A Structural Model of Segregation in Social Networks*

Angelo Mele†

March 31, 2015

Abstract

We propose a model of directed network formation with heterogeneous players that converges to a unique stationary equilibrium. Payoffs depend on direct links, but also link externalities. We show that under mild assumptions, the network formation is a potential game. The equilibrium characterization shows that the model nests the exponential random graph model as a special case. The estimation is complicated by an intractable likelihood, known up to a normalizing constant that is usually approximated using simulation methods. However, standard local simulation strategies may fail to converge to the correct distribution. Extending results from the graph limits and large deviations literature, we show that identification and simulation problems depend on the sign of the link externalities. Using the latter theoretical results, we propose a modification of the algorithms for simulations that allows a more efficient exploration of the likelihood. The posterior distribution of the structural parameters is estimated using an exchange algorithm that avoids evaluation of the intractable constant. We test the estimation strategy with artificial data, showing good performance for moderate lengths of simulations.

JEL Codes: D85, C15, C73
Keywords: Social Networks, Bayesian Estimation, Markov Chain Monte Carlo, Graph Limits

*I am grateful to Roger Koenker, Ron Laschever, Dan Bernhardt and George Deltas for their inputs at crucial stages of this work. I am grateful to the editor and five outstanding referees that provided invaluable suggestions and comments to previous versions. Thanks to Lingjiong Zhu for his inputs about graph limits and large deviations. I thank Federico Bandi, Alberto Bisin, Ethan Cole, Aureo de Paula, Steven Durlauf, Andrea Galeotti, Shweta Gaonkar, Sanjeev Goyal, Dan Karney, Darren Lubotsky, Antonio Mele, Luca Merlino, Ferdinando Monte, Tom Parker, Dennis O'Dea, Micah Pollak, Sergey Popov, Sudipta Sarangi, Giorgio Topa, Antonella Tutino and seminar participants at SED 2010, AddHealth Users 2010, Notre Dame, Carey Business School, UPenn, SMU, FED Board, UQAM, SBIES 2011, PET 2011, Conference on Networks in Cambridge, UK for helpful comments and suggestions. I am grateful to Andriy Norets and Matt Jackson for suggesting references on how to prove and bound convergence of the algorithms. Financial support from the R. Ferber Award, the R. W. Harbeson Memorial Dissertation Fellowship, and the NET Institute Summer Research Grant 2010 is gratefully acknowledged. All remaining errors are mine.

†*Address*: Johns Hopkins University - Carey Business School, 100 International Drive, Baltimore, MD 21202. *Email*: angelo.mele@jhu.edu

1 Introduction

Social networks are important determinants of individuals' socioeconomic performance. An increasing amount of evidence shows that the number and composition of social ties affects employment prospects, school performance, risky behavior, adoption of new technologies, diffusion of information and health outcomes.[1]

The literature on strategic models of network formation provides a framework to interpret the observed network as the equilibrium of a game.[2] However, the estimation and identification of strategic models is challenging. First, network formation models usually have multiple equilibria, because linking generates externalities that are not fully accounted for by individuals. Second, there is a curse of dimensionality: the number of possible network configurations increases exponentially with the number of players. Finally, the data available to the econometrician usually consist of a single network snapshot.

We develop a model of strategic network formation with heterogeneous players that combines ingredients from the strategic and random network formation literature.[3] We contribute to the economic literature on network formation on three dimensions. First, while most strategic models have multiple equilibria, we establish the existence of a unique stationary equilibrium, which characterizes the likelihood of observing a specific network structure in the data. As a consequence, we can estimate the structural parameters using only one observation of the network. Second, we propose a Bayesian Markov Chain Monte Carlo algorithm that drastically reduces the computational burden for estimating the posterior distribution of structural parameters. Finally, we analyze the model's behavior in large networks, to study identification and bound the algorithm's speed of convergence.

In the basic version of our model, players have preferences over network realizations and individual characteristics. The utility function includes payoffs from direct links, but also from linking externalities: reciprocated links, indirect links and popularity. In each period, a player is randomly drawn from the population and meets another individual, according to a random meeting technology. Upon meeting, the player has the opportunity to revise his link. Before updating his linking strategy, the player receives a random shock to his preferences, unobserved by the econometrician. The dynamic of the model follows a stochastic best-response dynamics, and generates a sequence of directed networks.[4]

The theoretical results provide conditions that guarantee the existence of a potential function, simplifying the analysis of the equilibria of the network formation game.[5] The set of Nash equilibria of the model corresponds to the maxima of the potential function. Under mild restrictions on preferences, meeting technology and standard assumptions on prefer-

[1] For example, see the contributions of Topa (2001); Laschever (2009); Cooley (2010); De Giorgi et al. (2010); Nakajima (2007); Bandiera and Rasul (2006); Conley and Udry (forthcoming); Golub and Jackson (2011); Acemoglu et al. (2011).

[2] See Jackson (2008), Jackson and Wolinsky (1996), Bala and Goyal (2000), Currarini et al. (2009), Currarini et al. (2010), De Marti and Zenou (2009), Echenique et al. (2006) for examples.

[3] See Jackson (2008) for a review of network formation models.

[4] The directed nature of the network is not essential to most of the results in this paper.

[5] See Monderer and Shapley (1996)

ences shocks, we show that the sequence of networks generated by the model is a Markov chain, and it converges to a unique stationary equilibrium distribution. The latter provides the likelihood of observing a specific network realization in the long-run.

We show that in the special case of linear-in-parameters utility functions, the stationary distribution belongs to the discrete exponential family, and our model coincides with some specifications of the exponential random graph model (ERGM).[6] Assuming that the network observed in the data is a realization from the stationary distribution, we can estimate the model using only one network observation.

Estimation of the posterior distribution for the structural preference parameters is computationally very demanding because of the curse of dimensionality: the likelihood is known up to a normalizing constant that cannot be computed even for very small networks.[7] Traditional MCMC techniques like Metropolis-Hastings and Gibbs samplers are infeasible for this model.

The statistical literature on likelihoods with intractable normalizing constants, suggests to use MCMC simulation methods to approximate the normalizing constant.[8] The usual approach is to simulate the network using a local MCMC sampler: at each iteration we randomly pick a link, and we update that link according to a Metropolis-Hastings ratio. We show that such simulation strategy has several convergence problems, using a mix of graph limits theory, large deviations theory and mean-field approximations for the exponential family.[9] While some of these problems were known to practitioners, we extend the approach of Diaconis and Chatterjee (2011) to show precise asymptotic results that allow us to analyze the identification, speed of convergence and feasibility of estimation for the model in some special cases.

We show that for large networks, the normalizing constant solves a variational problem in the space of probability density functions defined in the unit square. Furthermore, in the special case of *homogeneous players and (only) positive externalities* from link formation, this variational problem has closed-form solution. We prove that in such special case the model is asymptotically indistinguishable from a directed Erdos-Renyi model. This implies that for this special case, the identification breaks apart in the large n limit. Furthermore, we can simplify the network sampler by simulating the model as a matrix of Bernoulli variables.

On the other hand, if *at least one of the linking externalities is negative* and sufficiently

[6] See Butts (2009), Snijders (2002), Chandrasekhar and Jackson (2014), Diaconis and Chatterjee (2011), Wasserman and Pattison (1996), Caimo and Friel (2010).

[7] For a small network with 10 players, a state-of-the-art supercomputer may take several years to evaluate the constant at a single parameter value. This makes traditional optimization algorithms infeasible.

[8] See Besag (1974), Geyer and Thompson (1992), Snijders (2002)

[9] The analysis presented here extends to directed network the methodology developed by Diaconis and Chatterjee (2011). The main difficulty in the extension to directed networks is that the regularity conditions used in Diaconis and Chatterjee (2011) are not sufficient in the directed networks, and we need to provide different regularity conditions that guarantee compactness of the metric spaces in which the analysis is performed. We also provide a detailed analysis of the variational problem in special cases, following an approach similar to Radin and Yin (2013) and Aristoff and Zhu (2014). We provide the technical details in Appendix D.

large in magnitude, the variational problem cannot be solved at a directed Erdos-Renyi model. This means that our model does not degenerate into an independent-links model, and the linking externalities can be identified. This case is likely to be more relevant from an empirical point of view, since when we include covariates and there is homophily, some linking externalities will necessarily be negative. In this special case and in the general case the network simulation approach is the only way to approximate the normalizing constant and perform inference.[10]

We also extend the methodology of Bhamidi et al. (2011) to our directed network model, and show that in the case of only positive linking externalities we may be unable to run the simulations in feasible time or the model may converge to the wrong stationary distribution. If we simulate the model using a local MCMC sampler, there is a region of the parameter space in which the model converges to stationarity in exponential time, i.e. in order e^{n^2} steps. This happens because in this region of the parameters, the variational problem has two local maxima and the sampler can get stuck in the local maximum, never reaching the global maximum. We also show that the size of the region of exponentially slow convergence increases with the addition of higher order dependencies in the utility functions, e.g. utility from common connections, or cycles.[11]

We propose a modification of the standard local sampler for network simulations to improve the convergence problems, taking into account the previous theoretical analysis. We construct a network sampler that makes large steps, allowing the chain to jump across Nash equilibria more efficiently. In particular, it is crucial that the algorithm is allowed to make steps whose size is proportional to the number of players n.[12] If we consider a large step of fixed dimension (say 20 links per iteration), there is a large enough n for which such sampler becomes *local* with respect to the size of the network. We show evidence that such small modification implies convergence to the correct stationary distribution. These larger steps allow the sampler to jump to the global maximum of the variational problem, without getting stuck in the local maxima. However, the cost of such modification is an increase in the computational burden, that limits our technique to networks of moderate size ($n \approx 300/500$).

We estimate the posterior distribution of the parameters using an approximate exchange algorithm (Murray et al. (2006)) that samples from the posterior in two steps: first, it proposes a new parameter value as in a standard Metropolis-Hastings samplers; second, given the proposed parameter, it simulates the model using the modified network simulation algorithm. We prove that this approximate algorithm is ergodic and converges to the correct posterior distribution.

We test the algorithms with artificial data, showing that for moderate size networks, we

[10] An alternative approach consists of solving the variational problem directly, perhaps through some discretization and approximation. However, at the time of writing it is not clear whether such an approach is computationally less demanding than the simulation method. See He and Zheng (2013) and Mele (2015) for simple examples.

[11] The inclusion of triangles it is known in the ERGM literature to create problems for the convergence of the simulations. We provide complimentary results to Diaconis and Chatterjee (2011) that explain this phenomenon.

[12] In other words it is crucial that the size of the step is not $o(n)$.

have good convergence properties, even with a relatively moderate amount of simulations.

This paper contributes to the literature on empirical models of network formation in several dimensions. The challenges that lead to multiple equilibria and the curse of dimensionality have been addressed in different ways, e.g. modeling the network formation as a sequential process (Christakis et al. (2010)), restricting the type of externalities considered (Miyauchi (2012), or using subnetworks as the unit of analysis (Sheng (2012), Chandrasekhar and Jackson (2014)). Others have focused on the observable implications of homophily (Boucher (2013)) or modeled the network formation as a game with imperfect information (Leung (2014*b*)). Our model considers a sequential network formation process with complete information and restricts the preferences to guarantee the existence of a potential function. While the characterization using potential games has been considered in previous work (Jackson and Watts (2001), Gilles and Sarangi (2004), Butts (2009)), we show that this modeling strategy reduces the computational complexity of the simulations, because allows us to simulate changes in the potential levels, without keeping track of all the players.

Modeling the network formation externalities jointly with unobserved heterogeneity is challenging. Indeed, Graham (2014) provides frequentist inference for a model with unobserved heterogeneity, but rules out the network formation externalities that are crucial in our model. We abstract from unobserved heterogeneity, which can be included in our model with substantial additional computational effort. However, it is not clear whether it is possible to separately identify unobserved heterogeneity from externalities using a single observation of the network (Graham (2014)).

The literature considers identification in two settings. In the *many networks* asymptotics, the researcher observes multiple networks (Miyauchi (2012), Sheng (2012)). In the *large network* asymptotics the econometrician observes only one single network, perhaps large (Chandrasekhar and Jackson (2014), Graham (2014), Leung (2014*a*), DePaula et al. (2014)). Our model is trivially identified in the many networks framework, because the likelihood belongs to the exponential family. The case of large networks is more complicated and we provide an analysis for several special cases.

Our model generates a *dense* network, i.e. the probability of linking does not converge to zero as the number of players grows large (Diaconis and Chatterjee (2011), Lovasz (2012), Graham (2014)). Chandrasekhar and Jackson (2014) show that when we impose sparsity, estimation of structural parameters is simpler in many specifications. DePaula et al. (2014) show that sparsity is crucial for identification. In our model, we can impose a certain degree of sparsity by forcing a link externality to be negative, and we show that such model does not converge to an independent links model, thus guaranteeing identification of the link externalities. Badev (2013) extends our model to include both binary actions and network formation, with an application to smoking among teenagers. Hsieh and Lee (2012) and Goldsmith-Pinkham and Imbens (2013) consider similar models.

The paper proceeds as follows. Section 2 describes the model, the equilibrium characterization and the relation to ERGMs. In Section 3 we discuss the simulation methods for network sampling and posterior sampling, providing a detailed analysis of the model's behavior for large networks. In section 4 we show the estimation results with artificial data

and Section 5 concludes.

2 A Model of Network Formation

2.1 Setup

Let $\mathcal{I} = \{1, 2, ..., n\}$ be the set of agents, each identified by a vector of A (exogenous) characteristics $X_i = \{X_{i1}, ..., X_{iA}\}$, e.g. gender, wealth, age, location, etc. Let the matrix $X = \{X_1, X_2, ..., X_n\}$ collect the vectors of characteristics for the population and let $\mathcal{X}$ denote the set of all possible matrices X. Time is discrete.

The social network is represented as a $n \times n$ binary matrix $G \in \mathcal{G}$, where $\mathcal{G}$ is the set of all $n \times n$ binary matrices. The entry g_{ij} is equal to 1 if individual i forms a connection to individual j, and 0 otherwise; by convention $g_{ii} = 0$, for any i. The network G is *directed*, i.e. $g_{ij} = 1$ does not necessarily imply $g_{ji} = 1$.[13]

Let the *realization* of the network at time t be denoted as g^t and the *realization* of the link between i and j at time t be g^t_{ij}. The network including all the current links but g^t_{ij}, i.e. $g^t \backslash g^t_{ij}$, is denoted as g^t_{-ij}; while g^t_{-i} denotes the network matrix excluding the i-th row (i.e. all the links of player i).

2.1.1 Preferences

The utility of player i from a network g and population attributes $X = (X_1, ..., X_n)$ at parameter θ is given by

$$U_i(g, X; \theta) = \underbrace{\sum_{j=1}^{n} g_{ij} u_{ij}^{\theta_u}}_{\text{direct links}} + \underbrace{\sum_{j=1}^{n} g_{ij} g_{ji} m_{ij}^{\theta_m}}_{\text{mutual links}} + \underbrace{\sum_{j=1}^{n} g_{ij} \sum_{\substack{k=1 \\ k \neq i,j}}^{n} g_{jk} v_{ik}^{\theta_v}}_{\text{indirect links}} + \underbrace{\sum_{j=1}^{n} g_{ij} \sum_{\substack{k=1 \\ k \neq i,j}}^{n} g_{ki} w_{kj}^{\theta_w}}_{\text{popularity}} \quad (1)$$

where $u_{ij}^{\theta_u} \equiv u(X_i, X_j; \theta_u)$, $m_{ij}^{\theta_m} \equiv m(X_i, X_j; \theta_m)$, $v_{ij}^{\theta_v} \equiv v(X_i, X_j; \theta_v)$ and $w_{ij}^{\theta_w} \equiv w(X_i, X_j; \theta_w)$ are (bounded) real-valued functions of the attributes. The utility of the network is the sum of the net benefits received from each link. The total benefit from an *additional link* has four components.

When player i creates a link to agent j, he receives a *direct* net benefit $u_{ij}^{\theta_u}$ that includes both costs and benefits from the relationship. The net benefit can possibly be negative, e.g. when only homophily enters payoffs of direct links, the net utility $u_{ij}^{\theta_u}$ is positive if i and j belong to the same group, while it is negative when they are of different types.

Players value linking externalities, i.e. links formed by other players. A player receives additional utility $m_{ij}^{\theta_m}$ if the link is mutual; a connection has different value when the other party reciprocates.

Players value the composition of indirect connections. When i is deciding whether to

[13] The assumption of directed networks is not crucial to many of the results.

create a link to j, she observes j's connections and their socioeconomic characteristics. Each of j's links provides additional utility $v(X_i, X_k; \theta_v)$ to i. To be concrete, suppose there are only two types: A and B. In this model, an agent who has the opportunity to form an additional link, values a type-A individual with three links to type-B agents as a different good than a type-A individual with two type-A connections and one type-B connection.[14] In other words, individuals value both *exogenous* heterogeneity and *endogenous* heterogeneity: the former is determined by the socioeconomic characteristics of the agents, while the latter arises endogenously with the process of network formation. In the baseline version of the model we assume that only indirect links are valuable and they are perfect substitutes: individuals do not receive utility from two-links-away contacts.[15]

The fourth component corresponds to a *popularity effect*. If individual i links j, he automatically creates an indirect link for the all the agents that had a link to i. Thus i generates an externality (positive or negative) for each k that formed a link to him in previous periods. This externality makes i more or less popular.

2.1.2 Network Formation Process

The process of network formation follows a *stochastic best-response dynamics* (Blume (1993)), generating a Markov chain of networks. The main ingredients of this process are random meetings and utility maximization. The implicit assumption is that meetings are very frequent, and the players can revise their linking strategies often.

Meeting Technology. At the beginning of each period a player i is randomly selected from the population, and he meets individual j, according to a meeting technology. The *meeting process* is a stochastic sequence $m = \{m^t\}_{t=1}^{\infty}$ with support $\mathcal{I} \times \mathcal{I}$. The realizations of the meeting process are ordered pairs $m^t = \{i, j\}$, indicating which agent i should play and which link g_{ij} can be updated at period t.[16]

The probability that player i is randomly chosen from the population and meets agent j is defined as

$$\Pr\left(m^t = ij | g^{t-1}, X\right) = \rho\left(g^{t-1}, X_i, X_j\right) \tag{2}$$

where $\sum_{i=1}^{n}\sum_{j=1}^{n}\rho\left(g, X_i, X_j\right) = 1$ for any $g \in \mathcal{G}$. The meeting probability depends on the current network g (e.g. the existence of a common link between i and j) and the characteristics of the pair. This general formulation includes meeting technologies with

[14] A similar assumption is used in De Marti and Zenou (2009) where the agents' cost of linking depends on the racial composition of friends of friends. Their model is an extension of the connection model of Jackson and Wolinsky (1996), and the links are formed with mutual consent. The corresponding network is undirected.

[15] This benchmark model can be extended to incorporate additional utility components, as shown below.

[16] Several models incorporate a meeting technology in the network formation process. Jackson and Watts (2002) assume individuals meet randomly according to a discrete uniform distribution. Currarini et al. (2009) introduce a matching process that is biased towards individuals of the same type. Christakis et al. (2010) develop a dynamic model, where the sequence of meetings determines which players have the opportunity to form a link in each period.

a bias for same-type individuals as in Currarini et al. (2009). The simplest example of meeting technology is an i.i.d. discrete uniform process with $\rho\left(g^{t-1}, X_i, X_j\right) = \frac{1}{n(n-1)}$. An example with bias for same-type agents is $\rho\left(g^{t-1}, X_i, X_j\right) \propto \exp\left[-d\left(X_i, X_j\right)\right]$, where $d\left(\cdot,\cdot\right)$ is a distance function.

Utility Maximization. Conditional on the meeting $m^t = ij$, player i updates the link g_{ij} to maximize his current utility, taking the existing network g^t_{-ij} as given. We assume that the agents do not take into account the effect of their linking strategy on the future evolution of the network. The players have *complete information*, since they can observe the entire network and the individual attributes of all agents.[17] Before updating his link to j, individual i receives an idiosyncratic shock $\varepsilon \sim F\left(\varepsilon\right)$ to his preferences that the econometrician cannot observe. This shock models unobservables that could influence the utility of an additional link. Player i links agent j at time t if and only if it is a best response to the current network configuration, i.e. $g^t_{ij} = 1$ if and only if

$$U_i\left(g^t_{ij} = 1, g^{t-1}_{-ij}, X; \theta\right) + \varepsilon_{1t} \geq U_i\left(g^t_{ij} = 0, g^{t-1}_{-ij}, X; \theta\right) + \varepsilon_{0t}. \tag{3}$$

We assume that when the equality holds, the agent plays the status quo.[18] The network formation process generates a Markov chain of networks, with transition probabilities determined by the meeting process and agents' linking choices.

2.2 Equilibrium Analysis

We impose an additional assumption on the functional forms of the utility functions, which provides important equilibrium and identification restrictions. We assume that the utility $m^{\theta_m}_{ij}$ obtained from mutual links is symmetric, and that the utility of an indirect link $v^{\theta_v}_{ij}$ has the same functional form as the utility from the popularity effect $w^{\theta_v}_{ij}$.

ASSUMPTION 1 ***(Preferences)*** *The preferences satisfy the following restrictions*

$$\begin{aligned} m\left(X_i, X_j; \theta_m\right) &= m\left(X_j, X_i; \theta_m\right) \text{ for all } i, j \in \mathcal{I} \\ w\left(X_k, X_j; \theta_v\right) &= v\left(X_k, X_j; \theta_v\right) \text{ for all } k, j \in \mathcal{I} \end{aligned}$$

The symmetry in $m_{ij}(\theta_m)$ does not imply that a mutual link between i and j gives both the same utility. If i and j have a mutual link, they receive the same common utility component $(m_{ij}(\theta_m))$ but they may receive different payoffs from direct or indirect links. Two individuals with the same exogenous characteristics $X_i = X_j$ who form a mutual link receive the same $u_{ij}(\theta_u)$ and $m_{ij}(\theta_m)$, but they may have different payoffs from the additional link because of the composition of their indirect contacts and their popularity. Therefore, the

[17] More precisely, to make a decision about linking, the player needs to observe his in-links and the out-links of his friends.

[18] This assumption does not affect the main result and is relevant only when the distribution of the preference shocks is discrete.

first part of the assumption is necessary for identification of the utility from indirect links and popularity.

The second part of the assumption imposes an identifying restriction to the externality generated by i when creating a link to j: any individual k that has formed a link to i, has an additional indirect contact, i.e. j, who agent k values by an amount $w\left(X_k, X_j; (\theta_w)\right)$. When $w\left(X_k, X_j; (\theta_v)\right) = v\left(X_k, X_j; (\theta_v)\right)$, an individual i values his popularity effect as much as k values the indirect link to j, i.e., i internalizes the externality he creates.

Assumption 1[19] is the main ingredient that allows us to characterize the network formation as a potential game (see also Butts (2009) and Chandrasekhar and Jackson (2014) for similar characterizations).

PROPOSITION 1 ***(Existence of a Potential Function)*** *Under Assumption 1, the deterministic component of the incentives of any player in any state of the network are summarized by a* ***potential function****,* $Q : \mathcal{G} \times \mathcal{X} \to \mathbb{R}$

$$Q\left(g, X; \theta\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}u_{ij}(\theta_u) + \sum_{i=1}^{n}\sum_{j>i}^{n} g_{ij}g_{ji}m_{ij}(\theta_m) + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq i,j}}^{n} g_{ij}g_{jk}v_{ik}(\theta_v), \qquad (4)$$

and the network formation game is a Potential Game.

Proof. See Appendix A ■

The intuition for the result is simple.[20] Under the restrictions of Assumption 1, for any player i and any link g_{ij} we have

$$Q\left(g_{ij}, g_{-ij}, X; \theta\right) - Q\left(1 - g_{ij}, g_{-ij}, X; \theta\right) = U_i\left(g_{ij}, g_{-ij}, X; \theta\right) - U_i\left(1 - g_{ij}, g_{-ij}, X; \theta\right)$$

Consider two networks, $g = (g_{ij}, g_{-ij})$ and $g' = (1 - g_{ij}, g_{-ij})$, that differ only with respect to one link, g_{ij}, chosen by individual i: the difference in utility that agent i receives from the two networks, $U_i\left(g, X; \theta\right) - U_i\left(g', X; \theta\right)$, is exactly equal to the difference of the *potential* function evaluated at the two networks, $Q\left(g, X; \theta\right) - Q\left(g', X; \theta\right)$. That is, the potential is an aggregate function that summarizes both the state of the network and the deterministic incentives of the players in each state.

Characterizing the network formation as a potential game facilitates the analysis and the simulations. To compute the equilibria of the model, there is no need to keep track of each

[19] The first part of the assumption is a normalization of the utility function that allows identification for the utility of indirect links and popularity. The second part of the assumption is an identification restriction, that guarantees the model's coherency in the sense of Tamer (2003). In simple words, this part of the assumption guarantees that the system of conditional linking probabilities implied by the model generates a proper joint distribution of the network matrix.Similar restrictions are also encountered in spatial econometrics models (Besag, 1974) and in the literature on qualitative response models (Heckman, 1978; Amemiya, 1981)

[20] See Monderer and Shapley (1996) for definitions and properties of potential games.

player's behavior: the potential function contains all the relevant information.[21]

To analyze the long run behavior of the model, I impose more structure on the meeting technology.[22]

ASSUMPTION 2 ***(Meeting Process)*** *The meeting probability between i and j does not depend on the existence of a link between them, and each meeting has a positive probability of occurring, i.e.* $\rho(g^{t-1}, X_i, X_j) = \rho(g^{t-1}_{-ij}, X_i, X_j) > 0$ *for any* $ij \in \mathcal{I} \times \mathcal{I}$

The meeting process is such that any player can be chosen and any pair of agents can meet. This assumption guarantees that any equilibrium network can be reached with positive probability. For example, a discrete uniform distribution satisfies this assumption. The other restriction is for identification purposes: if we allow ρ to depend on the current link between i and j, then we cannot write the likelihood in closed form. Using data from a single network observation it is impossible to identify the function ρ unless we make very restricting assumptions.

A Nash equilibrium is a network in which any player has no profitable deviations from his current linking strategy, when randomly selected from the population. We can show that the set of Nash networks corresponds to the local maxima of the potential function. Suppose that the current network is a Nash network. As a consequence, if a player deviates from the current linking strategy, he receives less utility.[23] Since the change in utility for any agent is equivalent to the change in potential, any deviation from the Nash network must decrease the potential. It follows that the Nash network must be a local maximizer of the potential function over the set of networks that differ from the current network for at most one link.

In the absence of preference shocks, the consequences of assumptions 1 and 2 are that the model will evolve according to a Markov Chain, converging to one of the Nash networks with probability one (see formal details in Appendix A). Suppose a player is drawn from the meeting process. Such agent will play a best response to the current network configuration. Therefore, his utility cannot decrease. This holds for any player and any period. It follows that the potential is nondecreasing over time. Since there is a finite number of possible networks, in the long run, the sequence of networks must reach a local maximum of the potential, i.e., a Nash equilibrium.

[21] This property is key for the analysis of networks with many players: the usual check for existence of profitable deviations from the Nash equilibrium can be performed using the potential, instead of checking each player's possible deviation in sequence. The computation of all profitable deviations for each player involves $n(n-1)2^{n(n-1)}$ operations: each player has $n-1$ possible deviations, there are n players and a total of $2^{n(n-1)}$ possible network configurations. As it is shown below (Proposition 2), when the game is a potential game, the computation of all Nash equilibria is equivalent to finding the local maxima of the potential function. This corresponds to evalutating the potential function for all the $2^{n(n-1)}$ possible network structures. The latter task involves fewer operations by a factor of $n(n-1)$, thus decreasing the computational burden.

[22] Christakis et al. (2010) assume that individuals can meet only once and their links remain in place forever. This assumption is convenient when estimating a large network, but it does not allow the characterization of the stationary equilibrium.

[23] When the utility from the equilibrium and the deviation is the same, the agent plays the status quo, i.e., the Nash strategy.

The following standard parametric assumption on the shocks allows us to characterize the stationary distribution and transition probabilities.

ASSUMPTION 3 ***(Idiosyncratic Shocks)*** *The shock follows a Type I extreme value distribution, i.i.d. among links and across time.*

Under Assumptions 1-3, the network evolves as a Markov chain with transition probabilities given by the conditional choice probabilities and the probability law of the meeting process m^t. One can easily show that the sequence $[g^0, g^1, ..., g^t]$ is *irreducible* and *aperiodic*.[24] The following theorem summarizes the main theoretical result.

THEOREM 1 ***(Uniqueness and Characterization of Stationary Equilibrium)*** *Consider the network formation game with idiosyncratic shocks, under Assumptions 1-3.*

1. *There exists a unique stationary distribution* $\pi(g, X; \theta)$
2. *The stationary distribution* $\pi(g, X; \theta)$ *is*

$$\pi\left(g, X; \theta\right) = \frac{\exp\left[Q\left(g, X; \theta\right)\right]}{\sum_{\omega \in \mathcal{G}} \exp\left[Q\left(\omega, X; \theta\right)\right]}, \tag{5}$$

where $Q\left(g, X; \theta\right)$ *is the potential function (4).*

Proof. In Appendix A ■

The first part of the proposition follows directly from the irreducibility and aperiodicity of the Markov process generated by the network formation game. The uniqueness of the stationary distribution is crucial in estimation, since one does not need to worry about multiple equilibria. Furthermore, the stationary equilibrium characterizes the likelihood of observing a specific network configuration in the data. As a consequence, we can estimate the structural parameters from observations of only *one network at a specific point in time*, under the assumption that the observed network is drawn from the stationary equilibrium.

The second part of the proposition provides a closed-form solution for the stationary distribution. The latter can be interpreted as the probability of observing a specific network structure, when the network is observed in the long run. In the long run, the system of interacting agents will visit more often those states/networks that have high potential. Therefore a high proportion of the possible networks generated by the network formation game, will correspond to Nash networks.

The stationary distribution $\pi\left(g, X; \theta\right)$ includes a normalizing constant

$$c\left(\mathcal{G}, X; \theta\right) \equiv \sum_{\omega \in \mathcal{G}} \exp\left[Q\left(\omega, X; \theta\right)\right] \tag{6}$$

[24] Intuitively, since the meeting probability $\Pr\left(m^t = ij\right) > 0$ for all ij, there is always a positive probability of reaching a new network in which the link g_{ij} can be updated. The logistic shock assumption implies that there is always a positive probability of switching to another state of the network, thus eliminating absorbing states.

that guarantees that (5) is a proper probability distribution. Unfortunately, this normalizing constant greatly complicates estimation, since it cannot be evaluated exactly or approximated with precision. The details about how this problem is circumvented are presented in the empirical strategy section.

2.3 Relation with Exponential Random Graphs

The Exponential Random Graph Model (ERGM) is a statistical model of random network formation, with complex dependencies among links. This class of models posits that the probability of observing a specific network is proportional to an exponential function of a linear combination of network statistics. Exponential random graphs have been successfully used to fit social network data, providing a useful benchmark for alternative models. However, as any random network formation model, they lack the equilibrium micro-foundations of the strategic literature.[25]

PROPOSITION 2 ***(Exponential Family Likelihood)***
Let Assumptions 1-3 hold. If the utility functions are linear in parameters, the stationary distribution $\pi(g, X; \theta)$ *belongs to the* ***exponential family****, i.e., it can be written in the form*

$$\pi(g, X; \theta) = \frac{\exp\left[\theta' \mathbf{t}(g, X)\right]}{\sum_{\omega \in \mathcal{G}} \exp\left[\theta' \mathbf{t}(\omega, X)\right]}, \tag{7}$$

where $\theta = (\theta_u, \theta_m, \theta_v)'$ *is a (column) vector of parameters and* $\mathbf{t}(g, X)$ *is a (column) vector of canonical statistics.*

Proof. See Appendix A ■

The vector $\mathbf{t}(g, X) = (t_1(g, X), ..., t_K(g, X))$ is a vector of sufficient statistics for the network formation model. This vector may include the number of links, the number of whites-to-whites links, the number of male-to-female links and so on.

This likelihood is analogous to the one of exponential random graph models: we can interpret *some specifications* of ERGMs as the stationary equilibrium of a strategic game of network formation, where myopic agents follow a stochastic best response dynamics and utilities are linear functions of the parameters. Not all the ERGM specifications are necessarily compatible with our model, as explained in the extensions below.

[25] Frank and Strauss (1986) developed the theory of Markov random graphs. These are models of random network formation in which there is dependence among links: the probability that a link occurs depends on the existence of other links. Wasserman and Pattison (1996) generalized the Markov random graphs to more general dependencies, developing the Exponential Random graph models. Snijders (2002) reviews these models and the related estimation techniques.

2.4 Extensions and discussion

Additional utility components. It is possible to modify the baseline utility function (1) to include additional components. For example, one may be interested in studying preferences that include utility from cyclic triangles effects, i.e. individual i links to j, j connects to k and k links to i. The latter can be modeled as a component of the utility τ that varies with the characteristics of the three players involved in the relationships, i.e. $\tau(X_i, X_j, X_k; \theta_\tau)$ for all $i, j, k \in \mathcal{I}$. The utility is easily modified by including a term $\sum_{j=1}^{n} g_{ij} \sum_{k \neq i,j} g_{jk} g_{ki} \tau_{ijk}(\theta_\tau)$. However, to guarantee the existence of a potential function, we need to restrict τ in analogous way as in Assumption 1: the function τ must satisfy $\tau_{ijk}(\theta_\tau) = \tau_{i'j'k'}(\theta_\tau)$ for any i', j', k' permutation of i, j, k. The potential is easily computed as before, by adding the term $\frac{1}{3} \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} \sum_{k \neq i,j} g_{jk} g_{ki} \tau_{ijk}(\theta_\tau)$.

In general, it is possible to include additional utility components to (1) as long as we can find restrictions on the payoffs that guarantee the existence of a potential function. Some examples are provided in the proofs of Appendix D.

Undirected networks. The model is concerned about directed networks, but this is not essential to most of the characterizations. The results about the existence of the potential game, the existence and characterization of the stationary distribution and the relation with the ERGM model can be extended to undirected networks with minimal modifications (see Chandrasekhar and Jackson (2014)).[26] Most of the asymptotic and convergence results in the next section hold also for undirected networks (see Diaconis and Chatterjee (2011)).

Sparsity. The model generates *dense networks*, i.e. each player can potentially form all his $n-1$ links. This means that as $n \to \infty$ the unconditional probability of a link does not become vanishingly small (see Lovasz (2012)). Chandrasekhar and Jackson (2014) show that assuming sparsity reduces the computational complexity of estimation and it implies good statistical properties (e.g. consistency). In this model we can create sparsity with a simple modification, i.e. a parameter $\theta_p^{(n)} \to -\infty$ as $n \to \infty$.

