functions that are symmetric, i.e., $\alpha(x,y)=\alpha(y,x)$. [26]

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 [27].

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

[23]See Lovasz (2012), Borgs et al. (2008)

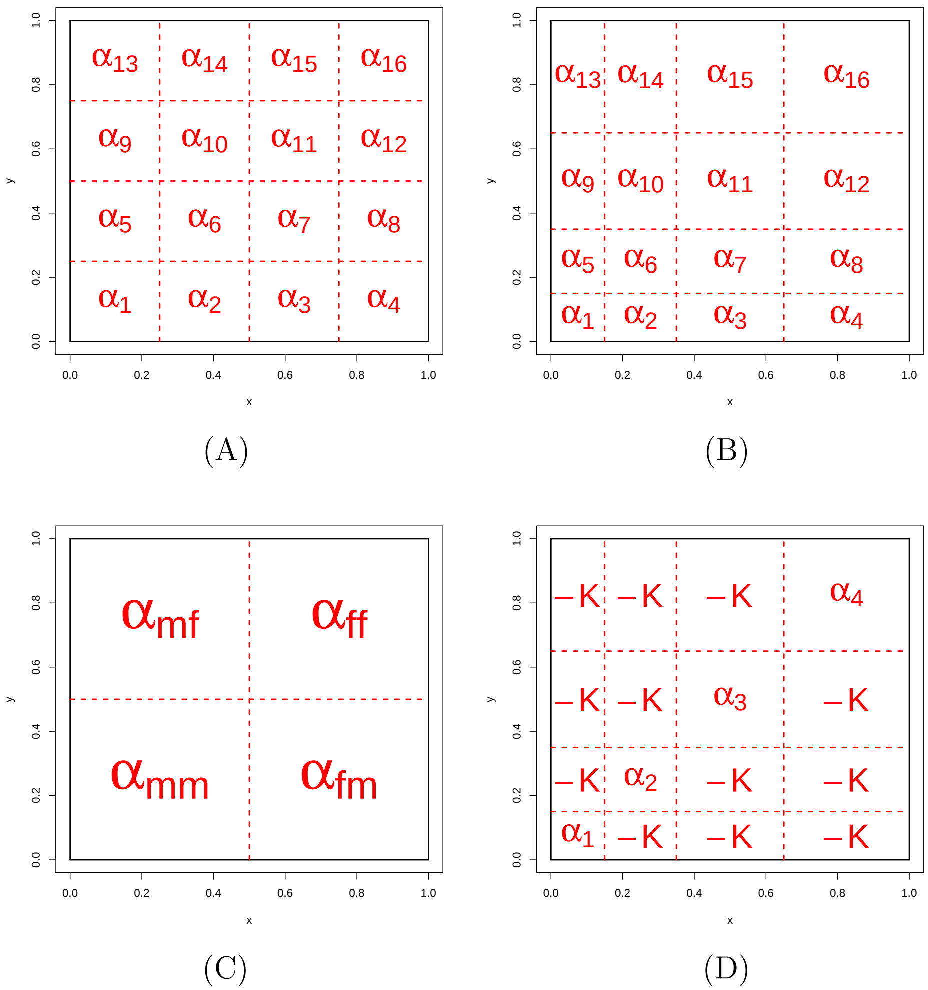
[24]See Chatterjee and Varadhan (2011), Chatterjee and Diaconis (2013)

[25]See Aristoff and Zhu (2014), Radin and Yin (2013) among others.

[26]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 $\mathbb{R}$.

[27]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 4.1. Examples of function $\alpha(x, y)$.



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.

We assume that $\alpha(x, y)$ is unifomly bounded in x and y.

ASSUMPTION 4.2. *Assume that*

$$\sup_{(x,y)\in[0,1]^2} |\alpha(x, y)| < \infty. \tag{4.2}$$

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, $i = \frac{n}{2} + 1, \frac{n}{2} + 2, \ldots, n$.[28] That means, $\alpha(x, y)$ takes three values

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

according to the three regions:

$$\left\{(x,y): 0<x,y<\tfrac{1}{2}\right\}, \tag{4.3}$$

$$\left\{(x,y): \tfrac{1}{2}<x,y<1\right\}, \tag{4.4}$$

$$\left\{(x,y): 0<x<\tfrac{1}{2}<y<1\right\} \bigcup \left\{(x,y): 0<y<\tfrac{1}{2}<x<1\right\}, \tag{4.5}$$

and these three regions correspond precisely to pairs: male-male, female-female, and male-female. This example is represented in Figure 4.1(C).

The work of Chatterjee and Diaconis (2013) show that the variational problem in (3.1) translates into an analogous variational problem for the graph limit.[29] For our model we can show that the variational problem for the graphon is

$$\begin{aligned}\psi(\alpha,\beta) = \sup_{h\in\mathcal{W}} \Bigg\{ \int_0^1\int_0^1 \alpha(x,y)h(x,y)dxdy + \frac{\beta}{2}\int_0^1\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz \\ - \frac{1}{2}\int_0^1\int_0^1 I(h(x,y))dxdy \Bigg\},\end{aligned} \tag{4.6}$$

where $\mathcal{W} := \{h : [0,1]^2 \to [0,1], h(x,y) = h(y,x), 0 \le x,y \le 1\}$, and we define the entropy function:

$$I(x) := x\log x + (1-x)\log(1-x), \qquad 0 \le x \le 1, \tag{4.7}$$

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

In essence the first line in (4.6) corresponds to the expected potential function in the continuum, while the second line in (4.6) corresponds to the entropy of the graphon $h(x,y)$.

4.1. 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 4.1. *Under Assumption 4.2 and for fixed network size n, the approximation error of the variational mean-field problem is bounded as*

$$C_3(\beta)n^{-1} \le \psi_n(\alpha,\beta) - \psi_n^{MF}(\boldsymbol{\mu}(\alpha,\beta)) \le C_1(\alpha,\beta)n^{-1/5}(\log n)^{1/5} + C_2(\alpha,\beta)n^{-1/2}, \tag{4.8}$$

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:

$$\begin{aligned} C_1(\alpha,\beta) &:= c_1\cdot(\|\alpha\|_\infty + |\beta|^4 + 1), \\ C_2(\alpha,\beta) &:= c_2\cdot(\|\alpha\|_\infty + |\beta| + 1)^{1/2}\cdot(1+|\beta|^2)^{1/2}, \\ C_3(\beta) &:= |\beta|, \end{aligned}$$

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

Proof. See Appendix. ∎

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

The constants in Theorem 4.1 are functions of the parameters α and β. 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. 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.

Remark 4.1. *It follows from the estimates in Theorem 4.1 that we can allow moderate sparsity in our model, in the sense that $|\alpha_{ij}|$ and $|\beta|$ can have moderate growth in n instead of being bounded, and the difference of ψ_n and ψ_n^{MF} goes to zero if $C_1(\alpha,\beta)$ grows slower than $n^{1/5}/(\log n)^{1/5}$ and $C_2(\alpha,\beta)$ grows slower than $n^{1/2}$ as $n \to \infty$. For example, if $|||\alpha|||_\infty = O(n^{\delta_1})$, $|\beta| = O(n^{\delta_2})$, where $\delta_1 < \frac{1}{5}$ and $\delta_2 < \frac{1}{20}$, 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. $|\beta| = \Omega(n)$, then ψ_n cannot be approximated by ψ_n^{MF}.*

We are interested in estimating the model in large networks. The result in Theorem 4.1 shows that the solution to the variational mean-field problem becomes exact as $n \to \infty$.

In addition, Chatterjee and Diaconis (2013) show that as $n \to \infty$ the log-constant of the ERGM converges to the solution of the variational problem (4.6), that is

$$\psi_n(\alpha,\beta) \to \psi(\alpha,\beta). \tag{4.9}$$

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

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

$$\psi_n^{MF}(\boldsymbol{\mu}(\alpha,\beta)) \to \psi(\alpha,\beta). \tag{4.10}$$

Proof. It follows directly from Theorem 4.1 and (4.9). ☐

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 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. **Approximation of the graph limit.** We can also exploit the graph limit result and provide an approximation of the log-constant. The variational formula for $\psi(\alpha,\beta)$ 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). For details, see Proposition A.1 in the Appendix.

The lower-bound in Proposition A.1 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).

5. Special cases

The general solution of the variational problem (4.6) 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.

5.1. **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.

PROPOSITION 5.1. *Let $0 = a_0 < a_1 < \cdots < a_M = 1$ be a given sequence. Assume that*

$$\alpha(x, y) = \alpha_{mm}, \qquad \text{if } a_{m-1} < x, y < a_m, \qquad m = 1, 2, \ldots, M. \tag{5.1}$$

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

$$\psi(\alpha, \beta; -\infty) = \sum_{m=1}^{M} (a_m - a_{m-1})^2 \sup_{0\leq x\leq 1} \left\{ \alpha_{mm} x + \frac{\beta}{2} x^2 - \frac{1}{2} I(x) \right\}. \tag{5.2}$$

Proof. See Appendix. □

The net benefit function $\alpha(x, y)$ assumed in the Proposition is shown in Figure 4.1(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 (2014) and Mele (2017). It can be shown that the solutions μ_m^*, where $m = 1, .., M$, are the fixed point of equations

$$\mu_m = \frac{\exp\left[\alpha_{mm} + \beta\mu_m\right]}{1 + \exp\left[\alpha_{mm} + \beta\mu_m\right]}, \tag{5.3}$$

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 (2014).

Chatterjee and Diaconis (2013) show that the network of each group corresponds to an Erdős-Rényi graph with probability of a link equal to μ_m^*.

5.2. **Two groups of equal size.** Consider a model with only two types, e.g. male and females, each group of measure $\frac{1}{2}$, as in Figure 4.1(C). We assume that players have the same preferences as in Example 2.1.There is a cost $c > 0$ to form a link among players of the same group, and a cost $C > c$ for a link among players of different type. In terms of Figure 4.1(C) we are assuming $\alpha_{mm} = \alpha_{ff}$ and $\alpha_{mf} = \alpha_{fm}$. In such model, the variational problem becomes a maximization in two variables, as shown in the next proposition.

PROPOSITION 5.2. *Let us assume that $\alpha(x,y)$ takes two values:*

$$\alpha(x,y) = \begin{cases} \alpha_1 = V - c & \text{if } 0 < x,y < \frac{1}{2} \text{ or } \frac{1}{2} < x,y < 1, \\ \alpha_2 = V - C & \text{if } 0 < x < \frac{1}{2} < y < 1 \text{ or } 0 < y < \frac{1}{2} < x < 1. \end{cases} \tag{5.4}$$

and let the function $F(u,v)$ be defined as

$$F(u,v) := \frac{\alpha_1}{2}u - \frac{1}{4}I(u) + \frac{\alpha_2}{2}v - \frac{1}{4}I(v) + \frac{\beta}{8}(u+v)^2 . \tag{5.5}$$

Then, the asymptotic normalizing constant is the solution of the following maximization problem

$$\psi(\alpha,\beta) = \sup_{0\le u,v\le 1} F(u,v). \tag{5.6}$$

Proof. See Appendix. □

The solution for the maximization (5.6) is much simpler than solving the general variational problem because it reduces an infinite-dimensional optimization problem to a two-dimensional one. In this special case, we are able to provide a more precise characterization of the maxima. Indeed, the analysis of (5.6) allows us to recover the graphon for this specification of the model.