3 Empirical Strategy

3.1 Computational Problem

Estimation and inference are complicated by the structure of the likelihood function, which is known up to the normalizing constant (6), i.e. $c(\mathcal{G}, X, \theta) = \sum_{\omega \in \mathcal{G}} \exp[Q(\omega, X, \theta)]$. To compute the latter constant at parameter vector θ for a network of n players, we would need to sum over all $2^{n(n-1)}$ possible network configurations the value of the potential function. For example, for a small network of $n = 10$ players, there are $2^{90} \backsimeq 10^{27}$ network configurations. A supercomputer that can compute 10^{12} potential functions in one second would take almost 40 million years to compute the constant.

[26] It is also possible to include binary actions (e.g. decision to smoke) into the model, as in Badev (2013).

Therefore standard maximum likelihood maximization routines are impractical. A standard Bayesian estimation approach would encounter the same challenges. Let $p(\theta)$ be the prior distribution, and let the likelihood function of the observed data (g, X) be the long-run stationary distribution of the model $\pi(g, X, \theta)$. The posterior distribution of θ is

$$p(\theta|g, X) = \frac{\pi(g, X, \theta)\, p(\theta)}{\int_\Theta \pi(g, X, \theta)\, p(\theta)\, d\theta}. \tag{8}$$

Using a standard Metropolis-Hastings algorithm to sample from this posterior, we would have to compute ratios

$$\begin{aligned} \alpha(\theta, \theta') &= \min\left\{1, \frac{p(\theta'|g, X)\, q_\theta(\theta|\theta')}{p(\theta|g, X)\, q_\theta(\theta'|\theta)}\right\} \\ &= \min\left\{1, \frac{\exp[Q(g, X, \theta')]}{\exp[Q(g, X, \theta)]} \frac{c(\mathcal{G}, X, \theta)}{c(\mathcal{G}, X, \theta')} \frac{p(\theta')}{p(\theta)} \frac{q_\theta(\theta|\theta')}{q_\theta(\theta'|\theta)}\right\}. \end{aligned}$$

The issue is the computation of the likelihood ratios, which involve evaluation of the constant that is infeasible.

3.2 Network simulations

This computational problem is common to many models in the statistical literature. The usual approach is to provide an approximation of the normalizing constant and the likelihood, using Markov Chain Monte Carlo simulation methods.[27] For a fixed parameter value θ, the algorithm simulates a Markov chain of networks whose unique invariant distribution is (5).

ALGORITHM 1 ***Metropolis-Hastings for Network Simulations***
Fix a parameter vector θ. At iteration r, with current network g_r

1. *Propose a network g' from a proposal distribution $g' \sim q_g(g'|g_r)$*
2. *Compute the Metropolis-Hastings ratio*

$$\alpha_{mh}(g_r, g') = \min\left\{1, \frac{\exp[Q(g', X, \theta)]}{\exp[Q(g_r, X, \theta)]} \frac{q_g(g_r|g')}{q_g(g'|g_r)}\right\} \tag{9}$$

3. *Update the network according to*

$$g_{r+1} = \begin{cases} g' & \text{with prob. } \alpha_{mh}(g_r, g') \\ g_r & \text{with prob. } 1 - \alpha_{mh}(g_r, g') \end{cases} \tag{10}$$

[27] The algorithm used in this paper is similar to the Metropolis-Hastings algorithm proposed in Snijders (2002).

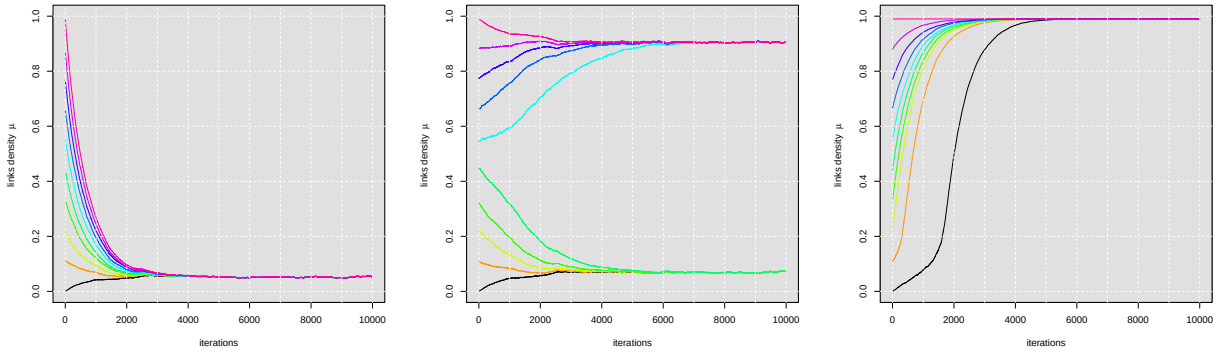
At each iteration of ALGORITHM 1 a random network g' is proposed, and the update is accepted with probability $\alpha_{mh}(g_r, g')$. The main advantage of this simulation strategy is that the acceptance ratio $\alpha_{mh}(g_r, g')$ does not contain the normalizing constant $c(\mathcal{G}, X, \theta)$ of the stationary distribution. Each quantity in the acceptance ratio can be computed exactly. The Metropolis-Hastings structure of the algorithm guarantees convergence. Standard results[28] show that the chain generated by the algorithm converges uniformly to the likelihood of the model.

The standard version of this algorithm is a *local sampler*: at each iteration, we select a random player i with probability $1/n$, we then select another player j with probability $1/(n-1)$, and we update the link g_{ij} according to the Metropolis-Hastings ratio (9).

However, this algorithm has several practical convergence problems. To be concrete, let's implement the local sampler in a special case with homogeneous players, that includes only direct utility and indirect utility.

$$\pi_n(g; \alpha, \beta) = \frac{\exp\left\{\left[\alpha \sum_{i=1}^n \sum_{j=1}^n g_{ij} + \beta \sum_{i=1}^n \sum_{j=1}^n \sum_{k \neq i}^n g_{ij} g_{jk}\right]\right\}}{c(\alpha, \beta, \mathcal{G}_n)} \tag{11}$$

Figure 1: Network simulations at different parameter values



(A) (B) (C)

Traceplots of simulations of model (11) using Algorithm 1 with local chains. The simulations are obtained for a network with $n = 100$ players, with parameters $\alpha = -3$ and $\beta = \{1/n, 3/n, 7/n\}$ (Panel (A), (B) and (C) respectively). Each simulation is started at 10 different starting networks, each corresponding to a directed Erdos-Reny network with probability of link $\mu = \{0, .111, .222, .333, .444, .555, .666, .777, .888, 1\}$.

We simulate this model using the local sampler just described. In Figure 1 we show the trace plot of algorithm 1 for three different parameter vectors: $\alpha = -3$ and $\beta = \{1/n, 3/n, 7/n\}$ (Panels (A), (B) and (C) respectively). We start the simulations at 10 different starting values, each corresponding to a directed Erdos-Renyi with probability of

[28] See Meyn and Tweedie (2009), Levin et al. (2008)

link $\mu = \{0, .111, .222, .333, .444, .555, .666, .777, .888, 1\}$. In the figures we show the link density of each iteration.[29] The network has $n = 100$ players.

The simulations (A) with parameters $(\alpha, \beta) = (-3, 1/n)$ converge to a very sparse network; while the simulations (C) with parameters $(\alpha, \beta) = (-3, 7/n)$ converge to a very dense network. On the other hand, when we consider simulations in (B) with parameters $(\alpha, \beta) = (-3, 3/n)$, we observe that the chains started at relatively dense networks converge to a very dense network with density of links $\mu_2 \approx 0.92$, while chains started at relatively sparse networks converge to a sparse network, with link density $\mu_1 \approx 0.07$.

This is a phenomenon that practitioners have encountered in the ERGM literature and in statistical physics models.[30] The model seems to put very large probability mass on few networks, an issue called degeneracy. In the next section we provide several theoretical results that explain the issue, and we propose a modification of the local sampler that decreases this problem.

3.3 Large network analysis

There are two ways to study the asymptotic properties of empirical network formation models. First, we can consider a sample of independent networks and study the properties of the model as the number of networks grows large (*many networks asymptotics*). Second, we can consider a single network observation, and a sequence of graphs whose number of players n grows large *(large networks asymptotics*). The former case is relatively standard and follows from the theory of exponential families under usual regularity conditions.[31] Identification of the parameters is also standard.

The latter case of large networks is relatively more complicated, but has recently gained attention in the literature.[32] We provide a detailed asymptotic analysis of the model in the homogeneous players case.[33]

Consider a sequence of directed graphs g_n, where the number of nodes $n \to \infty$. To consider such network limits, we re-scale the potential function, to avoid exploding terms as $n \to \infty$: each aggregate utility term is scaled by a factor $n^{v(H)}$, where $v(H)$ is the number of vertices involved in the utility term. For example, if we consider the direct utility of links, the building piece is a link g_{ij} and the corresponding aggregate term in the potential function is $\sum_i \sum_j g_{ij}$, i.e. the total number of links. Each link can be interpreted as a small network H_1 including only 2 vertices, so $v(H_1) = 2$ and we end up with a rescaled statistics $t(H_1, g) = n^{-2} \sum_i \sum_j g_{ij}$. If we are considering the indirect utility term $H_2 = g_{ij} g_{jk}$, we have $v(H_2) = 3$ and we rescale the aggregate term in the potential by n^3,

[29] The traceplot for the density of indirect links (the second network statistics) has similar pattern.

[30] See Snijders (2002), Butts (2009), Koskinen (2008) for examples.

[31] See Lehman (1983), Sheng (2012), Badev (2013).

[32] See Chandrasekhar and Jackson (2014), Graham (2014), Leung (2014*a*), DePaula et al. (2014) for recent contributions.

[33] The explanation that follows is relatively informal, and we leave the technical details about graph limits, large deviations and mean-field approximations in Appendix D.

i.e. $t(H_2, g) = n^{-3} \sum_i \sum_j \sum_k g_{ij} g_{jk}$. In summary

$$t(H_1, g) = \frac{1}{n^2} \sum_i \sum_j g_{ij} \qquad \text{and} \qquad t(H_2, g) = \frac{1}{n^3} \sum_i \sum_j \sum_k g_{ij} g_{jk}$$

When n becomes large, it is convenient to consider the population of nodes as a continuum interval $[0, 1]$; intuitively for large n the adjacency matrix is replaced by a function $h : [0, 1]^2 \to [0, 1]$ where $h(x, y)$ indicates the probability that there is a directed edge from x to y. The goal of the analysis is to characterize the behavior of the network statistics in the limit $n \to \infty$. For the terms $t(H_1, g)$ and $t(H_2, g)$ we have

$$\begin{aligned} t(H_1, g) &\to t(H_1, h) \equiv \int_{[0,1]^2} h(x, y) dx dy \\ t(H_2, g) &\to t(H_2, h) \equiv \int_{[0,1]^3} h(x, y) h(y, z) dx dy dz \end{aligned}$$

From the re-scaled potential function $\mathcal{T}(g)$ we can derive the potential function corresponding to the continuum of players $\mathcal{T}(h)$. If for a discrete n we have a re-scaled potential $\mathcal{T}(g)$

$$\mathcal{T}(g) = \alpha t(H_1, g) + \beta t(H_2, g) \tag{12}$$

when $n \to \infty$, the potential function becomes

$$\mathcal{T}(h) = \alpha t(H_1, h) + \beta t(H_2, h)$$

The previous ideas can be generalized. The potential is essentially a weighted sum of P aggregate utility terms that depend on subgraphs H_p of the network g. Each weight is a parameter θ_p to estimate, so $\mathcal{T}(g)$ and $\mathcal{T}(h)$ are defined as

$$\mathcal{T}(g) = \sum_{p=1}^{P} \theta_p t(H_p, g) \qquad \text{and} \qquad \mathcal{T}(h) = \sum_{p=1}^{P} \theta_p t(H_p, h)$$

and each term $t(H_p, h)$ is defined as

$$t(H_p, h) = \int_{[0,1]^{v(H_p)}} \prod_{(i,j) \in E(H_p)} h(x_i, x_j) dx_1 \cdots dx_{v(H_p)}$$

where $E(H_p)$ indicates the set of links included in the subgraph H_p. We can re-scale the potential of the model (11) with direct utility and indirect utility, to get the probability of observing g in the stationary equilibrium as

$$\begin{aligned} \pi_n(g; \alpha, \beta) &= \frac{\exp\left\{n^2 \left[\alpha \frac{\sum_{i=1}^n \sum_{j=1}^n g_{ij}}{n^2} + \beta \frac{\sum_{i=1}^n \sum_{j=1}^n \sum_{k \neq i}^n g_{ij} g_{jk}}{n^3}\right]\right\}}{c(\alpha, \beta, \mathcal{G}_n)} \\ &= \frac{\exp\left\{n^2 \left[\alpha t(H_1, g) + \beta t(H_2, g)\right]\right\}}{c(\alpha, \beta, \mathcal{G}_n)} = \exp\left\{n^2 \left[\mathcal{T}(g) - \psi_n\right]\right\} \end{aligned} \tag{13}$$

where the constant ψ_n is defined as

$$\psi_n = \frac{1}{n^2} \log \sum_{g \in \mathcal{G}_n} \exp\left[n^2 \mathcal{T}(g)\right] \tag{14}$$

Notice that the above model is equivalent to the original model (11) with parameter β re-scaled by n.[34] Let the entropy of the probability density $h(x, y)$ be denoted as $\mathcal{I}(h)$

$$\mathcal{I}(h) \equiv \int_0^1 \int_0^1 \left[h(x,y) \log h(x,y) + (1 - h(x,y)) \log(1 - h(x,y))\right] dxdy \tag{15}$$

The following theorem provides an asymptotic estimate of the normalizing constant, as the solution of a variational problem. This result extend the analogous result for undirected network provided in Diaconis and Chatterjee (2011).[35]

THEOREM 2 *(Asymptotic constant). If $\mathcal{T} : \mathcal{W} \to \mathbb{R}$ is a bounded continuous function and ψ_n and $\mathcal{I}$ are defined as in (14) and (15) respectively, then*

$$\psi \equiv \lim_{n \to \infty} \psi_n = \sup_{h \in \mathcal{W}} \{\mathcal{T}(h) - \mathcal{I}(h)\} \tag{16}$$

Proof. See Theorem 10 in Appendix D ■

The result in the theorem provides a consistent estimator for the log of the normalizing constant. This formulation is the asymptotic analogous of the variational representation of the discrete exponential family in mean parameterization, as shown in Wainwright and Jordan (2008). In general, the variational problem in (16) does not have a closed-form solution, but there are special cases in which the problem becomes tractable.

Using the characterization in Theorem 2, we can provide precise results about identification and convergence of the algorithms, for the special case of homogenous players and no covariates.[36]

THEOREM 3 *The model (13) has the following asymptotic behavior:*

1. *If $\beta \geq 0$, then the networks generated by the model are indistinguishable from a directed Erdos-Renyi graph with linking probability μ^* that solves*

$$\mu = \frac{\exp\left[\alpha + 2\beta\mu\right]}{1 + \exp\left[\alpha + 2\beta\mu\right]} \tag{17}$$

and satisfy $2\beta\mu(1 - \mu) < 1$, for almost all $\alpha \in \mathbb{R}$ and $\beta \geq 0$.

[34] This is important when we run the simulations using the usual ERGM form. For example, we need to use $\beta^o = \frac{\beta}{n}$ for simulations using the `ergm` package in the software R. The same is true for the replication routines of this paper.

[35] The extension to directed networks is not trivial, since the regularity conditions used by Diaconis and Chatterjee (2011) to guaranteed compactness of the space of graphons do not work in the directed case network. In addition, in directed networks the existence of homomorphisms is not guaranteed. See Appendix D for the technical details.

[36] We use the approach developed in Radin and Yin (2013) and Aristoff and Zhu (2014) to study the maximization problem implied by the simplified variational problem.

2. *If $\beta \geq 0$ there exists a continuous and monotone decreasing function $\zeta : (-\infty, -2] \to [2, \infty]$, such that for any $\alpha < -2$, if $\beta = \zeta(\alpha)$, the model generated graph is asymptotically indistinguishable from a mixture of directed Erdos-Renyi graphs with linking probabilities μ_1^* and μ_2^*, such that $\mu_1^* < 0.5 < \mu_2^*$ and both solve equation (17) and satisfy $2\beta\mu(1-\mu) < 1$.*

3. *Let $\beta \geq 0$. Then there exist two functions $S(\phi_1(\alpha))$ and $S(\phi_2(\alpha))$ that delimit a V-shaped region of the parameters (α, β). If $\beta \in (S(\phi_2(\alpha)), S(\phi_1(\alpha)))$ the variational problem (16) has two local maximizers, that correspond to directed Erdos-Renyi graphs with linking probability μ_1^* and μ_2^*, such that $\mu_1^* < 0.5 < \mu_2^*$ and both solve equation (17) and satisfy $2\beta\mu(1-\mu) < 1$. If $\beta \in (S(\phi_2(\alpha)), \zeta(\alpha))$ then μ_1^* is the global maximizer. If $\beta \in (\zeta(\alpha), S(\phi_1(\alpha)))$ then μ_2^* is the global maximizer.*

4. *If $\beta < 0$, then for any $\alpha \in \mathbb{R}$ there exists a positive constant $C(\alpha) > 0$, such that for $\beta < -C(\alpha)$ the model is asymptotically different from a directed Erdos-Renyi model.*

Proof. The first, second and third part follow from Theorem 11 and Theorem 19 in Appendix D. The third part is proven in Theorem 14 in Appendix D. ■

In Figure 2(A) we provide a visualization of the regions described in Theorem 3. The functions $S(\phi_1(\alpha))$ and $S(\phi_2(\alpha))$ delimit a V-shaped region, which contains the function $\zeta(\alpha)$. The functions ϕ_1 and ϕ_2 are defined in the proof of Theorem 11 in Appendix D.

The first part of Theorem 3 shows that identifcation can fail as $n \to \infty$, in a model with *positive externalities*. If $\beta > 0$, a realization of the model with parameters (α, β) such that $\beta \neq \zeta(\alpha)$, will be indistinguishable from the realization of a model with parameters $(\alpha', 0)$ where $\alpha' = \log \frac{\mu^*}{1-\mu^*}$ and μ^* is the unique solution of equation (17).[37]

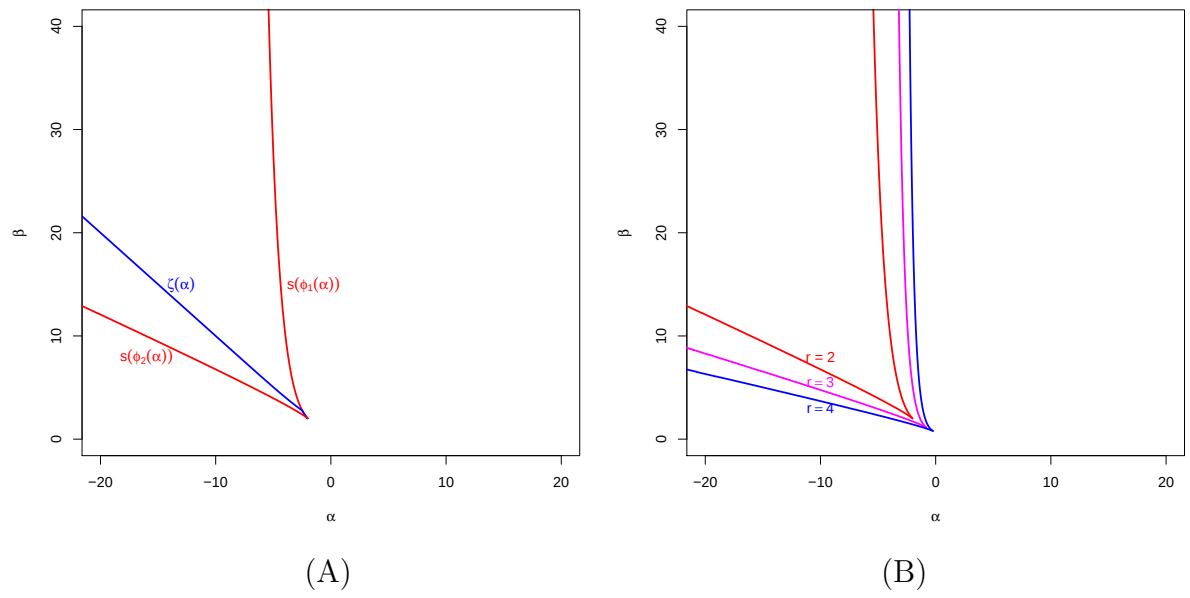
The second part shows that along the function $\beta = \zeta(\alpha)$ the model behaves as a mixture of Erdos-Renyi graphs for large population. This is also an identification problem. In particular, the parameters $(\alpha, \zeta(\alpha))$ can generate two completely different networks, one with link density $\mu_1^* < 0.5$ and one with link density $\mu_2^* > 0.5$.[38] Such problem was observed by practitioners (see Snijders (2002) for example) using simulation methods, and it was proven analytically for undirected networks in Diaconis and Chatterjee (2011). We extend their result to directed networks.

The third statement shows that when the parameters belong to the V-shaped region delimited by $S(\phi_1(\alpha))$ and $S(\phi_2(\alpha))$, we have two local maximizers of (16), but only one of them is global. If the externality is small enough (i.e. $\beta \in (S(\phi_2(\alpha)), \zeta(\alpha))$), the global maximizer corresponds to a *relatively sparse* network; if the externality is large enough (i.e. $\beta \in (\zeta(\alpha), S(\phi_1(\alpha)))$) the corresponding global maximizer is *relatively dense*.

[37] The model with homogeneous players and only positive externalities violates the condition of *expectation-identification* in Chandrasekhar and Jackson (2014). The condition requires that different parameters correspond to different expected network statistics. This is clearly violated in this special case.

[38] In the applied mathematics and physics literature, such sets of parameters are crucial because they generate a phase transition. See for example Radin and Yin (2013).

Figure 2: Visualization of the regions described in Theorem 3



(A) (B)

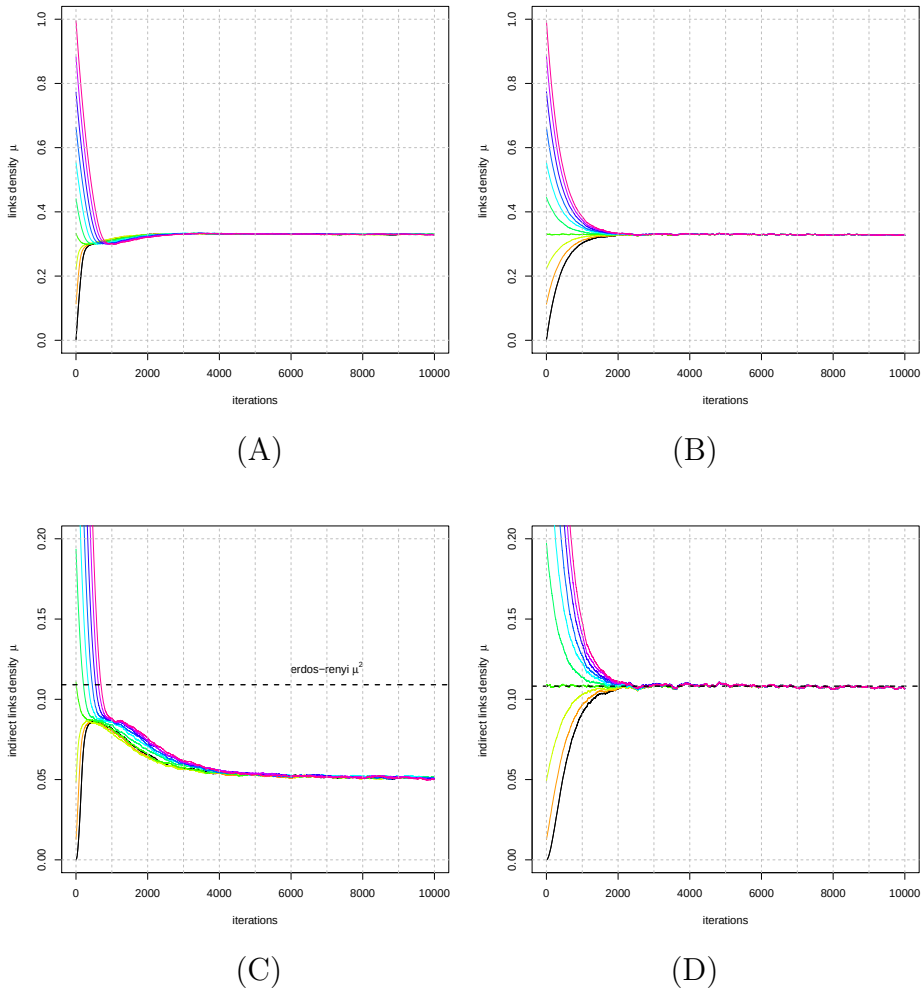
Panel (A) shows the functions $\zeta(\alpha)$, $S(\phi_1(\alpha))$ and $S(\phi_2(\alpha))$ described in Theorem 3. Panel (B) shows how the V-shaped region delimited by $S(\phi_1(\alpha))$ and $S(\phi_2(\alpha))$ change if we consider the model with direct utility and only one externality, i.e. a model with two parameters only. Here r defines the order of interdependencies of the second utility term (the externality): $r = 2$ corresponds to the original model in Theorem 3; $r = 3$ correspond to a model with direct links utility and utility from common connections (cyclic triangles); $r = 4$ corresponds to a model with direct links utility and utility from 4 common connections (e.g. 4-cycle). If we increase the order of dependencies the region increases. This is crucial for the discussion about convergence of the algorithms. The derivations and additional utility terms are considered in Appendix D.

The fourth statement provides sufficient conditions that guarantee the model does not converge to a trivial network with independent links. The problem of asymptotic identification is generated by positive externalities: a model with *sufficiently large negative externalities* generates graphs that do not converge asymptotically to directed Erdos-Renyi networks. In other words, the constant function $h(x, y) = \mu$ is not a solution of the variational problem (16) when $\beta < 0$ and sufficiently large in magnitude.

Figure 2(B) shows how the V-shaped region delimited by $S(\phi_1(\alpha))$ and $S(\phi_2(\alpha))$ change if we consider alternative externalities to the term $t(H_2, g)$. In the figure, r defines the order of interdependencies of the second utility term (the externality): $r = 2$ corresponds to the original model in Theorem 3; $r = 3$ correspond to a model with direct links utility and utility from common connections (cyclic triangles), i.e. $t(H_2, g) = n^{-3} \sum_i \sum_j \sum_k g_{ij} g_{jk} g_{ki}$; $r = 4$ corresponds to a model with direct links utility and externality from 4 connections, e.g. 4-cycle with $t(H_2, g) = n^{-4} \sum_i \sum_j \sum_k \sum_l g_{ij} g_{jk} g_{kl} g_{li}$. The general result is that if we increase the order of dependencies the size of the V-shaped region increases. We show below that this V-shaped region is related to the convergence problems of the algorithm for networks simulations. The derivations and analysis with several alternative utility terms are provided

in Appendix D.

Figure 3: Model with negative externalities does not converge to Erdos-Renyi graphs



The 4 panels compare simulations of model (13) with parameters $(\alpha, \beta) = (5, -10)$ with correspoding Erdos-Renyi model with same direct link density. The network has $n = 300$ players, and we run the simulation for 1500000 iterations, sampling every 150 iterations. In Panel (A) we show the convergence of the direct links density to $\mu = 0.3302742$ for model model (13). In Panel (B) we simulate the corresponding Erdos-Renyi model with parameter $\alpha = \log \frac{\mu}{1-\mu}$. If model (13) with parameters $(\alpha, \beta) = (5, -10)$ converges to an Erdos-Renyi model, then the density of indirect links should be $\mu^2 = 0.109081$, shown as the horizontal dashed line in Panel (C): this shows that the model is different from a trivial Erdos-Renyi model with independent direct links. The simulations for the correspoding Erdos-Renyi model are in Panel (D), showing convergence to μ^2.

The simulations in Figure 3 show evidence that the model with $\beta < 0$ does not converge to an Erdos-Renyi model in the large n limit. We start the simulations at 10 different starting

values, corresponding to Erdos-Renyi graphs with probability of linking μ equi-spaced on the unit interval, for a network of size $n = 300$.[39] In Figure 3(A) we report simulations for $(\alpha, \beta) = (5, -10)$, which converge to a network density of $\mu = 0.3302742$. Figure 3(B) shows the simulations of the corresponding Erdos-Renyi model with parameter $\alpha' = \log \frac{\mu}{1-\mu}$. If the model with $(\alpha, \beta) = (5, -10)$ converges to an Erdos-Renyi graph, then the density of indirect links should be $\mu^2 = 0.109081$. Figure 3(C) and (D) prove that this is not the case. Indeed in Figure 3(C) our model converges to a different density of indirect links, smaller than the corresponding Erdos-Renyi indirect link density in Figure 3(D). Figure 3(B) and (D) are shown to prove that this is not an artifact of the simulations or the sampler.

The result in Theorem 3 applies to more general models. Indeed, if we augment model (13) to include the effect of common links, i.e. cyclic triangles, we obtain a rescaled potential

$$T(g) = \alpha t(H_1, g) + \beta t(H_2, g) + \gamma t(H_3, g) \tag{18}$$

where the network statistics $t(H_1, g)$ and $t(H_2, g)$ are the same as in model (13) and $t(H_3, g) = n^{-3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \neq i}^{n} g_{ij} g_{jk} g_{ki}$. For such model we can provide a similar characterization of the asymptotic behavior.

THEOREM 4 *Let μ_0 be (uniquely) determined by*

$$6\gamma = \frac{2\mu_0 - 1}{\mu_0^2 (1 - \mu_0)^2}$$

and let α_0 and β_0 be defined as follows:

$$\beta_0 = \frac{1}{2\mu_0(1-\mu_0)} - 3\gamma\mu_0 \qquad \text{and} \qquad \alpha_0 = \log \frac{\mu_0}{1-\mu_0} - \frac{1}{(1-\mu_0)} + \frac{2\mu_0 - 1}{2(1-\mu_0)^2}$$

Then model (18) has the following asymptotic behavior.

1. *If $\beta \geq 0$ and $\gamma \geq 0$, then the networks generated by the model are indistinguishable from a directed Erdos-Renyi graph with linking probability μ^* that solves the following equation*

$$\mu = \frac{\exp\left[\alpha + 2\beta\mu + 3\gamma\mu^2\right]}{1 + \exp\left[\alpha + 2\beta\mu + 3\gamma\mu^2\right]} \tag{19}$$

   and satisfy $(2\beta\mu + 6\gamma\mu^2)(1 - \mu) < 1$, for almost all $\alpha \in \mathbb{R}$, $\beta \geq 0$ and $\gamma \geq 0$.

2. *If $\beta > 0$, for any $\gamma > 0$, there exists a continuous and monotone decreasing function $\zeta_\gamma : (-\infty, \alpha_0] \to [\beta_0, \infty]$, such that for any $\alpha < \alpha_0$, if $\beta = \zeta_\gamma(\alpha) > \max\{0, \beta_0\}$, the model generated graph is asymptotically indistinguishable from a mixture of directed Erdos-Renyi graphs with linking probabilities μ_1^* and μ_2^*, such that $\mu_1^* < 0.5 < \mu_2^*$ and both solve equation (19) and satisfy $(2\beta\mu + 6\gamma\mu^2)(1 - \mu) < 1$.*

[39] The theoretical results approximate networks of size $n > 50$ quite well.

3. *If $\beta > 0$, for any $\gamma > 0$ there exist two functions $S_\gamma(\phi_1(\alpha))$ and $S_\gamma(\phi_2(\alpha))$ that delimit a V-shaped region of the parameters (α, β). If $\beta \in (S_\gamma(\phi_2(\alpha)), S_\gamma(\phi_1(\alpha)))$ the variational problem (16) has two local maximizers, that correspond to directed Erdos-Renyi graphs with linking probability μ_1^* and μ_2^*, such that $\mu_1^* < 0.5 < \mu_2^*$ and both solve equation (17) and satisfy $2\beta\mu(1-\mu) < 1$. If $\beta \in (S_\gamma(\phi_2(\alpha)), \zeta_\gamma(\alpha))$ then μ_1^* is the global maximizer. If $\beta \in (\zeta_\gamma(\alpha), S_\gamma(\phi_1(\alpha)))$ then μ_2^* is the global maximizer.*

4. *If $\beta < 0$, then for any $\alpha \in \mathbb{R}$ and $\gamma \geq 0$ there exists a positive constant $C(\alpha, \gamma) > 0$, such that for $\beta < -C(\alpha, \gamma)$ the model is asymptotically different from a directed Erdos-Renyi model. Analogously, if $\gamma < 0$, then for any $\alpha \in \mathbb{R}$ and $\beta \geq 0$ there exists a positive constant $C(\alpha, \beta) > 0$, such that for $\gamma < -C(\alpha, \beta)$ the model is asymptotically different from a directed Erdos-Renyi model.*

Proof. The first, second and third statements follow from Theorem 12, Theorem 17 and Theorem 18 in Appendix D. The fourth statement is proven in Theorem 16 in Appendix D. ■

The theorem confirms the identification problem for models with only positive externalities. The last part of the theorem shows that this problem does not arise when *at least one of the externalities is negative and sufficiently large*. The generalization to additional externalities with alternative utility subgraphs is straightforward, but tedious.[40]

The main lesson from this analysis is that models with homogeneous players including *only positive externalities* converge asymptotically to trivial Erdos-Renyi models and are essentially ill-identified in the large n limit. However, as long as *at least one externality is negative*, the model does not degenerate into a trivial independent-links model. This is important, because when we have homophily in indirect links, some linking decision will generate negative externalities for some players. While we were not able to prove similar results for the more general model with heterogeneous players,[41] we conjecture that the sign of the linking externalities is crucial for identification in these class of models.

3.4 Convergence of network simulations

The graphs in Figure 1 show that network simulations with a local sampler may have convergence issues. Furthermore, the work of Bhamidi et al. (2011) provides conditions under which the simulations are infeasible, focusing on the undirected version of the exponential random graph model. While a general characterization of convergence is not tractable, it is possible to provide exact bounds for the special case with only positive externalities, following the same approach of Bhamidi et al. (2011). Consider our basic model with a meeting technology such that in each period a player i is selected with probability $1/n$, and meets an agent j with probability $1/(n-1)$. This model behaves as a random scan Gibbs sampler, and it is a local Markov chain, because it updates one link per iteration. More generally a chain is local if it updates only $o(n)$ links per iteration.

[40] Examples of additional externalities are shown in Appendix D.