PROPOSITION 5.3. *The specification of the model in Proposition 5.2 has the following properties*

1. *The solution (u^*,v^*) satisfies:*
$$u^* = \frac{e^{2\alpha_1+\beta(u^*+v^*)}}{1+e^{2\alpha_1+\beta(u^*+v^*)}}, \qquad v^* = \frac{e^{2\alpha_2+\beta(u^*+v^*)}}{1+e^{2\alpha_2+\beta(u^*+v^*)}}.$$
2. *If $\beta \le 2$, the maximization problem (5.6) has a unique solution (u^*,v^*). In addition, if $\alpha_1+\alpha_2+\beta = 0$, then the unique solution is given by $(u^*,v^*) = \left(\frac{e^{2\alpha_1+\beta}}{1+e^{2\alpha_1+\beta}}, \frac{e^{2\alpha_2+\beta}}{1+e^{2\alpha_2+\beta}}\right)$.*
3. *If $\alpha_1+\alpha_2+\beta = 0$ and $\beta > \frac{\left(1+e^{\alpha_1-\alpha_2}\right)^2}{2e^{\alpha_1-\alpha_2}}$ then the maximization problem (5.6) has two solutions (u^*,v^*) and $(1-v^*,1-u^*)$ with $F(u^*,v^*) = F(1-v^*,1-u^*)$.*

Proof. See Appendix. □

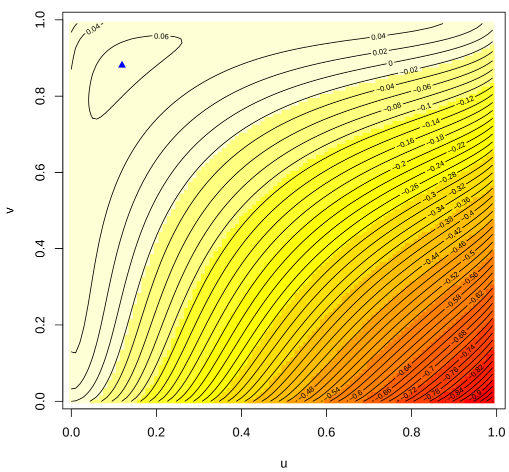
The proposition shows that the model has a simple solution, given by the solution of a system of two equations.

We also show that there exists a phase transition, when β is sufficiently large with respect to the difference $(\alpha_1-\alpha_2)$ and the parameters are on the plane $\alpha_1+\alpha_2+\beta = 0$. In such configuration of the parameters, there are two global maxima of $F(u,v)$ and the same set of parameters can generate either a very sparse or a very dense network. This points to an identification problem as also shown in Chatterjee and Diaconis (2013) and Mele (2017).

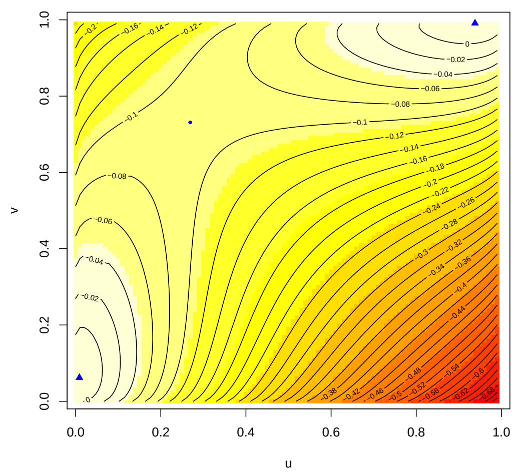
In essence this model (for large n) converges to a stochastic block model with probability of links within the same group equal to u^* and probability of links across groups equal to v^*.

To illustrate the results in Proposition 5.3 we show some examples in Figures 5.1 and

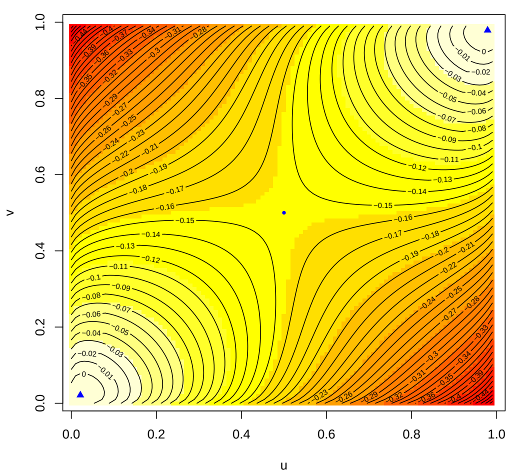
FIGURE 5.1. Examples of maxima characterized in Proposition 5.3 with $\beta = 4$



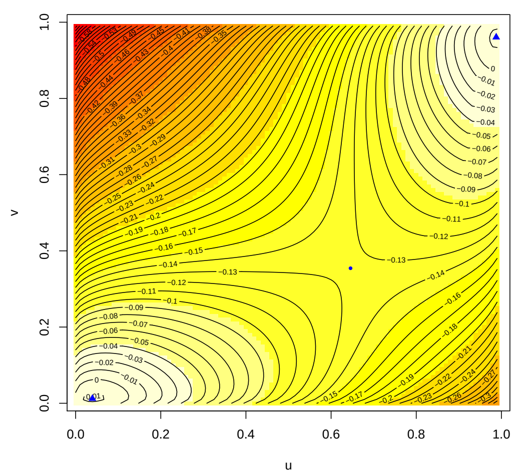
(A) $(\alpha_1, \alpha_2) = (-3, -1)$



(B) $(\alpha_1, \alpha_2) = (-2.5, -1.5)$



(C) $(\alpha_1, \alpha_2) = (-2, -2)$

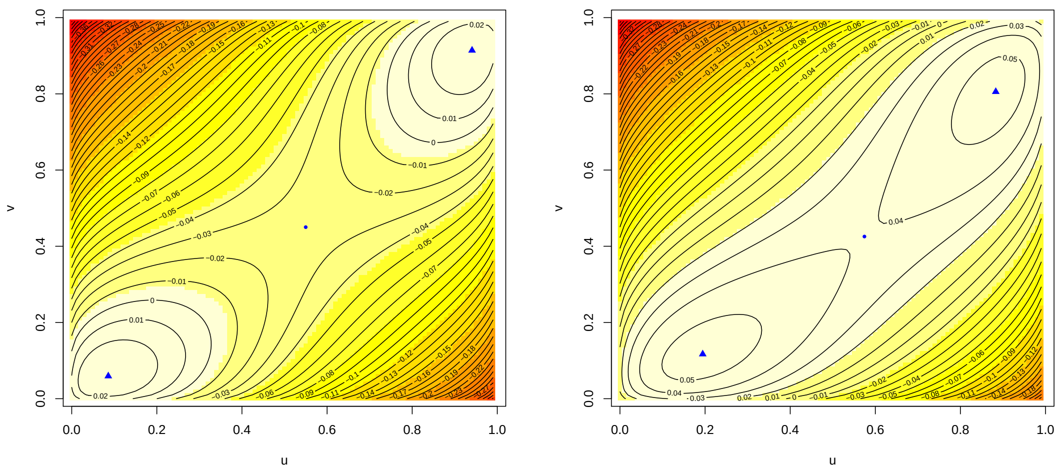


(D)$(\alpha_1, \alpha_2) = (-1.7, -2.3)$

The figures show the level curves of $F(u, v)$ for different vectors of parameters. In all the pictures $\beta = 4$. The global maxima are represented as blue triangles.

5.2. In Figure 5.1 we fix $\beta = 4$ and change the values of α_1 and α_2. When the difference $\alpha_1 - \alpha_2$ is relatively large, see Figure 5.1(A), there is a unique maximizer $(u^*, v^*) = (0.1192029, 0.8807971)$ indicated as the blue triangle. When we decrease the distance between α_1 and α_2 as in Figure 5.1(B), we obtain two global maximizer of the function $F(u, v)$: $(u^*, v^*) = (0.0088671, 0.0620063)$ and $(1 - v^*, 1 - u^*) = (0.9379937, 0.9911329)$. Notice that the first solution provides a very sparse network while the second global maximum is a dense network. This property of the model is also related to convergence problems of the simulation algorithms for ERGMs, as shown in Bhamidi et al. (2011) and Mele (2017). Not

FIGURE 5.2. More examples of maxima characterized in Proposition 5.3



(A) $(\alpha_1, \alpha_2, \beta) = (-1.4, -1.6, 3)$ (B) $(\alpha_1, \alpha_2, \beta) = (-1.1, -1.4, 2.5)$

The figures show the level curves of $F(u, v)$ for different vectors of parameters. The global maxima are represented as blue triangles.

surprisingly, when $\alpha_1 = \alpha_2$ we have a perfectly symmetric pair of solutions in Figure 5.1(C): $(u^*, v^*) = (0.0212494, 0.0212494)$ and $(1-v^*, 1-u^*) = (0.9787506, 0.9787506)$. Figure 5.1(D) provides an additional example with $|\alpha_1| < |\alpha_2|$: we show in the Appendix that what matters for the shape of the level curves is the absolute difference $|\alpha_1 - \alpha_2|$ and not the actual values of these parameters.

In Figure 5.2 we show how a smaller β implies that the two solutions (u^*, v^*) and $(1 - v^*, 1 - u^*)$ are less extreme and closer to each other. Indeed in Figure 5.2(A) we have $(u^*, v^*) = (0.0859306, 0.0592804)$ and $(1 - v^*, 1 - u^*) = (0.9407196, 0.9140694)$ and in Figure 5.2(B) it is $(u^*, v^*) = (0.1944159, 0.116957)$ and $(1 - v^*, 1 - u^*) = (0.883043, 0.8055841)$. As β increases, the maximizers are very close to zero and/or one.

6. A CONCENTRATION RESULT

The special cases in the previous section provide a simple characterization of the variational problem solution. However, the general variational problem does not have a closed form solution. Nonetheless, we can gain some additional insights by considering the behavior of the networks generated by our model.

In particular we want to characterize the probability that the network generated by the model belongs to the set of graphons that solve the variational problem (3.1).

The probability measure $\mathbb{P}_n$ of observing the network configuration g is given by

$$\mathbb{P}_n(g) = \frac{1}{Z_n} \exp\left\{ \sum_{i,j} \alpha_{ij} g_{ij} + \frac{\beta}{2n} \sum_{i,j,k} g_{ij} g_{jk} \right\}, \tag{6.1}$$

where $Z_n = e^{n^2 \psi_n(\alpha,\beta)}$ is the normalizer.

For any $\epsilon > 0$, we can define the set $\mathcal{A}_\epsilon$

(6.2)

$$\mathcal{A}_\epsilon := \Big\{ h \in \mathcal{W} : \int_0^1 \int_0^1 \alpha(x,y) h(x,y) dxdy + \frac{\beta}{2} \int_0^1 \int_0^1 \int_0^1 h(x,y) h(y,z) dxdydz$$
$$- \frac{1}{2} \int_0^1 \int_0^1 \left[h(x,y) \log h(x,y) + (1 - h(x,y)) \log(1 - h(x,y)) \right] dxdy \leq \psi(\alpha,\beta) - \epsilon \Big\}.$$

where $\psi(\alpha,\beta)$ is the solution of variational problem (3.1). The following proposition shows that the probability that the network g belongs to the set $\mathcal{A}_\epsilon$ under the probability measure $\mathbb{P}_n$ is vanishingly small as n grows large.

PROPOSITION 6.1. *There exists some $0 < \gamma < \epsilon$, so that for any sufficiently large n,*

$$\mathbb{P}_n(g \in \mathcal{A}_\epsilon) \leq e^{-\gamma n^2}. \tag{6.3}$$

Proof. See Appendix. □

In other words, the model generates networks that asymptotically converge to the solutions of the variational problem (3.1).

In the special case of two groups of equal size analyzed in the previous section, this suggests that our model converges to a stochastic block model in the large n limit. In the special case of extreme homophily, our model converges to a block-diagonal model.

7. Estimation Experiments in finite networks

We have performed simple Monte Carlo experiments to study the performance of our asymptotic approximation in finite networks. We compare the mean-field approximation with the standard simulation-based MCMC-MLE (Geyer and Thompson (1992), Snijders (2002)), and the Maximum Pseudo-Likelihood estimator (Besag (1974)).