[41] We are not aware of any result in the literature on graph limits that allows for covariates.

In this special case and for non-negative externalities, we can provide exact bounds to the number of iterations necessary to couple a Markov chain started at the empty network with one started at the full network.

THEOREM 5 *Consider the same model as in Theorem 4, with probability of meeting* $\rho_{ij} = 1/(n(n-1))$. *Let* μ_0, α_0, β_0, $S_\gamma(\phi_1(\alpha))$ *and* $S_\gamma(\phi_2(\alpha))$ *be defined as in Theorem 4. Fix any* $\gamma \geq 0$. *Then, for any* $\beta \geq 0$

1. *If* $\beta \in [S_\gamma(\phi_2(\alpha)), S_\gamma(\phi_1(\alpha))]$, *the model converges to stationarity in* e^{Cn^2} *steps, where* $C > 0$ *is a constant. This result extends to any local chain.*

2. *If* $\beta \notin [S_\gamma(\phi_2(\alpha)), S_\gamma(\phi_1(\alpha))]$, *the model converges to stationarity in* $Cn^2 \log n$ *steps, where* $C > 0$ *is a constant.*

Proof. Follows from Theorem 4 above, and Theorems 5 and 6 in Bhamidi et al. (2011) ■

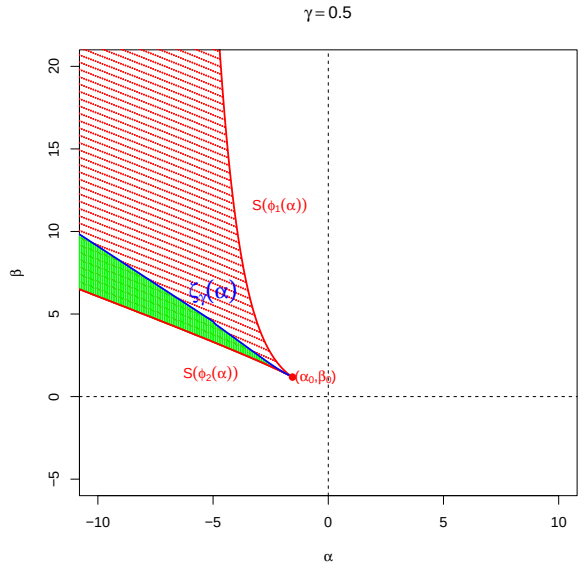
If we use a local sampler to simulate networks from the stationary distribution, we are not able to run the simulations to stationarity if parameters belong to the V-shaped region in Figure 4, delimited by functions $S(\phi_1(\alpha))$ and $S(\phi_2(\alpha))$. This happens because in that region, the stationary distribution has two local maxima and the chain may get trapped in one of them. Thus convergence is in exponential time, because once close to one of the local modes, the chain has probability e^{-Cn^2} to reach the other mode.

It is trivial to show that an increase in γ would increase α_0 and decrease β_0, thus increasing the area of exponentially slow convergence. For a visualization see proof of Theorem 17 in Appendix D.

When the convergence is in quadratic time (i.e. in order $n^2 \log n$ steps), the sampler is feasible for moderate size networks ($n < 500$). However, if we superimpose the statements of Theorem 5 with the previous Theorem 4 we realize that the region of quadratic convergence corresponds to regions in which the model behaves asymptotically as an Erdos-Renyi model. Therefore the sampler can be simplified drastically to simulate the model as a matrix of Bernoulli variables. This provides a good benchmark to check if alternative simulation strategies are correct.

The results on convergence and identification raise also another concern. In Figure 4 we highlight the V-shaped area of slow convergence, delimited by the functions $S_\gamma(\phi_1(\alpha))$ and $S(\phi_2(\alpha))$. Let's focus on the area below $\zeta_\gamma(\alpha)$, in green. When the parameters belong to this area, the model is asymptotically equivalent to a directed Erdos-Renyi graph with probability of linking $\mu_1^* < 0.5$. However, as stated in Theorem 4 (part 3), for each combination of (α, β) there are two local maxima of the variational problem in this area, $\mu_1^* < 0.5 < \mu_2^*$, which correspond to the two local modes of the stationary distribution. If we start the sampler from a very sparse graph (e.g. the empty network), it will converge to μ_1^* in quadratic time. However, if we start the sampler at a very dense graph (e.g. the complete network), it will converge to μ_2^* in quadratic time, getting trapped in the local maximum of the variational

Figure 4: Regions of fast and slow convergence, for given γ



The figure shows the V-shaped region of exponentially slow convergence for a local sampler. In this region, the variational problem (16) has two local solutions and the local sampler can get trapped in the local maximum, without ever visiting the global maximum.

problem. It can be shown that the probability of escaping the local maximum is of order e^{-Cn^2} (see proofs of Theorem 5 and 6 in Bhamidi et al. (2011)), which is essentially zero. Analogously, if we consider the region above $\zeta_\gamma(\alpha)$ and below $S_\gamma(\phi_1(\alpha))$, there are two local maximizers of the variational problem, $\mu_1^* < 0.5 < \mu_2^*$ for each combination of (α, β), and μ_2^* is the global maximizer. If we start the sampler from a very sparse graph (e.g. the empty network), it will converge to μ_1^* in quadratic time. If we start the sampler at a very dense graph (e.g. the complete network), it will converge to μ_2^* in quadratic time, getting trapped in the local maximum of the variational problem.

Therefore, for any $\gamma > 0$, at any combination of (α, β) inside the V-shaped region delimited by the functions $S_\gamma(\phi_1(\alpha))$ and $S(\phi_2(\alpha))$, our local sampler could fail to converge to the correct asymptotic network density.

A non-local sampler.

The previous theoretical results indicate the local chain property of the sampler creates convergence problems. We can this modify our sampler to take this analysis into account. Essentially the modification allows the sampler to make *larger steps*, in particular steps that are not $o(n)$. The local chain selects a link g_{ij} with probability $1/(n(n-1))$, proposing to swap the value to $1 - g_{ij}$.

The modified algorithm, has several large steps. First, with probability p_r, the sampler selects a player i at random (with probability $1/n$) and proposes to swap all his links, i.e.

$g_{ij} = 1 - g_{ij}$ for each $j = 1, ..., n$. Second, with probability p_c, the sampler selects a player i at random (with probability $1/n$) and proposes to swap all the links pointing at i, i.e. $g_{ji} = 1 - g_{ji}$ for each $j = 1, ..., n$. Third, with probability p_f, the sampler selects uniformly at random $\lceil \lambda n \rceil$ links, where $\lambda \in (0, 1)$, and proposes to swap all of them. Notice that this step size is a function of n, and in particular is not $o(n)$. The crucial ingredient is to make the length of the step a function of n. The parameter λ is under control of the researcher: higher values allow larger steps and increase the computational cost of sampling. Lastly, with probability p_{inv} the sampler proposes to invert the adjacency matrix. The goal of this large step is to provide a way to jump across modes of the stationary distribution, when it is bimodal. [42]

Using this sampler, we reproduce the simulation in Figure 1. We know that the local chain can get trapped in local maxima of the variational problem. If we simulate model (13) with parameters $(\alpha, \beta) = (-3, 3)$, we obtain Figure 5(A). While Theorem 3 states that the simulations should converge to the sparse network density $\mu_1 \approx 0.07$, we observe that the local sampler converges to a dense network with $\mu_2 \approx 0.93$, if started at dense networks. In other words, when started at a dense network (say the full network), the sampler gets trapped in a local maximum of the variational problem, with density $\mu_2 \approx 0.93$. Figure 5(B) shows that our modified sampler does not have this problem, and also the chains started at dense network converge to the correct (sparse) network density. This simple modification gets rid of the exponentially slow convergence of the local algorithm. More generally, these large steps allow the sampler to escape local maxima of the potential function.

3.5 Posterior Estimation

We estimate the posterior distribution of the structural parameters using an approximate version of the *exchange algorithm* (see Murray et al. (2006)). The approximate algorithm uses a double Metropolis-Hastings step to avoid the computation of the normalizing constant $c(\mathcal{G}, X, \theta)$ in the likelihood, as in Liang (2010).[43] Several authors have proposed similar algorithms in the related literature on Exponential Random Graphs Models (ERGM).[44]

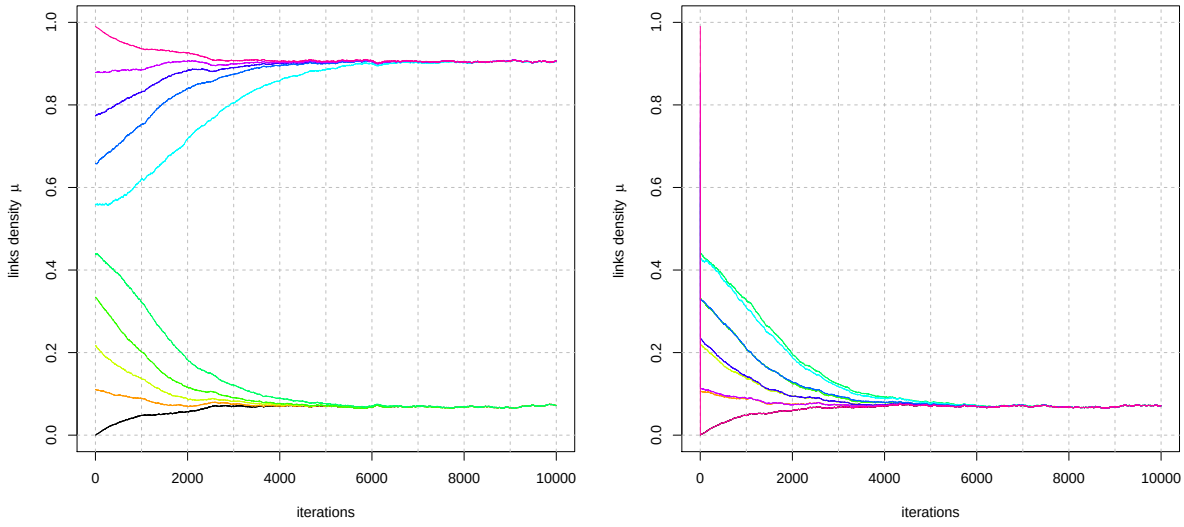
The idea of the algorithm is to sample from an augmented distribution using an auxiliary variable. At each iteration, the algorithm proposes a new parameter vector θ', drawn from a

[42] We have seen that this is the case in the homogeneous player case, for many parameter values.

[43] This improvement comes with a possible cost: the algorithm may produce MCMC chains of parameters that have very poor mixing properties (Caimo and Friel, 2010) and high autocorrelation. We partially correct for this problem by carefully calibrating the proposal distribution. In this paper we use a random walk proposal. Alternatively one could update the parameters in blocks or use recent random block techniques as in Chib and Ramamurthy (2009) to improve convergence and mixing.

[44] Caimo and Friel (2010) use the exchange algorithm to estimate ERGM. They improve the mixing of the sampler using the snooker algorithm. Koskinen (2008) proposes the Linked Importance Sampler Auxiliary variable (LISA) algorithm, which uses importance sampling to provide an estimate of the acceptance probability. Another variation of the algorithm is used in Liang (2010).

Figure 5: Local sampler versus Modified sampler



(A) (B)

Comparison of network samplers for model (13), with parameters $(\alpha, \beta) = (-3, 3)$. Panel (A) shows the simulation using the local-chain sampler, which converges to two different link densities ($\mu_1 \approx 0.07$ and $\mu_2 \approx 0.93$). However, we know from Theorem 3, that the correct simulation should converge to the sparse network density. So the local chain fails to sample correctly if we start it at a dense network, because it gets trapped at a local maximum of the stationary distribution. Panel (B) shows the simulation using the modified algorithm. We use $p_r = p_f = p_{inv} = 0.01$. The simulations converge to the correct link density for any starting value, therefore our modified algorithm provides a better sampler for the model.

suitable proposal distribution $q_\theta(\theta'|\theta)$; in the second step, it samples a network g' (the auxiliary variable) from the likelihood $\pi(g', X, \theta')$; finally, the proposed parameter is accepted with a probability $\alpha_{ex}(\theta, \theta')$, such that the Markov chain of parameters generated by these update rules, has the posterior (8) as unique invariant distribution.

The result in Lemma 1 in Appendix B shows that choosing the observed network as initial network for the simulations guarantees that the *approximate* and the *exact* exchange algorithm have the same acceptance ratio, for any length R of the network simulations. Therefore, the proof of convergence to the correct posterior only needs to show the convergence of the proposal distribution, i.e. convergence of the network simulations to the stationary equilibrium of the model (see details in Appendix B).

ALGORITHM 2 ***(APPROXIMATE EXCHANGE ALGORITHM)***
Fix the number of simulations R. At each iteration t, with current parameter $\theta_t = \theta$ and network data g:

1. *Propose a new parameter θ' from a distribution $q_\theta(\cdot|\theta)$,*

$$\theta' \sim q_\theta(\cdot|\theta). \tag{20}$$

2. *Start **ALGORITHM 1** at the observed network g, iterating for R steps using parameter θ' and collect the last simulated network g'*

$$g' \sim \mathcal{P}_{\theta'}^{(R)}(g'|g). \tag{21}$$

3. *Update the parameter according to*

$$\theta_{t+1} = \begin{cases} \theta' & \text{with prob. } \alpha_{ex}(\theta, \theta', g', g) \\ \theta & \text{with prob. } 1 - \alpha_{ex}(\theta, \theta', g', g) \end{cases}$$

where

$$\alpha_{ex}(\theta, \theta', g', g) = \min\left\{1, \frac{\exp\left[Q(g', X, \theta)\right]}{\exp\left[Q(g, X, \theta)\right]} \frac{p(\theta')}{p(\theta)} \frac{q_\theta(\theta|\theta')}{q_\theta(\theta'|\theta)} \frac{\exp\left[Q(g, X, \theta')\right]}{\exp\left[Q(g', X, \theta')\right]}\right\}. \tag{22}$$

The main advantage of this algorithm is that all quantities in the acceptance ratio (22) can be evaluated: there are no integrals or normalizing constants to compute. This simple modification of the original Metropolis-Hastings scheme makes estimation feasible.

The sampler is likely to accept proposals that move towards high density regions of the posterior, but it is likely to reject proposals that move towards low density regions of the posterior. The formal statement about convergence is contained in the following theorem.

THEOREM 6 *(Ergodicity of the Approximate Exchange Algorithm). The approximate exchange algorithm is ergodic, and it converges to the correct posterior distribution.*

1. (Convergence)*) Let $\widetilde{P}_R^{(s)}(\theta_0, \cdot)$ be the s-th step transition of the approximate exchange algorithm, when the auxiliary network is sampled using R steps of the network simulation algorithm and the initial parameter of the simulation is θ_0. Let $\|\cdot\|_{TV}$ be the total variation distance and $p(\cdot|g, X)$ the posterior distribution.*
Then, for any $\epsilon > 0$ there exist $R_0 \in \mathbb{N}$ and $S_0 \in \mathbb{N}$ such that for any $R > R_0$ and $s > S_0$ and any initial parameter vector $\theta_0 \in \Theta$

$$\left\| \widetilde{P}_R^{(s)}(\theta_0, \cdot) - p(\cdot|g, X) \right\|_{TV} \leq \epsilon \tag{23}$$

2. (WLLN) *A Weak Law of Large Numbers holds: for any initial parameter vector $\theta_0 \in \Theta$ and any bounded integrable function $h(\cdot)$*

$$\frac{1}{S} \sum_{s=1}^{S} h(\theta_s) \xrightarrow{P} \int_\Theta h(\theta) p(\theta|g, X) \, d\theta \tag{24}$$

Proof. In Appendix B. ■

The theorem states that the algorithm produces good samples as long as the number of steps of the network simulation algorithm is big enough and the algorithm is run for a sufficient number of iterations.

In general, for a fixed number of network simulations R, the samples generated by the algorithm will converge to a posterior that is "close" to the correct posterior. As $R \to \infty$ the algorithm converges to the exact exchange algorithm of Murray et al. (2006), producing exact samples from the posterior distribution. However, an higher value of R would increase the computational cost and result in a higher rejection rate for the proposed parameters. The results in the next section provide some practical guidance on setting a suitable R, without compromising computational efficiency.

4 Simulation results

The performance of the estimation method is tested using artificial data. All the computations with artificial data are performed in a standard desktop Dell Precision T7620 with 2 Intel Xeon CPUs E5-2697 v2 with 12 Dual core processors at 2.7GHZ each and 64GB of RAM. For replication purposes, there is a package in Github at https://github.com/meleangelo/netnew.[45]

Ideally, we want to compare the results of the approximate exchange algorithm with the exact algorithm. This is feasible for a special case, where preferences depend only on direct and mutual links (i.e. excluding friends of friends and popularity effects).

$$Q(g, \alpha, \beta) = \alpha \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} + \beta \sum_{i=1}^{n} \sum_{j>i1}^{n} g_{ij} g_{ji} \tag{25}$$

For this model, described by equation (25), we can show that the constant is

$$c(\theta) = \left(1 + 2e^{\alpha} + e^{2\alpha+\beta}\right)^{\frac{n(n-1)}{2}}$$

thus we can compute the exact likelihood and we can perform inference using the *exact* Metropolis-Hastings sampler. We then compare the results of the exact algorithm with the approximate exchange algorithm.

The results of the simulations are shown in Table 1. The data were generated by parameters $(\alpha, \beta) = (-2.0, 0.5)$. The number of network simulations per each proposed parameter are $R = \{1000, 5000, 10000, 50000, 100000, 1000000, 10000000\}$. We run each algorithm for $S = 10000$ parameters iterations, and we use the output to measure the

[45] In all estimation exercises we use independent normal priors $\mathcal{N}(0, 10)$. The proposal of the exchange algorithm is a random walk $\mathcal{N}(0, \Sigma)$. We repeat the estimation twice: the first time we use a diagonal Σ; in the second round, we use the covariance from the first round as baseline. In all simulations the probability of large steps is 0.001 and a large step updates $0.1n$ links.

Table 1: Convergence of estimated posteriors, model (25)

$n = 100$	Exact Metropolis		R=1000		R=5000		R=10000		R=50000		R=100000					
	α	β	α	β	α	β	α	β	α	β	α	β				
mean	-1.923	0.286	-1.915	0.286	-1.925	0.285	-1.922	0.275	-1.921	0.284	-1.919	0.286				
median	-1.923	0.288	-1.919	0.296	-1.923	0.292	-1.921	0.273	-1.921	0.287	-1.919	0.286				
std. dev.	0.034	0.114	0.105	0.263	0.054	0.141	0.042	0.123	0.034	0.115	0.034	0.111				
mcse	0.000	0.002	0.007	0.033	0.001	0.005	0.001	0.004	0.000	0.003	0.000	0.003				
pctile 2.5%	-1.992	0.058	-2.115	-0.257	-2.034	-0.007	-2.006	0.034	-1.987	0.039	-1.985	0.069				
pctile 97.5%	-1.857	0.506	-1.705	0.767	-1.820	0.553	-1.842	0.514	-1.853	0.505	-1.851	0.512				
KS	NA	NA	0.275	0.205	0.114	0.057	0.066	0.057	0.032	0.015	0.060	0.022				
KL	NA	NA	0.041	0.027	0.013	0.186	0.039	0.075	0.040	0.062	0.006	0.088				

$n = 200$	Exact Metropolis		R=1000		R=5000		R=10000		R=50000		R=100000		R=1mil		R=10mil	
	α	β	α	β	α	β	α	β	α	β	α	β	α	β	α	β
mean	-1.988	0.463	-1.975	0.463	-1.964	0.463	-1.979	0.465	-1.989	0.455	-1.988	0.463	-1.988	0.466	-1.987	0.459
median	-1.989	0.467	-1.974	0.509	-1.968	0.468	-1.978	0.465	-1.989	0.454	-1.989	0.464	-1.988	0.467	-1.987	0.457
std. dev.	0.017	0.061	0.048	0.275	0.042	0.113	0.033	0.073	0.019	0.053	0.017	0.059	0.017	0.057	0.016	0.055
mcse	0	0.003	0.002	0.075	0.002	0.012	0.001	0.005	0	0.002	0	0.003	0	0.003	0	0.002
pctile 2.5%	-2.021	0.335	-2.071	-0.21	-2.044	0.186	-2.042	0.32	-2.024	0.353	-2.024	0.339	-2.024	0.358	-2.017	0.354
pctile 97.5%	-1.954	0.572	-1.89	0.889	-1.872	0.687	-1.921	0.614	-1.949	0.56	-1.955	0.571	-1.955	0.582	-1.955	0.577
KS	NA	NA	0.343	0.34	0.381	0.135	0.25	0.057	0.067	0.105	0.015	0.039	0.039	0.045	0.062	0.086
KL	NA	NA	0.1	0.178	0.105	0.079	0.129	0.099	0.05	0.05	0.041	0.058	0.092	0.044	0.173	0.234
time	0.124s		14.539s		21.808s		30.451s		100.761s		193.722s		1762.202s		17370.945s	

$n = 500$	Exact Metropolis		R=1000		R=5000		R=10000		R=50000		R=100000		R=1mil		R=10mil	
	α	β	α	β	α	β	α	β	α	β	α	β	α	β	α	β
mean	-2.018	0.551	-1.941	0.337	-2.014	0.562	-2.017	0.561	-2.017	0.552	-2.018	0.552	-2.018	0.55	-2.018	0.55
median	-2.018	0.552	-1.922	0.369	-2.012	0.562	-2.019	0.562	-2.016	0.553	-2.018	0.552	-2.018	0.551	-2.018	0.55
std. dev.	0.007	0.024	0.071	0.218	0.045	0.107	0.036	0.074	0.016	0.036	0.012	0.028	0.007	0.022	0.007	0.022
mcse	0	0	0.005	0.047	0.002	0.011	0.001	0.005	0	0.001	0	0	0	0	0	0
pctile 2.5%	-2.032	0.501	-2.074	-0.106	-2.105	0.335	-2.085	0.424	-2.05	0.479	-2.041	0.497	-2.032	0.508	-2.031	0.507
pctile 97.5%	-2.004	0.596	-1.838	0.666	-1.931	0.755	-1.942	0.707	-1.988	0.621	-1.994	0.606	-2.004	0.596	-2.005	0.592
KS	NA	NA	0.743	0.703	0.408	0.363	0.341	0.31	0.229	0.117	0.107	0.066	0.027	0.034	0.033	0.032
KL	NA	NA	0.466	0.196	0.121	0.081	0.081	0.019	0.026	0.009	0.02	0.008	0.061	0.036	0.049	0.041
time	0.187s		87.344s		95.831s		105.955s		181.413s		275.357s		2010.322s		19319.663s	

$n = 1000$	Exact Metropolis		R=1000		R=5000		R=10000		R=50000		R=100000		R=1mil		R=10mil	
	α	β	α	β	α	β	α	β	α	β	α	β	α	β	α	β
mean	-2.001	0.481	-1.986	0.456	-1.974	0.459	-1.995	0.486	-1.999	0.479	-2.001	0.479	-2.001	0.481	-2.002	0.481
median	-2.001	0.48	-1.991	0.501	-1.979	0.461	-1.993	0.479	-2.001	0.48	-2.001	0.479	-2.001	0.48	-2.002	0.482
std. dev.	0.003	0.011	0.081	0.247	0.05	0.091	0.031	0.07	0.017	0.037	0.011	0.026	0.004	0.012	0.003	0.012
mcse	0	0	0.007	0.06	0.002	0.007	0.001	0.004	0	0.001	0	0	0	0	0	0
pctile 2.5%	-2.008	0.459	-2.148	-0.007	-2.047	0.284	-2.057	0.361	-2.029	0.404	-2.024	0.425	-2.01	0.457	-2.008	0.459
pctile 97.5%	-1.995	0.503	-1.835	0.813	-1.825	0.642	-1.938	0.64	-1.966	0.55	-1.979	0.529	-1.993	0.505	-1.995	0.504
KS	NA	NA	0.506	0.484	0.63	0.47	0.502	0.355	0.351	0.268	0.271	0.216	0.078	0.018	0.027	0.036
KL	NA	NA	0.3	0.261	0.528	0.041	0.297	0.083	0.45	0.21	0.137	0.047	0.021	0.031	0.034	0.014
time	0.234s		364.761s		371.563s		381.172s		459.578s		556.330s		2304.228s		19730.772	

Kolmogorov-Smirnov distance and the Kullback-Leibler divergence between the posterior estimated with the exact metropolis sampler $p(\theta|g,X)$ and the posterior estimated with the approximated algorithm with R network simulations $p_R(\theta|g,X)$

$$\begin{aligned} KS &= \sup_{\theta_i \in \Theta_i} \left| \int_{-\infty}^{\theta_i} p_R(\theta_i|g,X) - \int_{-\infty}^{\theta_i} p(\theta_i|g,X) \right| \\ KL &= \int_{\Theta_i} \log \left[\frac{p_R(\theta_i|g,X)}{p(\theta_i|g,X)} \right] p_R(\theta_i|g,X) d\theta_i \end{aligned}$$

The table reports posterior mean, median, standard deviation, Monte Carlo standard errors for the posterior mean (mcse), 95% credibility intervals, Kolmogorov-Smirnov statistics, Kullback-Leibler divergence and time for computation.

The exact Metropolis-Hastings is reported in the first column of the table. The approximate exchange algorithm works very well for small to moderate networks. For a small network with $n = 100$ players, a reasonable degree of accuracy can be reached with as low as $R = 5000$ network simulations per parameter. Simulations from over-dispersed starting values converge to the same posterior distribution. Convergence is quite fast to the high density region of the posterior.[46]

Let's now consider a model with homogeneous players where there is no utility from reciprocated links, but only from indirect connections and popularity, i.e.

$$Q(g,\alpha,\beta) = \alpha \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} + \beta \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk} \tag{26}$$

The estimated parameters for this model are shown in Table 2 and 3. We generate the network data using different parameter vectors. The first panel correspond to parameters $(\alpha,\beta) = (-3, 1/n)$. This is a model that generates a sparse network and the likelihood has a unique mode. The second panel shows estimates for a model with parameters $(\alpha,\beta) = (-3, 3/n)$, with a variational problem with two local solutions that generates problems of convergence with a local sampler. The last panel is a model with negative externalities $(\alpha,\beta) = (5, -10/n)$ that does not converge to an Erdos-Renyi model. We also simulated a model with parameters $(\alpha,\beta) = (-3, 7/n)$. However, if we solve the variational problem with these parameters, we can show that the solution is an Erdos-Renyi model with probability of linking $\mu^* = 1$, i.e. the full network. Therefore a model with parameters $\alpha = -3$, for any $\beta > 7/n$ would also generate a full network. Any attempt to estimate β with data consisting of a full network is futile.

The estimates using the non-local sampler are precise for a moderate amount of network simulations. Clearly, estimates in Table 2 are less precise than the ones in Table 3, since the number of links is smaller. Because of our modified network sampler, there is no need to

[46] This result is common with the class of exchange algorithms. See Caimo and Friel (2010), Atchade and Wang (forthcoming) for examples. Computations can be faster if we embed sparse matrix algebra routines in the codes. The results in Table 1 are obtained with codes that do no use sparse matrix algebra, thus representing a worst case scenario in computational time.

Table 2: Estimated structural parameters for model (26), $n = 100$

true parameters $(\alpha, \beta) = (-3, 0.01)$								
$n = 100$	$R = 1000$		$R = 10000$		$R = 100000$		$R = 1000000$	
true $= (-3, .01)$	α	β	α	β	α	β	α	β
mean	-2.7450	-0.0233	-2.9095	-0.0035	-2.9407	-0.0009	-2.9206	-0.0025
median	-2.7575	-0.0185	-2.9171	-0.0021	-2.9498	0.0003	-2.9288	-0.0014
std. dev.	0.4782	0.0460	0.2032	0.0201	0.1860	0.0183	0.1916	0.0189
mcse	0.0975	0.0010	0.0111	0.0001	0.0094	0.0001	0.0104	0.0001
pctile 2.5%	-3.6862	-0.1208	-3.2660	-0.0462	-3.2614	-0.0400	-3.2468	-0.0412
pctile 97.5%	-1.7789	0.0545	-2.4916	0.0305	-2.5452	0.0303	-2.5158	0.0297
time (secs)	25.3800		236.2100		2485.5200		24658.1500	

true parameters $(\alpha, \beta) = (-3, 0.03)$								
$n = 100$	$R = 1000$		$R = 10000$		$R = 100000$		$R = 1000000$	
true $= (-3, 0.03)$	α	β	α	β	α	β	α	β
mean	-2.6075	0.0002	-2.7578	0.0124	-2.7618	0.0126	-2.7720	0.0134
median	-2.6425	0.0036	-2.7804	0.0140	-2.7812	0.0140	-2.7917	0.0148
std.dev.	0.4396	0.0306	0.1757	0.0122	0.1663	0.0116	0.1671	0.0116
mcse	0.0819	0.0004	0.0075	0.0000	0.0073	0.0000	0.0080	0.0000
pctile 2.5%	-3.4144	-0.0682	-3.0320	-0.0150	-3.0185	-0.0132	-3.0165	-0.0129
pctile 97.5%	-1.6856	0.0526	-2.3671	0.0299	-2.4054	0.0299	-2.3897	0.0297
time (secs)	27.3900		256.2500		2647.7600		26277.7000	

true parameters $(\alpha, \beta) = (5, -0.1)$								
$n = 100$	R=1000		R=10000		R=100000		R=1000000	
true $= (5, -0.1)$	α	β	α	β	α	β	α	β
mean	4.8397	-0.0964	4.8722	-0.0968	4.8743	-0.0968	4.8856	-0.0970
median	4.8265	-0.0963	4.8674	-0.0968	4.8682	-0.0968	4.8846	-0.0970
std. dev.	0.4031	0.0067	0.1550	0.0026	0.1188	0.0018	0.1137	0.0018
mcse	0.0427	0.0000	0.0064	0.0000	0.0041	0.0000	0.0039	0.0000
pctile 2.5%	4.0677	-0.1101	4.5688	-0.1020	4.6493	-0.1005	4.6645	-0.1006
pctile 97.5%	5.6615	-0.0836	5.1707	-0.0920	5.1202	-0.0933	5.1156	-0.0936
time (secs)	49.3300		433.5100		4254.5800		41218.3700	

have a large number of network simulations. The reason is that the non-local sampler can jump quickly to the correct mode(s) of the likelihood: once it reaches an area close to the global maximum, convergence is in quadratic time, since it will reject jumps to local maxima of the variational problem that are not global maxima. While the number of iterations may be lower, the computational time of each iteration is higher, because the large steps are computationally expensive. In Table 2, the precision gain from additional network simulations is negligible when $R > 10000$. Notice that the computational time is higher for the model with negative externality. The reason is that the equilibrium network generated at the true parameters $(\alpha, \beta) = (5, -10/n)$ is denser than the ones in the previous panels, and therefore the large steps are more computationally expensive than for the other two models.

In the next table we consider a model where players are homogeneous and they receive utility from direct links, reciprocated links, indirect links and popularity. The potential function of such model is defined as

$$Q(g, \alpha, \beta, \gamma) = \alpha \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} + \beta \sum_{i=1}^{n} \sum_{j>i}^{n} g_{ij} g_{ji} + \gamma \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk} \tag{27}$$

Table 3: Estimated structural parameters for model (26), $n = 200$

true parameters $(\alpha, \beta) = (-3, 0.005)$								
$n = 200$	$R = 1000$		$R = 10000$		$R = 100000$		$R = 1000000$	
true $= (-3, .005)$	α	β	α	β	α	β	α	β
mean	-2.6694	-0.0112	-2.9707	0.0042	-3.0529	0.0083	-3.0603	0.0086
median	-2.7045	-0.0086	-2.9972	0.0056	-3.0675	0.0089	-3.0784	0.0095
std. dev.	0.5631	0.0254	0.1841	0.0083	0.1137	0.0053	0.1113	0.0052
mcse	0.1341	0.0003	0.0089	0.0000	0.0035	0.0000	0.0032	0.0000
pctile 2.5%	-3.7109	-0.0665	-3.2574	-0.0148	-3.2202	-0.0035	-3.2244	-0.0036
pctile 97.5%	-1.5002	0.0332	-2.5689	0.0159	-2.8044	0.0159	-2.8007	0.0158
time (secs)	173.7800		1651.4400		16248.9400		149962.1800	

true parameters $(\alpha, \beta) = (-3, 0.015)$								
$n = 200$	$R = 1000$		$R = 10000$		$R = 100000$		$R = 1000000$	
true $= (-3, 0.015)$	α	β	α	β	α	β	α	β
mean	-2.4770	-0.0033	-2.7773	0.0075	-2.8601	0.0104	-2.8518	0.0101
median	-2.5002	-0.0019	-2.8042	0.0083	-2.8785	0.0111	-2.8703	0.0108
std. dev.	0.5828	0.0200	0.1627	0.0055	0.1012	0.0035	0.1028	0.0035
mcse	0.1012	0.0001	0.0078	0.0000	0.0028	0.0000	0.0028	0.0000
pctile 2.5%	-3.6184	-0.0474	-3.0206	-0.0054	-3.0026	0.0024	-2.9961	0.0020
pctile 97.5%	-1.2515	0.0346	-2.4080	0.0148	-2.6267	0.0150	-2.6149	0.0149
time (secs)	190.5800		1783.1900		17496.5900		161462.5300	

true parameters $(\alpha, \beta) = (5, -0.05)$								
$n = 200$	R=1000		R=10000		R=100000		R=1000000	
true $= (5, -0.05)$	α	β	α	β	α	β	α	β
mean	5.0734	-0.0504	5.0782	-0.0503	5.0528	-0.0501	5.0477	-0.0501
median	5.0475	-0.0502	5.0718	-0.0503	5.0539	-0.0501	5.0478	-0.0501
std. dev.	0.4791	0.0039	0.1535	0.0012	0.0713	0.0005	0.0644	0.0005
mcse	0.0765	0.0000	0.0068	0.0000	0.0017	0.0000	0.0011	0.0000
pctile 2.5%	4.1775	-0.0587	4.7926	-0.0529	4.9129	-0.0512	4.9220	-0.0511
pctile 97.5%	6.0765	-0.0431	5.3987	-0.0480	5.1904	-0.0490	5.1781	-0.0491
time (secs)	361.5600		2996.3300		29001.6900		257572.3700	

The data are generated by parameters $(\alpha, \beta, \gamma) = (-2.00, 0.50, 0.01)$. The pattern of Table 4 is similar to the previous analysis: the increase in precision for $R > 10000$ is minimal with respect to the increased cost of sampling networks.