We implemented our variational approximation in the R package `mfergm`, available in Github.[30] We follow the machine learning literature and use an iterative algorithm that is guaranteed to converge to a local maximum of the mean-field problem.[31] Indeed, taking the first order conditions of the mean-field problem 3.3 with respect to each μ_{ij} we obtain

$$\mu_{ij} = \frac{\exp\left[2\alpha_{ij} + \frac{\beta}{n}\sum_{k=1}^n (\mu_{jk} + \mu_{ki})\right]}{1 + \exp\left[2\alpha_{ij} + \frac{\beta}{n}\sum_{k=1}^n (\mu_{jk} + \mu_{ki})\right]}, \quad i,j = 1, ..., n. \tag{7.1}$$

The algorithm starts from an initialized matrix $\boldsymbol{\mu}^{(\mathbf{0})}$ and iteratively applies the update (7.1) to each entry of the matrix. After updating all entries, the objective function is re-evaluated.

[30] See https://github.com/meleangelo/mfergm

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

Since the problem is concave in each μ_{ij}, this iterative method is guaranteed to find a local maximum of (3.3).

We re-initialize the algorithm several times to get a better approximation: in the Monte Carlo exercise below we restarted the approximation 5 times.[32] Notice that this step is easily parallelizable, while the standard MCMC method for estimation of ERGMs is an intrinsically sequential algorithm.

We test our approximation technique using artificial network data. Each network is generated using a 10 million run of the Metropolis-Hastings sampler implemented in the `ergm` command in R.

We report results for networks with 50, 100 and 200 nodes. The results are summarized by the median and several percentiles of the estimated parameters in the 1000 simulations.[33]

In Table 7.1 we generate data from the vector of parameters $(\alpha_1, \alpha_2, \beta) = (-2, 1, 2)$. While all three methods seem to work well, the Mean-Field approximation gives more robust estimates. At the same time, as it is well known, the mean-field can be biased. The MCMC-MLE and MPLE estimates are very similar. One possible reason is that the MCMC-MLE default starting value for the simulations is the MPLE estimate.

Similar results are shown in Table 7.2. The true parameter vector in this table is $(\alpha_1, \alpha_2, \beta) = (-2, 1, 3)$. The mean field approximation is more robust, giving a smaller range of estimates. However, it is clear that it is affected by some bias.[34]

The computational complexity of the mean-field iterative algorithm is of order n^2, while it is well known that the simulation methods used in the MCMC-MLE may have complexity of order e^{n^2} for some parameter vector (Bhamidi et al. (2011), Chatterjee and Diaconis (2013), Mele (2017)).

8. Conclusions

In this paper we have developed a model of strategic network formation with heterogeneous players and we have shown that our model is a potential game. In each period, a pair of players are randomly matched and they decide whether to form or delete a link. We show that this dynamic (Markovian) model converges in the long-run to a stationary distribution that coincides with an exponential random graph (ERGM). As a consequence it inherits all the challenges of estimation of the ERGMs.

In particular, the likelihood of the model is proportional to a normalizing constant that is infeasible to compute exactly. The standard estimation strategy in the ERGM literature bypasses the evaluation of the normalizing constant, and provides an approximated likelihood using Markov Chain Monte Carlo simulations. However, such algorithms have convergence

[32] This fixed number of restarts could be suboptimal. It seems reasonable to increase the number of restarts as the network grows larger.

[33] In few cases the optimization routine in R did not converge: we do not include those results in the tables.

[34] One practical improvement with respect to the bias problem consists of increasing the number of restarts: this would provide a better approximation at every likelihood evaluation. A more brute-force approach is to use a grid search algorithm instead of a standard optimization routine. Finally, as shown in the Figures above, in many cases the objective function is relatively flat, and standard gradient-based methods may be unable to compute numerical derivatives in some cases.

TABLE 7.1. Monte Carlo estimates, comparison of three methods. True parameter vector is $(\alpha_1, \alpha_2, \beta) = (-2, 1, 2)$

$n = 50$	MCMC-MLE			MEAN-FIELD			MPLE		
	α_1	α_2	β	α_1	α_2	β	α_1	α_2	β
median	-1.960	1.005	1.569	-1.992	1.000	2.005	-1.956	0.999	1.527
5th pctile	-2.496	0.697	-2.320	-2.056	0.891	1.968	-2.320	0.796	-3.458
25th pctile	-2.141	0.897	0.023	-2.009	0.981	1.997	-2.108	0.912	-0.040
75th pctile	-1.788	1.103	3.142	-1.965	1.020	2.024	-1.798	1.100	2.887
95th pctile	-1.500	1.294	6.227	-1.900	1.066	2.101	-1.513	1.236	4.427
$n = 100$	MCMC-MLE			MEAN-FIELD			MPLE		
	α_1	α_2	β	α_1	α_2	β	α_1	α_2	β
median	-1.985	1.001	1.892	-1.983	1.008	2.015	-1.979	0.998	1.808
0.05	-2.306	0.817	-0.208	-2.021	0.931	1.997	-2.152	0.896	0.153
0.25	-2.099	0.936	1.071	-1.997	0.998	2.004	-2.058	0.955	1.174
0.75	-1.882	1.067	2.741	-1.967	1.021	2.044	-1.909	1.049	2.417
0.95	-1.725	1.186	4.498	-1.918	1.046	2.237	-1.788	1.123	3.125
$n = 200$	MCMC-MLE			MEAN-FIELD			MPLE		
	α_1	α_2	β	α_1	α_2	β	α_1	α_2	β
median	-1.997	1.007	1.988	-1.981	1.002	2.045	-1.981	1.001	1.815
0.05	-2.573	0.834	-2.816	-2.016	0.920	2.003	-2.131	0.945	0.061
0.25	-2.181	0.950	0.409	-1.990	0.977	2.011	-2.044	0.979	1.183
0.75	-1.839	1.062	3.732	-1.963	1.011	2.140	-1.916	1.023	2.413
0.95	-1.555	1.177	7.729	-1.920	1.026	2.583	-1.812	1.057	3.270

Results of 1000 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.

problems and may converge exponentially slow (Bhamidi et al. (2011), Chatterjee and Diaconis (2013),Mele (2017)).

We provide alternative approximations that rely on a variational representation of the ERGM normalization constant. When the types of players are finite, we show that we can approximate the variational problem with a mean-field algorithm. We compute exact bounds for the approximation error and prove that our mean-field approximation is asymptotically exact. Therefore, our method delivers the exact value of the log-likelihood as the number of players grows large. We also provide additional approximations that make use of the graph limits for our ERGM stationary model.

TABLE 7.2. Monte Carlo estimates, comparison of three methods. True parameter vector is $(\alpha_1, \alpha_2, \beta) = (-2, 1, 3)$

$n = 50$	MCMC-MLE			MEAN-FIELD			MPLE		
	α_1	α_2	β	α_1	α_2	β	α_1	α_2	β
median	-1.954	1.006	2.601	-1.941	1.031	3.032	-1.961	1.004	2.635
0.05	-2.375	0.714	0.447	-2.024	0.878	2.947	-2.238	0.829	-0.125
0.25	-2.104	0.906	1.787	-1.971	0.996	3.001	-2.090	0.921	1.647
0.75	-1.821	1.098	3.504	-1.904	1.065	3.097	-1.808	1.087	3.440
0.95	-1.553	1.253	5.324	-1.803	1.132	3.381	-1.577	1.219	4.351

$n = 100$	MCMC-MLE			MEAN-FIELD			MPLE		
	α_1	α_2	β	α_1	α_2	β	α_1	α_2	β
median	-2.022	0.987	3.116	-1.849	1.081	3.387	-1.997	1.001	2.980
0.05	-2.499	0.681	1.578	-2.388	0.887	3.012	-2.100	0.920	2.537
0.25	-2.156	0.904	2.644	-1.896	0.998	3.113	-2.040	0.966	2.817
0.75	-1.911	1.072	3.786	-1.785	1.129	3.699	-1.958	1.038	3.144
0.95	-1.649	1.251	5.667	-1.676	2.159	4.143	-1.891	1.094	3.358

$n = 200$	MCMC-MLE			MEAN-FIELD			MPLE		
	α_1	α_2	β	α_1	α_2	β	α_1	α_2	β
median	-2.018	1.000	3.090	-1.883	0.967	3.493	-1.988	1.000	2.916
0.05	-2.902	0.760	-2.010	-2.009	0.819	3.022	-2.092	0.955	1.994
0.25	-2.279	0.925	1.591	-1.925	0.912	3.185	-2.030	0.982	2.585
0.75	-1.784	1.069	4.852	-1.839	1.024	3.933	-1.941	1.017	3.239
0.95	-1.281	1.237	9.355	-1.760	1.078	4.278	-1.859	1.045	3.609

Results of 1000 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.

We characterize the mean-field approximation for several special cases. First, when there is extreme homophily, i.e. if the cost of linking across groups is extremely high, then we show that we can approximate the normalizing constant of the ERGM with the sum of independent maximization problems solutions, one for each type. Second, if we have only two groups of equal size, we show that the model exhibits a phase transition. If the net benefits of linking between groups and across groups are not too different, then the model may generate either very sparse networks or very dense networks.

Finally we show that the networks generated by our model concentrate around the solution of the mean-field approximation. This means that in the special cases of extreme homophily and two types of equal size, our model converges to a stochastic block model. We are not able to provide a complete characterization of the general case.

We perform a simple Monte Carlo exercise to compare our approximation and the standard estimation methods for ERGMs in finite networks. We show that our method provides reliable estimates and it is more robust than the other methods, while exhibiting some bias.

References

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 (2014), On the phase transition curve in a directed exponential random graph model. Working Paper.

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

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.

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.

Braun, Michael and Jon McAuliffe (2010), 'Variational inference for large-scale models of discrete choice', *Journal of the American Statistical Association* **105**(489).

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 (2014), Nonlinear large deviations. Working Paper.

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. Homomorphisms and Limits.
URL: *http://www.sciencedirect.com/science/article/pii/S0195669811000655*

Christakis, Nicholas, James Fowler, Guido W. Imbens and Karthik Kalyanaraman (2010), An empirical model for strategic network formation. Harvard University.

Currarini, Sergio, Matthew O. Jackson and Paolo Pin (2009), 'An economic model of friendship: Homophily, minorities, and segregation', *Econometrica* **77**(4), 1003–1045.

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.

DePaula, Aureo, Seth Richards-Shubik and Elie Tamer (2014), Identification of preferences in network formation games. working paper.

Frank, Ove and David Strauss (1986), 'Markov graphs', *Journal of the American Statistical Association* **81**, 832–842.

Galichon, Alfred, Khai Chiong and Matthew Shum (2016), 'Duality in dynamic discrete choice models', *Quantitative Economics* **7**(1), 83–115.

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.

Grimmer, Justin (2011), 'An introduction to bayesian inference via variational approximations', *Political Analysis* **19**(1), 32–47.

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.

Jaakkola, Tommi (2000), *Advanced mean field methods: theory and practice*, MIT PRESS, chapter Tutorial on Variational Approximation Methods.

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

Leung, Michael (2014), Two-step estimation of network-formation models with incomplete information. working paper.

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

Mele, Angelo (2017), A structural model of dense network formation. forthcoming, Econometrica.

Menzel, Konrad (2016), Strategic network formation with many agents. Working Paper.

Miyauchi, Yuhei (2012), Structural estimation of a pairwise stable network formation with nonnegative externality.

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.

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 Katherine Faust (1994), *Social Network Analysis: Methods and Applications*, Cambridge University Press.

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* .

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: Proofs

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

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

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

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

and therefore,

$$\psi(\alpha, 0) = \frac{1}{2}\iint_{[0,1]^2} \log(1+e^{2\alpha(x,y)})dxdy. \tag{A.3}$$

A.1. **Proof of Theorem 4.1.** 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

$$f_i := \frac{\partial f}{\partial x_i}, \qquad f_{ij} := \frac{\partial^2 f}{\partial x_i \partial x_j}, \tag{A.4}$$

and let

$$a := \|f\|, \qquad b_i := \|f_i\|, \qquad c_{ij} := \|f_{ij}\|. \tag{A.5}$$

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

$$\sum_{i=1}^N (f_i(x) - d_i)^2 \le N\epsilon^2. \tag{A.6}$$

Let us define

$$F := \log \sum_{x\in\{0,1\}^N} e^{f(x)}, \tag{A.7}$$

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

$$I(x) := \sum_{i=1}^N [x_i \log x_i + (1-x_i)\log(1-x_i)]. \tag{A.8}$$

In the proof we extend Theorem 1.5 in Chatterjee and Dembo (2014) that we reproduce in Theorem A.1 to help the reader:

THEOREM A.1 (Chatterjee and Dembo (2014))**.** *For any $\epsilon > 0$,*

$$\sup_{x\in[0,1]^N}\{f(x)-I(x)\}-\frac{1}{2}\sum_{i=1}^{N}c_{ii}\le F\le \sup_{x\in[0,1]^N}\{f(x)-I(x)\}+\mathcal{E}_1+\mathcal{E}_2, \tag{A.9}$$

where

$$\mathcal{E}_1 := \frac{1}{4}\left(N\sum_{i=1}^{N}b_i^2\right)^{1/2}\epsilon+3N\epsilon+\log|\mathcal{D}(\epsilon)|, \tag{A.10}$$

and

$$\begin{aligned}\mathcal{E}_2 \;&:= 4\left(\textstyle\sum_{i=1}^{N}(ac_{ii}+b_i^2)+\frac{1}{4}\sum_{i,j=1}^{N}(ac_{ij}^2+b_ib_jc_{ij}+4b_ic_{ij})\right)^{1/2}\\ &\quad+\tfrac{1}{4}\left(\textstyle\sum_{i=1}^{N}b_i^2\right)^{1/2}\left(\textstyle\sum_{i=1}^{N}c_{ii}^2\right)^{1/2}+3\textstyle\sum_{i=1}^{N}c_{ii}+\log 2.\end{aligned} \tag{A.11}$$

We will use the Theorem A.1 to derive the lower and upper bound 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

$$Z_n := \sum_{x_{ij}\in\{0,1\},x_{ij}=x_{ji},1\le i<j\le n} e^{\sum_{1\le i,j\le n}\alpha_{ij}x_{ij}+\frac{\beta}{2n}\sum_{1\le i,j,k\le n}x_{ij}x_{jk}}, \tag{A.12}$$

be the normalizing factor and also define

$$\begin{aligned}L_n := \sup_{x_{ij}\in[0,1],x_{ij}=x_{ji},1\le i<j\le n}&\left\{\frac{1}{n^2}\sum_{i,j}\alpha_{ij}x_{ij}+\frac{\beta}{2n^3}\sum_{i,j,k}x_{ij}x_{jk}\right.\\ &\left.-\frac{1}{n^2}\sum_{1\le i<j\le n}[x_{ij}\log x_{ij}+(1-x_{ij})\log(1-x_{ij})]\right\}.\end{aligned} \tag{A.13}$$

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

$$f(x)=\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}x_{ij}+\frac{\beta}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}x_{ij}x_{jk}. \tag{A.14}$$

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

$$\begin{aligned}a=\|f\| &\le \sum_{i=1}^{n}\sum_{j=1}^{n}|\alpha_{ij}|+\sum_{i=1}^{n}\sum_{k=1}^{n}\frac{1}{2}|\beta|\\ &\le n^2\left[\int_{[0,1]^2}|\alpha(x,y)|dxdy+\frac{1}{2}|\beta|+1\right].\end{aligned} \tag{A.15}$$

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

$$T(x) := \frac{1}{n^{k-2}} \sum_{q \in [n]^k} \prod_{\{\ell,\ell'\} \in E} x_{q_\ell q_{\ell'}}, \tag{A.16}$$

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

$$\left\| \frac{\partial T}{\partial x_{ij}} \right\| \leq 2|E|, \tag{A.17}$$

and

$$\left\| \frac{\partial^2 T}{\partial x_{ij} \partial x_{i'j'}} \right\| \leq \begin{cases} 4|E|(|E|-1)n^{-1} & \text{if } |\{i,j,i',j'\}| = 2 \text{ or } 3, \\ 4|E|(|E|-1)n^{-2} & \text{if } |\{i,j,i',j'\}| = 4. \end{cases} \tag{A.18}$$

Therefore, by (A.17), we can compute that

$$b_{(ij)} = \left\| \frac{\partial f}{\partial x_{ij}} \right\| \leq 2 \sup_{0 \leq x,y \leq 1} |\alpha(x,y)| + 2|\beta|. \tag{A.19}$$

By (A.18), we can also compute that

$$\begin{aligned} c_{(i,j)(i'j')} &= \left\| \frac{\partial^2 f}{\partial x_{ij} \partial x_{i'j'}} \right\| \\ &\leq \begin{cases} 4|\beta| n^{-1} & \text{if } |\{i,j,i',j'\}| = 2 \text{ or } 3, \\ 4|\beta| n^{-2} & \text{if } |\{i,j,i',j'\}| = 4. \end{cases} \end{aligned} \tag{A.20}$$

Next, we compute that

$$\frac{\partial f}{\partial x_{ij}} = 2\alpha_{ij} + \frac{\partial}{\partial x_{ij}} \frac{\beta}{2n} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n x_{ij} x_{jk}. \tag{A.21}$$

Let T be defined as

$$T(x) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n x_{ij} x_{jk}. \tag{A.22}$$

Then, we have

$$\frac{\partial f}{\partial x_{ij}} = 2\alpha_{ij} + \frac{\beta}{2} \frac{\partial T}{\partial x_{ij}}. \tag{A.23}$$

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 (A.6) (with $f = T$) so that

$$|\tilde{\mathcal{D}}(\epsilon)| \leq \exp\left\{ \frac{\tilde{C}_1 2^4 3^4 n}{\epsilon^4} \log \frac{\tilde{C}_2 2^4 3^4}{\epsilon^4} \right\} = \exp\left\{ \frac{C_1 n}{\epsilon^4} \log \frac{C_2}{\epsilon^4} \right\}, \tag{A.24}$$

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

$$\mathcal{D}(\epsilon) := \left\{ 2\alpha_{ij} + \frac{\beta}{2} d : d \in \tilde{\mathcal{D}}(2\epsilon/\beta), 1 \le i \le j \le n \right\}. \tag{A.25}$$

Hence, $\mathcal{D}(\epsilon)$ satisfies the criterion (A.6) and

$$|\mathcal{D}(\epsilon)| \le \frac{1}{2} n(n+1) |\tilde{\mathcal{D}}(2\epsilon/\beta)| \le \frac{1}{2} n(n+1) \exp\left\{ \frac{C_1 \beta^4 n}{2^4 \epsilon^4} \log \frac{C_2 \beta^4}{2^4 \epsilon^4} \right\}. \tag{A.26}$$

Therefore,

$$\begin{aligned} \mathcal{E}_1 &= \frac{1}{4} \left(\binom{n}{2} \sum_{1 \le i < j \le n} b_{(ij)}^2 \right)^{1/2} \epsilon + 3 \binom{n}{2} \epsilon + \log |\mathcal{D}(\epsilon)| \\ &\le \left[\frac{1}{4} (2\|\alpha\|_\infty + 2|\beta|) + 3 \right] \binom{n}{2} \epsilon + \log \left(\frac{1}{2} n(n+1) \right) + \frac{C_1 \beta^4 n}{2^4 \epsilon^4} \log \frac{C_2 \beta^4}{2^4 \epsilon^4} \\ &\le C_1(\alpha, \beta) n^2 \epsilon + \frac{C_1(\alpha, \beta) n}{\epsilon^4} \log \frac{C_1(\alpha, \beta)}{\epsilon^4} \\ &= C_1(\alpha, \beta) n^{9/5} (\log n)^{1/5}, \end{aligned} \tag{A.27}$$

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

$$C_1(\alpha, \beta) := c_1 (\|\alpha\|_\infty + |\beta|^4 + 1), \tag{A.28}$$

where $c_1 > 0$ is some universal constant [35].

[35]It follows from (A.27) that we can choose $C_1(\alpha, \beta)$ such that $C_1(\alpha, \beta) \ge \max\{\tilde{c}_1 \|\alpha\|_\infty + \tilde{c}_2 |\beta| + \tilde{c}_3, \tilde{c}_4 \beta^4\}$, where $\tilde{c}_1$, $\tilde{c}_2$, $\tilde{c}_3$, $\tilde{c}_4 > 0$ are some universal constants. Note that $\max\{\tilde{c}_1 \|\alpha\|_\infty + \tilde{c}_2 |\beta| + \tilde{c}_3, \tilde{c}_4 \beta^4\} \le \tilde{c}_1 \|\alpha\|_\infty + \tilde{c}_2 |\beta| + \tilde{c}_3 + \tilde{c}_4 \beta^4 \le c_1 (\|\alpha\|_\infty + |\beta|^4 + 1)$ for some universal constant $c_1 > 0$. Thus, we can take $C_1(\alpha, \beta) := c_1 (\|\alpha\|_\infty + |\beta|^4 + 1)$.

We can also compute that

(A.29)

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

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

$$C_2(\alpha,\beta) := c_2\left(\|\alpha\|_\infty + |\beta| + 1\right)^{1/2}(1+|\beta|^2)^{1/2}, \tag{A.30}$$

where $c_2 > 0$ is some universal constant.

Finally, to get lower bound, notice that

$$\frac{1}{2}\sum_{1\le i<j\le n} c_{(ij)(ij)} = \frac{1}{2}\binom{n}{2}4|\beta|n^{-1} \le C_3(\beta)n, \tag{A.31}$$

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

A.2. Statement and Proof of Proposition A.1.

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

$$\alpha(x,y) = \alpha_{ml}, \qquad \text{if } a_{m-1} < x < a_m \text{ and } a_{l-1} < y < a_l, \text{ where } 1 \leq m, l \leq M. \tag{A.32}$$

Then, we have

(A.33)

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

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

$$u_{ij}(x) = \frac{1}{a_j - a_{j-1}} \int_{a_{j-1}}^{a_j} h(x,y) dy, \qquad \text{for any } a_{i-1} < x < a_i. \tag{A.34}$$

We can compute that

$$\iint_{[0,1]^2} \alpha(x,y) h(x,y) dx dy = \sum_{i=1}^{M} \sum_{j=1}^{M} (a_j - a_{j-1}) \int_{a_{i-1}}^{a_i} \alpha_{ij} u_{ij}(x) dx. \tag{A.35}$$

Moreover,

$$\begin{aligned}
\frac{\beta}{2} \int_0^1 \int_0^1 \int_0^1 h(x,y) h(y,z) dx dy dz &= \frac{\beta}{2} \int_0^1 \left(\int_0^1 h(x,y) dy \right)^2 dx \\
&= \frac{\beta}{2} \sum_{i=1}^{M} \int_{a_{i-1}}^{a_i} \left(\sum_{j=1}^{M} (a_j - a_{j-1}) u_{ij}(x) \right)^2 dx.
\end{aligned} \tag{A.36}$$

By Jensen's inequality, we can also compute that

$$
\begin{aligned}
&\frac{1}{2}\int_0^1\int_0^1 I(h(x,y))dxdy \\
&=\frac{1}{2}\sum_{i=1}^M\int_{a_{i-1}}^{a_i}\left[\sum_{j=1}^M\int_{a_{j-1}}^{a_j} I(h(x,y))dy\right]dx \\
&=\frac{1}{2}\sum_{i=1}^M\int_{a_{i-1}}^{a_i}\left[\sum_{j=1}^M(a_j-a_{j-1})\frac{1}{a_j-a_{j-1}}\int_{a_{j-1}}^{a_j} I(h(x,y))dy\right]dx \\
&\geq\frac{1}{2}\sum_{i=1}^M\int_{a_{i-1}}^{a_i}\left[\sum_{j=1}^M(a_j-a_{j-1})I\left(\frac{1}{a_j-a_{j-1}}\int_{a_{j-1}}^{a_j} h(x,y)dy\right)\right]dx \\
&=\frac{1}{2}\sum_{i=1}^M\int_{a_{i-1}}^{a_i}\sum_{j=1}^M(a_j-a_{j-1})I(u_{ij}(x))dx
\end{aligned}
\tag{A.37}
$$