Finally, we estimate a simple model with heterogeneous players. There is only one binary covariate X and the players receive utility from direct links, and indirect links and popularity. The covariate is generated as a Bernoulli variable with $P(X_i = 1) = 0.3$. The utility from indirect links/popularity is positive if both i and k belong to type-1; and it is negative if they belong to different types. The potential of this model is

$$\begin{aligned} Q(g, \alpha, \beta, \gamma) = \alpha \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij} \quad &+ \quad \beta \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} g_{ij}g_{jk}\mathbb{1}_{\{X_i = X_k = 1\}} \\ &+ \quad \gamma \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} g_{ij}g_{jk}\mathbb{1}_{\{X_i \neq X_k\}} \end{aligned} \tag{28}$$

and the data are generated with parameters $(2, 11/n, -5/n)$.

The estimation results in Table 5 for $n = 100$. The estimates are again very precise for a

Table 4: Estimated structural parameters for model (27), n=100

	true parameters $(\alpha, \beta, \gamma) = (-2.00, 0.50, 0.01)$											
$n = 100$	R=1000			R=10000			R=100000			R=1000000		
	α	β	γ	α	β	γ	α	β	γ	α	β	γ
mean	-1.9321	0.5098	0.0074	-2.1182	0.5168	0.0133	-2.1034	0.5115	0.0129	-2.0938	0.5196	0.0126
median	-1.9756	0.5080	0.0089	-2.1382	0.5214	0.0139	-2.1251	0.5134	0.0136	-2.1066	0.5207	0.0131
std. dev.	0.4677	0.2330	0.0135	0.1899	0.0997	0.0054	0.1877	0.0894	0.0054	0.1832	0.0882	0.0053
mcse	0.0967	0.0209	0.0001	0.0121	0.0037	0.0000	0.0110	0.0028	0.0000	0.0132	0.0027	0.0000
pctile 2.5%	-2.7241	0.0459	-0.0224	-2.4259	0.3186	0.0014	-2.4002	0.3341	0.0012	-2.3871	0.3416	0.0013
pctile 97.5%	-0.9492	0.9699	0.0300	-1.7149	0.7115	0.0216	-1.6983	0.6896	0.0213	-1.7014	0.6894	0.0211
time (secs)		42			355			3545			35806	

Table 5: Estimated structural parameters for model (28), $n = 100$

	true parameters $(\alpha, \beta, \gamma) = (2.00, 0.11, -0.05)$								
$n = 100$	R=1000			R=10000			R=100000		
	α	β	γ	α	β	γ	α	β	γ
mean	2.2144	0.1056	-0.0530	2.0718	0.1053	-0.0503	2.0636	0.1052	-0.0501
median	2.0828	0.1064	-0.0506	2.0443	0.1054	-0.0498	2.0396	0.1055	-0.0497
std. dev.	0.8575	0.0227	0.0155	0.3348	0.0084	0.0061	0.2723	0.0066	0.0049
mcse	0.4485	0.0002	0.0001	0.0711	0.0000	0.0000	0.0487	0.0000	0.0000
pctile 2.5%	0.8456	0.0598	-0.0881	1.4803	0.0875	-0.0636	1.5959	0.0914	-0.0608
pctile 97.5%	4.1495	0.1473	-0.0286	2.8115	0.1222	-0.0397	2.6445	0.1174	-0.0418
time (secs)		97			314			2913	

moderate amount of simulations.

5 Conclusions

We developed an empirical model of network formation with heterogeneous players, that converges to a unique stationary equilibrium. The payoffs depend on direct connections, but also link externalities, e.g. reciprocated links, indirect links, popularity, common connections, etc. The inclusion of these externalities generates complex interdependencies among links, and therefore the likelihood does not factorize into independent components. Indeed, the likelihood is intractable because of a normalizing constant that is infeasible to compute.

We show that for large networks, the constant can be estimated as the solution of a variational problem. In the special case of homogeneous players the variational problem is tractable and provides very precise guidance about identification, convergence to stationarity and the role of externalities.

The general variational problem is intractable and we need some form of approximation. In this paper, we considered an approximation through sampling, using a Markov chain Monte Carlo method to approximate the likelihood and the posterior distribution of the parameters. Sampling is not the only alternative: we could approximate the variational problem using a deterministic technique. Some preliminary attempts in this direction are provided in He and Zheng (2013) and Mele (2015), using (structural) mean-field approximations for the exponential family (see Wainwright and Jordan (2008) and Bishop (2006)). An

alternative approach is provided in Chandrasekhar and Jackson (2014), by imposing sparsity, which implies good statistical properties of the estimators and improves the tractability of the model.

In the development of a model of empirical network formation, we also need to consider how modeling unobserved heterogeneity affects our results. Graham (2014) includes unobserved heterogeneity in a model with heterogeneous agents, but excludes the link externalities that are central to our model. We can include unobserved heterogeneity in our model, with substantial increase in computational burden. However, it is not clear that we can separately identify externalities and unobserved heterogeneity using only one network realization.

References

Acemoglu, D., M. Dahleh, I. Lobel and A. Ozdaglar (2011), 'Bayesian learning in social networks', *Review of Economic Studies* **forthcoming**.

Amemiya, Takeshi (1981), 'Qualitative response models: A survey', *Journal of Economic Literature* **19**(4), 1483–1536.

Andrieu, C. and G. O. Roberts (2009), 'The pseudo-marginal approach for efficient monte carlo computations', *Annals of Statistics* **37**(2), 697–725.

Aristoff, David and Lingjiong Zhu (2014), On the phase transition curve in a directed exponential random graph model.

Atchade, Yves and Jing Wang (forthcoming), 'Bayesian inference of exponential random graph models for large social networks', *Communications in Statistics - Simulation and Computation* .

Athey, Susan and Guido Imbens (2007), 'Discrete choice models with multiple unobserved choice characteristics', *International Economic Review* **48**(4), 1159–1192.

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

Bala, Venkatesh and Sanjeev Goyal (2000), 'A noncooperative model of network formation', *Econometrica* **68**(5), 1181–1229.

Bandiera, Oriana and Imran Rasul (2006), 'Social networks and technology adoption in northern mozambique', *Economic Journal* **116**(514), 869–902.

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.

Boeckner, Derek (2013), Directed graph limits and directed threshold graphs. PhD Thesis.

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

Boucher, Vincent (2013), Structural homophily.

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

Caimo, Alberto and Nial Friel (2010), 'Bayesian inference for exponential random graph models', *Social Networks* **forthcoming**.

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

Chatterjee, Sourav and S.R.S. Varadhan (2011), 'The large deviation principle for the erdosrnyi random graph', *European Journal of Combinatorics* **32**(7), 1000 – 1017. Homomorphisms and Limits.
URL: *http://www.sciencedirect.com/science/article/pii/S0195669811000655*

Chib, Siddartha and Srikanth Ramamurthy (2009), 'Tailored random block', *Journal of Econometrics* .

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

Conley, Timothy and Christopher Udry (forthcoming), 'Learning about a new technology: Pineapple in ghana', *American Economic Review* .

Cooley, Jane (2010), Desegregation and the achievement gap: Do diverse peers help? working paper.

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

Currarini, Sergio, Matthew O. Jackson and Paolo Pin (2010), 'Identifying the roles of race-based choice and chance in high school friendship network formation', *the Proceedings of the National Academy of Sciences* **107**(11), 48574861.

De Giorgi, Giacomo, Michele Pellizzari and Silvia Redaelli (2010), 'Identification of social interactions through partially overlapping peer groups', *American Economic Journal: Applied Economics* **2**(2).

De Marti, Joan and Yves Zenou (2009), Ethnic identity and social distance in friendship formation, Cepr discussion papers, C.E.P.R. Discussion Papers.

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

Diaconis, Persi and D. Stroock (1991), 'Geometric bounds for eigenvalues of markov chains', *Annals of Applied Probability* **1**(1), 36–61.

Diaconis, Persi and Sourav Chatterjee (2011), Estimating and understanding exponential random graph models.

Echenique, Federico, Roland G. Fryer and Alex Kaufman (2006), 'Is school segregation good or bad?', *American Economic Review* **96**(2), 265–269.

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

Gelman, A., G. O. Roberts and W. R. Gilks (1996), 'Efficient metropolis jumping rules', *Bayesian Statistics* **5**, 599–608.

Gelman, A., J. Carlin, H. Stern and D. Rubin (2003), *Bayesian Data Analysis, Second Edition*, Chapman & Hall/CRC.

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.

Gilles, Robert P. and Sudipta Sarangi (2004), Social network formation with consent, Discussion paper, Tilburg University, Center for Economic Research.

Goldsmith-Pinkham, Paul and Guido W. Imbens (2013), 'Social networks and the identification of peer effects', *Journal of Business and Economic Statistics* **31**(3), 253–264.

Golub, Benjamin and Matthew Jackson (2011), 'Network structure and the speed of learning: Measuring homophily based on its consequences', *Annals of Economics and Statistics* .

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

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

Heckman, James J. (1978), 'Dummy endogenous variables in a simultaneous equation system', *Econometrica* **46**(4), 931–959.

Hsieh, Chih-Sheng and Lung-Fei Lee (2012), A structural modeling approach for network formation and social interactions with applications to students' friendship choices and selectivity on activities.

Jackson, Matthew and Alison Watts (2001), 'The existence of pairwise stable networks', *Seoul Journal of Economics* **14**(3), 299–321.

Jackson, Matthew and Allison Watts (2002), 'The evolution of social and economic networks', *Journal of Economic Theory* **106**(2), 265–295.

Jackson, Matthew and Asher Wolinsky (1996), 'A strategic model of social and economic networks', *Journal of Economic Theory* **71**(1), 44–74.

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

Koskinen, Johan H. (2008), The linked importance sampler auxiliary variable metropolis hastings algorithm for distributions with intractable normalising constants. MelNet Social Networks Laboratory Technical Report 08-01, Department of Psychology, School of Behavioural Science, University of Melbourne, Australia.

Laschever, Ron (2009), The doughboys network: Social interactions and labor market outcomes of world war i veterans. working paper.

Lehman, E. L. (1983), *Theory of Point Estimation*, Wiley and Sons.

Leung, Michael (2014*a*), A random-field approach to inference in large models of network formation. working paper.

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

Levin, David. A, Yuval Peres and Elizabeth L. Wilmer (2008), *Markov Chains and Mixing Times*, American Mathematical Society.

Liang, Faming (2010), 'A double metropolis-hastings sampler for spatial models with intractable normalizing constants', *Journal of Statistical Computing and Simulation* **forthcoming**.

Liang, Faming, Chuanhai Liu and Raymond J. Carroll (2010), *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*, John Wiley and Sons, Ltd.

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

Lovasz, Laszlo and Balazs Szegedy (2007), 'Szemeredi's lemma for the analyst', *GAFA Geometric And Functional Analysis* **17**(1), 252–270.

Mele, Angelo (2015), Approximate variational inference for a model of social interactions. Working Paper.

Meyn, Sean and Richard L. Tweedie (2009), *Markov Chains and Stochastic Stability*, Cambridge University Press.

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.

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

Nakajima, Ryo (2007), 'Measuring peer effects on youth smoking behavior', *Review of Economic Studies* **74**(3), 897–935.

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

Robert, Christian P. and George Casella (2005), *Monte Carlo Statistical Methods (Springer Texts in Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Rossi, P., R. McCulloch and G. Allenby (1996), 'The value of purchase history data in target marketing', *Marketing Science* **15**(4), 321–340.

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).

Tamer, Elie (2003), 'Incomplete simultaneous discrete response model with multiple equilibria', *The Review of Economic Studies* **70**(1), 147–165.

Topa, Giorgio (2001), 'Social interactions, local spillovers and unemployment', *Review of Economic Studies* **68**(2), 261–295.

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

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

A Proofs

Proof of Proposition 1
The potential is a function Q from the space of actions to the real line such that $Q\left(g_{ij}, g_{-ij}, X\right) - Q\left(g'_{ij}, g_{-ij}, X\right) = U_i\left(g_{ij}, g_{-ij}, X\right) - U_i\left(g'_{ij}, g_{-ij}, X\right)$, for any ij.[47] A simple computation shows that, for any ij

$$\begin{aligned} Q\left(g_{ij}=1, g_{-ij}, X\right) - Q\left(g_{ij}=0, g_{-ij}, X\right) &= u_{ij} + g_{ji} m_{ij} + \sum_{\substack{k=1 \\ k\neq i,j}}^{n} g_{jk} v_{ik} + \sum_{\substack{k=1 \\ k\neq i,j}}^{n} g_{ki} v_{kj} \\ &= U_i\left(g_{ij}=1, g_{-ij}, X\right) - U_i\left(g_{ij}=0, g_{-ij}, X\right) \end{aligned}$$

therefore Q is the potential of the network formation game.

Proof of Proposition 2
The proof consists of showing that $Q\left(g, X\right)$ can be written in the form $\theta' \mathbf{t}\left(g, X\right)$. Consider the first part of the potential

$$\begin{aligned} \sum_i \sum_j g_{ij} u_{ij} &= \sum_i \sum_j g_{ij} \sum_{p=1}^{P} \theta_{up} H_{up}\left(X_i, X_j\right) \\ &= \sum_{p=1}^{P} \theta_{up} \sum_i \sum_j g_{ij} H_{up}\left(X_i, X_j\right) \\ &\equiv \sum_{p=1}^{P} \theta_{up} t_{up}\left(g, X\right) \\ &= \theta'_u \mathbf{t}_u\left(g, X\right) \end{aligned}$$

where $t_{up}\left(g, X\right) \equiv \sum_i \sum_j g_{ij} H_{up}\left(X_i, X_j\right)$, $\theta_u = \left(\theta_{u1}, ..., \theta_{uP}\right)'$ and $\mathbf{t}_u\left(g, X\right) = \left(t_{u1}\left(g, X\right), ..., t_{uP}\left(g, X\right)\right)'$. Analogously define $\theta_m = \left(\theta_{m1}, \theta_{m2}, ..., \theta_{mL}\right)'$ and $\mathbf{t}_m\left(g, X\right) = \left(t_{m1}\left(g, X\right), t_{m2}\left(g, X\right), ..., t_{mL}\left(g, X\right)\right)'$ and $\theta_v = \left(\theta_{v1}, \theta_{v2}, ..., \theta_{vS}\right)'$ and $\mathbf{t}_v\left(g, X\right) = \left(t_{v1}\left(g, X\right), t_{v2}\left(g, X\right), ..., t_{vS}\left(g, X\right)\right)'$. It follows that

$$\begin{aligned} \sum_i \sum_{j>i} g_{ij} g_{ji} m_{ij} &= \sum_i \sum_{j>i} g_{ij} g_{ji} \sum_{l=1}^{L} \theta_{ml} H_{ml}\left(X_i, X_j\right) \\ &= \sum_{l=1}^{L} \theta_{ml} \sum_i \sum_{j>i} g_{ij} g_{ji} H_{ml}\left(X_i, X_j\right) \\ &= \sum_{l=1}^{L} \theta_{ml} t_{ml}\left(g, X\right) \\ &= \theta'_m \mathbf{t}_m\left(g, X\right) \end{aligned}$$

[47] For more details and definitions see Monderer and Shapley (1996).

and

$$\begin{aligned}\sum_i \sum_j g_{ij} \sum_{k \neq i,j} g_{jk} v_{ij} &= \sum_i \sum_j g_{ij} \sum_{k \neq i,j} g_{jk} \sum_{s=1}^{S} \theta_{vs} H_{vs}(X_i, X_k) \\ &= \sum_{s=1}^{S} \theta_{vs} \sum_i \sum_j g_{ij} \sum_{k \neq i,j} g_{jk} H_{vs}(X_i, X_k) \\ &= \sum_{s=1}^{S} \theta_{vs} t_{vs}(g, X) \\ &= \theta_v' \mathbf{t}_v(g, X)\end{aligned}$$

Therefore $Q(g, X)$ can be written in the form $\theta' \mathbf{t}(g, X)$, where $\theta = (\theta_u, \theta_m, \theta_v)'$ and $\mathbf{t}(g, X) = [\mathbf{t}_u(g, X), \mathbf{t}_m(g, X), \mathbf{t}_v(g, X)]'$

$$\begin{aligned}Q(g, X) &= \theta_u' \mathbf{t}_u(g, X) + \theta_m' \mathbf{t}_m(g, X) + \theta_v' \mathbf{t}_v(g, X) \\ &= \theta' \mathbf{t}(g, X)\end{aligned}$$

and the stationary distribution is

$$\pi(g, X) = \frac{\exp[\theta' \mathbf{t}(g, X)]}{\sum_{\omega \in \mathcal{G}} \exp[\theta' \mathbf{t}(\omega, X)]}.$$

Model without preference shocks: characterization of Nash networks

It is helpful to consider a *special case* of the model, in which there are no preference shocks: the characterization of equilibria and long run behavior for such model provides intuition about the dynamic properties of the full structural model.
Let $\mathcal{N}(g)$ be the set of networks that differ from g by only one element of the matrix, i.e.

$$\mathcal{N}(g) \equiv \{g' : g' = (g'_{ij}, g_{-ij}), \text{ for all } g'_{ij} \neq g_{ij}, \text{ for all } i, j \in \mathcal{I}\}. \tag{29}$$

A Nash network is defined as a network in which any player has no profitable deviations from his current linking strategy, when randomly selected from the population. The following results characterize the set of the pure-strategy Nash equilibria and the long run behavior of the model with no shocks.

PROPOSITION 3 ***(Model without Shocks: Equilibria and Long Run)***
Consider the model without idiosyncratic preference shocks. Under Assumptions 1 and 2:

1. *There exists at least one pure-strategy Nash equilibrium network*

2. *The set $\mathcal{NE}(\mathcal{G}, X, U)$ of all pure-strategy Nash equilibria of the network formation game is completely characterized by the local maxima of the potential function.*

$$\mathcal{NE}(\mathcal{G}, X, U) = \left\{ g^* : g^* = \arg \max_{g \in \mathcal{N}(g^*)} Q(g, X) \right\} \tag{30}$$

3. *Any pure-strategy Nash equilibrium is an absorbing state.*

4. *As $t \to \infty$, the network converges to one of the Nash networks with probability 1.*

Proof. 1) The existence of Nash equilibria follows directly from the fact that the network formation game is a potential game with finite strategy space. (see Monderer and Shapley (1996) for details)
2) The set of Nash equilibria is defined as the set of g^* such that, for every i and for every $g_{ij} \neq g^*_{ij}$

$$U_i\left(g^*_{ij}, g^*_{-ij}, X\right) \geq U_i\left(g_{ij}, g^*_{-ij}, X\right)$$

Therefore, since Q is a potential function, for every $g_{ij} \neq g^*_{ij}$

$$Q\left(g^*_{ij}, g^*_{-ij}, X\right) \geq Q\left(g_{ij}, g^*_{-ij}, X\right)$$

Therefore g^* is a maximizer of Q. The converse is easily checked by the same reasoning.
3) Suppose $g^t = g^*$. Since this is a Nash equilibrium, no player will be willing to change her linking decision when her turn to play comes. Therefore, once the chain reaches a Nash equilibrium, it cannot escape from that state.
4) The probability that the potential will increase from t to $t+1$ is

$$Pr\left[Q\left(g^{t+1}, X\right) \geq Q\left(g^t, X\right)\right] =$$

$$= \sum_i \sum_j Pr\left(m^{t+1} = ij\right) \underbrace{Pr\left[U_i\left(g^{t+1}_{ij}, g^t_{-ij}, X\right) \geq U_i\left(g^t_{ij}, g^t_{-ij}, X\right) \middle| m^{t+1} = ij\right]}_{=1 \text{ because agents play Best Response, conditioning on } m^{t+1}}$$

$$= \sum_i \sum_j \rho_{ij} = 1.$$

By part 3) of the proposition, a Nash network is an absorbing state of the chain. Therefore any probability distribution that puts probability 1 on a Nash network is a stationary distribution. For any initial network, the chain will converge to one of the stationary distributions. It follows that in the long run the model will be in a Nash network, i.e. for any $g^0 \in \mathcal{G}$

$$\lim_{t \to \infty} Pr\left[g^t \in NE \middle| g^0\right] = 1.$$

■

Proof of Theorem 1

1. The sequence of networks $[g^0, g^1, ...]$ generated by the network formation game is a markov chain. Inspection of the transition probability proves that the chain is irreducible and aperiodic, therefore it is ergodic. The existence of a unique stationary distribution then follows from the ergodic theorem (see Gelman et al. (1996) for details).

2. A sufficient condition for stationarity is the *detailed balance* condition. In our case this requires

$$P_{gg'}\pi_g = P_{g'g}\pi_{g'} \tag{31}$$

where

$$\begin{aligned} P_{gg'} &= \Pr\left(g^{t+1} = g' \middle| g^t = g\right) \\ \pi_g &= \pi\left(g^t = g\right) \end{aligned}$$

Notice that the transition from g to g' is possible if these networks differ by only one element g_{ij}. Otherwise the transition probability is zero and the detailed balance condition is satisfied. Let's consider the nonzero probability transitions, with $g = (1, g_{-ij})$ and $g' = (0, g_{-ij})$. Define $\Delta Q \equiv Q(1, g_{-ij}, X) - Q(0, g_{-ij}, X)$.

$$\begin{aligned} P_{gg'}\pi_g &= \Pr\left(m^t = ij\right) Pr\left(g_{ij} = 0 \middle| g_{-ij}\right) \frac{\exp\left[Q\left(1, g_{-ij}, X\right)\right]}{\sum_{\omega\in\mathcal{G}} \exp\left[Q\left(\omega, X\right)\right]} \\ &= \rho\left(g_{-ij}, X_i, X_j\right) \times \frac{1}{1+\exp\left[\Delta Q\right]} \times \frac{\exp\left[Q\left(1, g_{-ij}, X\right) + Q\left(0, g_{-ij}, X\right) - Q\left(0, g_{-ij}, X\right)\right]}{\sum_{\omega\in\mathcal{G}} \exp\left[Q\left(\omega, X\right)\right]} \\ &= \rho\left(g_{-ij}, X_i, X_j\right) \times \frac{1}{1+\exp\left[\Delta Q\right]} \times \frac{\exp\left[Q\left(1, g_{-ij}, X\right) - Q\left(0, g_{-ij}, X\right)\right] \exp\left[Q\left(0, g_{-ij}, X\right)\right]}{\sum_{\omega\in\mathcal{G}} \exp\left[Q\left(\omega, X\right)\right]} \\ &= \rho\left(g_{-ij}, X_i, X_j\right) \frac{\exp\left[\Delta Q\right]}{1+\exp\left[\Delta Q\right]} \frac{\exp\left[Q\left(0, g_{-ij}, X\right)\right]}{\sum_{\omega\in\mathcal{G}} \exp\left[Q\left(\omega, X\right)\right]} \\ &= \Pr\left(m^t = ij\right) \Pr\left(g_{ij} = 1 \middle| g_{-ij}\right) \frac{\exp\left[Q\left(0, g_{-ij}, X\right)\right]}{\sum_{\omega\in\mathcal{G}} \exp\left[Q\left(\omega, X\right)\right]} \\ &= P_{g'g}\pi_{g'} \end{aligned}$$

So the distribution (5) satisfies the detailed balance condition. Therefore it is a stationary distribution for the network formation model. From part 1) of the proposition, we know that the process is ergodic and it has a unique stationary distribution. Therefore $\pi(g, X)$ is also the unique stationary distribution.

B Computational Details

B.1 Network Simulation

The algorithm used to simulate the network (**ALGORITHM 1**) produces samples from the stationary equilibrium of the model.

1. The network simulation algorithm satisfies the detailed balance condition for the stationary distribution 5. Indeed for any given θ

$$\begin{aligned}
\Pr\left(g'|g,X,\theta\right)\pi\left(g,X,\theta\right) &= q_g\left(g'|g\right)\min\left\{1,\frac{\exp\left[Q\left(g',X,\theta\right)\right]}{\exp\left[Q\left(g,X,\theta\right)\right]}\frac{q_g\left(g|g'\right)}{q_g\left(g'|g\right)}\right\}\frac{\exp\left[Q\left(g,X,\theta\right)\right]}{c\left(\mathcal{G},X,\theta\right)} \\
&= \min\left\{q_g\left(g'|g\right)\frac{\exp\left[Q\left(g,X,\theta\right)\right]}{c\left(\mathcal{G},X,\theta\right)},\frac{\exp\left[Q\left(g',X,\theta\right)\right]}{c\left(\mathcal{G},X,\theta\right)}q_g\left(g|g'\right)\right\} \\
&= q_g\left(g|g'\right)\min\left\{\frac{q_g\left(g'|g\right)}{q_g\left(g|g'\right)}\frac{\exp\left[Q\left(g,X,\theta\right)\right]}{c\left(\mathcal{G},X,\theta\right)},\frac{\exp\left[Q\left(g',X,\theta\right)\right]}{c\left(\mathcal{G},X,\theta\right)}\right\} \\
&= q_g\left(g|g'\right)\min\left\{\frac{q_g\left(g'|g\right)}{q_g\left(g|g'\right)}\frac{\exp\left[Q\left(g,X,\theta\right)\right]}{\exp\left[Q\left(g',X,\theta\right)\right]},1\right\}\frac{\exp\left[Q\left(g',X,\theta\right)\right]}{c\left(\mathcal{G},X,\theta\right)} \\
&= \Pr\left(g|g',X,\theta\right)\pi\left(g',X,\theta\right)
\end{aligned}$$

   This concludes the proof.

2. The algorithm generates a Markov Chain of network with finite state space. The chain is irreducible and aperiodic and therefore it is uniformly ergodic (see Theorem 4.9, page 52 in Levin et al. (2008)).

3. The bound to the convergence rate used in the text was derived by Diaconis and Stroock (1991), for reversible finite chains.

The algorithm has a very useful property that can be exploited in the posterior simulation to reduce the computational burden. Adapting the suggestion in Liang (2010), define $\mathcal{P}_{\theta'}^{(R)}(g'|g)$ as the transition probability of a Markov chain that generates g' with R Metropolis-Hastings updates of the network simulation algorithm, starting at the observed network g and using the proposed parameter θ'. Then,

$$\mathcal{P}_{\theta'}^{(R)}(g'|g) = \mathcal{P}_{\theta'}(g^1|g)\mathcal{P}_{\theta'}(g^2|g^1)\cdots\mathcal{P}_{\theta'}(g'|g^{R-1}), \tag{32}$$

where $\mathcal{P}_{\theta'}(g^j|g^i) = q_g(g^j|g^i)\alpha_{mh}(g^i,g^j)$ is the transition probability of the network simulation algorithm above. Since the Metropolis-Hastings algorithm satisfies the detailed balance condition, we can prove the following

LEMMA 1 *Simulate a network g' from the stationary distribution $\pi(\cdot, X, \theta')$ using a Metropolis-Hastings algorithm starting at the network g observed in the data. Then*

$$\frac{\mathcal{P}_{\theta'}^{(R)}(g|g')}{\mathcal{P}_{\theta'}^{(R)}(g'|g)} = \frac{\exp\left[Q(g, X, \theta')\right]}{\exp\left[Q(g', X, \theta')\right]} \tag{33}$$

for all R, $g, g' \in \mathcal{G}$ and for any $\theta' \in \Theta$.

Proof. Let $\mathcal{P}_{\theta'}^{(R)}(g'|g)$ be defined as in (32). This is the transition probability of the chain that generates g' with R Metropolis-Hastings updates, starting at the observed network g and using the proposed parameter θ'. Notice that the Metropolis-Hastings algorithm satisfies the detailed balance for $\pi(g, X, \theta')$, therefore we have

$$\begin{aligned}
\mathcal{P}_{\theta'}^{(R)}(g|g')\pi(g', X, \theta') &= \mathcal{P}_{\theta'}(g_{R-1}|g')\mathcal{P}_{\theta'}(g_{R-2}|g_{R-1}) \cdots \mathcal{P}_{\theta'}(g|g_1)\pi(g', X, \theta') \\
&= \mathcal{P}_{\theta'}(g_1|g)\mathcal{P}_{\theta'}(g_2|g_1) \cdots \mathcal{P}_{\theta'}(g'|g_{R-1})\pi(g, X, \theta') \\
&= \mathcal{P}_{\theta'}^{(R)}(g'|g)\pi(g, X, \theta')
\end{aligned}$$

It follows that

$$\begin{aligned}
\frac{\mathcal{P}_{\theta'}^{(R)}(g|g')}{\mathcal{P}_{\theta'}^{(R)}(g'|g)} &= \frac{\pi(g, X, \theta')}{\pi(g', X, \theta')} \\
&= \frac{\exp\left[Q(g, X, \theta')\right]}{\exp\left[Q(g', X, \theta')\right]} \frac{c(\mathcal{G}, X, \theta')}{c(\mathcal{G}, X, \theta')} \\
&= \frac{\exp\left[Q(g, X, \theta')\right]}{\exp\left[Q(g', X, \theta')\right]}.
\end{aligned}$$

This concludes the proof. ■

One should notice that as long as the algorithm is started from the network g observed in the data (which is assumed to be a draw from the stationary equilibrium of the model), the equality in (33) is satisfied for any R.

The approximate exchange algorithm presented in this paper removes the requirement of exact sampling by exploiting the property of the stationary equilibrium characterization, described in Lemma 1.

B.2 Posterior Simulation

In this section I provide the technical details for the algorithm proposed in the empirical part of the paper. The first set of results show that the exchange algorithm generate (approximate) samples from the posterior distribution (8).

The original exchange algorithm developed in Murray et al. (2006) is slightly different from the one used here. The main modification is in Step 2: the original algorithm requires an *exact* sample from the stationary equilibrium of the model.

ALGORITHM 3 ***(EXACT EXCHANGE ALGORITHM)***
Start at current parameter $\theta_t = \theta$ and network data g.

1. *Propose a new parameter vector θ'*

$$\theta' \sim q_\theta(\cdot|\theta) \tag{34}$$

2. *Draw an exact sample network g' from the likelihood*

$$g' \sim \pi\left(\cdot|X, \theta'\right) \tag{35}$$

3. *Compute the acceptance ratio*

$$\begin{aligned} \alpha_{ex}\left(\theta, \theta', g', g\right) &= \min\left\{1, \frac{\exp\left[Q(g', X, \theta)\right]}{\exp\left[Q(g, X, \theta)\right]} \frac{p\left(\theta'\right)}{p\left(\theta\right)} \frac{q_\theta\left(\theta|\theta'\right)}{q_\theta\left(\theta'|\theta\right)} \frac{\exp\left[Q(g, X, \theta')\right]}{\exp\left[Q(g', X, \theta')\right]} \frac{c(\theta)c(\theta')}{c(\theta)c(\theta')}\right\} \\ &= \min\left\{1, \frac{\exp\left[Q(g', X, \theta)\right]}{\exp\left[Q(g, X, \theta)\right]} \frac{p\left(\theta'\right)}{p\left(\theta\right)} \frac{q_\theta\left(\theta|\theta'\right)}{q_\theta\left(\theta'|\theta\right)} \frac{\exp\left[Q(g, X, \theta')\right]}{\exp\left[Q(g', X, \theta')\right]}\right\} \end{aligned} \tag{36}$$

4. *Update the parameter according to*

$$\theta_{t+1} = \begin{cases} \theta' & \textit{with prob. } \alpha_{ex}\left(\theta, \theta', g', g\right) \\ \theta & \textit{with prob. } 1 - \alpha_{ex}\left(\theta, \theta', g', g\right) \end{cases} \tag{37}$$

The difference between this algorithm and the approximate one is in step 2. The exact and approximate algorithms use the same acceptance ratio $\alpha_{ex}\left(\theta, \theta', g', g\right)$, a consequence of LEMMA 1. Indeed the acceptance ratio for the approximate algorithm is

$$\widetilde{\alpha}_{ex}\left(\theta, \theta', g', g\right) = \min\left\{1, \frac{\exp\left[Q(g', X, \theta)\right]}{\exp\left[Q(g, X, \theta)\right]} \frac{p\left(\theta'\right)}{p\left(\theta\right)} \frac{q_\theta\left(\theta|\theta'\right)}{q_\theta\left(\theta'|\theta\right)} \frac{\mathcal{P}_{\theta'}^{(R)}(g|g')}{\mathcal{P}_{\theta'}^{(R)}(g'|g)}\right\} \tag{38}$$

$$= \min\left\{1, \frac{\exp\left[Q(g', X, \theta)\right]}{\exp\left[Q(g, X, \theta)\right]} \frac{p\left(\theta'\right)}{p\left(\theta\right)} \frac{q_\theta\left(\theta|\theta'\right)}{q_\theta\left(\theta'|\theta\right)} \frac{\exp\left[Q(g, X, \theta')\right]}{\exp\left[Q(g', X, \theta')\right]}\right\} \tag{39}$$

$$= \alpha_{ex}\left(\theta, \theta', g', g\right) \tag{40}$$

This result implies that to prove the convergence of the approximate algorithm to the exact algorithm, there is no need to prove convergence of $\widetilde{\alpha}_{ex}\left(\theta, \theta', g', g\right)$ to $\alpha_{ex}\left(\theta, \theta', g', g\right)$. The convergence of step 2 of the algorithm is sufficient.

B.2.1 Preliminary Lemmas for THEOREM 6

The convergence of the approximate exchange algorithm to the correct posterior distribution is proven in 4 steps.

1. First we prove that the exact exchange algorithm converges to the correct posterior (LEMMA 2)
2. Second, we prove that the approximate algorithm has a stationary distribution and it is ergodic (LEMMA 3, similar to the one in Liang 2010)
3. Third, we prove that the transition kernel of the approximate and exact algorithms are arbitrarily close for a large enough number of network simulations (LEMMA 4)
4. Fourth, we combine previous results to prove that the approximate algorithm converges to the correct posterior

A similar proof strategy is contained in Liang et al. (2010) and Andrieu and Roberts (2009).

Let $Q\left(d\vartheta|\theta\right) = q_{\theta}\left(\vartheta|\theta\right)\nu\left(d\vartheta\right)$. The transition kernel of the exact exchange algorithm can be written as

$$\begin{aligned} P\left(\theta, d\vartheta\right) &= \left[\sum_{g' \in \mathcal{G}} \pi\left(g', \vartheta\right) \alpha_{ex}\left(\theta, \vartheta, g', g\right)\right] Q\left(\theta, d\vartheta\right) \\ &+ \delta_{\theta}\left(d\vartheta\right)\left\{1 - \int_{\Theta}\left[\sum_{g' \in \mathcal{G}} \pi\left(g', \vartheta\right) \alpha_{ex}\left(\theta, \vartheta, g', g\right)\right] Q\left(\theta, d\vartheta\right)\right\} \end{aligned}$$

and the transition kernel of the approximate exchange algorithm can be written as

$$\begin{aligned} \widetilde{P}_{R}\left(\theta, d\vartheta\right) &= \left[\sum_{g' \in \mathcal{G}} \mathcal{P}_{\vartheta}^{(R)}\left(g'|g\right) \alpha_{ex}\left(\theta, \vartheta, g', g\right)\right] Q\left(\theta, d\vartheta\right) \\ &+ \delta_{\theta}\left(d\vartheta\right)\left\{1 - \int_{\Theta}\left[\sum_{g' \in \mathcal{G}} \mathcal{P}_{\vartheta}^{(R)}\left(g'|g\right) \alpha_{ex}\left(\theta, \vartheta, g', g\right)\right] Q\left(\theta, d\vartheta\right)\right\} \end{aligned}$$

Let $\eta\left(\theta\right)$ be the average rejection probability for the approximate algorithm, i.e.

$$\eta\left(\theta\right) := 1 - \int_{\Theta}\left[\sum_{g' \in \mathcal{G}} \mathcal{P}_{\vartheta}^{(R)}\left(g'|g\right) \alpha_{ex}\left(\theta, \vartheta, g', g\right)\right] Q\left(\theta, d\vartheta\right) \tag{41}$$

The next lemma proves that the transition kernel satisfies the detailed balance condition for the posterior distribution. For any pair of parameters $(\theta, \vartheta) \in \Theta$ we have

$$P\left[\theta, \vartheta|g, X\right] p\left(\theta|g, X\right) = \Pr\left[\theta|\vartheta, g, X\right] p\left(\vartheta|g, X\right) \tag{42}$$

The detailed balance condition is sufficient condition for the Markov chain generated by the algorithm to have stationary distribution the posterior (8) (for details see Robert and Casella (2005) or Gelman et al. (2003)).

LEMMA 2 *The exchange algorithm produces a Markov chain with invariant distribution (8).*

Proof. Define $\mathcal{Z} \equiv \int_{\Theta} \pi\left(g|X,\theta\right) p\left(\theta\right) d\theta$. In the algorithm the probability $\Pr\left[\vartheta|\theta,g,X\right]$ of transition to θ_j, given the current parameter θ and the observed data (g,X), can be computed as

$$\Pr\left[\vartheta|\theta,g,X\right] = q_{\theta}\left(\vartheta|\theta\right) \frac{\exp\left[Q(g',X,\vartheta)\right]}{c(\mathcal{G},X,\vartheta)} \alpha_{ex}\left(\theta,\vartheta,g',g\right). \tag{43}$$

This is the probability $q_{\theta}\left(\vartheta|\theta\right)$ of proposing ϑ times the probability of generating the new network g' from the model's stationary distribution, $\frac{\exp[Q(g',X,\vartheta)]}{c(\mathcal{G},X,\vartheta)}$ and accepting the proposed parameter $\alpha_{ex}\left(\theta,\vartheta,g',g\right)$. Therefore the left-hand side of (42) can be written as

$$\begin{aligned}
\Pr\left[\vartheta|\theta,g,X\right] p\left(\theta|g,X\right) &= q_{\theta}\left(\vartheta|\theta\right) \frac{\exp\left[Q(g',X,\vartheta)\right]}{c(\mathcal{G},X,\vartheta)} \alpha_{ex}\left(\theta,\vartheta,g',g\right) p\left(\theta|g,X\right) \\
&= q_{\theta}\left(\vartheta|\theta\right) \frac{\exp\left[Q(g',X,\vartheta)\right]}{c(\mathcal{G},X,\vartheta)} \alpha_{ex}\left(\theta,\vartheta,g',g\right) \frac{\frac{\exp[Q(g,X,\theta)]}{c(\mathcal{G},X,\theta)} p\left(\theta\right)}{\mathcal{Z}} \\
&= q_{\theta}\left(\vartheta|\theta\right) \frac{\exp\left[Q(g',X,\vartheta)\right]}{c(\mathcal{G},X,\vartheta)} \\
&\times \min\left\{1, \frac{\exp\left[Q(g',X,\theta)\right]}{\exp\left[Q(g,X,\theta)\right]} \frac{p\left(\vartheta\right)}{p\left(\theta\right)} \frac{q_{\theta}\left(\theta|\vartheta\right)}{q_{\theta}\left(\vartheta|\theta\right)} \frac{\exp\left[Q(g,X,\vartheta)\right]}{\exp\left[Q(g',X,\vartheta)\right]}\right\} \\
&\times \frac{\frac{\exp[Q(g,X,\theta)]}{c(\mathcal{G},X,\theta)} p\left(\theta\right)}{\mathcal{Z}}
\end{aligned}$$

$$= \min\left\{q_{\theta}\left(\vartheta|\theta\right) \frac{\exp\left[Q(g',X,\vartheta)\right]}{c(\mathcal{G},X,\vartheta)} \frac{\exp\left[Q(g,X,\theta)\right]}{c(\mathcal{G},X,\theta)} \frac{p\left(\theta\right)}{\mathcal{Z}}, q_{\theta}\left(\theta|\vartheta\right) \frac{\exp\left[Q(g',X,\theta\right]}{c(\mathcal{G},X,\theta)} \frac{\exp\left[Q(g,X,\vartheta)\right]}{c(\mathcal{G},X,\vartheta)} \frac{p\left(\vartheta\right)}{\mathcal{Z}}\right\}$$

$$\begin{aligned}
&= q_{\theta}\left(\theta|\vartheta\right) \frac{\exp\left[Q(g',X,\theta)\right]}{c(\mathcal{G},X,\theta)} \frac{\exp\left[Q(g,X,\vartheta)\right]}{c(\mathcal{G},X,\vartheta)} \frac{p\left(\vartheta\right)}{\mathcal{Z}} \times \\
&\times \min\left\{1, \frac{\exp\left[Q(g',X,\vartheta)\right]}{\exp\left[Q(g,X,\vartheta)\right]} \frac{p\left(\theta\right)}{p\left(\vartheta\right)} \frac{q_{\theta}\left(\vartheta|\theta\right)}{q_{\theta}\left(\theta|\vartheta\right)} \frac{\exp\left[Q(g,X,\theta)\right]}{\exp\left[Q(g',X,\theta)\right]}\right\} \\
&= q_{\theta}\left(\theta|\vartheta\right) \frac{\exp\left[Q(g',X,\theta)\right]}{c(\mathcal{G},X,\theta)} \alpha(\vartheta,\theta,g',g) \frac{\exp\left[Q(g,X,\vartheta)\right]}{c(\mathcal{G},X,\vartheta)} \frac{p\left(\vartheta\right)}{\mathcal{Z}} \\
&= q_{\theta}\left(\theta|\vartheta\right) \frac{\exp\left[Q(g',X,\theta)\right]}{c(\mathcal{G},X,\theta)} \alpha(\vartheta,\theta,g',g) p\left(\vartheta|g,X\right) \\
&= \Pr\left[\theta|\vartheta,g,X\right] p\left(\vartheta|g,X\right)
\end{aligned}$$

The latter step proves the detailed balance for a generic network g'. Since the condition is satisfied for any network g', detailed balance follows from summing over all possible networks.

■

LEMMA 3 *(The approximate algorithm is ergodic)*
Assume the exact exchange algorithm is ergodic and that for any $\vartheta \in \Theta$

$$\frac{\mathcal{P}_{\vartheta}^{(R)}(g'|g)}{\pi(g',\vartheta)} > 0 \quad \textit{for any } g' \in \mathcal{G} \tag{44}$$

Then for any $R \in \mathbb{N}$ *such that for any* $\theta \in \Theta$, $\rho(\theta) > 0$, *the transition kernel of the approximate algorithm* $\widetilde{P}_R$ *is also irreducible and aperiodic, and there exists a stationary distribution* $\widetilde{p}(\theta)$ *such that*

$$\lim_{s\to\infty} \left\| \widetilde{P}_R^{(s)}(\theta_0,\cdot) - \widetilde{p}(\theta) \right\|_{TV} = 0 \tag{45}$$

Proof. The exact algorithm with transition kernel P is an irreducible and aperiodic Markov chain. To prove that the approximate algorithm with transition kernel $\widetilde{P}_R$ defines an ergodic Markov chain, it is sufficient to prove that the set of accessible states of P are also included in those of $\widetilde{P}_R$. The proof proceeds by induction.
Formally, we need to show that for any $s \in \mathbb{N}$, $\theta \in \Theta$ and $A \in \mathcal{B}(\Theta)$ such that $P^{(s)}(\theta, A) > 0$, implies $\widetilde{P}_R^{(s)}(\theta, A) > 0$.
Notice that for any $\theta \in \Theta$ and $A \in \mathcal{B}(\Theta)$,

$$\begin{aligned}
\widetilde{P}_R(\theta, A) &= \int_A \left[\sum_{g' \in \mathcal{G}} \mathcal{P}_{\vartheta}^{(R)}(g'|g)\, \alpha_{ex}(\theta, \vartheta, g', g) \right] q_\theta(\vartheta|\theta)\, d\vartheta + \mathbb{I}(\theta \in A)\, \eta(\theta) \\
&\geq \int_A \left[\sum_{g' \in \mathcal{G}} \min\left\{ 1, \frac{\mathcal{P}_{\vartheta}^{(R)}(g'|g)}{\pi(g',\vartheta)} \right\} \pi(g',\vartheta)\, \alpha_{ex}(\theta, \vartheta, g', g) \right] q_\theta(\vartheta|\theta)\, d\vartheta + \mathbb{I}(\theta \in A)\, \eta(\theta) > 0
\end{aligned}$$

where the last inequality comes from $\frac{\mathcal{P}_{\vartheta}^{(R)}(g'|g)}{\pi(g',\vartheta)} > 0$ for any $g' \in \mathcal{G}$ and $\vartheta \in \Theta$.
This proves that the statement is true when $s = 1$. By induction we assume that it is true up to $s = n \geq 1$ and for some $\theta \in \Theta$ chose $A \in \mathcal{B}(\Theta)$ such that $P^{(n+1)}(\theta, A) > 0$ and assume that

$$\int_\Theta \widetilde{P}_R^{(n)}(\theta, d\vartheta)\, \widetilde{P}_R(\vartheta, A) = 0$$

This implies that $\widetilde{P}_R(\vartheta, A) = 0$, $\widetilde{P}_R^{(n)}(\theta,\cdot)$-a.s.; by the induction assumption at $s = 1$ it follows that $P(\vartheta, A) = 0$, $\widetilde{P}_R^{(n)}(\theta,\cdot)$-a.s.
From this and the induction assumption at $s = n$, $P(\vartheta, A) = 0$, $P^{(n)}(\theta,\cdot)$-a.s. (assume not, then $P(\vartheta, A) > 0$, $P^{(n)}(\theta,\cdot)$-a.s. which by induction would imply $\widetilde{P}_R(\vartheta, A) > 0$, which is a contradiction). The latter step contradicts $P^{(n+1)}(\theta, A) > 0$ and the result follows. ■

The next step consists of proving that the transition kernel of the approximate algorithm

$\widetilde{P}_R(\theta,\vartheta)$ and the exact algorithm $P(\theta,\vartheta)$ are arbitrarily close for a large enough number of network simulations R. Formally we prove a statement which is equivalent to proving convergence in total variation norm.[48]

LEMMA 4 *(Convergence of the exact and approximate transition kernels)*
Let $\epsilon \in (0,1]$. *There exists a number of simulations* $R_0 \in \mathbb{N}$ *such that for any function* $\phi : \Theta \to [-1,1]$ *and any* $R > R_0$,

$$\left|\widetilde{P}_R\phi(\theta) - P\phi(\theta)\right| < 2\epsilon \tag{46}$$

Proof. The transition of the exchange algorithm is

$$\begin{aligned} P(\phi(\theta),\phi(\vartheta)) &= \int_\Theta \phi(\vartheta)\left[\sum_{g'\in\mathcal{G}} \pi(g',\vartheta)\,\alpha_{ex}(\theta,\vartheta,g',g)\right] q_\theta(\vartheta|\theta)\,d\vartheta \\ &+ \phi(\theta)\left[1 - \int_\Theta \left[\sum_{g'\in\mathcal{G}} \pi(g',\vartheta)\,\alpha_{ex}(\theta,\vartheta,g',g)\right] q_\theta(\vartheta|\theta)\,d\vartheta\right] \end{aligned}$$

while the transition kernel for the approximate algorithm is

$$\begin{aligned} \widetilde{P}_R(\phi(\theta),\phi(\vartheta)) &= \int_\Theta \phi(\vartheta)\left[\sum_{g'\in\mathcal{G}} \mathcal{P}_\vartheta^{(R)}(g'|g)\,\alpha_{ex}(\theta,\vartheta,g',g)\right] q_\theta(\vartheta|\theta)\,d\vartheta \\ &+ \phi(\theta)\left[1 - \int_\Theta \left[\sum_{g'\in\mathcal{G}} \mathcal{P}_\vartheta^{(R)}(g'|g)\,\alpha_{ex}(\theta,\vartheta,g',g)\right] q_\theta(\vartheta|\theta)\,d\vartheta\right] \end{aligned}$$

and therefore the difference is

$$\begin{aligned} S &= P(\phi(\theta),\phi(\vartheta)) - \widetilde{P}_R(\phi(\theta),\phi(\vartheta)) \\ &= \int_\Theta \phi(\vartheta)\left[\sum_{g'\in\mathcal{G}} \left[\pi(g',\vartheta) - \mathcal{P}_\vartheta^{(R)}(g'|g)\right]\alpha_{ex}(\theta,\vartheta,g',g)\right] q_\theta(\vartheta|\theta)\,d\vartheta \\ &- \phi(\theta)\int_\Theta \left[\sum_{g'\in\mathcal{G}} \left[\pi(g',\vartheta) - \mathcal{P}_\vartheta^{(R)}(g'|g)\right]\alpha_{ex}(\theta,\vartheta,g',g)\right] q_\theta(\vartheta|\theta)\,d\vartheta \end{aligned}$$

Consider the quantity

$$\begin{aligned} S_0 &= \int_\Theta \left[\sum_{g'\in\mathcal{G}} \left[\pi(g',\vartheta) - \mathcal{P}_\vartheta^{(R)}(g'|g)\right]\alpha_{ex}(\theta,\vartheta,g',g)\right] q_\theta(\vartheta|\theta)\,d\vartheta \\ &\leq \int_\Theta \left[\sum_{g'\in\mathcal{G}} \left|\pi(g',\vartheta) - \mathcal{P}_\vartheta^{(R)}(g'|g)\right|\alpha_{ex}(\theta,\vartheta,g',g)\right] q_\theta(\vartheta|\theta)\,d\vartheta \end{aligned}$$

[48] See Levin et al. (2008), proposition 4.5, page 49.

and since $\alpha_{ex}(\theta, \vartheta, g', g) \leq 1$ for any $(\theta, \vartheta) \in \Theta \times \Theta$ and $(g', g) \in \mathcal{G} \times \mathcal{G}$, we have

$$\begin{aligned} S_0 &\leq \int_\Theta \left[\sum_{g' \in \mathcal{G}} \left| \pi(g', \vartheta) - \mathcal{P}_\vartheta^{(R)}(g'|g) \right| \right] q_\theta(\vartheta|\theta) \, d\vartheta \\ &= \int_\Theta \left[2 \sup_{g' \in \mathcal{G}} \left| \pi(g', \vartheta) - \mathcal{P}_\vartheta^{(R)}(g'|g) \right| \right] q_\theta(\vartheta|\theta) \, d\vartheta \end{aligned}$$

The convergence of the network simulation algorithm implies that for any $\varepsilon > 0$, there exists an $R_0(\vartheta, \varepsilon) \in \mathbb{N}$ such that for any $R > R_0(\vartheta, \varepsilon)$ and for any $g \in \mathcal{G}$

$$2 \sup_{g' \in \mathcal{G}} \left| \pi(g', \vartheta) - \mathcal{P}_\vartheta^{(R)}(g'|g) \right| \leq \varepsilon$$

Pick $R_0(\varepsilon) = \max_{\vartheta \in \Theta} \{R_0(\vartheta, \varepsilon)\}$. Then for any $\epsilon \in (0, 1]$, there is an $R_0(\epsilon) \in \mathbb{N}$ such that for any $R > R_0(\epsilon)$ and for any $g \in \mathcal{G}$

$$S_0 \leq \int_\Theta \epsilon q_\theta(\vartheta|\theta) \, d\vartheta = \epsilon$$

This implies that

$$|S| \leq |2S_0| = 2\epsilon \tag{47}$$

■

The next theorem is the main result for the convergence. It states that the approximate exchange algorithm converges to the correct posterior distribution, provided that the number of network simulations and parameter samples are big enough.

B.2.2 Proof of THEOREM 6

. **Proof.** The main idea is to decompose the total variation in two components

$$\begin{aligned} \left\| \widetilde{P}_R^{(s)}(\theta_0, \cdot) - p(\cdot|g, X) \right\|_{TV} &= \left\| \widetilde{P}_R^{(s)}(\theta_0, \cdot) - P^{(s)}(\theta_0, \cdot) + P^{(s)}(\theta_0, \cdot) - p(\cdot|g, X) \right\|_{TV} \\ &\leq \left\| \widetilde{P}_R^{(s)}(\theta_0, \cdot) - P^{(s)}(\theta_0, \cdot) \right\|_{TV} + \left\| P^{(s)}(\theta_0, \cdot) - p(\cdot|g, X) \right\|_{TV} \end{aligned}$$

and prove that each component converges. We will use the same idea, but rewrite the total variation in a more convenient form.[49] For any function $\phi : \Theta \to [-1, 1]$ we have

$$\begin{aligned} \left| \widetilde{P}_R^{(s)} \phi(\theta_0) - p(\phi) \right| &= \left| \widetilde{P}_R^{(s)} \phi(\theta_0) - P^{(s)} \phi(\theta_0) + P^{(s)} \phi(\theta_0) - p(\phi) \right| \\ &\leq \left| \widetilde{P}_R^{(s)} \phi(\theta_0) - P^{(s)} \phi(\theta_0) \right| + \left| P^{(s)} \phi(\theta_0) - p(\phi) \right| \end{aligned}$$

[49] See Levin et al. (2008), proposition 4.5, page 49.

The second component converges because the exact exchange algorithm is ergodic, as stated in Lemma. For any $\varepsilon > 0$ there is number of simulation steps $s(\theta_0, \varepsilon)$, such that for any $s \geq s(\theta_0, \varepsilon)$

$$\left| P^{(s)} \phi(\theta_0) - p(\phi) \right| \leq \varepsilon \tag{48}$$

For the remaining of the proof, I will set $s_0 := s(\theta_0, \varepsilon)$. I use the telescoping sum decomposition in Andrieu and Roberts (2009) (page 15, adapted from last formula)

$$\begin{aligned} \left| \widetilde{P}_R^{(s_0)} \phi(\theta_0) - P^{(s_0)} \phi(\theta_0) \right| &= \left| \sum_{l=0}^{s_0-1} \left[P^{(l)} \widetilde{P}_R^{(s_0-l)} \phi(\theta_0) - P^{(l+1)} \widetilde{P}_R^{(s_0-(l+1))} \phi(\theta_0) \right] \right| \\ &= \left| \sum_{l=0}^{s_0-1} P^{(l)} \left(\widetilde{P}_R - P \right) \widetilde{P}_R^{(s_0-(l+1))} \phi(\theta_0) \right| \end{aligned}$$

Now we can apply s_0 times the result of LEMMA 4 (as in Liang et al. (2010) and Andrieu and Roberts (2009)) to prove that there exists an $R_0(\theta_0, \varepsilon) \in \mathbb{N}$ such that for any $R > R_0(\theta_0, \varepsilon)$

$$\left| \widetilde{P}_R^{(s_0)} \phi(\theta_0) - P^{(s_0)} \phi(\theta_0) \right| \leq 2 s_0 \varepsilon \tag{49}$$

this implies

$$\left| \widetilde{P}_R^{(s)} \phi(\theta_0) - p(\phi) \right| \leq (2 s_0 + 1) \varepsilon \tag{50}$$

We conclude the proof by choosing $\varepsilon = \epsilon / (2 s_0 + 1)$.

This proves that the approximate exchange algorithm is ergodic, therefore the law of large number holds, and the second part of the theorem is proven. ■

C Unobserved heterogeneity

It is possible to incorporate unobserved heterogeneity or random coefficients in the model. However this would significantly increase the computational cost of estimation. The simplest way to introduce unobserved heterogeneity is to model the preference shock ε_{ij} as incorporating individual random effects. The decision of the player to form a link is modified as follows

$$U_i(g_{ij} = 1, g_{-ij}, X) + \eta_i + \eta_j + \nu_{ij1} \geq U_i(g_{ij} = 0, g_{-ij}, X) + \eta_i + \nu_{ij0} \tag{51}$$

where ν_{ij} is an i.i.d. shock with logistic distribution and the vector $\eta = \{\eta_1, ..., \eta_n\}$ is drawn at time 0 from a known distribution $W(\eta)$. In this formulation, we assume that the players observe the random effect η but the econometrician does not. Notice that the random effect of player i cancels out, while the choice of linking j is conditional on the random effect of player j (which is present only when the link is formed).

Conditioning on the realization of the vector $\eta \in \Upsilon$, the potential function is modified as follows

$$\mathcal{Q}(g, X, \theta; \eta) = Q(g, X, \theta) + \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} \eta_j \tag{52}$$

To compute the unconditional likelihood we need to integrate out the unobserved vector η to obtain

$$\pi(g, X, \theta) = \int_{\Upsilon} \frac{\exp\left[\mathcal{Q}(g, X, \theta; \eta)\right]}{\sum_{\omega \in \mathcal{G}} \exp\left[\mathcal{Q}(\omega, X, \theta; \eta)\right]} dW(\eta) \tag{53}$$

The integral above can be computed using Monte Carlo techniques, as it is standard in the empirical industrial organization literature or labor economics. However, the model does not allow standard Monte Carlo, because of the normalizing constant.

A more feasible strategy is to use data augmentation and Markov Chain Monte Carlo methods as in the discrete choice literature (Rossi et al. (1996), Athey and Imbens (2007)). Conditioning on the realization of the unobserved component η, we can use the exchange algorithm to sample from the posterior distribution of θ. Conditioning on the proposed θ we can use a Metropolis-Hastings step to sample the unobserved component η.

Given an initial (θ, η) at simulation s, we propose a new θ' and use the exchange algorithm to accept or reject the proposal. Given the new value of θ_{s+1}, we propose a new vector of unobserved components η' and accept using a Metropolis-Hastings step. The probability of η, conditioning on (θ, g, X) is

$$\Pr(\eta | g, X, \theta) = \frac{W(\eta)\, \pi(g, X, \theta; \eta)}{\pi(g, X, \theta)} \tag{54}$$

The Metropolis-Hastings step proceeds by proposing a new η' from a distribution $q_\eta(\eta'|\eta)$, which is accepted with probability

$$\alpha_\eta(\eta, \eta', g, \theta_s) = \left\{1, \frac{W(\eta')\, \pi(g, X, \theta; \eta')\, q_\eta(\eta|\eta')}{W(\eta)\, \pi(g, X, \theta; \eta)\, q_\eta(\eta'|\eta)}\right\} \tag{55}$$

Similar ideas apply to random coefficients. However, as discussed in Graham (2014), when we observe only one network in the data, it is not possible to separately identify the linking externalities and the unobserved heterogeneity.

The main cost of these extensions is the increased computational burden, which may be substantial.

D Large networks analysis and convergence

In this paper, we developed a network formation game model, which results in an equilibrium network similar to a directed ERGM. The probability of observing network g is given by (notice that $g_{ij} = 1$ does not imply $g_{ji} = 1$, because it is a directed network)

$$\pi_n(g) = \frac{\exp\left[\sum_{i=1}^n \sum_{j=1}^n g_{ij} u_{ij} + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n g_{ij} g_{ji} m_{ij} + \sum_{i=1}^n \sum_{j=1}^n \sum_{k \neq i,j}^n g_{ij} g_{jk} v_{ik}\right]}{c(\mathcal{G}_n)}$$

where the functions $u_{ij} = u(X_i, X_j, \theta_u)$, $m_{ij} = m(X_i, X_j, \theta_m)$ and $v_{ik} = v(X_i, X_k, \theta_v)$ are function of vectors of covariates $X_i's$ and parameters $\theta = (\theta_u, \theta_m, \theta_v)$. To simplify, we

will assume that all this functions are constants, so that we do not consider the covariates. Hence, the probability of observing network g with parameters α, β, γ

$$\pi_n(g;\alpha,\beta,\gamma) = \frac{\exp\left[\alpha \sum_{i=1}^n \sum_{j=1}^n g_{ij} + \frac{\beta}{2} \sum_{i=1}^n \sum_{j=1}^n g_{ij} g_{ji} + \gamma^o \sum_{i=1}^n \sum_{j=1}^n \sum_{k\neq i}^n g_{ij} g_{jk}\right]}{c(\alpha,\beta,\gamma,\mathcal{G}_n)}$$

To apply the analysis of Diaconis and Chatterjee (2011), we rescale the terms as

$$\pi_n(g;\alpha,\beta,\gamma) = \frac{\exp\left\{n^2\left[\alpha \frac{\sum_{i=1}^n \sum_{j=1}^n g_{ij}}{n^2} + \frac{\beta}{2} \frac{\sum_{i=1}^n \sum_{j=1}^n g_{ij} g_{ji}}{n^2} + \gamma \frac{\sum_{i=1}^n \sum_{j=1}^n \sum_{k\neq i}^n g_{ij} g_{jk}}{n^3}\right]\right\}}{c(\alpha,\beta,\gamma,\mathcal{G}_n)} \tag{56}$$

Notice that γ needs to be rescaled (i.e. divided by n) when we run the simulations using the usual ERGM form, i.e. $\gamma^o = \frac{\gamma}{n}$ for simulations using the `ergm` package in the software R.

In the formula above, the term $\frac{\sum_{i=1}^n \sum_{j=1}^n g_{ij}}{n^2}$ is the directed edge density of the network, the term $\frac{\sum_{i=1}^n \sum_{j=1}^n g_{ij} g_{ji}}{n^2}$ is the reciprocity density, while $\frac{\sum_{i=1}^n \sum_{j=1}^n \sum_{k\neq i}^n g_{ij} g_{jk}}{n^3}$ is the density of directed two-paths (in our model the latter is intepreted as popularity or indirect links effect).

In this appendix we provide the technical results about the graph limits, large deviations and mean-field approximations of the model. In the exposition for graph limits and large deviations we report some results for undirected networks from Chatterjee and Varadhan (2011) and Diaconis and Chatterjee (2011), for completeness.

D.1 A crash course on graph limits

Most of this brief digression follows the overview in Diaconis and Chatterjee (2011), focusing on directed graphs. For a more detailed introduction to graph limits, see Lovasz (2012), Borgs et al. (2008), and Lovasz and Szegedy (2007). Most of the theory is developed for dense graphs, but there are several results for sparse graphs. The model presented here generates a dense graph, therefore we present only the relevant theory.

Consider a sequence of simple directed graphs G_n, where the number of nodes n tends to infinity. Let $|hom(H,G)|$ denote the number of homomorphisms of simple directed graph H into G. An homomorphism is an arc-preserving map from the set of vertices $V(H)$ of H to the set of vertices $V(G)$ of G.[50] For the graph limits we are interested in the *homomorphism densities* of the form

$$t(H,G) = \frac{|hom(H,G)|}{|V(G)|^{|V(H)|}}$$

Intuitively, $t(H,G)$ is the probability that a random mapping $V(H) \to V(G)$ is a homomorphism. We are interested in the behavior of $t(H,G_n)$ when $n \to \infty$. In particular we

[50] An important difference between homomorphisms for undirected graphs and directed graphs is that in the latter class of models, the existence of homomorphisms is not guaranteed. See Lovasz (2012) for some additional details.

want to characterize the limit object $t(H)$, for any simple graph H. The work of Lovasz, (see Lovasz (2012) for an extensive overview) provides the limit object for this problem. Let $h \in \mathcal{W}$ be a function in the space $\mathcal{W}$ of all measurable functions $h : [0,1]^2 \to [0,1]$. This slightly differs from the original paper of Diaconis and Chatterjee (2011) because we are considering directed graphs, therefore we do not require the function h to be symmetric. For comparison with the original formulation, let $\mathcal{W}_o$ denote the set of all measurable functions $h : [0,1]^2 \to [0,1]$ such that $h(x,y) = h(y,x)$.

The existence of such limit objects and the characterization for directed graphs is contained in Boeckner (2013) and extends the usual formulation for undirected graphs. If H is a simple directed graph with k vertices (i.e. $V(H) = \{1, 2, ..., k\}$) the limit object for $t(H, G_n)$ is

$$t(H,h) = \int_{[0,1]^k} \prod_{(i,j)\in E(H)} h(x_i, x_j) dx_1 \cdots dx_k$$

where $E(H)$ is the set of directed edges of H. For example, if we are interested in homorphisms of a directed edge, the homomorphism density is

$$t(H,G) = \frac{|hom(H,G)|}{|V(G)|^{|V(H)|}} = \frac{\sum_i \sum_j g_{ij}}{n^2}$$

and the limit object is

$$t(H,h) = \int_{[0,1]^k} \prod_{(i,j)\in E(H)} h(x_i, x_j) dx_1 \cdots dx_k = \int_0^1 \int_0^1 h(x,y) dx dy$$

If we are interested in the indirect links as in our model, we have

$$t(H,G) = \frac{|hom(H,G)|}{|V(G)|^{|V(H)|}} = \frac{\sum_i \sum_j \sum_k g_{ij} g_{jk}}{n^3}$$

with limit object

$$t(H,h) = \int_{[0,1]^k} \prod_{(i,j)\in E(H)} h(x_i, x_j) dx_1 \cdots dx_k = \int_0^1 \int_0^1 \int_0^1 h(x,y) h(y,z) dx dy dz$$

A sequence of graphs $\{G_n\}_{n\geq 1}$ converges to h if for every simple directed graph H

$$\lim_{n\to\infty} t(H, G_n) = t(H, h)$$

The intuitive interpretation of this theory is simple: when n becomes large, we rescale the vertices to a continuum interval $[0,1]$; and $h(x,y)$ is the probability that there is a directed edge from x to y. The limit object $h \in \mathcal{W}$ is called *graphon*. For any finite graph G with vertex set $\{1, ..., n\}$ we can always define the graph limit representation f^G as

$$f^G(x,y) = \begin{cases} 1 \text{ if } (\lceil nx \rceil, \lceil ny \rceil) \text{ is a directed edge of } G \\ 0 \text{ otherwise} \end{cases}$$

where the symbol $\lceil a \rceil$ indicates the ceiling of a, i.e. the smallest integer greater than or equal to a.

To study convergence in the space $\mathcal{W}$ of the functions h, we need to define a metric. We use the *cut distance*

$$d_\square(f,g) \equiv \sup_{S,T \subseteq [0,1]} \left| \int_{S \times T} [f(x,y) - g(x,y)] \, dxdy \right|$$

where f and g are functions in $\mathcal{W}$. However, there is some non-trivial complication in the topology induced by the cut metric. To solve this complication, the usual approach is to work with a suitably defined quotient space $\widetilde{\mathcal{W}}$. We introduce an equivalence relation in $\mathcal{W}$: $f \sim g$ if $f(x,y) = g_\sigma(x,y) = g(\sigma x, \sigma y)$ for some measure preserving bijection $\sigma : [0,1] \to [0,1]$. We will use $\widetilde{h}$ to denote the equivalence class of h in $(\mathcal{W}, d_\square)$. Since $d_\square$ is invariant under σ, we can define a distance on the quotient space $\widetilde{\mathcal{W}}$ as

$$\delta_\square\left(\widetilde{f}, \widetilde{g}\right) \equiv \inf_\sigma d_\square(f, g_\sigma) = \inf_\sigma d_\square(f_\sigma, g) = \inf_{\sigma_1, \sigma_2} d_\square\left(f_{\sigma_1}, g_{\sigma_2}\right)$$

This makes $\left(\widetilde{\mathcal{W}}, \delta_\square\right)$ a metric space. with several nice properties: it is compact and the homomorphism densities $t(H,h)$ are continuous functions on it. We associate f^G to any finite graph G and we have $\widetilde{G} = \tau f^G = \widetilde{f^G} \in \widetilde{\mathcal{W}}$, where τ is a mapping, $\tau : f \to \widetilde{f}$. For completeness, we prove the compactness of the metric space, which is crucial for some of the following proofs.

LEMMA 5 *The metric space* $\left(\widetilde{\mathcal{W}}, \delta_\square\right)$ *is compact.*

Proof. The proof follows similar steps as in Theorem 5.1 of Lovasz and Szegedy (2007). For every function $h \in \mathcal{W}$ and a partition $\mathcal{P} = \{P_1, P_2, ..., P_k\}$ of $[0,1]$ into measurable sets, we define $h_\mathcal{P} : [0,1]^2 \to [0,1]$ to be the stepfunction obtained from h by replacing its value at $(x,y) \in P_i \times P_j$ by the average of h over $P_i \times P_j$.

Let h_1, h_2, ... be a sequence of functions in $\mathcal{W}$. We need to construct a subsequence that has limit in $\widetilde{\mathcal{W}}$. According to Lemmas 3.1.20 and 3.1.21 in Boeckner (2013), we can create a partition $\mathcal{P}_{n,k} = \{P_{1,n,k}, ..., P_{m_k,n,k}\}$ of $[0,1]$ for every n and k. This partition corresponds to a step-function $h_{n,k} = h_{\mathcal{P}_{n,k}} \in \mathcal{W}$, such that:

1. $\delta_\square(h_n, h_{n,k}) \leq 1/k$
2. $|\mathcal{P}_{n,k}| = m_k$ (where m_k only depends on k)
3. the partition $\mathcal{P}_{n,k+1}$ refines the partition $\mathcal{P}_{n,k}$ for every k

Notice that since $\delta_\square(h_n, h_{n,k}) \leq 1/k$, we can re-arrange the range of $h_{n,k}$ so that all the steps of the function are intervals. Select a subsequence of h_n such that the length of the i-th interval $P_{i,n,1}$ of $h_{n,1}$ converges for every i as $n \to \infty$; and the value $h_{n,1}$ on $P_{i,n,1} \times P_{j,n,1}$ also converges for every i and j as $n \to \infty$. Hence, the sequence $h_{n,1}$ converges to a limit almost everywhere. Let's call the limit U_1: notice that U_1 is also a step-function with m_1 steps (that are themselves intervals). We can repeat this procedure for $k = 2, 3, ...$ We obtain subsequences for which $h_{n,k} \to U_k$ almost everywhere, and U_k is a step-function with m_k steps.