Hence, by (A.35), (A.36), (A.37), we get

(A.38)

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

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

$$
h(x,y)=u_{ij},\qquad \text{if } a_{i-1}<x<a_i \text{ and } a_{j-1}<y<a_j, \text{ where } 1\leq i,j\leq M,
\tag{A.39}
$$

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:

$$
\begin{aligned}
\psi(\alpha,\beta)\geq\sup_{\substack{0\leq u_{ij}\leq 1\\ u_{ij}=u_{ji},1\leq i,j\leq M}}&\sum_{i=1}^M(a_i-a_{i-1})\left\{\sum_{j=1}^M(a_j-a_{j-1})\alpha_{ij}u_{ij}\right. \\
&\left.+\frac{\beta}{2}\left(\sum_{j=1}^M(a_j-a_{j-1})u_{ij}\right)^2-\frac{1}{2}\sum_{j=1}^M(a_j-a_{j-1})I(u_{ij})\right\}.
\end{aligned}
\tag{A.40}
$$

□

A.3. **Proof of Proposition 5.1.** First, observe that

$$\begin{aligned}
&\psi(\alpha,\beta;-\infty) \\
&= \sup_{h\in\mathcal{W}^-}\Bigg\{\sum_{i=1}^M \alpha_i \iint_{[a_{i-1},a_i]^2} h(x,y)dxdy + \frac{\beta}{2}\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz \\
&\qquad\qquad - \frac{1}{2}\sum_{i=1}^M \iint_{[a_{i-1},a_i]^2} I(h(x,y))dxdy\Bigg\} \\
&= \sup_{h\in\mathcal{W}^-}\Bigg\{\sum_{i=1}^M \alpha_i \iint_{[a_{i-1},a_i]^2} h(x,y)dxdy + \frac{\beta}{2}\sum_{i=1}^M\int_{a_{i-1}}^{a_i}\left(\int_{a_{i-1}}^{a_i} h(x,y)dy\right)^2 dx \\
&\qquad\qquad - \frac{1}{2}\sum_{i=1}^M \iint_{[a_{i-1},a_i]^2} I(h(x,y))dxdy\Bigg\} \\
&= \sum_{i=1}^M \sup_{\substack{h:[a_{i-1},a_i]^2\to[0,1]\\ h(x,y)=h(y,x)}}\Bigg\{\alpha_i \iint_{[a_{i-1},a_i]^2} h(x,y)dxdy + \frac{\beta}{2}\int_{a_{i-1}}^{a_i}\left(\int_{a_{i-1}}^{a_i} h(x,y)dy\right)^2 dx \\
&\qquad\qquad - \frac{1}{2}\iint_{[a_{i-1},a_i]^2} I(h(x,y))dxdy\Bigg\},
\end{aligned} \tag{A.41}$$

where

$$\mathcal{W}^- := \left\{h\in\mathcal{W} : h(x,y)=0 \text{ for any } (x,y)\notin\bigcup_{i=1}^M [a_{i-1},a_i]^2\right\}. \tag{A.42}$$

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

$$\psi(\alpha,\beta;-\infty) \geq \sum_{i=1}^M (a_i-a_{i-1})^2 \sup_{0\leq x\leq 1}\left\{\alpha_i x + \frac{\beta}{2}x^2 - \frac{1}{2}I(x)\right\}. \tag{A.43}$$

By Jensen's inequality

$$\begin{aligned}
\psi(\alpha,\beta;-\infty) &\leq \sum_{i=1}^M \sup_{\substack{h:[a_{i-1},a_i]^2\to[0,1]\\ h(x,y)=h(y,x)}}\Bigg\{\alpha_i\int_{a_{i-1}}^{a_i}\left(\int_{a_{i-1}}^{a_i} h(x,y)dy\right)dx \\
&\qquad + \frac{\beta}{2}\int_{a_{i-1}}^{a_i}\left(\int_{a_{i-1}}^{a_i} h(x,y)dy\right)^2 dx \\
&\qquad\qquad - \frac{1}{2}(a_i-a_{i-1})\int_{a_{i-1}}^{a_i} I\left(\frac{1}{a_i-a_{i-1}}\int_{a_{i-1}}^{a_i} h(x,y)dy\right)dx\Bigg\} \\
&\leq \sum_{i=1}^M (a_i-a_{i-1})^2 \sup_{0\leq x\leq 1}\left\{\alpha_i x + \frac{\beta}{2}x^2 - \frac{1}{2}I(x)\right\}.
\end{aligned} \tag{A.44}$$

A.4. **Characterization of the variational problem.** The variational problem for the graphons is

$$\psi(\alpha,\beta)=\sup_{h\in\mathcal{W}}\left\{\int_0^1\int_0^1\alpha(x,y)h(x,y)dxdy+\frac{\beta}{2}\int_0^1\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz\right.$$
$$\left.-\frac{1}{2}\int_0^1\int_0^1\left[h(x,y)\log h(x,y)+(1-h(x,y))\log(1-h(x,y))\right]dxdy\right\}. \tag{A.45}$$

PROPOSITION A.2. *The optimal graphon h that solves the variational problem (A.45) satisfies the Euler-Lagrange equation:*

$$2\alpha(x,y)+\beta\int_0^1 h(x,y)dx+\beta\int_0^1 h(x,y)dy=\log\left(\frac{h(x,y)}{1-h(x,y)}\right). \tag{A.46}$$

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

COROLLARY 1. *If $\alpha(x,y)$ is not a constant function, then the optimal graphon h that solves the variational problem (A.45) is not a constant function.*

Proof. If the optimal graphon h is a constant function, then (A.46) 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 A.3. *Assume that $\beta<0$. If h is a graphon that satisfies the Euler-Lagrange equation (A.46), then h is a local maximizer of the variational problem (A.45).*

Proof. Let us define

$$\Lambda[h]:=\int_0^1\int_0^1\alpha(x,y)h(x,y)dxdy+\frac{\beta}{2}\int_0^1\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz$$
$$-\frac{1}{2}\int_0^1\int_0^1\left[h(x,y)\log h(x,y)+(1-h(x,y))\log(1-h(x,y))\right]dxdy. \tag{A.47}$$

Let h satisfies (A.46) and for any symmetric function g and $\epsilon>0$ sufficiently small, we can compute that

$$\begin{aligned}\Lambda[h+\epsilon g]-\Lambda[h]&\\ &=\epsilon^2\left[\frac{\beta}{2}\int_0^1\left(\int_0^1 g(x,y)dy\right)^2dx-\frac{1}{4}\int_0^1\int_0^1 I''(h(x,y))g^2(x,y)dxdy\right]+O(\epsilon^3)\\ &=\epsilon^2\left[\frac{\beta}{2}\int_0^1\left(\int_0^1 g(x,y)dy\right)^2dx-\frac{1}{4}\int_0^1\int_0^1 \frac{g^2(x,y)}{h(x,y)(1-h(x,y))}dxdy\right]+O(\epsilon^3),\end{aligned} \tag{A.48}$$

and since $\beta<0$, we conclude that h is a local maximizer in (A.45). □

A.5. **Proof of Proposition 5.2.** For the convenience of the notations, let us define

$$u(x) := 2\int_0^{\frac{1}{2}} h(x,y)dy, \qquad 0 < x < \frac{1}{2}, \tag{A.49}$$

$$w(x) := 2\int_{\frac{1}{2}}^1 h(x,y)dy, \qquad \frac{1}{2} < x < 1, \tag{A.50}$$

$$v_1(x) := 2\int_0^{\frac{1}{2}} h(x,y)dy, \qquad \frac{1}{2} < x < 1, \tag{A.51}$$

$$v_2(x) := 2\int_{\frac{1}{2}}^1 h(x,y)dy, \qquad 0 < x < \frac{1}{2}. \tag{A.52}$$

We can compute that

$$\begin{aligned}\iint_{[0,1]^2} \alpha(x,y)h(x,y)dxdy = \frac{\alpha_1}{2}\int_0^{\frac{1}{2}} u(x)dx + \frac{\alpha_1}{2}\int_{\frac{1}{2}}^1 w(x)dx \\ + \frac{\alpha_2}{2}\int_{\frac{1}{2}}^1 v_1(x)dx + \frac{\alpha_2}{2}\int_0^{\frac{1}{2}} v_2(x)dx.\end{aligned} \tag{A.53}$$

Moreover,

$$\begin{aligned}&\frac{\beta}{2}\int_0^1\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz \\ &= \frac{\beta}{2}\int_0^1\left(\int_0^1 h(x,y)dy\right)^2 dx \\ &= \frac{\beta}{8}\int_0^{\frac{1}{2}}(u(x)+v_2(x))^2dx + \frac{\beta}{8}\int_{\frac{1}{2}}^1 (v_1(x)+w(x))^2dx.\end{aligned} \tag{A.54}$$

We can also compute that by Jensen's inequality,

$$\begin{aligned}\frac{1}{2}\int_0^1\int_0^1 I(h(x,y))dxdy &= \frac{1}{2}\int_0^{\frac{1}{2}}\int_0^{\frac{1}{2}} I(h(x,y))dydx + \frac{1}{2}\int_0^{\frac{1}{2}}\int_{\frac{1}{2}}^1 I(h(x,y))dydx \\ &\quad + \frac{1}{2}\int_{\frac{1}{2}}^1\int_0^{\frac{1}{2}} I(h(x,y))dydx + \frac{1}{2}\int_{\frac{1}{2}}^1\int_{\frac{1}{2}}^1 I(h(x,y))dydx \\ &\geq \frac{1}{4}\int_0^{\frac{1}{2}} I(u(x))dx + \frac{1}{4}\int_0^{\frac{1}{2}} I(v_2(x))dx \\ &\quad + \frac{1}{4}\int_{\frac{1}{2}}^1 I(v_1(x))dx + \frac{1}{4}\int_{\frac{1}{2}}^1 I(w(x))dx\end{aligned} \tag{A.55}$$

where $I(x) := x\log x + (1-x)\log(1-x)$.