We know that for every $k < l$, the partition $\mathcal{P}_{n,l}$ is a refinement of partition $\mathcal{P}_{n,k}$. As a consequence, the partition into the steps of $h_{n,l}$ is a refinement of the partition into the steps of $h_{n,k}$. Clearly, the same relation must hold for U_l and U_k, i.e. the partition into the steps of U_l is a refinement of the partition into the steps of U_k. By construction of $h_\mathcal{P}$, the function $h_{n,k}$ can be obtained from $h_{n,l}$ by averaging its value over each step. As a consequence, the same holds for U_l and U_k.

It is shown in the proof of Lemma 3.1.21 in Boeckner (2013) that if we pick a random point (X, Y) uniformly over $[0, 1]^2$ the sequence $U_1(X, Y)$, $U_2(X, Y), ...$ is martingale, and each element of the sequence is bounded. Using the Martingale Convergence Theorem we can show that the sequence $U_1(X, Y)$, $U_2(X, Y), ...$ converges almost everywhere. We define this limit U.

The rest of the proof is the same as in Theorem 5.1 of Lovasz and Szegedy (2007). Fix an $\varepsilon > 0$. Then there exists a $k > 3/\varepsilon$, which we denote as K, such that $\|U - U_k\|_1 < \varepsilon/3$. Fix $k = K$: then there is an N, such that for all $n \geq N$ we have $\|U_k - h_{n,k}\|_1 < \varepsilon/3$. Then we finally have

$$\begin{aligned} \delta_\square(U, h_n) &\leq \delta_\square(U, U_k) + \delta_\square(U_k, h_{n,k}) + \delta_\square(h_{n,k}, h_n) \\ &\leq \|U - U_k\|_1 + \|U_k - h_{n,k}\|_1 + \delta_\square(h_{n,k}, h_n) \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \end{aligned}$$

As a consequence $h_n \to U$ in the metric space $\left(\widetilde{\mathcal{W}}, \delta_\square\right)$. ■

D.2 A crash course on large deviations for random graphs

D.2.1 Undirected graphs (Original Chatterjee and Varadhan (2011) formulation)

Chatterjee and Varadhan (2011) developed a large deviation principle for the undirected Erdos-Renyi graph. Let $G(n, p)$ indicate the the random *undirected* graph with n vertices where each link is formed independently with probability p. Define a function $I_p : [0, 1] \to \mathbb{R}$

$$I_p(u) \equiv \frac{1}{2} u \log \frac{u}{p} + \frac{1}{2}(1 - u) \log \frac{1 - u}{1 - p} \tag{57}$$

whose domain is easily extended to $\mathcal{W}_o$ as

$$\begin{aligned} I_p(h) &= \int_0^1 \int_0^1 I_p\left(h(x,y)\right) dxdy \\ &= \frac{1}{2} \int_0^1 \int_0^1 \left[h(x,y) \log \frac{h(x,y)}{p} + (1-h(x,y)) \log \frac{1-h(x,y)}{p} \right] dxdy \end{aligned} \tag{58}$$

Analogously we can define I_p on $\widetilde{\mathcal{W}_o}$ as $I_p(\widetilde{h}) \equiv I_p(h)$. The graph $G(n,p)$ induces a probability distribution $P_{n,p}$ on $\mathcal{W}_o$, because we can use the map $G \to f^G$; and it induces a probability distribution $\widetilde{P}_{n,p}$ on $\widetilde{\mathcal{W}_o}$ according to the map $G \to f^G \to \widetilde{f}^G = \widetilde{G}$. Chatterjee and Varadhan (2011) state a large deviation principle for the Erdos Renyi random graph in both spaces $(\mathcal{W}_o, d_\square)$ and $(\widetilde{\mathcal{W}_o}, \delta_\square)$.

We report the main result of Chatterjee and Varadhan (2011) for completeness.

THEOREM 7 *(Large deviation principle for Erdos-Renyi graph, Chatterjee and Varadhan (2011)). For each fixed $p \in (0,1)$, the sequence $\widetilde{P}_{n,p}$ obeys a large deviation principle in the space $(\widetilde{\mathcal{W}_o}, \delta_\square)$ with rate function $I_p(h)$ defined in (58). For any closed set $\widetilde{F} \subseteq \widetilde{\mathcal{W}}$*

$$\limsup_{n\to\infty} \frac{1}{n^2} \log \widetilde{P}_{n,p}(\widetilde{F}) \leq - \inf_{\widetilde{h} \in \widetilde{F}} I_p(\widetilde{h})$$

and for any open set $\widetilde{U} \subseteq \widetilde{\mathcal{W}}$,

$$\liminf_{n\to\infty} \frac{1}{n^2} \log \widetilde{P}_{n,p}(\widetilde{U}) \geq - \inf_{\widetilde{h} \in \widetilde{U}} I_p(\widetilde{h})$$

D.2.2 Directed graphs

First, we consider the extension of Theorem 7 to *directed* Erdos-Renyi graphs. Let $G_d(n,p)$ indicate the random *directed* graph with n vertices where each arc is formed independently with probability p. Define a function $\mathcal{I}_p : [0,1] \to \mathbb{R}$

$$\mathcal{I}_p(u) \equiv u \log \frac{u}{p} + (1-u) \log \frac{1-u}{1-p} \tag{59}$$

whose domain is easily extended to $\mathcal{W}$ as

$$\begin{aligned} \mathcal{I}_p(h) &= \int_0^1 \int_0^1 I_p\left(h(x,y)\right) dxdy \\ &= \int_0^1 \int_0^1 \left[h(x,y) \log \frac{h(x,y)}{p} + (1-h(x,y)) \log \frac{1-h(x,y)}{p} \right] dxdy \end{aligned} \tag{60}$$

Analogously we can define $\mathcal{I}_p$ on $\widetilde{\mathcal{W}}$ as $\mathcal{I}_p(\widetilde{h}) \equiv \mathcal{I}_p(h)$. Chatterjee and Varadhan (2011) (see their Lemma 2.1) prove that this function is lower semicontinuous on $\widetilde{\mathcal{W}}$ under the metric $\delta_\square$.

The graph $G_d(n,p)$ induces a probability distribution $\mathcal{P}_{n,p}$ on $\mathcal{W}$, because we can use the map $G \to f^G$; and it induces a probability distribution $\widetilde{\mathcal{P}}_{n,p}$ on $\widetilde{\mathcal{W}}$ according to the map $G \to f^G \to \widetilde{f}^G = \widetilde{G}$. The large deviation principle for this case is presented in the following theorem.

THEOREM 8 *(Large deviation principle for directed Erdos-Renyi graph) For each fixed $p \in (0,1)$, the sequence $\widetilde{\mathcal{P}}_{n,p}$ obeys a large deviation principle in the space $(\widetilde{\mathcal{W}}, \delta_\square)$ with rate function $\mathcal{I}_p(h)$ defined in (60). For any closed set $\widetilde{F} \subseteq \widetilde{\mathcal{W}}$*

$$\limsup_{n\to\infty} \frac{1}{n^2} \log \widetilde{\mathcal{P}}_{n,p}(\widetilde{F}) \leq - \inf_{\widetilde{h}\in\widetilde{F}} \mathcal{I}_p(\widetilde{h})$$

and for any open set $\widetilde{U} \subseteq \widetilde{\mathcal{W}}$,

$$\liminf_{n\to\infty} \frac{1}{n^2} \log \widetilde{\mathcal{P}}_{n,p}(\widetilde{U}) \geq - \inf_{\widetilde{h}\in\widetilde{U}} \mathcal{I}_p(\widetilde{h})$$

Proof. The proof follows the same steps as in the original theorem for undirected graphs in Chatterjee and Varadhan (2011), but substituting the new rate function in (60). For the upper bound, we define $p_{i,j}$ as in the original paper, but we do not require symmetry. We use slightly different regularity conditions, as provided in Boeckner (2013), because of the directed nature of the graph. In particular we use Lemmas 3.1.14, 3.1.20 and 3.1.21 in Boeckner (2013). With these small changes, Lemma 2.4, 2.5 and 2.6 in Chatterjee and Varadhan (2011) hold. The proof follows the same steps as in the undirected case. For the lower bound, the proof is identical, without the requirement of simmetry. ■

D.3 Undirected ERGM (Chatterjee and Diaconis 2013)

Let $T : \widetilde{\mathcal{W}}_o \to \mathbb{R}$ be a bounded continuous function in space $(\widetilde{\mathcal{W}}_o, \delta_\square)$. For a given n the probability function for the graphs is given by

$$\pi_n(G) = \exp\left\{n^2 \left[T(\widetilde{G}) - \psi_n\right]\right\}$$

where $\widetilde{G}$ is defined on $\widetilde{\mathcal{W}}_o$ according to the map $G \to f^G \to \widetilde{f}^G = \widetilde{G}$, and ψ_n is a constant defined as

$$\psi_n = \frac{1}{n^2} \log \sum_{G\in\mathcal{G}_n} \exp\left\{n^2 \left[T(\widetilde{G})\right]\right\} \tag{61}$$

The rescaling by n^2 is necessary to guarantee that the limits for $n \to \infty$ converge to some non-trivial quantity. We are interested in finding the value of ψ_n as $n \to \infty$. We define a rate function

$$I(u) \equiv \frac{1}{2} u \log u + \frac{1}{2}(1-u)\log(1-u) \tag{62}$$

which we extend to $\widetilde{\mathcal{W}_o}$ as

$$I(\widetilde{h}) \equiv \frac{1}{2}\int_0^1 \int_0^1 I(h(x,y))dxdy$$

$$\begin{aligned} I(\widetilde{h}) &\equiv \frac{1}{2}\int_0^1 \int_0^1 I(h(x,y))dxdy \\ &= \frac{1}{2}\int_0^1 \int_0^1 \left[h(x,y)\log h(x,y) + (1-h(x,y))\log(1-h(x,y))\right] dxdy \end{aligned} \tag{63}$$

THEOREM 9 *(Theorem 3.1 for ERGM in Chatterjee-Diaconis 2013). If $T : \widetilde{\mathcal{W}_o} \to \mathbb{R}$ is a bounded continuous function and ψ_n and I are defined as in (61) and (63) respectively, then*

$$\psi \equiv \lim_{n\to\infty} \psi_n = \sup_{\widetilde{h}\in\widetilde{\mathcal{W}_o}} \left\{ T(\widetilde{h}) - I(\widetilde{h}) \right\}$$

D.4 Directed ERGM

Let $\mathcal{T} : \widetilde{\mathcal{W}} \to \mathbb{R}$ be a bounded continuous function in space $(\widetilde{\mathcal{W}}, \delta_\square)$. In our model $\mathcal{T}$ corresponds to the potential function Q of the network formation game after rescaling some of the utility components (see below for details and examples). In what follows, we omit the dependence on the parameters to simplify notation. For a given n, the probability of observing network G is given by

$$\pi_n(G) = \exp\left\{ n^2 \left[\mathcal{T}(\widetilde{G}) - \psi_n \right] \right\}$$

where $\widetilde{G}$ is defined on $\widetilde{\mathcal{W}}$ according to the map $G \to f^G \to \widetilde{f}^G = \widetilde{G}$, and ψ_n is a normalization constant defined as

$$\psi_n = \frac{1}{n^2} \log \sum_{G\in\mathcal{G}_n} \exp\left\{ n^2 \left[\mathcal{T}(\widetilde{G}) \right] \right\} \tag{64}$$

This is the same as the stationary distribution of our model, after some re-scaling of the utility functions. The rescaling by n^2 is necessary to guarantee that the limits for $n \to \infty$

converge to some non-trivial quantity. We are interested in finding the value of ψ_n as $n \to \infty$, using the same line of reasoning in Theorem 3.1 of Diaconis and Chatterjee (2011). We define a rate function

$$\mathcal{I}(u) \equiv u \log u + (1-u)\log(1-u) \tag{65}$$

which we extend to $\widetilde{\mathcal{W}}$ as

$$\begin{aligned} \mathcal{I}(\widetilde{h}) &\equiv \int_0^1 \int_0^1 I(h(x,y))dxdy \\ &= \int_0^1 \int_0^1 \left[h(x,y)\log h(x,y) + (1-h(x,y))\log(1-h(x,y))\right] dxdy \end{aligned} \tag{66}$$

THEOREM 10 *(Asymptotic log-constant for Directed ERGM). If $\mathcal{T} : \widetilde{\mathcal{W}} \to \mathbb{R}$ is a bounded continuous function and ψ_n and $\mathcal{I}$ are defined as in (64) and (66) respectively, then*

$$\psi \equiv \lim_{n\to\infty} \psi_n = \sup_{\widetilde{h}\in\widetilde{\mathcal{W}}} \left\{ \mathcal{T}(\widetilde{h}) - \mathcal{I}(\widetilde{h}) \right\} \tag{67}$$

Proof. The proof of this result follows closely the proof of Theorem 3.1 in Diaconis and Chatterjee (2011), with minimal changes. Let $\widetilde{A}$ denote a Borel set $\widetilde{A} \subseteq \widetilde{\mathcal{W}}$. For each n let $\widetilde{A}_n$ be the (finite) set

$$\widetilde{A}_n \equiv \left\{ \widetilde{h} \in \widetilde{A} \text{ such that } \widetilde{h} = \widetilde{G} \text{ for some } G \in \mathcal{G}_n \right\}$$

Let $\mathcal{P}_{n,p}$ be the probability distribution of the directed random graph $G_d(n,p)$ defined above. We have

$$|\widetilde{A}_n| = 2^{n(n-1)}\mathcal{P}_{n,1/2}(\widetilde{A}_n) = 2^{n(n-1)}\mathcal{P}_{n,1/2}(\widetilde{A})$$

We can use the result in Theorem 8 to show that for a closed subset $\widetilde{F}$ of $\widetilde{\mathcal{W}}$ we have

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n^2} \log \widetilde{\mathcal{P}}_{n,1/2}(\widetilde{F_n}) &= \limsup_{n\to\infty} \frac{1}{n^2} \left[\log |\widetilde{F}_n| - n(n-1)\log 2 \right] \\ &= \limsup_{n\to\infty} \frac{1}{n^2} \log |\widetilde{F}_n| - \log 2 \\ &\leq -\inf_{\widetilde{h}\in\widetilde{F}} \mathcal{I}_{1/2}(\widetilde{h}) \end{aligned}$$

Therefore we obtain

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n^2} \log |\widetilde{F}_n| &\leq \log 2 - \inf_{\widetilde{h}\in\widetilde{F}} \mathcal{I}_{1/2}(\widetilde{h}) \\ &= \inf_{\widetilde{h}\in\widetilde{F}} \mathcal{I}(\widetilde{h}) \end{aligned}$$

Similarly for an open subset $\widetilde{U}$ of $\widetilde{\mathcal{W}}$ we have

$$\begin{aligned}\lim \inf_{n\to\infty} \frac{1}{n^2} \log |\widetilde{U}_n| &\geq \log 2 - \inf_{\widetilde{h}\in\widetilde{U}} \mathcal{I}_{1/2}(\widetilde{h}) \\ &= \inf_{\widetilde{h}\in\widetilde{U}} \mathcal{I}(\widetilde{h})\end{aligned}$$

The rest of the proof is equivalent to the undirected case (see proof of Theorem 3.1 in Diaconis and Chatterjee (2011). ■

The result of Theorem 10 shows that as n grows large we can compute the normalizing constant of the ERGM as the result of a variational problem. The main issue is that the variational problem does not have a closed-form solution for most cases. However, there are some special cases in which the solution can be computed explicitly. Let's consider a model with utility from directed links and friends of friends. Using the notation developed above, we are considering a model with function $\mathcal{T}$

$$\mathcal{T}(\widetilde{G}) = \theta_1 \frac{\sum_{i=1}^n \sum_{j=1}^n g_{ij}}{n^2} + \theta_2 \frac{\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n g_{ij} g_{jk}}{n^3} \tag{68}$$

For any $h \in \mathcal{W}$ we can define

$$\mathcal{T}(h) = \theta_1 t(H_1, h) + \theta_2 t(H_2, h)$$

where the limit objects are

$$t(H_1, h) = \int \int_{[0,1]^2} h(x, y) dx dy$$

and

$$t(H_2, h) = \int \int \int_{[0,1]^3} h(x, y) h(y, z) dx dy dz$$

We will assume that $\theta_2 > 0$. In this case there is an explicit solution of the variational problem. The following theorem provides a characterization of the variational problem along the same lines of Radin and Yin (2013) and Aristoff and Zhu (2014).

THEOREM 11 *Let $\theta_2 > 0$ and $\mathcal{T}$ be defined as in (68) above. Then*

$$\lim_{n\to\infty} \psi_n = \psi = \sup_{\mu\in[0,1]} \left\{\theta_1 \mu + \theta_2 \mu^2 - \mu \log \mu - (1-\mu) \log(1-\mu)\right\}$$

1. *If $\theta_2 \leq 2$, the maximization problem has a unique maximizer $\mu^* \in [0, 1]$*
2. *If $\theta_2 > 2$ and $\theta_1 \geq -2$ then there is a unique maximizer $\mu^* > 0.5$*

3. *If $\theta_2 > 2$ and $\theta_1 < -2$, then there is a V-shaped region of the parameters such that*

   (a) *inside the V-shaped region, the maximization problem has two local maximizers $\mu_1^* < 0.5 < \mu_2^*$*

   (b) *outside the V-shaped region, the maximization problem has a unique maximizer μ^**

4. *For any θ_1 inside the V-shaped region, there exists a $\theta_2 = q(\theta_1)$, such that the two maximizers are both global, i.e. $\ell(\mu_1^*) = \ell(\mu_2^*)$.*

Proof. We need to use the Holder inequality: if p, q are such that $1/p + 1/q = 1$, then for any measurable functions f, g defined on the same domain

$$\int f(x)g(x)dx \leq \left(\int f(x)^p dx\right)^{\frac{1}{p}} \left(\int g(x)^q dx\right)^{\frac{1}{q}}$$

In particular we have in our case

$$\begin{aligned}
t(H_2, h) &= \int\int\int_{[0,1]^3} h(x,y)h(y,z)dxdydz \\
&\leq \left(\int\int\int_{[0,1]^3} h(x,y)^2 dxdydz\right)^{\frac{1}{2}} \left(\int\int\int_{[0,1]^3} h(y,z)^2 dxdydz\right)^{\frac{1}{2}} \\
&= \left(\int\int_{[0,1]^2} h(x,y)^2 \left[\int_{[0,1]} dz\right] dxdy\right)^{\frac{1}{2}} \left(\int\int_{[0,1]^2} h(y,z)^2 \left[\int_{[0,1]} dx\right] dydz\right)^{\frac{1}{2}} \\
&= \left(\int\int_{[0,1]^2} h(x,y)^2 dxdy\right)^{\frac{1}{2}} \left(\int\int_{[0,1]^2} h(y,z)^2 dydz\right)^{\frac{1}{2}} \\
&= \left(\int\int_{[0,1]^2} h(x,y)^2 dxdy\right)^{\frac{1}{2}} \left(\int\int_{[0,1]^2} h(x,y)^2 dxdy\right)^{\frac{1}{2}} \\
&= \int\int_{[0,1]^2} h(x,y)^2 dxdy
\end{aligned}$$

We have assumed that $\theta_2 > 0$. Given the results of the Holder's inequality we can say that

$$\begin{aligned}
\mathcal{T}(h) &= \theta_1 t(H_1, h) + \theta_2 t(H_1, h) \\
&= \theta_1 \int\int_{[0,1]^2} h(x,y)dxdy + \theta_2 \int\int\int_{[0,1]^3} h(x,y)h(y,z)dxdydz \\
&\leq \theta_1 \int\int_{[0,1]^2} h(x,y)dxdy + \theta_2 \int\int_{[0,1]^2} h(x,y)^2 dxdy
\end{aligned}$$

Suppose $h(x,y) = \mu$ is a constant. Then the equality holds and if $\mu \in [0,1]$ solves the variational problem

$$\lim_{n\to\infty} \psi_n(\theta) = \psi(\theta) = \sup_{\mu\in[0,1]} \theta_1\mu + \theta_2\mu^2 - \mu\log\mu - (1-\mu)\log(1-\mu)$$

then $h(x,y) = \mu$ is the limit graphon.

To show that this is the only solution, let's consider the maximization problem again. For $h(x,y)$ to be a solution, we need

$$\mathcal{T}(h) \;=\; \theta_1 \int\int_{[0,1]^2} h(x,y)dxdy + \theta_2 \int\int_{[0,1]^2} h(x,y)^2 dxdy$$

In other words, the Holder inequality must hold with equality, i.e. we need

$$\begin{aligned} t(H_2,h) &= \int\int\int_{[0,1]^3} h(x,y)h(y,z)dxdydz \\ &= \int\int_{[0,1]^2} h(x,y)^2 dxdy \end{aligned}$$

This implies that

$$h(x,y) = h(y,z)$$

for almost all (x,y,z). In particular, we have that given x and y, $\mu = h(x,y) = h(y,z)$ for any $z \in [0,1]$ because the left-hand-side does not depend on z. Given y and z, we have $\mu' = h(y,z) = h(x,y)$ for any $x \in [0,1]$ because the left-hand-side does not depend on x. For $x = y$ and $z = y$ we have $\mu = h(y,y) = h(y,y) = \mu'$. In addition, we have $h(x,y) = h(y,x) = \mu = h(x,z)$. It follows that $h(x,y) = \mu$ almost everywhere.

It follows that $\mathcal{T}(h) = \theta_1\mu + \theta_2\mu^2$ and $I(\mu) = \mu\log\mu + (1-\mu)\log(1-\mu)$, so we get

$$\lim_{n\to\infty} \psi_n = \psi = \sup_{\mu\in[0,1]} \left\{\theta_1\mu + \theta_2\mu^2 - \mu\log\mu - (1-\mu)\log(1-\mu)\right\}$$

We can now characterize the maximization problem above, to obtain the rest of the results. The analysis follows the same steps of Radin and Yin (2013), Aristoff and Zhu (2014). The first order and second order conditions are

$$\ell'(\mu,\theta_1,\theta_2) = \theta_1 + 2\theta_2\mu - \log\frac{\mu}{1-\mu} \tag{69}$$

$$\ell''(\mu,\theta_1,\theta_2) = 2\theta_2 - \frac{1}{\mu(1-\mu)} \tag{70}$$

Let's study the concavity of $\ell(\mu;\theta_1,\theta_2)$. We have that $\ell''(\mu,\theta_1,\theta_2) \le 0$ when

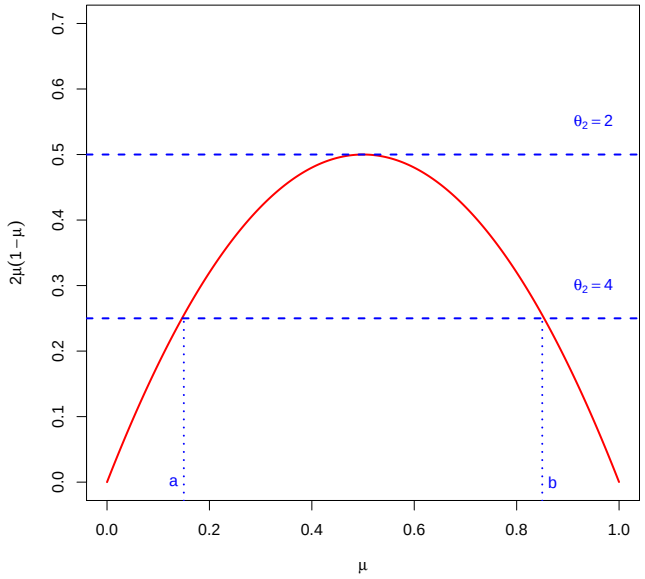
$$\theta_2 \le \frac{1}{2\mu(1-\mu)}$$

Notice that $2 \leq \frac{1}{2\mu(1-\mu)} \leq \infty$ for any $\mu \in [0,1]$; and $\frac{1}{2\mu(1-\mu)} = 2$ if $\mu = 0.5$. As a consequence, the function $\ell(\mu; \theta_1, \theta_2)$ is concave on the whole interval $[0,1]$ for $\theta_2 \leq 2$.

When $\theta_2 > 2$, the second derivative can be positive or negative, with inflection points denoted as a and b: notice that $a < 0.5 < b$.[51]

Consider the first derivative $\ell'(\mu, \theta_1, \theta_2)$. For $\theta_2 \leq 2$, the derivative is decreasing for any μ, because $\ell''(\mu, \theta_1, \theta_2) \leq 0$ for any $\mu \in [0,1]$.

For $\theta_2 > 2$ then (see picture of parabola), it is decreasing in $[0, a)$, increasing in (a, b) and decreasing in $(b, 1]$.



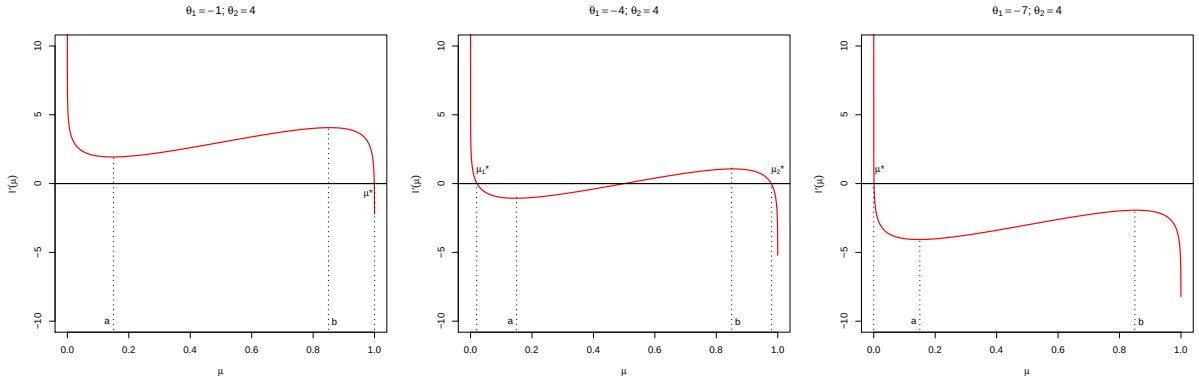
The function $\ell(\mu, \theta_1, \theta_2)$ is bounded and continuous for any θ and $\mu \in [0,1]$, and we could find the interior maximizers by studying the first and second derivative. If we consider the case $\theta_2 \leq 2$, the derivative $\ell'(\mu, \theta_1, \theta_2)$ is decreasing on the whole interval $[0,1]$. It is easy to show that $\ell'(0) = \infty$ and $\ell'(1) = -\infty$. Therefore, when $\theta_2 \leq 2$, there is only one maximizer μ^* that solves $\ell'(\mu, \theta_1, \theta_2) = 0$.

If $\theta_2 > 2$, then we have 3 possible cases. We know that in this region $\ell'(\mu, \theta_1, \theta_2)$ is decreasing in $[0, a)$, increasing in (a, b) and decreasing in $(b, 1]$.

1. If $\ell'(a, \theta_1, \theta_2) \geq 0$, then there is a unique maximizer $\mu^* > b$
2. If $\ell'(b, \theta_1, \theta_2) \leq 0$, then there is a unique maximizer $\mu^* < a$
3. If $\ell'(a, \theta_1, \theta_2) < 0 < \ell'(b, \theta_1, \theta_2)$, then there are 2 local maximizers $\mu_1^* < a < b < \mu_2^*$

[51] This is because, when $\theta_2 > 2$, we have $\ell''(\mu, \theta_1, \theta_2) \leq 0$ when $\theta_2 \leq \frac{1}{2\mu(1-\mu)}$ or $2\mu(1-\mu) \leq \frac{1}{\theta_2}$. The equality is realized at two intersections of the horizontal line $1/\theta_2$ with the parabola $2\mu(1-\mu)$. We call the intersections $\frac{1}{\theta_2} = 2\mu(1-\mu)$, respectively a and b.

The three cases are shown in the following pictures, where we plot $\ell'(\mu, \theta_1, \theta_2)$ against μ for several values of θ_1 and for a fixed $\theta_2 = 4 > 2$.



We indicate the maximizer with μ^* when it is unique, and with μ_1^*, μ_2^* when there are two.

Let's consider the first case, with $\ell'(a, \theta_1, \theta_2) \geq 0$. To compute $\ell'(a, \theta_1, \theta_2)$, notice that $\theta_2 = \frac{1}{2a(1-a)}$. Substituting in $\ell'(a, \theta_1, \theta_2)$ we obtain

$$\ell'(a, \theta_1, \theta_2) = \theta_1 + \frac{1}{1-a} - \log \frac{a}{1-a}$$

and analogously for $\theta_2 = \frac{1}{2b(1-b)}$ we have

$$\ell'(b, \theta_1, \theta_2) = \theta_1 + \frac{1}{1-b} - \log \frac{b}{1-b}$$

So $\ell'(a, \theta_1, \theta_2) \geq 0$ implies

$$\theta_1 \geq \log \frac{a}{1-a} - \frac{1}{1-a}$$

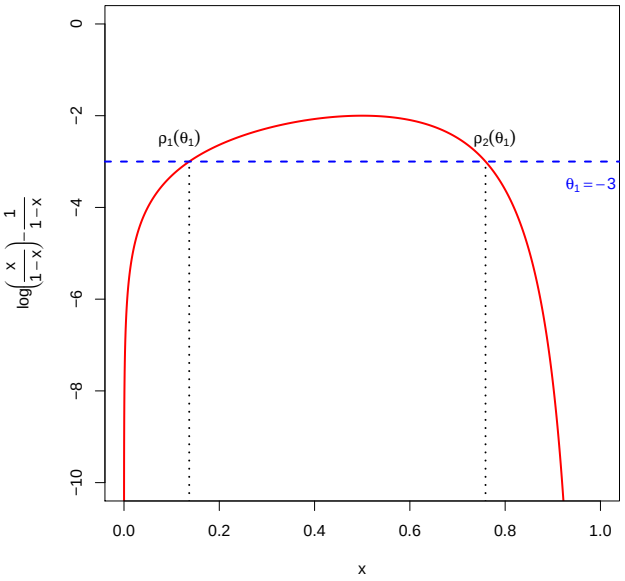
The function $\log \frac{a}{1-a} - \frac{1}{1-a}$ has a maximum at -2 and therefore we have [52]

$$\ell'(a, \theta_1, \theta_2) \geq 0 \Leftrightarrow \theta_1 \geq -2$$

When the above condition is satisfied, there is a unique maximizer, $\mu^* > b$, as shown in the picture on the left.

When $\theta_1 < -2$ it is easier to draw a picture of the function $\log \frac{a}{1-a} - \frac{1}{1-a}$, shown below.

[52] Taking derivative $\frac{1}{a} + \frac{1}{1-a} - \frac{1}{(1-a)^2} = 0$, we obtain the maximizer $a^* = 0.5$. The function is increasing in $[0, 0.5)$ and decreasing in $(0.5, 1]$. The maximum is therefore at -2.



Notice that when $\theta_1 < -2$ there are two intersections of the function and the horizontal line $y = \theta_1$ (in the picture $\theta_1 = -3$). We denote the intersections $\phi_1(\theta_1)$ and $\phi_2(\theta_1)$. By construction, we know that $a < 0.5 < b$. By looking at the picture, it is clear that $\ell'(a, \theta_1, \theta_2) > 0$ if $a < \phi_1(\theta_1)$ and $\ell'(a, \theta_1, \theta_2) < 0$ if $a > \phi_1(\theta_1)$. Analogously, we have $\ell'(b, \theta_1, \theta_2) > 0$ if $b > \phi_2(\theta_1)$ and $\ell'(b, \theta_1, \theta_2) < 0$ if $b < \phi_2(\theta_1)$.

For any $\theta_1 < -2$, there exist $\phi_1(\theta_1)$ and $\phi_2(\theta_1)$ which are the intersection of the function $y = \log\left(\frac{x}{1-x}\right) - \frac{1}{1-x}$ with the line $y = \theta_1$. Since the function is continuous, monotonic increasing in $[0, 0.5)$ and monotonic decreasing in $(0.5, 1]$ it follows that $\phi_1(\theta_1)$ and $\phi_2(\theta_1)$ are both continuous in θ_1. In addition, $\phi_1(\theta_1)$ is increasing in θ_1 and $\phi_2(\theta_1)$ is decreasing in θ_1. It's trivial to show that when θ_1 decreases, $\phi_1(\theta_1)$ converges to 0 while $\phi_2(\theta_1)$ converges to 1.

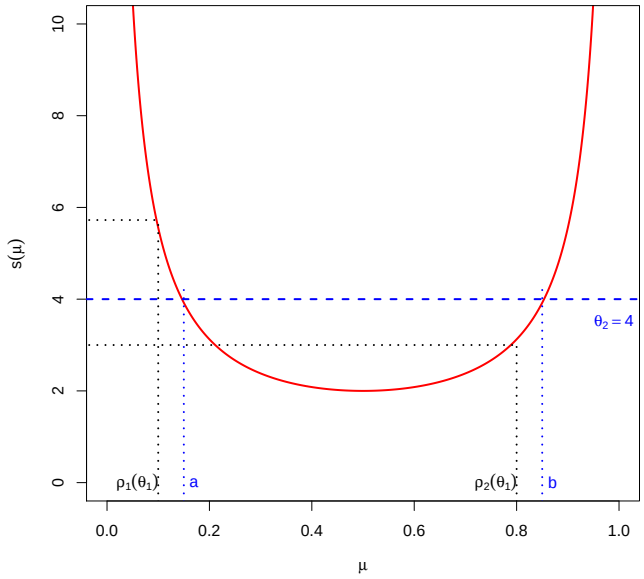
Consider the case in which $\ell'(a, \theta_1, \theta_2) < 0 < \ell'(b, \theta_1, \theta_2)$ with two maximizers. Define the function

$$s(\mu) \equiv \frac{1}{2\mu(1-\mu)}$$

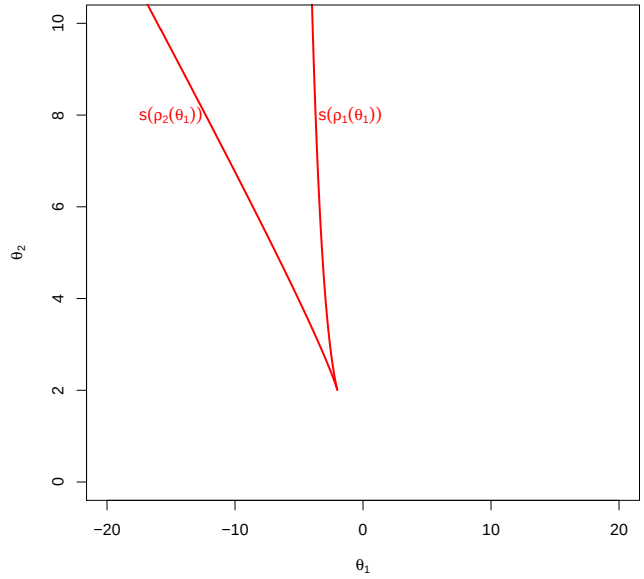
Since $\ell'(a, \theta_1, \theta_2) < 0$ we have $a > \phi_1(\theta_1)$, which implies $s(a) < s(\phi_1(\theta_1))$. Therefore $\theta_2 < s(\phi_1(\theta_1)) = \frac{1}{2\phi_1(\theta_1)(1-\phi_1(\theta_1))}$.

Since $\ell'(b, \theta_1, \theta_2) > 0$ we have $b > \phi_2(\theta_1)$, which implies $s(b) > s(\phi_2(\theta_1))$. Therefore $\theta_2 > s(\phi_2(\theta_1)) = \frac{1}{2\phi_2(\theta_1)(1-\phi_2(\theta_1))}$.

Notice that $s(\phi_1(\theta_1)) > s(\phi_2(\theta_1))$ for any (θ_1, θ_2) in this region of the parameters (see picture below).

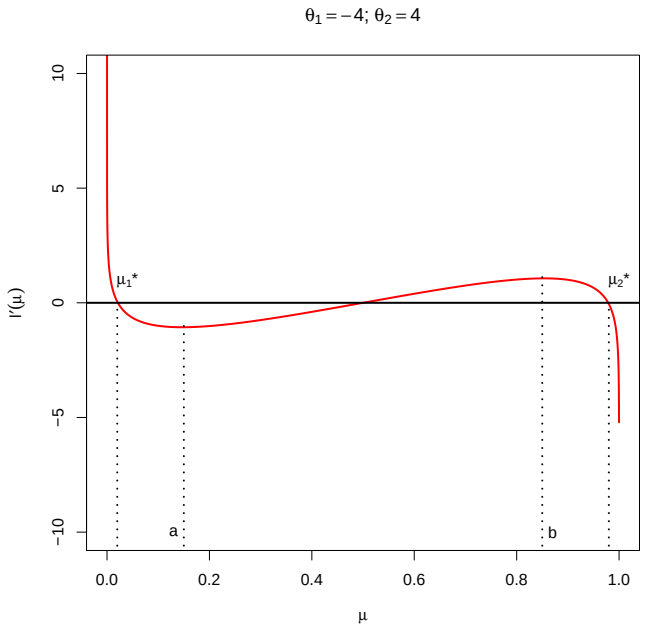


The areas are shown in the following picture



Within the V-shaped region there are 2 solutions to the maximization problem, i.e. two local maxima. Also, it is trivial to show that there exists a function q, such that for $\theta_2 = q(\theta_1)$ both solutions are global maxima. Indeed, the two local maxima are both global maxima if $\ell(\mu_2^*, \theta_1, \theta_2) - \ell(\mu_1^*, \theta_1, \theta_2) = 0$. The latter difference is negative when μ_1^* is the global maximizer, while it is positive when μ_2^* is the global maximizer. Therefeore for a given value of θ_1 there must be a unique θ_2 such that $s(\phi_1(\theta_1)) > \theta_2 > s(\phi_2(\theta_1))$ such that both μ_1^* and

μ_2^* are global maximizer. Let's indicate this value of $\theta_2 = q(\theta_1)$.



Notice that the difference $\ell(\mu_2^*, \theta_1, \theta_2) - \ell(\mu_1^*, \theta_1, \theta_2)$, corresponds to the difference between the positive and negative areas between $\mu_1^*$and μ_2^* in the graph above, i.e. (let $\widehat{\mu}$ indicate the intersection of $\ell'(\mu, \theta_1, \theta_2)$ and the x-axis between $\mu_1^*$and μ_2^*)

$$\begin{aligned}
\ell(\mu_2^*, \theta_1, \theta_2) - \ell(\mu_1^*, \theta_1, \theta_2) &= \int_0^{\mu_2^*} \ell'(\mu, \theta_1, \theta_2) d\mu - \int_0^{\mu_1^*} \ell'(\mu, \theta_1, \theta_2) d\mu \\
&= \int_0^{\mu_1^*} \ell'(\mu, \theta_1, \theta_2) d\mu + \int_{\mu_1^*}^{\widehat{\mu}} \ell'(\mu, \theta_1, \theta_2) d\mu \\
&+ \int_{\widehat{\mu}}^{\mu_2^*} \ell'(\mu, \theta_1, \theta_2) d\mu - \int_0^{\mu_1^*} \ell'(\mu, \theta_1, \theta_2) d\mu \\
&= \int_{\mu_1^*}^{\widehat{\mu}} \ell'(\mu, \theta_1, \theta_2) d\mu + \int_{\widehat{\mu}}^{\mu_2^*} \ell'(\mu, \theta_1, \theta_2) d\mu
\end{aligned}$$

When this difference is equal to zero, it means that the positive area and the negative area are equivalent and they cancel each other out. If we increase θ_1, then the curve $\ell'(\mu, \theta_1, \theta_2)$ will shift upwards and the negative area will decrease, therefore we have to decrease θ_2 to counterbalance this effect. The opposite happens when we decrease θ_1. Therefore, $q(\theta_1)$ is a downward-sloping curve and it is continuous because of the continuity of $\ell'(\mu, \theta_1, \theta_2)$. This completes the proof. ■

This theoretical result is confirmed by simulations.

It turns out that there is a more general result. If the homomorphism density $t(H_2, G)$ associated with the parameter θ_2 is such that the resulting variational problem can be shown to be

$$\psi = \sup_{\mu \in [0,1]} \ell(\mu, \alpha, \beta) = \sup_{\mu \in [0,1]} \{\alpha\mu + \beta\mu^r - \mu \ln \mu - (1-\mu)\ln(1-\mu)\}$$

where we assume $r \geq 2$, then the same characterization applies, as shown in the next theorem. For example, this is the case if we consider

$$t(H_2, G) = \frac{\sum_i \sum_j \sum_k g_{ij} g_{jk} g_{ki}}{n^3}$$

with $r = 3$; or if we consider

$$t(H_2, G) = \frac{\sum_i \sum_j \sum_k \sum_l g_{ij} g_{jk} g_{kl} g_{li}}{n^4}$$

with $r = 4$.

The next Lemma, provides conditions under which the network statistics can be upper-bounded by the power of the graphon. For practical purposes this condition is necessary to be able to re-write the variational problem as a calculus problem, as shown in the Theorems below.

LEMMA 6 *For the following homomorphism densities:*

$$t(H, G) = \frac{\sum_i \sum_j \sum_k g_{ij} g_{jk} g_{ki}}{n^3} \tag{71}$$

$$t(H, G) = \frac{\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n g_{ij} g_{jk}}{n^3} \tag{72}$$

$$t(H, G) = \frac{\sum_i \sum_j \sum_k \sum_l g_{ij} g_{jk} g_{kl} g_{li}}{n^4} \tag{73}$$

$$t(H, G) = \frac{1}{n^r} \sum_{1 \leq i, j_1, j_2, .., j_r \leq n} g_{ij_1} g_{j_1 j_2} \cdot \cdot g_{j_r i} \tag{74}$$

$$t(H, G) = \frac{1}{n^{r-1}} \sum_{1 \leq i, j_1, j_2, .., j_r \leq n} g_{ij_1} g_{ij_2} \cdot \cdot g_{ij_r} \tag{75}$$

the following property holds

$$t(H, h) \leq \int_0^1 \int_0^1 h(x, y)^{e(H)} dx dy$$

where $e(H)$ is the number of directed links included in the subgraph H.

Proof. For the homomorphism density (71) the value $e(H) = 3$ and the limit object is

$$t(H,h) = \int_{[0,1]^3} h(x,y)h(y,z)h(z,x)dxdydz$$

Using the Holder inequality and some algebra, we obtain

$$\begin{aligned}
t(H,h) &= \int_{[0,1]^3} h(x,y)h(y,z)h(z,x)dxdydz \\
&\leq \left(\int_{[0,1]^3} h(x,y)^3 dxdydz\right)^{\frac{1}{3}} \left(\int_{[0,1]^3} h(y,z)^3 dxdydz\right)^{\frac{1}{3}} \left(\int_{[0,1]^3} h(z,x)^3 dxdydz\right)^{\frac{1}{3}} \\
&= \left(\int_{[0,1]^2} h(x,y)^3 dxdy \int_0^1 dz\right)^{\frac{1}{3}} \left(\int_{[0,1]^2} h(y,z)^3 dydz \int_0^1 dx\right)^{\frac{1}{3}} \left(\int_{[0,1]^2} h(z,x)^3 dxdz \int_0^1 dy\right)^{\frac{1}{3}} \\
&= \left(\int_{[0,1]^2} h(x,y)^3 dxdy\right)^{\frac{1}{3}} \left(\int_{[0,1]^2} h(y,z)^3 dydz\right)^{\frac{1}{3}} \left(\int_{[0,1]^2} h(z,x)^3 dxdz\right)^{\frac{1}{3}} \\
&= \left(\int_{[0,1]^2} h(x,y)^3 dxdy\right)^{\frac{1}{3}} \left(\int_{[0,1]^2} h(x,y)^3 dxdy\right)^{\frac{1}{3}} \left(\int_{[0,1]^2} h(x,y)^3 dxdy\right)^{\frac{1}{3}} \\
&= \int_{[0,1]^2} h(x,y)^3 dxdy = \int_0^1 \int_0^1 h(x,y)^{e(H)} dxdy
\end{aligned}$$

For the homomorphism density in (72), $e(H) = 2$ and using Holder inequality we get

$$\begin{aligned}
t(H,h) &= \int\int\int_{[0,1]^3} h(x,y)h(y,z)dxdydz \\
&\leq \left(\int\int\int_{[0,1]^3} h(x,y)^2 dxdydz\right)^{\frac{1}{2}} \left(\int\int\int_{[0,1]^3} h(y,z)^2 dxdydz\right)^{\frac{1}{2}} \\
&= \left(\int\int_{[0,1]^2} h(x,y)^2 \left[\int_{[0,1]} dz\right] dxdy\right)^{\frac{1}{2}} \left(\int\int_{[0,1]^2} h(y,z)^2 \left[\int_{[0,1]} dx\right] dydz\right)^{\frac{1}{2}} \\
&= \left(\int\int_{[0,1]^2} h(x,y)^2 dxdy\right)^{\frac{1}{2}} \left(\int\int_{[0,1]^2} h(y,z)^2 dydz\right)^{\frac{1}{2}} \\
&= \left(\int\int_{[0,1]^2} h(x,y)^2 dxdy\right)^{\frac{1}{2}} \left(\int\int_{[0,1]^2} h(x,y)^2 dxdy\right)^{\frac{1}{2}} \\
&= \int\int_{[0,1]^2} h(x,y)^2 dxdy
\end{aligned}$$

For the homomorphism density in (74), $e(H) = r$ and using Holder inequality we get

$$\begin{aligned}
t(H,h) &= \int_{[0,1]^r} h(x_i,x_{j_1})h(x_{j_1},x_{j_2})\cdots h(x_{j_r},x_i)dx_i dx_{j_1}\cdots dx_{j_r} \\
&\leq \left(\int_{[0,1]^r} h(x_i,x_{j_1})^r dx_i dx_{j_1}\cdots dx_{j_r}\right)^{\frac{1}{r}} \left(\int_{[0,1]^r} h(x_{j_1},x_{j_2})^r dx_i dx_{j_1}\cdots dx_{j_r}\right)^{\frac{1}{r}} \\
&\quad \cdots \left(\int_{[0,1]^r} h(x_{j_r},x_i)^r dx_i dx_{j_1}\cdots dx_{j_r}\right)^{\frac{1}{r}} \\
&= \left(\int_{[0,1]^2} h(x_i,x_{j_1})^r dx_i dx_{j_1} \int_{[0,1]^{r-2}} dx_{j_2}\cdots dx_{j_r}\right)^{\frac{1}{r}} \\
&\quad \times \left(\int_{[0,1]^2} h(x_{j_1},x_{j_2})^r dx_{j_1} dx_{j_2} \int_{[0,1]^{r-2}} dx_i dx_{j_3}\cdots dx_{j_r}\right)^{\frac{1}{r}} \\
&\quad \cdots \left(\int_{[0,1]^2} h(x_{j_r},x_i)^r dx_{j_r} dx_i \int_{[0,1]^{r-2}} dx_{j_1}\cdots dx_{j_{r-1}}\right)^{\frac{1}{r}} \\
&= \left(\int_{[0,1]^2} h(x,y)^r dxdy\right)^{\frac{1}{r}} \left(\int_{[0,1]^2} h(x,y)^r dxdy\right)^{\frac{1}{r}} \cdots \left(\int_{[0,1]^2} h(x,y)^r dxdy\right)^{\frac{1}{r}} \\
&= \int_{[0,1]^2} h(x,y)^r dxdy = \int_0^1\int_0^1 h(x,y)^{e(H)} dxdy
\end{aligned}$$

For the homomorphism density in (75), $e(H) = r$ and using Holder inequality we get

$$\begin{aligned}
t(H,h) &= \int_{[0,1]^r} h(x_i,x_{j_1})h(x_i,x_{j_2})\cdots h(x_i,x_{j_r})dx_i dx_{j_1}\cdots dx_{j_r} \\
&\leq \left(\int_{[0,1]^r} h(x_i,x_{j_1})^r dx_i dx_{j_1}\cdots dx_{j_r}\right)^{\frac{1}{r}} \left(\int_{[0,1]^r} h(x_i,x_{j_2})^r dx_i dx_{j_1}\cdots dx_{j_r}\right)^{\frac{1}{r}} \\
&\quad \cdots \left(\int_{[0,1]^r} h(x_i,x_{j_r})^r dx_i dx_{j_1}\cdots dx_{j_r}\right)^{\frac{1}{r}} \\
&= \left(\int_{[0,1]^2} h(x_i,x_{j_1})^r dx_i dx_{j_1} \int_{[0,1]^{r-2}} dx_{j_2}\cdots dx_{j_r}\right)^{\frac{1}{r}} \\
&\quad \times \left(\int_{[0,1]^2} h(x_i,x_{j_2})^r dx_i dx_{j_2} \int_{[0,1]^{r-2}} dx_{j_1} dx_{j_3}\cdots dx_{j_r}\right)^{\frac{1}{r}} \\
&\quad \cdots \left(\int_{[0,1]^2} h(x_i,x_{j_r})^r dx_i dx_{j_r} \int_{[0,1]^{r-2}} dx_{j_1}\cdots dx_{j_{r-1}}\right)^{\frac{1}{r}} \\
&= \left(\int_{[0,1]^2} h(x,y)^r dxdy\right)^{\frac{1}{r}} \left(\int_{[0,1]^2} h(x,y)^r dxdy\right)^{\frac{1}{r}} \cdots \left(\int_{[0,1]^2} h(x,y)^r dxdy\right)^{\frac{1}{r}} \\
&= \int_{[0,1]^2} h(x,y)^r dxdy = \int_0^1\int_0^1 h(x,y)^{e(H)} dxdy
\end{aligned}$$

■

The following theorem uses the result of the Lemma 6 above, to show that the variational problem can be solved explicitly as a one-variable calculus problem in special cases. This result is very useful in studying the behavior of the model as the number of players grows large and it provides a way to characterize the convergence of the sampling algorithms according to the same argument of Bhamidi et al. (2011) (see more detail below).

THEOREM 12 *Let $\beta > 0$. For the following models*

$$\mathcal{T}(G) = \alpha\frac{\sum_{i=1}^n\sum_{j=1}^n g_{ij}}{n^2} + \beta\frac{\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n g_{ij}g_{jk}}{n^3}$$

$$\mathcal{T}(G) = \alpha\frac{\sum_{i=1}^n\sum_{j=1}^n g_{ij}}{n^2} + \beta\frac{\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n g_{ij}g_{jk}g_{ki}}{n^3}$$

$$\mathcal{T}(G) = \alpha\frac{\sum_{i=1}^n\sum_{j=1}^n g_{ij}}{n^2} + \beta\frac{\sum_{1\leq i,j_1,j_2,..,j_r\leq n} g_{ij_1}g_{j_1j_2}\cdot\cdot g_{j_ri}}{n^r}$$

$$\mathcal{T}(G) = \alpha\frac{\sum_{i=1}^n\sum_{j=1}^n g_{ij}}{n^2} + \beta\frac{\sum_{1\leq i,j_1,j_2,..,j_r\leq n} g_{ij_1}g_{ij_2}\cdot\cdot g_{ij_r}}{n^{r-1}}$$

the log-partitition asymptotic variational problem becomes a calculus problem. Let $\ell(\mu, \alpha, \beta)$ *be the following function*

$$\ell(\mu, \alpha, \beta) = \alpha\mu + \beta\mu^r - \mu \log \mu - (1-\mu)\log(1-\mu)$$

Then, as $n \to \infty$*, the log-partition is the solution of the following*

$$\lim_{n\to\infty} \psi_n(\theta) = \psi(\theta) = \sup_{\mu\in[0,1]} \ell(\mu, \alpha, \beta)$$

For the following model with $\beta > 0$ *and* $\gamma > 0$

$$\mathcal{T}(G) = \alpha \frac{\sum_{i=1}^n \sum_{j=1}^n g_{ij}}{n^2} + \beta \frac{\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n g_{ij} g_{jk}}{n^3} + \gamma \frac{\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n g_{ij} g_{jk} g_{ki}}{n^3}$$

the log-partitition asymptotic variational problem is

$$\lim_{n\to\infty} \psi_n(\theta) = \psi(\theta) = \sup_{\mu\in[0,1]} \left\{\alpha\mu + \beta\mu^2 + \gamma\mu^3 - \mu \log \mu - (1-\mu)\log(1-\mu)\right\}$$

Proof. Consider the first model. We have assumed that $\beta > 0$. Given the results of the Holder's inequality in Lemma 6 we can say that

$$\begin{aligned} \mathcal{T}(h) &= \alpha t(H_1, h) + \beta t(H_2, h) \\ &\leq \alpha \int\int_{[0,1]^2} h(x,y)dxdy + \beta \int\int_{[0,1]^2} h(x,y)^2 dxdy \end{aligned}$$

Suppose $h(x, y) = \mu$ is a constant. Then the equality holds and if $\mu \in [0, 1]$ solves the variational problem

$$\lim_{n\to\infty} \psi_n(\theta) = \psi(\theta) = \sup_{\mu\in[0,1]} \alpha\mu + \beta\mu^2 - \mu \log \mu - (1-\mu)\log(1-\mu)$$

then $h(x, y) = \mu$ is the limit graphon.

To show that this is the only solution, let's consider the maximization problem again. For $h(x, y)$ to be a solution, we need

$$\mathcal{T}(h) = \alpha \int\int_{[0,1]^2} h(x,y)dxdy + \beta \int\int_{[0,1]^2} h(x,y)^2 dxdy$$

In other words, the Holder inequality must hold with equality, i.e. we need

$$\begin{aligned} t(H_2, h) &= \int\int\int_{[0,1]^3} h(x,y)h(y,z)dxdydz \\ &= \int\int_{[0,1]^2} h(x,y)^2 dxdy \end{aligned}$$

This implies that

$$h(x,y) = h(y,z)$$

for almost all (x,y,z). In particular, we have that given x and y, $\mu = h(x,y) = h(y,z)$ for any $z \in [0,1]$ because the left-hand-side does not depend on z. Given y and z, we have $\mu' = h(y,z) = h(x,y)$ for any $x \in [0,1]$ because the left-hand-side does not depend on x. For $x = y$ and $z = y$ we have $\mu = h(y,y) = h(y,y) = \mu'$. In addition, we have $h(x,y) = h(y,x) = \mu = h(x,z)$. It follows that $h(x,y) = \mu$ almost everywhere.

It follows that $\mathcal{T}(h) = \alpha\mu + \beta\mu^2$ and $I(\mu) = \mu\log\mu + (1-\mu)\log(1-\mu)$, so we get

$$\lim_{n\to\infty} \psi_n = \psi = \sup_{\mu\in[0,1]} \left\{\alpha\mu + \beta\mu^2 - \mu\log\mu - (1-\mu)\log(1-\mu)\right\}$$

The proof for the remaining models follows similar steps and reasoning and it is omitted for brevity. ■

The next theorem contains a complete characterization of the maximization problem considered in the previous theorem.

THEOREM 13 *Assume that $\beta > 0$ and $r \geq 2$. If the variational problem can be shown to be*

$$\lim_{n\to\infty} \psi_n(\theta) = \psi(\theta) = \sup_{\mu\in[0,1]} \{\alpha\mu + \beta\mu^r - \mu\log\mu - (1-\mu)\log(1-\mu)\}$$

then we have

1. *If $\beta \leq \frac{r^{r-1}}{(r-1)^r}$, the maximization problem has a unique maximizer $\mu^* \in [0,1]$*
2. *If $\beta > \frac{r^{r-1}}{(r-1)^r}$ and $\alpha \geq \log(r-1) - \frac{r}{r-1}$ then there is a unique maximizer $\mu^* > 0.5$*
3. *If $\beta > \frac{r^{r-1}}{(r-1)^r}$ and $\alpha < \log(r-1) - \frac{r}{r-1}$, then there is a V-shaped region of parameters such that*
   (a) *inside the V-shaped region, the maximization problem has two local maximizers $\mu_1^* < 0.5 < \mu_2^*$*
   (b) *outside the V-shaped region, the maximization problem has a unique maximizer μ^**
4. *For any α inside the V-shaped region, there exists a $\beta = \zeta(\alpha)$, such that the two maximizers are both global, i.e. $\ell(\mu_1^*) = \ell(\mu_2^*)$.*

Proof. The first and second order conditions are

$$\begin{aligned}
\ell'(\mu,\alpha,\beta) &= \alpha + \beta r\mu^{r-1} - \ln\left(\frac{\mu}{1-\mu}\right) \\
\ell''(\mu,\alpha,\beta) &= \beta r(r-1)\mu^{r-2} - \frac{1}{\mu(1-\mu)}
\end{aligned}$$

The function $\ell(\mu, \alpha, \beta)$ is concave if $\ell''(\mu, \alpha, \beta) < 0$, i.e. when

$$\beta < \frac{1}{r(r-1)\mu^{r-1}(1-\mu)} \equiv s(\mu)$$

The function $s(\mu)$ has a minimum at $\frac{r}{r-1}$, where $s(\frac{r}{r-1}) = \frac{r^{r-1}}{(r-1)^r}$; it is decreasing i, i.e. $\ell(\mu_1^*) = \ell(\mu_2^*)$n the interval $\left[0, \frac{r}{r-1}\right)$ and increasing in the interval $\left(\frac{r}{r-1}, 1\right]$. Therefore the function $\ell(\mu, \alpha, \beta)$ is concave on the whole interval $[0, 1]$ if $\beta < \frac{r^{r-1}}{(r-1)^r}$. [53] In this region, there is a unique maximizer μ^* of $\ell(\mu, \alpha, \beta)$.

If $\beta > \frac{r^{r-1}}{(r-1)^r}$ there are three possible cases. We know that in this region the second derivative $\ell''(\mu, \alpha, \beta)$ can be positive or negative, with inflection points denoted as a and b, found by solving the equation $\beta = s(\mu)$. An example for $r = 3$ and $\beta = 4$ is shown in the figure below (notice that we are plotting the function $1/s(\mu)$ against the line $1/\beta$).

[53] Consider the function $1/s(\mu) = r(r-1)\mu^{r-1}(1-\mu) = r(r-1)(\mu^{r-1} - \mu^r)$. This function has derivative

$$\frac{\partial[1/s(\mu)]}{\partial \mu} = r(r-1)^2\mu^{r-2} - r^2(r-1)\mu^{r-1} = r(r-1)\mu^{r-2}[(r-1) - r\mu]$$

$$\frac{\partial^2[1/s(\mu)]}{\partial \mu \partial \mu} = r(r-1)^2(r-2)\mu^{r-3} - r^2(r-1)^2\mu^{r-2} = r(r-1)^2\mu^{r-3}[(r-2) - r\mu]$$

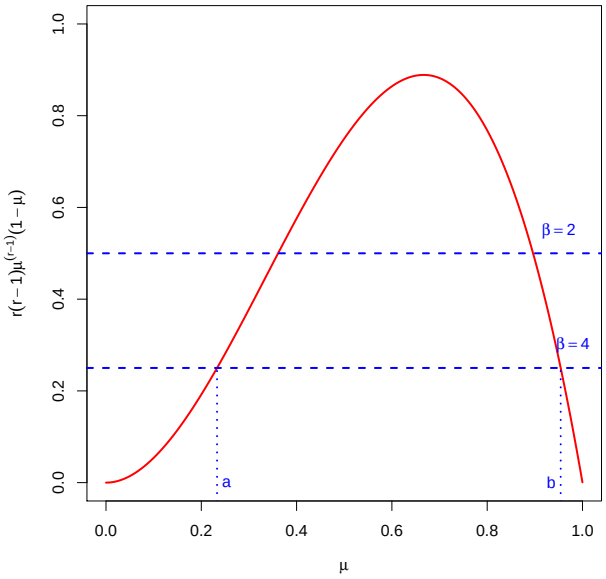
So solving the FOCs we obtain the maximizer of $1/s(\mu)$

$$r(r-1)\mu^{r-2}[(r-1) - r\mu] = 0 \Leftrightarrow \mu = \frac{r-1}{r}$$

and the maximum is

$$1/s\left(\frac{r-1}{r}\right) = r(r-1)\left(\frac{r-1}{r}\right)^{r-1}(1 - \frac{r-1}{r}) = \frac{(r-1)^r}{r^{r-1}}$$

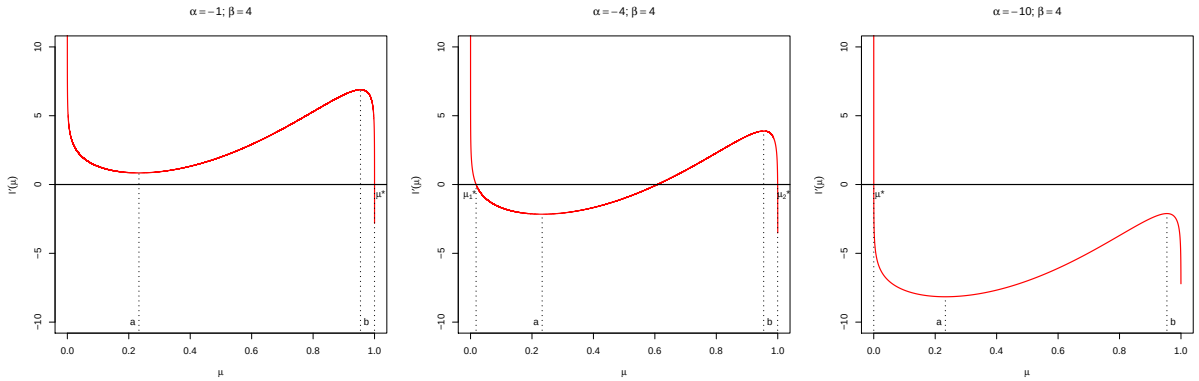
Therefore the minimum of $s(\mu)$ is $\frac{r^{r-1}}{(r-1)^r}$.



In particular, the first derivative $\ell'(\mu, \alpha, \beta)$ is decreasing in $[0, a)$, increasing in (a, b) and decreasing in $(b, 1]$.

1. If $\ell'(a, \alpha, \beta) \geq 0$, then there is a unique maximizer $\mu^* > b$
2. If $\ell'(b, \alpha, \beta) \leq 0$, then there is a unique maximizer $\mu^* < a$
3. If $\ell'(a, \alpha, \beta) < 0 < \ell'(b, \alpha, \beta)$, then there are 2 local maximizers $\mu_1^* < a < b < \mu_2^*$

The three cases are shown in the following pictures, where we plot $\ell'(\mu, \alpha, \beta)$ against μ for several values of α and for a fixed $\beta = 4$. In the pictures $r = 3$.



We indicate the maximizer with μ^* when it is unique, and with μ_1^*, μ_2^* when there are two.

Let's consider the first case, with $\ell'(a, \alpha, \beta) \geq 0$. To compute $\ell'(a, \alpha, \beta)$, notice that $\beta = s(a) = \frac{1}{r(r-1)a^{r-1}(1-a)}$. Substituting in $\ell'(a, \alpha, \beta)$ we obtain

$$\ell'(a,\alpha,\beta) = \alpha + \frac{1}{(r-1)(1-a)} - \log\frac{a}{1-a}$$

and analogously for $\beta = s(b) = \frac{1}{r(r-1)b^{r-1}(1-b)}$we have

$$\ell'(b,\alpha,\beta) = \alpha + \frac{1}{(r-1)(1-b)} - \log\frac{b}{1-b}$$

So $\ell'(a,\alpha,\beta) \geq 0$ implies

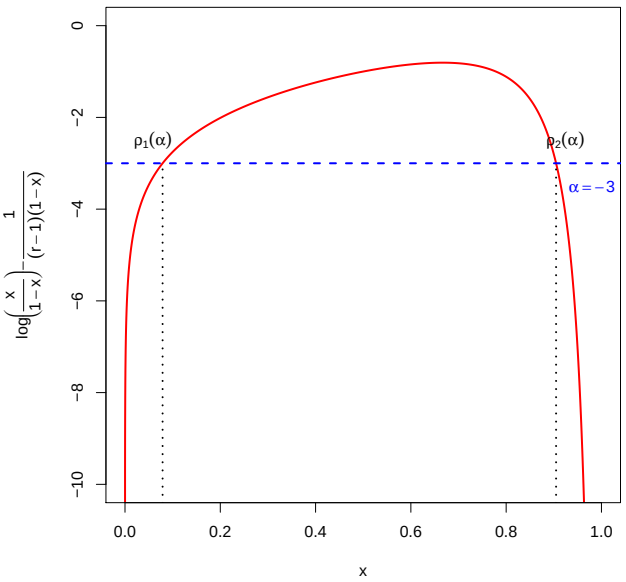
$$\alpha \geq \log\frac{a}{1-a} - \frac{1}{(r-1)(1-a)}$$

The function $\log\frac{a}{1-a} - \frac{1}{(r-1)(1-a)}$ has a maximum at $\log(r-1) - \frac{r}{r-1}$ and therefore we have [54]

$$\ell'(a,\alpha,\beta) \geq 0 \Leftrightarrow \theta_1 \geq \log(r-1) - \frac{r}{r-1}$$

When the above condition is satisfied, there is a unique maximizer, $\mu^* > b$, as shown in the picture on the left.

When $\theta_1 < \log(r-1) - \frac{r}{r-1}$ it is easier to draw a picture of the function $\log\frac{a}{1-a} - \frac{1}{(r-1)(1-a)}$, shown below.



Notice that when $\theta_1 < \log(r-1) - \frac{r}{r-1}$ there are two intersections of the function and the horizontal line $y = \alpha$ (in the picture $\alpha = -3$). We denote the intersections $\phi_1(\alpha)$ and

[54] Taking derivative $\frac{1}{a} + \frac{1}{1-a} - \frac{1}{(r-1)(1-a)^2} = 0$, we obtain the maximizer $a^* = \frac{r-1}{r}$. The function is increasing in $[0, \frac{r-1}{r})$ and decreasing in $(\frac{r-1}{r}, 1]$. The maximum is therefore at $\log(r-1) - \frac{r}{r-1}$.

$\phi_2(\alpha)$. By construction, we know that $a < 0.5 < b$. By looking at the picture, it is clear that $\ell'(a,\alpha,\beta) > 0$ if $a < \phi_1(\alpha)$ and $\ell'(a,\alpha,\beta) < 0$ if $a > \phi_1(\alpha)$. Analogously, we have $\ell'(b,\alpha,\beta) > 0$ if $b > \phi_2(\alpha)$ and $\ell'(b,\alpha,\beta) < 0$ if $b < \phi_2(\alpha)$.

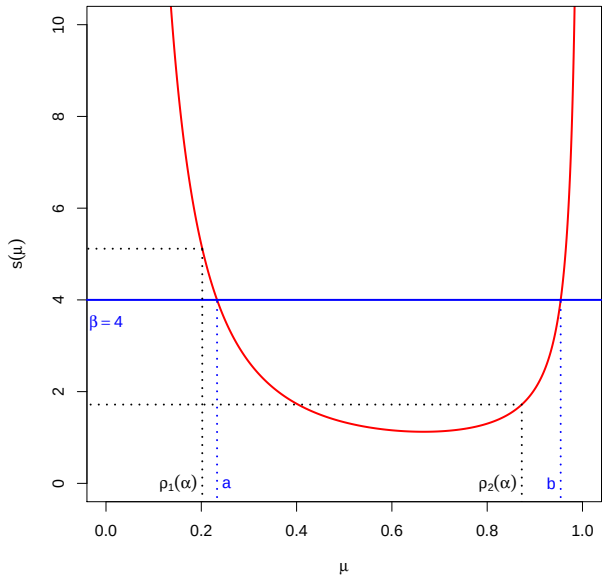
For any $\alpha < -2$, there exist $\phi_1(\alpha)$ and $\phi_2(\alpha)$ which are the intersection of the function $y = \log\left(\frac{x}{1-x}\right) - \frac{1}{(r-1)(1-x)}$ with the line $y = \alpha$. Since the function is continuous, monotonic increasing in $[0, \frac{r-1}{r})$ and monotonic decreasing in $(\frac{r-1}{r}, 1]$ it follows that $\phi_1(\alpha)$ and $\phi_2(\alpha)$ are both continuous in α. In addition, $\phi_1(\alpha)$ is increasing in α and $\phi_2(\alpha)$ is decreasing in α. It's trivial to show that when α decreases, $\phi_1(\alpha)$ converges to 0 while $\phi_2(\alpha)$ converges to 1.

Consider the case in which $\ell'(a,\alpha,\beta) < 0 < \ell'(b,\alpha,\beta)$ with two maximizers of $\ell(\mu,\alpha,\beta)$. Consider the function $s(\mu)$ defined above.

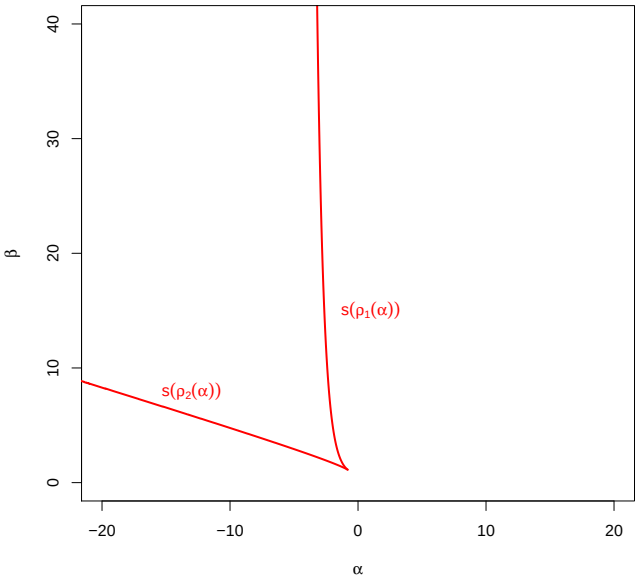
Since $\ell'(a,\alpha,\beta) < 0$ we have $a > \phi_1(\alpha)$, which implies $s(a) < s(\phi_1(\alpha))$. Therefore $\beta < s(\phi_1(\alpha,\beta)) = \frac{1}{r(r-1)\phi_1(\alpha)^{r-1}(1-\phi_1(\alpha))}$.