Therefore,

$$\begin{aligned}
\psi(\alpha,\beta) \leq \sup_{u,v_1,v_2,w} \Bigg\{ & \int_0^{\frac{1}{2}} \left[\frac{\alpha_1}{2}u(x) - \frac{1}{4}I(u(x))\right] dx + \int_{\frac{1}{2}}^1 \left[\frac{\alpha_1}{2}w(x) - \frac{1}{4}I(w(x))\right] dx \\
& + \int_{\frac{1}{2}}^1 \left[\frac{\alpha_2}{2}v_1(x) - \frac{1}{4}I(v_1(x))\right] dx + \int_0^{\frac{1}{2}} \left[\frac{\alpha_2}{2}v_2(x) - \frac{1}{4}I(v_2(x))\right] dx \\
& + \frac{\beta}{8}\int_0^{\frac{1}{2}} (u(x)+v_2(x))^2 dx + \frac{\beta}{8}\int_{\frac{1}{2}}^1 (v_1(x)+w(x))^2 dx \Bigg\} \\
\leq \frac{1}{2} & \sup_{0\leq u,v_2\leq 1} \left\{ \frac{\alpha_1}{2}u - \frac{1}{4}I(u) + \frac{\alpha_2}{2}v_2 - \frac{1}{4}I(v_2) + \frac{\beta}{8}(u+v_2)^2 \right\} \\
& + \frac{1}{2}\sup_{0\leq w,v_1\leq 1} \left\{ \frac{\alpha_1}{2}w - \frac{1}{4}I(w) + \frac{\alpha_2}{2}v_1 - \frac{1}{4}I(v_1) + \frac{\beta}{8}(w+v_1)^2 \right\} \\
= & \sup_{0\leq u,v\leq 1} \left\{ \frac{\alpha_1}{2}u - \frac{1}{4}I(u) + \frac{\alpha_2}{2}v - \frac{1}{4}I(v) + \frac{\beta}{8}(u+v)^2 \right\}.
\end{aligned} \tag{A.56}$$

On the other hand, by restricting the supremum over the graphons $h(x,y)$ that takes the constant values u and v, i.e.,

$$h(x,y) = \begin{cases} u & \text{if } 0 < x,y < \frac{1}{2} \text{ or } \frac{1}{2} < x,y < 1, \\ v & \text{if } 0 < x < \frac{1}{2} < y < 1 \text{ or } 0 < y < \frac{1}{2} < x < 1, \end{cases} \tag{A.57}$$

and optimize over all the possible values $0 \leq u,v \leq 1$, we get the lower bound:

$$\psi(\alpha,\beta) \geq \sup_{0\leq u,v\leq 1} \left\{ \frac{\alpha_1}{2}u - \frac{1}{4}I(u) + \frac{\alpha_2}{2}v - \frac{1}{4}I(v) + \frac{\beta}{8}(u+v)^2 \right\}. \tag{A.58}$$

Since the upper bound (A.56) and the lower bound (A.58) match, the proof is complete.

Remark A.2. *When $\alpha_1 = \alpha_2$, the variational problem in equation 5.6 becomes*

$$\psi(\alpha,\beta) = \sup_{0\leq u,v\leq 1} \left\{ \frac{\alpha_1}{2}(u+v) - \frac{1}{4}I(u) - \frac{1}{4}I(v) + \frac{\beta}{8}(u+v)^2 \right\}. \tag{A.59}$$

By Jensen's inequality

$$\begin{aligned}
\psi(\alpha,\beta) &\leq \sup_{0\leq u,v\leq 1} \left\{ \frac{\alpha_1}{2}(u+v) - \frac{1}{2}I\left(\frac{u+v}{2}\right) + \frac{\beta}{8}(u+v)^2 \right\} \\
&= \sup_{0\leq x\leq 1} \left\{ \alpha_1 x - \frac{1}{2}I(x) + \frac{\beta}{2}x^2 \right\}.
\end{aligned} \tag{A.60}$$

On the other hand,

$$\psi(\alpha,\beta) \geq \sup_{0\leq u=v\leq 1}\left\{\frac{\alpha_1}{2}u - \frac{1}{4}I(u) + \frac{\alpha_2}{2}v - \frac{1}{4}I(v) + \frac{\beta}{8}(u+v)^2\right\}$$
$$= \sup_{0\leq x\leq 1}\left\{\alpha_1 x - \frac{1}{2}I(x) + \frac{\beta}{2}x^2\right\}. \tag{A.61}$$

Hence, when $\alpha_1 = \alpha_2$, we recover Theorem 6.4. in Chatterjee and Diaconis (2013):

$$\psi(\alpha,\beta) = \sup_{0\leq x\leq 1}\left\{\alpha_1 x - \frac{1}{2}I(x) + \frac{\beta}{2}x^2\right\}. \tag{A.62}$$

The solution to the optimization problem (A.62) has been fully characterized in e.g. Radin and Yin (2013) and Aristoff and Zhu (2014).

A.6. **Proof of Proposition 5.3.** Let us consider the two-dimensional optimization problem:

$$\psi(\alpha,\beta) = \sup_{0\leq u,v\leq 1}\left\{\frac{\alpha_1}{2}u - \frac{1}{4}I(u) + \frac{\alpha_2}{2}v - \frac{1}{4}I(v) + \frac{\beta}{8}(u+v)^2\right\}. \tag{A.63}$$

At optimality, we have

$$\frac{\alpha_1}{2} - \frac{1}{4}I'(u^*) + \frac{\beta}{4}(u^*+v^*) = 0,$$
$$\frac{\alpha_2}{2} - \frac{1}{4}I'(v^*) + \frac{\beta}{4}(u^*+v^*) = 0.$$

Let $\gamma = u^* + v^*$, then, we can compute that

$$u^* = \frac{e^{2\alpha_1+\beta\gamma}}{1+e^{2\alpha_1+\beta\gamma}}, \qquad v^* = \frac{e^{2\alpha_2+\beta\gamma}}{1+e^{2\alpha_2+\beta\gamma}}. \tag{A.64}$$

And γ satisfies the equation:

$$\frac{e^{2\alpha_1+\beta\gamma}}{1+e^{2\alpha_1+\beta\gamma}} + \frac{e^{2\alpha_2+\beta\gamma}}{1+e^{2\alpha_2+\beta\gamma}} = \gamma. \tag{A.65}$$

Let us define:

$$F(u,v) := \frac{\alpha_1}{2}u - \frac{1}{4}I(u) + \frac{\alpha_2}{2}v - \frac{1}{4}I(v) + \frac{\beta}{8}(u+v)^2. \tag{A.66}$$

At (u^*, v^*), the Hessian matrix is given by
(A.67)

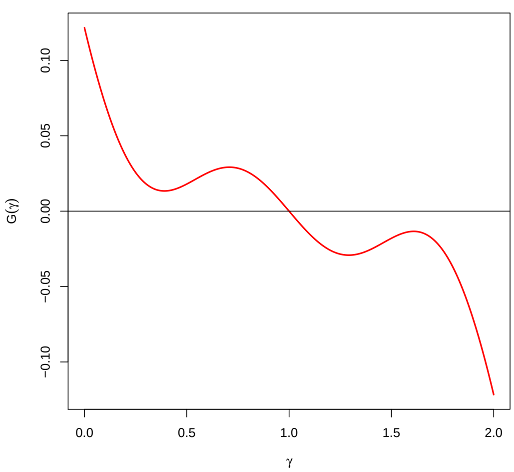
$$\mathcal{H}(u^*,v^*) = \begin{bmatrix} -\frac{1}{4}I''(u^*) + \frac{\beta}{4} & \frac{\beta}{4} \\ \frac{\beta}{4} & -\frac{1}{4}I''(v^*) + \frac{\beta}{4} \end{bmatrix} = \begin{bmatrix} -\frac{1}{4u^*(1-u^*)} + \frac{\beta}{4} & \frac{\beta}{4} \\ \frac{\beta}{4} & -\frac{1}{4v^*(1-v^*)} + \frac{\beta}{4} \end{bmatrix}.$$

It follows that when $\beta < 0$, (u^*, v^*) always gives a local maximum. To analyze the maximization problem, let us also define a function G as follows:

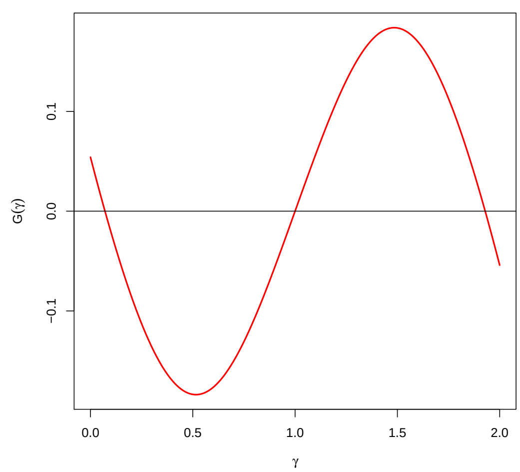
$$G(\gamma) := \frac{e^{2\alpha_1+\beta\gamma}}{1+e^{2\alpha_1+\beta\gamma}} + \frac{e^{2\alpha_2+\beta\gamma}}{1+e^{2\alpha_2+\beta\gamma}} - \gamma. \tag{A.68}$$

Then we must have $G(-\infty) = \infty$ and $G(+\infty) = -\infty$ which implies the existence of γ as a function of α_1, α_2, and β. The function $G(\gamma)$ is shown in Figures A.1 and A.2, corresponding to the examples in Figures 5.1 and 5.2 in the main text. We can also compute that

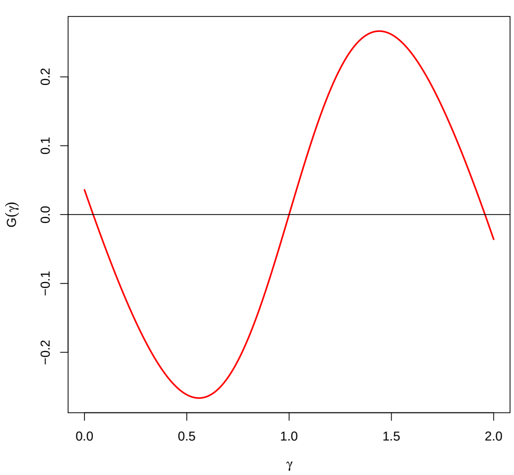
FIGURE A.1. Examples of maxima characterized in Proposition 5.3 with $\beta = 4$



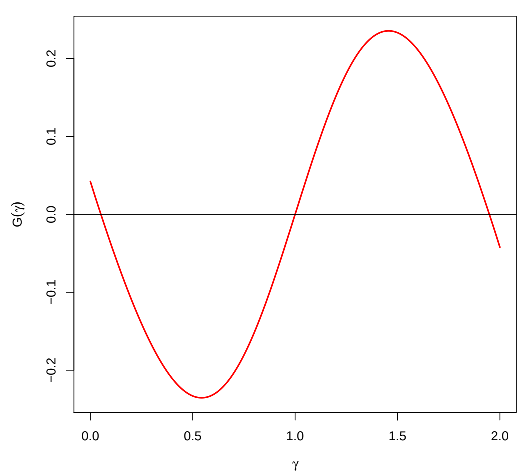
(A) $(\alpha_1, \alpha_2) = (-3, -1)$



(B) $(\alpha_1, \alpha_2) = (-2.5, -1.5)$



(C) $(\alpha_1, \alpha_2) = (-2, -2)$



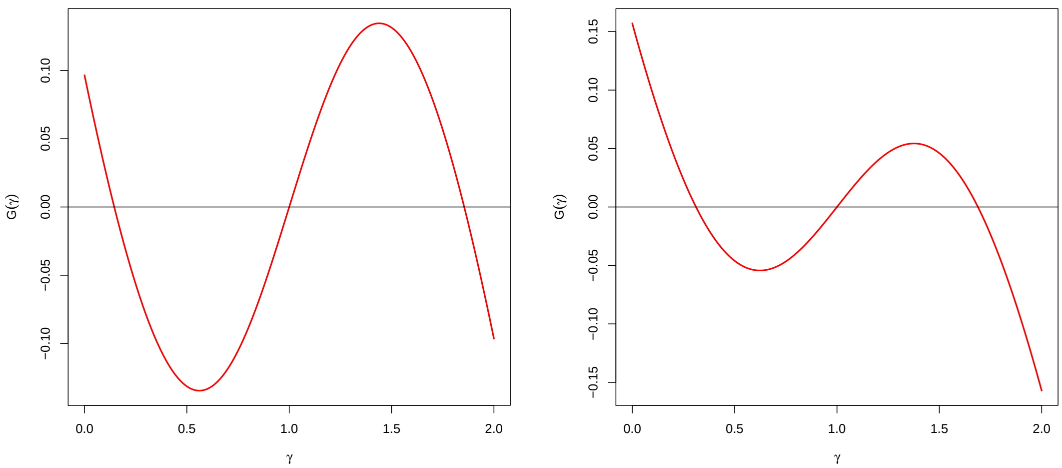
(D)$(\alpha_1, \alpha_2) = (-1.7, -2.3)$

The figures show function $G(\gamma)$ for different vectors of parameters, corresponding to Figure 5.1 in the main text. In all the pictures $\beta = 4$.

$$G'(\gamma) = \frac{\beta e^{2\alpha_1+\beta\gamma}}{(1+e^{2\alpha_1+\beta\gamma})^2} + \frac{\beta e^{2\alpha_2+\beta\gamma}}{(1+e^{2\alpha_2+\beta\gamma})^2} - 1. \tag{A.69}$$

When $\beta < 0$, $G'(\gamma) < 0$, and the γ is unique, which implies that (u^*, v^*) is unique and there is no phase transition.

FIGURE A.2. More examples of maxima characterized in Proposition 5.3



(A) $(\alpha_1, \alpha_2, \beta) = (-1.4, -1.6, 3)$ (B) $(\alpha_1, \alpha_2, \beta) = (-1.1, -1.4, 2.5)$

The figures show function $G(\gamma)$ for different vectors of parameters, corresponding to Figure 5.2 in the main text.

Note that for $\beta \geq 0$, since $\frac{x}{(1+x)^2} \leq \frac{1}{4}$ for any $x \geq 0$, we get:

$$G'(\gamma) \leq \frac{\beta}{4} + \frac{\beta}{4} - 1 \leq 0, \tag{A.70}$$

for $\beta \leq 2$. Hence, we conclude that whenever $\beta \leq 2$, the optimizer (u^*, v^*) is unique and there is no phase transition.

In addition, if $\alpha_1 + \alpha_2 + \beta = 0$ then $\gamma = 1$ is a root of $G(\gamma) = 0$ and hence the unique optimizer (u^*, v^*) is given by

$$(u^*, v^*) = \left(\frac{e^{2\alpha_1+\beta}}{1+e^{2\alpha_1+\beta}}, \frac{e^{2\alpha_2+\beta}}{1+e^{2\alpha_2+\beta}} \right). \tag{A.71}$$

To show this, notice that when $\alpha_1 + \alpha_2 + \beta = 0$, we have

$$\begin{aligned} G(\gamma) &= \frac{e^{2\alpha_1+\beta\gamma}}{1+e^{2\alpha_1+\beta\gamma}} + \frac{e^{2\alpha_2+\beta\gamma}}{1+e^{2\alpha_2+\beta\gamma}} - \gamma \\ &= \frac{e^{2\alpha_1-\alpha_1\gamma-\alpha_2\gamma}}{1+e^{2\alpha_1-\alpha_1\gamma-\alpha_2\gamma}} + \frac{e^{2\alpha_2-\alpha_1\gamma-\alpha_2\gamma}}{1+e^{2\alpha_2-\alpha_1\gamma-\alpha_2\gamma}} - \gamma. \end{aligned}$$

If we substitute $\gamma = 1$ we obtain

$$\begin{aligned} G(\gamma) &= \frac{e^{\alpha_1-\alpha_2}}{1+e^{\alpha_1-\alpha_2}} + \frac{e^{\alpha_2-\alpha_1}}{1+e^{\alpha_2-\alpha_1}} \frac{e^{\alpha_1-\alpha_2}}{e^{\alpha_1-\alpha_2}} - 1 \\ &= \frac{e^{\alpha_1-\alpha_2}}{1+e^{\alpha_1-\alpha_2}} + \frac{1}{1+e^{\alpha_1-\alpha_2}} - 1 = 0. \end{aligned}$$

Therefore $\gamma = 1$ is a root.

For the rest, we will always assume that $\beta > 2$. Phase transition occurs at $(\alpha_1, \alpha_2, \beta)$ when there exist $\gamma_1 \neq \gamma_2$ with $G(\gamma_1) = G(\gamma_2) = 0$ and $(u^*(\gamma_1), v^*(\gamma_1)) \neq (u^*(\gamma_2), v^*(\gamma_2))$ and moreover:

$$F(u^*(\gamma_1), v^*(\gamma_1)) = F(u^*(\gamma_2), v^*(\gamma_2)). \tag{A.72}$$

We can gain some insights on the phase transition by first looking at the Hessian of the maximization problem

$$\mathcal{H}(u,v) = \begin{bmatrix} -\frac{1}{4u(1-u)} + \frac{\beta}{4} & \frac{\beta}{4} \\ \frac{\beta}{4} & -\frac{1}{4v(1-v)} + \frac{\beta}{4} \end{bmatrix}. \tag{A.73}$$

To get a local maximum at (u, v), we need

$$h_{uu} < 0\,, \tag{A.74}$$

$$h_{uu}h_{vv} - (h_{uv})^2 > 0\,. \tag{A.75}$$

The condition (A.74) reads:

$$-\frac{1}{4u(1-u)} + \frac{\beta}{4} < 0\,, \tag{A.76}$$

which is equivalent to:

$$u(1-u) < \frac{1}{\beta}. \tag{A.77}$$

For the second condition (A.75), it is equivalent to:

$$\left[-\frac{1}{4u(1-u)} + \frac{\beta}{4}\right]\left[-\frac{1}{4v(1-v)} + \frac{\beta}{4}\right] - \frac{\beta^2}{16} > 0, \tag{A.78}$$

which is equivalent to:

$$\frac{1}{16u(1-u)v(1-v)} - \frac{\beta}{16u(1-u)} - \frac{\beta}{16v(1-v)} > 0. \tag{A.79}$$

We can re-arrange the expression and it reduces to

$$\frac{1}{u(1-u)v(1-v)} > \beta\left[\frac{1}{u(1-u)} + \frac{1}{v(1-v)}\right] = \beta\left[\frac{v(1-v)+u(1-u)}{u(1-u)v(1-v)}\right]. \tag{A.80}$$

Hence, we conclude that the condition (A.75) is equivalent to:

$$v(1-v) + u(1-u) < \frac{1}{\beta}. \tag{A.81}$$

Therefore, if the condition (A.75) holds, then the condition (A.74) is automatically satisfied.

Let's define the function

$$\eta(u,v) \equiv v(1-v) + u(1-u)\,.$$

The condition for positive determinant (A.75) is thus given by:

$$\eta(u,v) < \frac{1}{\beta}.$$

This is a region of the unit square: it is defined as the area outside the level curve $\eta(u,v) = 1/\beta$, i.e. the set $\{u, v \in [0,1]^2 : \eta(u,v) < \frac{1}{\beta}\}$.

We have shown before that if $\alpha_1 + \alpha_2 + \beta = 0$ then $\gamma = 1$ is a root of $G(\gamma) = 0$. To have two stationary points (other than $\gamma = 1$), since $G(-\infty) = \infty$, $G(\infty) = -\infty$, $G(1) = 0$ and $G(\gamma)$ is smooth in γ, it suffices to have $G'(1) > 0$. We will show later with additional arguments that, indeed when $G'(1) > 0$, we will have two local maxima. Note that

$$G'(1) = \frac{\beta e^{\alpha_1-\alpha_2}}{(1+e^{\alpha_1-\alpha_2})^2} + \frac{\beta e^{\alpha_2-\alpha_1}}{(1+e^{\alpha_2-\alpha_1})^2} - 1, \tag{A.82}$$

and we can compute that

$$\begin{aligned}
G'(1) &= \frac{\beta e^{\alpha_1-\alpha_2}}{(1+e^{\alpha_1-\alpha_2})^2} + \frac{\beta e^{\alpha_2-\alpha_1}}{(1+e^{\alpha_2-\alpha_1})^2} - 1 \\
&= \frac{\beta e^{\alpha_1-\alpha_2}}{(1+e^{\alpha_1-\alpha_2})^2} + \frac{\beta}{(1+e^{\alpha_1-\alpha_2})(1+e^{\alpha_2-\alpha_1})} - 1 \\
&= \frac{\beta}{(1+e^{\alpha_1-\alpha_2})}\left[\frac{e^{\alpha_1-\alpha_2}}{(1+e^{\alpha_1-\alpha_2})} + \frac{1}{(1+e^{\alpha_2-\alpha_1})}\right] - 1 \\
&= \frac{\beta}{(1+e^{\alpha_1-\alpha_2})}\left[\frac{e^{\alpha_1-\alpha_2}}{(1+e^{\alpha_1-\alpha_2})} + \frac{e^{\alpha_1-\alpha_2}}{(1+e^{\alpha_1-\alpha_2})}\right] - 1 \\
&= \frac{2\beta e^{\alpha_1-\alpha_2}}{(1+e^{\alpha_1-\alpha_2})^2} - 1.
\end{aligned}$$

Hence $G'(1) > 0$ if and only if

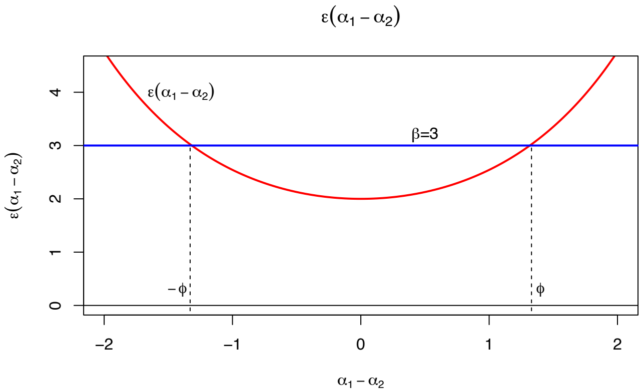
$$\beta > \frac{(1+e^{\alpha_1-\alpha_2})^2}{2e^{\alpha_1-\alpha_2}} \equiv \epsilon(\alpha_1 - \alpha_2). \tag{A.83}$$

Before we proceed, let us notice that $\frac{(1+e^{\alpha_1-\alpha_2})^2}{2e^{\alpha_1-\alpha_2}} \geq 2$. Thus if the above condition holds, then we are automatically in the regime $\beta > 2$.

Now we have a condition that guarantees that there exist two stationary points γ_1 and γ_2, such that

$$\gamma_1 < 1 < \gamma_2\,.$$

Notice that the function $\epsilon(\alpha_1 - \alpha_2)$ has a minimum of 2, which implies that any $\beta < 2$ would have a unique maximum, which is consistent with our conclusion before.



Let $\phi > 0$ be the value of $\alpha_1 - \alpha_2$ such that $\epsilon(\phi) = \beta$ and $\epsilon(-\phi) = \beta$. Then, if $|\alpha_1 - \alpha_2| < \phi$ the function F has two local stationary points in addition to $\gamma = 1$.

Next, let us prove that if $\gamma^* < 1$ is a solution, then $2 - \gamma^* > 1$ is also a solution.

Indeed, let us notice that

$$G(\gamma) = 0 \qquad \text{if and only if} \qquad \frac{e^{2\alpha_1+\beta\gamma}}{1+e^{2\alpha_1+\beta\gamma}} + \frac{e^{2\alpha_2+\beta\gamma}}{1+e^{2\alpha_2+\beta\gamma}} - \gamma = 0\,. \tag{A.84}$$

Under the assumption $\alpha_1 + \alpha_2 + \beta = 0$, we have

$$2\alpha_1 + 2\beta = -2\alpha_2, \qquad 2\alpha_2 + 2\beta = -2\alpha_1, \tag{A.85}$$

and thus we can compute that

$$\begin{aligned}
G(2-\gamma) &= \frac{e^{2\alpha_1+2\beta-\beta\gamma}}{1+e^{2\alpha_1+2\beta-\beta\gamma}} + \frac{e^{2\alpha_2+2\beta-\beta\gamma}}{1+e^{2\alpha_2+2\beta-\beta\gamma}} - (2-\gamma) \\
&= \frac{e^{-(2\alpha_2+\beta\gamma)}}{1+e^{-(2\alpha_2+\beta\gamma)}} + \frac{e^{-(2\alpha_1+\beta\gamma)}}{1+e^{-(2\alpha_1+\beta\gamma)}} - (2-\gamma) \\
&= \left[1 - \frac{e^{2\alpha_1+\beta\gamma}}{1+e^{2\alpha_1+\beta\gamma}}\right] + \left[1 - \frac{e^{2\alpha_2+\beta\gamma}}{1+e^{2\alpha_2+\beta\gamma}}\right] - 2 + \gamma \\
&= 1 - \frac{e^{2\alpha_1+\beta\gamma}}{1+e^{2\alpha_1+\beta\gamma}} + 1 - \frac{e^{2\alpha_2+\beta\gamma}}{1+e^{2\alpha_2+\beta\gamma}} - 2 + \gamma \\
&= \frac{e^{2\alpha_1+\beta\gamma}}{1+e^{2\alpha_1+\beta\gamma}} + \frac{e^{2\alpha_2+\beta\gamma}}{1+e^{2\alpha_2+\beta\gamma}} - \gamma,
\end{aligned}$$

which yields that $G(2-\gamma) = 0$ if and only if $G(\gamma) = 0$.

By the same reasoning and the previous algebra, it follows that if (u^*, v^*) is a stationary point, then also $(1 - v^*, 1 - u^*)$ is a stationary point. With some algebra we can show that $F(u, v) = F(1 - v, 1 - u)$. Indeed, we have

$$
\begin{aligned}
F(1-v,1-u) &= \frac{\alpha_1}{2}(1-v)-\frac{1}{4}I(1-v)+\frac{\alpha_2}{2}(1-u)-\frac{1}{4}I(1-u)+\frac{\beta}{8}(1-u+1-v)^2 \\
&= \frac{\alpha_1}{2}-\frac{\alpha_1}{2}v-\frac{1}{4}I(v)+\frac{\alpha_2}{2}-\frac{\alpha_2}{2}u-\frac{1}{4}I(u)+\frac{\beta}{8}(2-(u+v))^2 \\
&= \frac{\alpha_1}{2}-\frac{\alpha_1}{2}v-\frac{1}{4}I(v)+\frac{\alpha_2}{2}-\frac{\alpha_2}{2}u-\frac{1}{4}I(u)+\frac{\beta}{8}(4-4(u+v)+(u+v)^2) \\
&= \left(\frac{\alpha_1}{2}+\frac{\alpha_2}{2}+\frac{\beta}{2}\right)-\frac{\alpha_1}{2}v-\frac{1}{4}I(v)-\frac{\alpha_2}{2}u-\frac{1}{4}I(u)-\frac{\beta}{2}(u+v)+\frac{\beta}{8}(u+v)^2\,,
\end{aligned}
$$

and by using $\frac{\alpha_1}{2}+\frac{\alpha_2}{2}+\frac{\beta}{2}=0$, we have

$$
\begin{aligned}
F(1-v,1-u) &= -\frac{\alpha_1}{2}v-\frac{1}{4}I(v)-\frac{\alpha_2}{2}u-\frac{1}{4}I(u)-\frac{\beta}{2}(u+v)+\frac{\beta}{8}(u+v)^2 \\
&= -\frac{\alpha_1}{2}v-\frac{1}{4}I(v)-\frac{\alpha_2}{2}u-\frac{1}{4}I(u)-\frac{\beta}{2}u-\frac{\beta}{2}v+\frac{\beta}{8}(u+v)^2 \\
&= v\left(-\frac{\alpha_1}{2}-\frac{\beta}{2}\right)+u\left(-\frac{\alpha_2}{2}-\frac{\beta}{2}\right)-\frac{1}{4}I(u)-\frac{1}{4}I(v)+\frac{\beta}{8}(u+v)^2 \\
&= \frac{\alpha_2}{2}v+\frac{\alpha_1}{2}u-\frac{1}{4}I(u)-\frac{1}{4}I(v)+\frac{\beta}{8}(u+v)^2 \\
&= F(u,v)\,.
\end{aligned}
$$

Therefore it follows that when (u^*,v^*) is a maximizer, so it is $(1-v^*,1-u^*)$ and $F(u^*,v^*)=F(1-v^*,1-u^*)$.

The condition for positive definiteness for the Hessian matrix is $\eta(u,v)<\frac{1}{\beta}$. Notice that in our derivation we have assumed that $\beta>2$ and hence $\beta>0$, because we know that when $\beta\leq 2$ we have a unique maximizer (see above).

Assume $\alpha_1+\alpha_2+\beta=0$ and $\beta>\frac{(1+e^{\alpha_1-\alpha_2})^2}{2e^{\alpha_1-\alpha_2}}$, then there exists some $\gamma^*<1<2-\gamma^*$ that give two set of stationary points (u^*,v^*) and $(1-v^*,1-u^*)$ such that $F(u^*,v^*)=F(1-v^*,1-u^*)$. Since we know the maximum exists, to show that we can find two maxima (u^*,v^*) and $(1-v^*,1-u^*)$, it suffices to show that $\gamma^*=1$ corresponds to a saddle point. This is the case if the Hessian matrix has both positive and negative eigenvalues, which is the case if $\eta(u^*(1),v^*(1))>\frac{1}{\beta}$, which is equivalent to:

$$
\eta(u^*(1),v^*(1))=\frac{2}{(1+e^{\alpha_1-\alpha_2})(1+e^{\alpha_2-\alpha_1})}>\frac{1}{\beta}\,. \tag{A.86}
$$

It is easy to check that

$$
\frac{(1+e^{\alpha_1-\alpha_2})(1+e^{\alpha_2-\alpha_1})}{2}=\frac{(1+e^{\alpha_1-\alpha_2})^2}{2e^{\alpha_1-\alpha_2}}.
$$

Thus the condition (A.86) is equivalent to the assumption $\beta>\frac{(1+e^{\alpha_1-\alpha_2})^2}{2e^{\alpha_1-\alpha_2}}$.

¿From the equation (A.64), it is clear that $u^*(\gamma)$ is monotonic in γ and thus $u^*(\gamma_1)\neq u^*(\gamma_2)$ for any $\gamma_1\neq\gamma_2$.

Moreover, we need to show that $(u^*,v^*)\neq(1-v^*,1-u^*)$ so that we get two distinct global maximizers. This is simple since $u^*+v^*=\gamma^*\neq 1$.

To summarize, if $\alpha_1 + \alpha_2 + \beta = 0$, and

$$\beta > \frac{(1 + e^{\alpha_1 - \alpha_2})^2}{2e^{\alpha_1 - \alpha_2}},$$

then, there are two global maximizers. This completes the proof.

We can also decouple the solution to the optimization problem (5.6) so that the optimal u and v satisfy two independent equations as follows.

Remark A.3. *The optimal (u, v) in the optimization problem (5.6) satisfies:*

$$\frac{\alpha_1}{2} - \frac{1}{4}\log\left(\frac{u}{1-u}\right) + \frac{\beta}{4}(u+v) = 0, \tag{A.87}$$

$$\frac{\alpha_2}{2} - \frac{1}{4}\log\left(\frac{v}{1-v}\right) + \frac{\beta}{4}(u+v) = 0. \tag{A.88}$$

Equating the equations (A.87) and (A.88), we get:

$$\frac{\alpha_1}{2} - \frac{1}{4}\log\left(\frac{u}{1-u}\right) = \frac{\alpha_2}{2} - \frac{1}{4}\log\left(\frac{v}{1-v}\right), \tag{A.89}$$

which implies that

$$v = \frac{u}{u + e^{2(\alpha_1 - \alpha_2)}(1-u)}, \qquad u = \frac{v}{v + e^{2(\alpha_2 - \alpha_1)}(1-v)}. \tag{A.90}$$

Substituting into (A.87) and (A.88), we conclude that the optimal (u, v) satisfies:

$$\frac{\alpha_1}{2} - \frac{1}{4}\log\left(\frac{u}{1-u}\right) + \frac{\beta}{4}\left[u + \frac{u}{u + e^{2(\alpha_1 - \alpha_2)}(1-u)}\right] = 0, \tag{A.91}$$

$$\frac{\alpha_2}{2} - \frac{1}{4}\log\left(\frac{v}{1-v}\right) + \frac{\beta}{4}\left[\frac{v}{v + e^{2(\alpha_2 - \alpha_1)}(1-v)} + v\right] = 0. \tag{A.92}$$

A.7. **Proof of Proposition 6.1.** The probability measure $\mathbb{P}_n$ is defined as follows. For any configuration of the n-node graph X, let $X_{ij} = X_{ij} \in \{0, 1\}$, for $i \neq$ and $X_{ii} = 0$ and $X_{ij} = 1$ if there is an edge between node i and j and $X_{ij} = 0$ otherwise. Then, under $\mathbb{P}_n$, the probability of observing the configuration X is given by

$$\mathbb{P}_n(X) = \frac{1}{Z_n}\exp\left\{\sum_{i,j}\alpha_{ij}X_{ij} + \beta\sum_{i,j,k}X_{ij}X_{jk}\right\}, \tag{A.93}$$

where $Z_n = e^{n^2\psi_n(\alpha,\beta)}$ is the normalizer.

We have shown that $\psi_n(\alpha,\beta) \to \psi(\alpha,\beta)$, where

$$\begin{aligned}\psi(\alpha,\beta) = \sup_{h\in\mathcal{W}} \Bigg\{ &\int_0^1\int_0^1 \alpha(x,y)h(x,y)dxdy + \frac{\beta}{2}\int_0^1\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz \\ &- \frac{1}{2}\int_0^1\int_0^1 \left[h(x,y)\log h(x,y) + (1-h(x,y))\log(1-h(x,y))\right] dxdy \Bigg\}.\end{aligned} \tag{A.94}$$

For any $\epsilon > 0$, let

$$\begin{aligned}\mathcal{A}_\epsilon := \Bigg\{ h\in\mathcal{W} : &\int_0^1\int_0^1 \alpha(x,y)h(x,y)dxdy + \frac{\beta}{2}\int_0^1\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz \\ &- \frac{1}{2}\int_0^1\int_0^1 \left[h(x,y)\log h(x,y) + (1-h(x,y))\log(1-h(x,y))\right] dxdy \le \psi(\alpha,\beta)-\epsilon \Bigg\}.\end{aligned} \tag{A.95}$$

Note that

$$\mathbb{P}_n(X\in\mathcal{A}_\epsilon) = \frac{1}{Z_n}\sum_{X\in\mathcal{A}_\epsilon} \exp\left\{\sum_{i,j}\alpha_{ij}X_{ij} + \beta\sum_{i,j,k}X_{ij}X_{jk}\right\}. \tag{A.96}$$

We have shown that $\lim_{n\to\infty}\frac{1}{n^2}\log Z_n = \psi(\alpha,\beta)$. By the same approach, we can show that

$$\lim_{n\to\infty}\frac{1}{n^2}\log\sum_{X\in\mathcal{A}_\epsilon} \exp\left\{\sum_{i,j}\alpha_{ij}X_{ij} + \beta\sum_{i,j,k}X_{ij}X_{jk}\right\} = \psi_{\mathcal{A}_\epsilon}(\alpha,\beta), \tag{A.97}$$

where

$$\begin{aligned}\psi_{\mathcal{A}_\epsilon}(\alpha,\beta) = \sup_{h\in\mathcal{A}_\epsilon} \Bigg\{ &\int_0^1\int_0^1 \alpha(x,y)h(x,y)dxdy + \frac{\beta}{2}\int_0^1\int_0^1\int_0^1 h(x,y)h(y,z)dxdydz \\ &- \frac{1}{2}\int_0^1\int_0^1 \left[h(x,y)\log h(x,y) + (1-h(x,y))\log(1-h(x,y))\right] dxdy \Bigg\}.\end{aligned} \tag{A.98}$$

Hence, we conclude that

$$\limsup_{n\to\infty}\frac{1}{n^2}\log\mathbb{P}_n(X\in\mathcal{A}_\epsilon) \le \psi_{\mathcal{A}_\epsilon}(\alpha,\beta) - \psi(\alpha,\beta) \le -\epsilon. \tag{A.99}$$

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