Since $\ell'(b,\alpha,\beta) > 0$ we have $b > \phi_2(\alpha)$, which implies $s(b) > s(\phi_2(\alpha))$. Therefore $\beta > s(\phi_2(\alpha)) = \frac{1}{r(r-1)\phi_2(\alpha)^{r-1}(1-\phi_2(\alpha))}$.

Notice that $s(\phi_1(\alpha)) > s(\phi_2(\alpha))$ for any (α,β) in this region of the parameters (see picture below for an example with $\beta = 4$, $\alpha = -2$, and $r = 3$).



The areas are shown in the following picture



and the rest of the proof follows. The existence of $\zeta(\alpha)$ is shown using similar argument as in the proof of Theorem 11, so it is omitted for brevity. ■

The next result is analogous to Theorem 6.3 in Diaconis and Chatterjee (2011), adapted to the directed network model. It shows that not all the specifications of the model generate directed Erdos-Renyi networks. We show this by focusing on a special case.

THEOREM 14 *Consider the model with re-scaled potential* $\mathcal{T}(G)$ *and with* $\beta < 0$,

$$\mathcal{T}(G) = \alpha \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}}{n^2} + \beta \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} g_{ij}g_{jk}}{n^3}$$

Then for any value of α*, there exists a positive constant* $C(\alpha)$ *such that for* $\beta < -C(\alpha)$*, the variational problem is not solved at a constant graphon.*

Proof. Fix the value of α and let $p = \frac{e^{\alpha}}{1+e^{\alpha}}$, and $\lambda = -\beta$. For any h we have

$$
\begin{aligned}
\mathcal{T}(h)-\mathcal{I}(h) &= \alpha\int h(x,y)dxdy+\beta\int h(x,y)h(y,z)dxdydz \\
&\quad -\int h(x,y)\ln h(x,y)+(1-h(x,y))\ln(1-h(x,y))dxdy \\
&= \alpha\int h(x,y)dxdy+\beta\int h(x,y)h(y,z)dxdydz \\
&\quad +\int h(x,y)\ln(1+e^{\alpha})dxdy-\int h(x,y)\ln(1+e^{\alpha})dxdy \\
&\quad -\int h(x,y)\ln h(x,y)+(1-h(x,y))\ln(1-h(x,y))dxdy \\
&= \beta\int h(x,y)h(y,z)dxdydz+\int h(x,y)\ln p dxdy+\int h(x,y)\ln(1-p)dxdy \\
&\quad -\int h(x,y)\ln h(x,y)+(1-h(x,y))\ln(1-h(x,y))dxdy \\
&= \beta\int h(x,y)h(y,z)dxdydz+\int h(x,y)\ln p dxdy+\int h(x,y)\ln(1-p)dxdy \\
&\quad +\int \ln(1-p)dxdy-\int \ln(1-p)dxdy \\
&\quad -\int h(x,y)\ln h(x,y)+(1-h(x,y))\ln(1-h(x,y))dxdy \\
&= \beta\int h(x,y)h(y,z)dxdydz+\ln(1-p) \\
&\quad -\int h(x,y)\ln\frac{h(x,y)}{p}+(1-h(x,y))\ln\frac{1-h(x,y)}{1-p}dxdy \\
&= -\lambda t(H_2,h)+\ln(1-p)-\mathcal{I}_p(h)
\end{aligned}
$$

We have assumed that $\beta<0$. Assume that the quantity $\mathcal{T}(h)-\mathcal{I}(h)$ is maximized at a constant graphon $h(x,y)=\mu$. As a consequence, μ *minimizes* the function

$$\lambda t(H_2,h)+\mathcal{I}_p(h)=\lambda\mu^2+\mathcal{I}_p(\mu)$$

Since μ is the graphon that maximizes $\mathcal{T}(h)-\mathcal{I}(h)$, then we have that for any $x\in[0,1]$, the following holds: $\lambda\mu^2+\mathcal{I}_p(\mu)\leq\lambda x^2+\mathcal{I}_p(x)$. The first order conditions for minimization give

$$v(x)=\frac{d}{dx}\left[\lambda x^2+\mathcal{I}_p(x)\right]=2\lambda x+\ln\frac{x}{1-x}-\ln\frac{p}{1-p}$$

Notice that $v(0)=-\infty$ and $v(1)=+\infty$, therefore μ must be an interior minimum. By solving the first order conditions

$$2\lambda\mu+\ln\frac{\mu}{1-\mu}-\ln\frac{p}{1-p}=0$$

it is easy to see that there exists a function $c(\lambda)$ such that

$$\mu = \frac{\exp\left[-2\lambda\mu + \ln\frac{p}{1-p}\right]}{1+\exp\left[-2\lambda\mu + \ln\frac{p}{1-p}\right]} \leq c(\lambda)$$

So we get$\mu \leq c(\lambda)$, where $c(\lambda)$ is a function such that

$$\lim_{\lambda\to\infty} c(\lambda) = 0$$

and therefore it follows that

$$\lim_{\lambda\to\infty} \min_{x\in[0,1]} \lambda x^2 + \mathcal{I}_p(x) = \mathcal{I}_p(0) = \ln\frac{1}{1-p}$$

We will now show that there exists a graphon $\nu(x,y)$ which is not a constant and gives a lower value of the expression above.

Let $\nu(x,y)$ be the function

$$\nu(x,y) = \begin{cases} p & \text{if } x \in [0,.5] \text{and } y \in [.5,1] \\ 0 & \text{otherwise} \end{cases}$$

It follows that for almost all (x,y,z) triplets, $\nu(x,y)\nu(y,z) = 0$ and thus, $t(H_2,\nu) = 0$. If we compute the value of $\mathcal{I}_p(\nu)$ we obtain

$$\begin{aligned}
\mathcal{I}_p(\nu) &= \int_{\left[\frac{1}{2},1\right]\times\left[0,\frac{1}{2}\right]} 0\ln\frac{0}{p} + \ln\frac{1}{1-p}dxdy \\
&+ \int_{\left[0,\frac{1}{2}\right]\times\left[0,\frac{1}{2}\right]} 0\ln\frac{0}{p} + \ln\frac{1}{1-p}dxdy \\
&+ \int_{\left[0,\frac{1}{2}\right]\times\left[\frac{1}{2},1\right]} p\ln\frac{p}{p} + (1-p)\ln\frac{1-p}{1-p}dxdy \\
&+ \int_{\left[\frac{1}{2},1\right]\times\left[\frac{1}{2},1\right]} 0\ln\frac{0}{p} + \ln\frac{1}{1-p}dxdy \\
&= \frac{3}{4}\ln\frac{1}{1-p}
\end{aligned}$$

Therefore we have shown that for λ large enough (i.e. for β negative and large enough), $\mathcal{T}(\nu) - \mathcal{I}(\nu) \geq \mathcal{T}(\mu) - \mathcal{I}(\mu)$. So, given a value for α, there exists a $C(\alpha)$ large enough, such that for any $\beta < -C(\alpha)$ a constant graphon is not solution to the variational problem. ■

This result extends to models with two parameters and higher order dependencies, as shown in the next theorem

THEOREM 15 *For the models in the first part of Theorem 12, the result of Theorem 14 hold.*

Proof. The proof is equivalent to the proof of Theorem 14, replacing μ^2 with μ^r, where r is the order of dependence of the second homomorphism density $t(H_2, h)$.

■

THEOREM 16 *Consider the model with re-scaled potential $\mathcal{T}(G)$ and with $\beta < 0$,*

$$\mathcal{T}(G) = \alpha \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij}}{n^2} + \beta \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk}}{n^3} + \gamma \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} g_{ij} g_{jk} g_{ki}}{n^3} \tag{76}$$

Then for any value of $\alpha \in \mathbb{R}$ and $\gamma > 0$, there exists a positive constant $C(\alpha, \gamma) > 0$ such that for $\beta < -C(\alpha, \gamma)$, the variational problem is not solved at a constant graphon. Analogously, if $\gamma < 0$, then for any value of $\alpha \in \mathbb{R}$ and $\beta > 0$, there exists a positive constant $C(\alpha, \beta) > 0$ such that for $\gamma < C(\alpha, \gamma)$, the variational problem is not solved at a constant graphon.

Proof. Fix the value of α and $\gamma > 0$. Let $p = \frac{e^\alpha}{1+e^\alpha}$, and $\lambda = -\beta$. For any h we have

$$
\begin{aligned}
\mathcal{T}(h) - \mathcal{I}(h) &= \alpha \int h(x,y)dxdy + \beta \int h(x,y)h(y,z)dxdydz + \gamma \int h(x,y)h(y,z)h(z,x)dxdydz \\
&\quad - \int h(x,y)\ln h(x,y) + (1-h(x,y))\ln(1-h(x,y))dxdy \\
&= \alpha \int h(x,y)dxdy + \beta \int h(x,y)h(y,z)dxdydz + \gamma \int h(x,y)h(y,z)h(z,x)dxdydz \\
&\quad + \int h(x,y)\ln(1+e^{\alpha})dxdy - \int h(x,y)\ln(1+e^{\alpha})dxdy \\
&\quad - \int h(x,y)\ln h(x,y) + (1-h(x,y))\ln(1-h(x,y))dxdy \\
&= \beta \int h(x,y)h(y,z)dxdydz + \gamma \int h(x,y)h(y,z)h(z,x)dxdydz \\
&\quad + \int h(x,y)\ln p dxdy + \int h(x,y)\ln(1-p)dxdy \\
&\quad - \int h(x,y)\ln h(x,y) + (1-h(x,y))\ln(1-h(x,y))dxdy \\
&= \beta \int h(x,y)h(y,z)dxdydz + \gamma \int h(x,y)h(y,z)h(z,x)dxdydz \\
&\quad + \int h(x,y)\ln p dxdy + \int h(x,y)\ln(1-p)dxdy \\
&\quad + \int \ln(1-p)dxdy - \int \ln(1-p)dxdy \\
&\quad - \int h(x,y)\ln h(x,y) + (1-h(x,y))\ln(1-h(x,y))dxdy \\
&= \beta \int h(x,y)h(y,z)dxdydz + \gamma \int h(x,y)h(y,z)h(z,x)dxdydz + \ln(1-p) \\
&\quad - \int h(x,y)\ln\frac{h(x,y)}{p} + (1-h(x,y))\ln\frac{1-h(x,y)}{1-p}dxdy \\
&= \beta t(H_2,h) + \gamma t(H_3,h) + \ln(1-p) - \mathcal{I}_p(h)
\end{aligned}
$$

We have assumed that $\beta < 0$. Assume that the quantity $\mathcal{T}(h) - \mathcal{I}(h)$ is maximized at a constant graphon $h(x,y) = \mu$. As a consequence, μ *maximizes* the function

$$\beta t(H_2,h) + \gamma t(H_3,h) - \mathcal{I}_p(h) = \beta\mu^2 + \gamma\mu^3 - \mathcal{I}_p(\mu)$$

Since μ is the graphon that maximizes $\mathcal{T}(h) - \mathcal{I}(h)$, then we have that for any $x \in [0,1]$, the following holds: $\beta\mu^2 + \gamma\mu^3 - \mathcal{I}_p(\mu) \geq \beta x^2 + \gamma x^3 - \mathcal{I}_p(x)$. The first order conditions for maximization give

$$v(x) = \frac{d}{dx}\left[\beta x^2 + \gamma x^3 - \mathcal{I}_p(x)\right] = 2\beta x + 3\gamma x^2 - \ln\frac{x}{1-x} + \ln\frac{p}{1-p}$$

Notice that $v(0) = +\infty$ and $v(1) = -\infty$, therefore μ must be an interior maximum. By solving the first order conditions

$$2\beta\mu + 3\gamma\mu^2 - \ln\frac{\mu}{1-\mu} + \ln\frac{p}{1-p} = 0$$

it is easy to see that there exists a function $c(\beta,\gamma)$ such that

$$\mu = \frac{\exp\left[2\beta\mu + 3\gamma\mu^2 - \ln\frac{p}{1-p}\right]}{1+\exp\left[2\beta\mu + 3\gamma\mu^2 - \ln\frac{p}{1-p}\right]} \leq c(\beta,\gamma)$$

So we get $\mu \leq c(\beta,\gamma)$, and $c(\beta,\gamma)$ is a function such that

$$\lim_{\beta\to-\infty} c(\beta,\gamma) = 0$$

and therefore, it follows that

$$\lim_{\beta\to-\infty}\min_{x\in[0,1]} \beta x^2 + \gamma x^3 - \mathcal{I}_p(x) = -\mathcal{I}_p(0) = -\ln\frac{1}{1-p}$$

We will now show that there exists a graphon $\nu(x,y)$ which is not a constant and gives a lower value of the expression above.

Let $\nu(x,y)$ be the function

$$\nu(x,y) = \begin{cases} p & \text{if } x \in \left[0,\frac{1}{2}\right] \text{and } y \in \left[\frac{1}{2},1\right] \\ 0 & \text{otherwise} \end{cases}$$

It follows that for almost all (x,y,z) triplets, $\nu(x,y)\nu(y,z) = 0$ and $\nu(x,y)\nu(y,z)\nu(z,x) = 0$. As a consequence $t(H_2,\nu) = 0$ and $t(H_3,\nu) = 0$. If we compute the value of $\mathcal{I}_p(\nu)$ we obtain

$$\begin{aligned}
\mathcal{I}_p(\nu) &= \int_{\left[\frac{1}{2},1\right]\times\left[0,\frac{1}{2}\right]} 0\ln\frac{0}{p} + \ln\frac{1}{1-p}dxdy \\
&+ \int_{\left[0,\frac{1}{2}\right]\times\left[0,\frac{1}{2}\right]} 0\ln\frac{0}{p} + \ln\frac{1}{1-p}dxdy \\
&+ \int_{\left[0,\frac{1}{2}\right]\times\left[\frac{1}{2},1\right]} p\ln\frac{p}{p} + (1-p)\ln\frac{1-p}{1-p}dxdy \\
&+ \int_{\left[\frac{1}{2},1\right]\times\left[\frac{1}{2},1\right]} 0\ln\frac{0}{p} + \ln\frac{1}{1-p}dxdy \\
&= \frac{3}{4}\ln\frac{1}{1-p}
\end{aligned}$$

Therefore we have shown that for $\beta < 0$ large enough in magnitude, $\mathcal{T}(\nu) - \mathcal{I}(\nu) \geq \mathcal{T}(\mu) - \mathcal{I}(\mu)$. So, given a value of $\alpha \in \mathbb{R}$ and $\gamma > 0$, there exists a positive constant $C(\alpha, \gamma) > 0$, such that for $\beta < -C(\alpha, \gamma)$ a constant graphon is not solution to the variational problem (67) for the model in 76). The proof for $\gamma < 0$ follows the same steps. ■

THEOREM 17 *Fix parameter $\gamma > 0$. Let the variational problem be described as*

$$\lim_{n\to\infty} \psi_n(\theta) = \psi(\theta) = \sup_{\mu\in[0,1]} \left\{\alpha\mu + \beta\mu^2 + \gamma\mu^3 - \mu\log\mu - (1-\mu)\log(1-\mu)\right\}$$

Let μ_0 be (uniquely) determined by

$$6\gamma = \frac{2\mu_0 - 1}{\mu_0^2(1-\mu_0)^2}$$

and letα_0,β_0 be defined as follows:

$$\begin{aligned} \beta_0 &= \frac{1}{2\mu_0(1-\mu_0)} - 3\gamma\mu_0 \\ \alpha_0 &= \log\frac{\mu_0}{1-\mu_0} - \frac{1}{(1-\mu_0)} + \frac{2\mu_0 - 1}{2(1-\mu_0)^2} \end{aligned}$$

1. *If $\beta \leq \beta_0$, the maximization problem has a unique maximizer $\mu^* \in [0, 1]$*
2. *If $\beta > \beta_0$ and $\alpha \geq \alpha_0$ then there is a unique maximizer $\mu^* > 0.5$*
3. *If $\beta > \beta_0$ and $\alpha < \alpha_0$, then there are two functions $S_\gamma(\phi_1(\alpha))$ and $S_\gamma(\phi_2(\alpha))$ that define a V-shaped region of parameters (α, β) such that*
   (a) *inside the V-shaped region, the maximization problem has two local maximizers $\mu_1^* < 0.5 < \mu_2^*$*
   (b) *outside the V-shaped region, the maximization problem has a unique maximizer μ^**
4. *For any $\alpha < \alpha_0$ inside the V-shaped region, there exists a function $\beta = \zeta_\gamma(\alpha)$, such that $S_\gamma(\phi_1(\alpha)) < \zeta_\gamma(\alpha) < S_\gamma(\phi_2(\alpha))$ and the two maximizers are both global.*

Proof. Fix $\gamma > 0$ and consider the function

$$\ell_\gamma(\mu, \alpha, \beta) = \alpha\mu + \beta\mu^2 + \gamma\mu^3 - \mu\log\mu - (1-\mu)\log(1-\mu)$$

For the moment we do not constrain β to be positive. The first and second order derivatives w.r.t. μ are

$$\begin{aligned} \ell'_\gamma(\mu, \alpha, \beta) &= \alpha + 2\beta\mu + 3\gamma\mu^2 - \ln\left(\frac{\mu}{1-\mu}\right) \\ \ell''_\gamma(\mu, \alpha, \beta) &= 2\beta + 6\gamma\mu - \frac{1}{\mu(1-\mu)} \end{aligned}$$

The function $\ell_\gamma(\mu, \alpha, \beta)$ is concave if $\ell''_\gamma(\mu, \alpha, \beta) < 0$, i.e. when

$$2\beta + 6\gamma\mu < \frac{1}{\mu(1-\mu)} \equiv s(\mu)$$

The function $s(\mu)$ is decreasing in $[0, .5)$ and increasing in $(.5, 1]$, and it has a minimum at $\mu = .5$, where $s(0.5) = 4$.

Let μ_0 be the value of μ at which the line $2\beta + 6\gamma\mu$ is tangent to $s(\mu)$, defined as the solution of

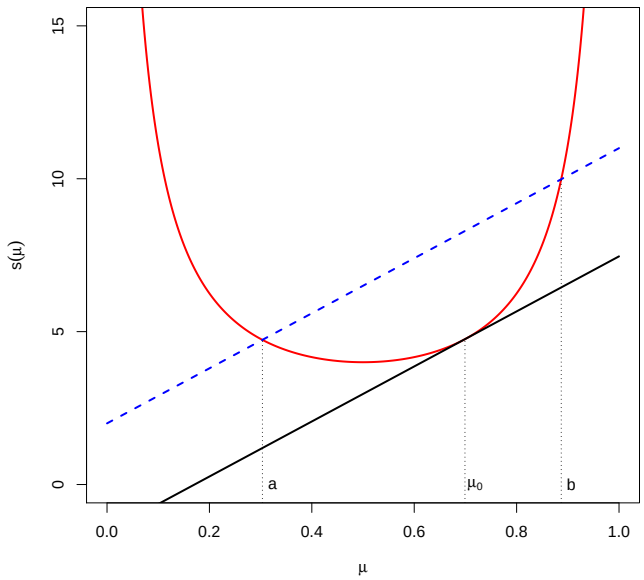
$$6\gamma = \frac{2\mu - 1}{\mu^2(1-\mu)^2}$$

Notice that μ_0 is unique, since the right-hand-side of the equation is a monotone increasing function. Given μ_0, we can find β_0 by solving

$$\beta_0 = \frac{1}{2}\left[-6\gamma\mu_0 + \frac{1}{\mu_0(1-\mu_0)}\right]$$

Therefore the function $\ell_\gamma(\mu, \alpha, \beta)$ is concave on the whole interval $[0, 1]$ if $\beta \leq \beta_0$. In this region, there is a unique maximizer μ^* of $\ell_\gamma(\mu, \alpha, \beta)$.

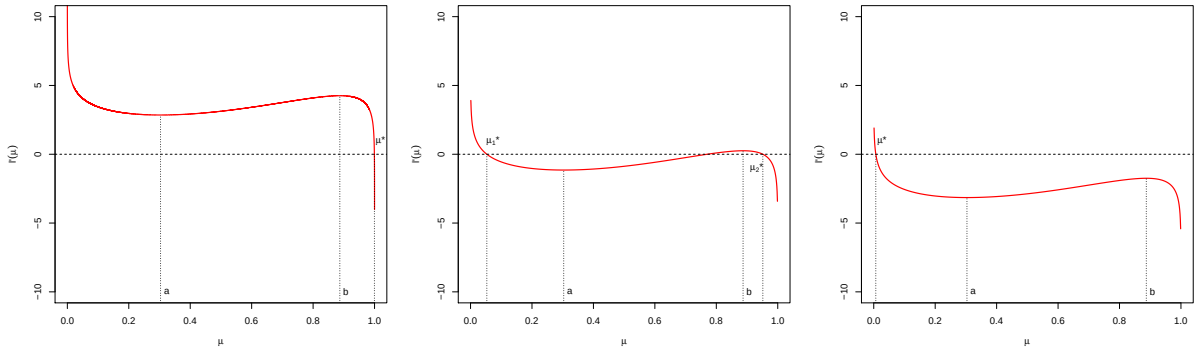
If $\beta > \beta_0$ the line $2\beta + 6\gamma\mu$ has two intersections with $s(\mu)$, and there are three possible cases. We know that in this region the second derivative $\ell''_\gamma(\mu, \alpha, \beta)$ can be positive or negative, with inflection points denoted as a and b, found by solving the equation $2\beta + 6\gamma\mu = s(\mu)$. In the picture below, we plot $s(\mu)$ (in red), the line $2\beta + 6\gamma\mu$ (blue dashed) that define the points a and b, and the tangent line (black solid) that defines μ_0.



By looking at the picture is clear that the first derivative $\ell'_\gamma(\mu, \alpha, \beta)$ is decreasing for $\mu \in [0, a)$, increasing in $\mu \in (a, b)$ and decreasing in $\mu \in (b, 1]$.

1. If $\ell'_\gamma(a, \alpha, \beta) \geq 0$, then there is a unique maximizer $\mu^* > b$
2. If $\ell'_\gamma(b, \alpha, \beta) \leq 0$, then there is a unique maximizer $\mu^* < a$
3. If $\ell'_\gamma(a, \alpha, \beta) < 0 < \ell'_\gamma(b, \alpha, \beta)$, then there are 2 local maximizers $\mu_1^* < a < b < \mu_2^*$

The three cases are shown in the following pictures, where we plot $\ell'_\gamma(\mu, \alpha, \beta)$ against μ for several values of α and for a fixed $\beta = 1$ and $\gamma = 1.5$



We indicate the maximizer with μ^* when it is unique, and with μ_1^*, μ_2^* when there are two.

Let's consider the first case, with $\ell'_\gamma(a, \alpha, \beta) \geq 0$. To compute $\ell'_\gamma(a, \alpha, \beta)$, notice that

$$\beta = \frac{1}{2a(1-a)} - \frac{2\mu_0 - 1}{2\mu_0^2(1-\mu_0)^2}a$$

Substituting in $\ell'_\gamma(a, \alpha, \beta)$ we obtain

$$\begin{aligned} \ell'_\gamma(a, \alpha, \beta) &= \alpha + \frac{a}{a(1-a)} - \frac{2\mu_0 - 1}{\mu_0^2(1-\mu_0)^2}a^2 + \frac{2\mu_0 - 1}{2\mu_0^2(1-\mu_0)^2}a^2 - \log\frac{a}{1-a} \\ &= \alpha + \frac{1}{(1-a)} - \frac{2\mu_0 - 1}{2\mu_0^2(1-\mu_0)^2}a^2 - \log\frac{a}{1-a} \end{aligned}$$

and analogously we have for b

$$\ell'_\gamma(b, \alpha, \beta) = \alpha + \frac{1}{(1-b)} - \frac{2\mu_0 - 1}{2\mu_0^2(1-\mu_0)^2}b^2 - \log\frac{b}{1-b}$$

Notice that we can write $\ell'_\gamma(a, \alpha, \beta) = \alpha + \eta(a)$, where $\eta(a) = \frac{1}{(1-a)} - \frac{2\mu_0-1}{2\mu_0^2(1-\mu_0)^2}a^2 - \log\frac{a}{1-a}$. Consider the derivative of $\eta(a)$

$$\begin{aligned} \eta'(a) &= \frac{1}{(1-a)^2} - \frac{2\mu_0 - 1}{\mu_0^2(1-\mu_0)^2}a - \frac{1}{a(1-a)} \\ &= a\left[\frac{2a-1}{a^2(1-a)^2} - \frac{2\mu_0 - 1}{\mu_0^2(1-\mu_0)^2}\right] \end{aligned}$$

We know that the function $\mathfrak{h}(a) = \frac{2a-1}{a^2(1-a)^2}$ is monotone increasing, with $\mathfrak{h}(0) = -\infty$ and $\mathfrak{h}(1) = \infty$. Therefore the minimum of $\eta(a)$ is found at $a = \mu_0$, where we have

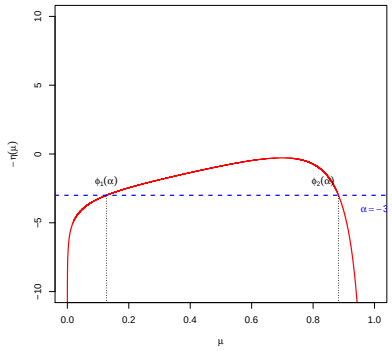
$$\eta(\mu_0) = \frac{1}{(1-\mu_0)} - \frac{2\mu_0 - 1}{2(1-\mu_0)^2} - \log\frac{\mu_0}{1-\mu_0}$$

This means that $\ell'_\gamma(a, \alpha, \beta) \geq 0$ only if

$$\alpha \geq \alpha_0 \quad = \quad -\eta(\mu_0) = \log\frac{\mu_0}{1-\mu_0} - \frac{1}{(1-\mu_0)} + \frac{2\mu_0 - 1}{2(1-\mu_0)^2}$$

When the above condition is satisfied, there is a unique maximizer, $\mu^* > b$, as shown in the picture on the left.

When $\alpha < \alpha_0$ and $\beta > \beta_0$, we have $\ell'_\gamma(a, \alpha, \beta) < 0 < \ell'(b, \alpha, \beta)$. We draw a picture of $-\eta(\mu)$ to help with the reasoning



Notice that when $\alpha < \alpha_0$ there are two intersections of the function and the horizontal line $y = \alpha$ (in the picture $\alpha = -3$). We denote the intersections $\phi_1(\alpha)$ and $\phi_2(\alpha)$. By construction, we know that $a < 0.5 < b$. By looking at the picture, it is clear that $\ell'_\gamma(a, \alpha, \beta) > 0$ if $a < \phi_1(\alpha)$ and $\ell'_\gamma(a, \alpha, \beta) < 0$ if $a > \phi_1(\alpha)$. Analogously, we have $\ell'_\gamma(b, \alpha, \beta) > 0$ if $b > \phi_2(\alpha)$ and $\ell'_\gamma(b, \alpha, \beta) < 0$ if $b < \phi_2(\alpha)$.

For any $\alpha < \alpha_0$, there exist $\phi_1(\alpha)$ and $\phi_2(\alpha)$ which are the intersections of the function $-\eta(\mu)$ with the line α. Since the function is continuous, monotonic increasing in $[0, \mu_0)$ and monotonic decreasing in $(\mu_0, 1]$ it follows that $\phi_1(\alpha)$ and $\phi_2(\alpha)$ are both continuous in α. In addition, $\phi_1(\alpha)$ is increasing in α and $\phi_2(\alpha)$ is decreasing in α. It's trivial to show that when α decreases, $\phi_1(\alpha)$ converges to 0 while $\phi_2(\alpha)$ converges to 1.

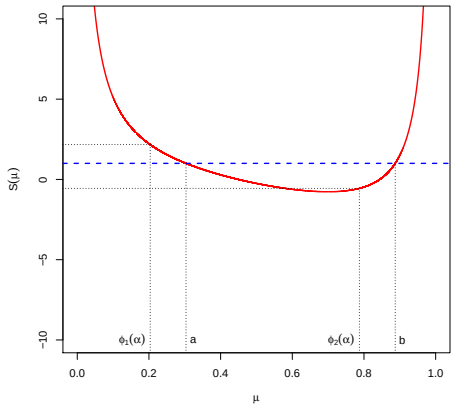
Consider the case in which $\ell'_\gamma(a, \alpha, \beta) < 0 < \ell'_\gamma(b, \alpha, \beta)$ with two maximizers of $\ell_\gamma(\mu, \alpha, \beta)$. Consider the function

$$S(\mu) = \frac{1}{2\mu(1-\mu)} - \frac{2\mu_0 - 1}{2\mu_0^2(1-\mu_0)^2}\mu$$

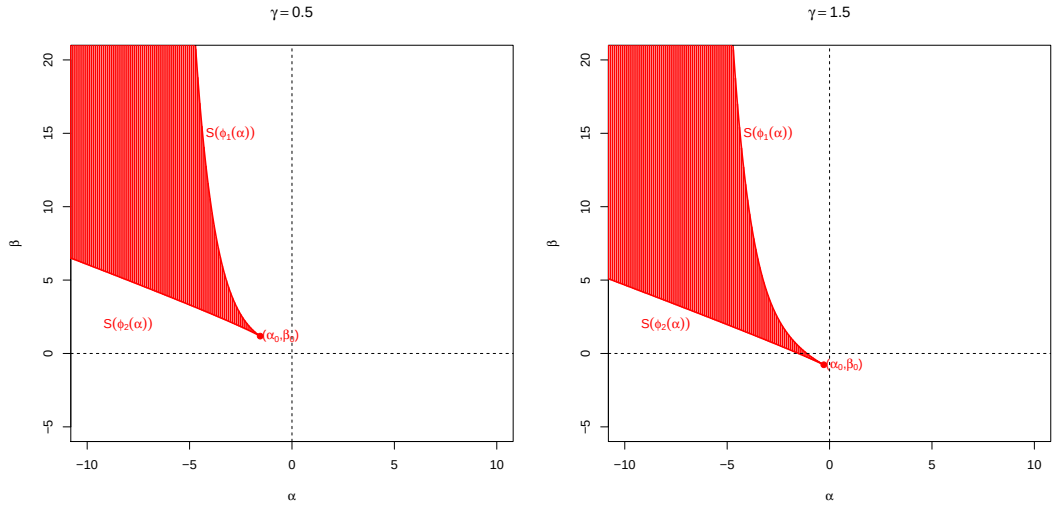
Since $\ell'_\gamma(a, \alpha, \beta) < 0$ we have $a > \phi_1(\alpha)$, which implies $S(a) < S(\phi_1(\alpha))$. Therefore $\beta < S(\phi_1(\alpha)) = \frac{1}{2\phi_1(\alpha)(1-\phi_1(\alpha))} - \frac{2\mu_0-1}{2\mu_0^2(1-\mu_0)^2}\phi_1(\alpha)$.

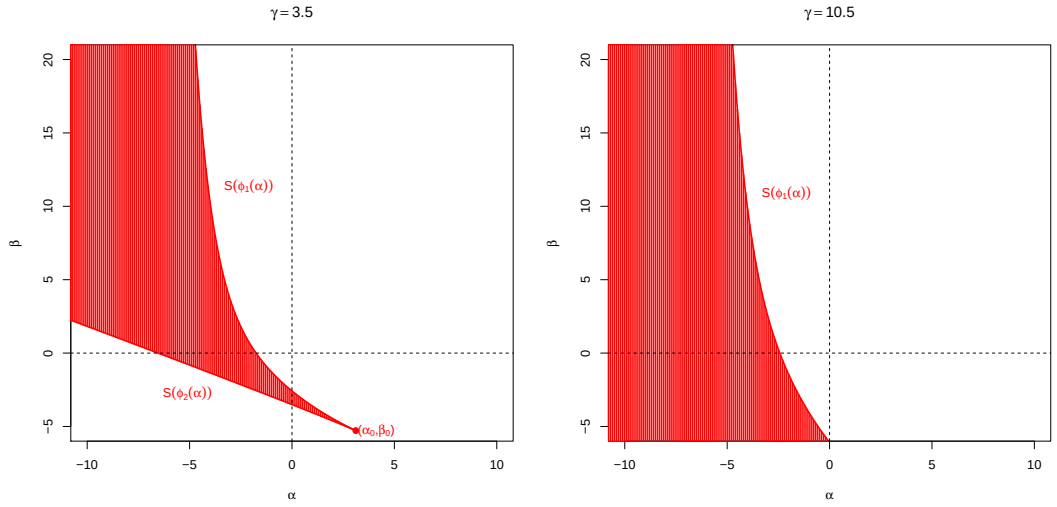
Since $\ell'_\gamma(b, \alpha, \beta) > 0$ we have $b > \phi_2(\alpha)$, which implies $S(b) > S(\phi_2(\alpha))$. Therefore $\beta > S(\phi_2(\alpha)) = \frac{1}{2\phi_2(\alpha)(1-\phi_2(\alpha))} - \frac{2\mu_0-1}{2\mu_0^2(1-\mu_0)^2}\phi_2(\alpha)$.

Notice that $S(\phi_1(\alpha)) > S(\phi_2(\alpha))$ for any (α, β) in this region of the parameters (see picture below for an example with $\beta = 1$, $\alpha = -3$, and $\gamma = 1.5$).



In the following pictures we show the function $S(\phi_1(\alpha))$ and $S(\phi_2(\alpha))$ in the (α, β) space, for a given $\gamma > 0$. Notice that for our models, we are only interested in the part of the graph where $\beta > 0$. The graphs show that when we increase the value of γ the area in which the model has multiple local maxima increases.





The existence of $\zeta_\gamma(\alpha)$ is shown using similar argument as in the proof of Theorem 11, so it is omitted for brevity. ■

The last set of results extends the analysis of sampling algorithms in Bhamidi et al. (2011) to directed graphs. In particular, the solution to the variational problems in the previous theorems provides a characterization for the convergence of the MCMC samplers commonly used to simulate samples of ERGMs from the model. The set of parameters that lie within the V-shaped region, correspond to what Bhamidi et al. (2011) call the *low temperature phase*. The set of parameters lying outside the V-shaped region correspond to the *high temperature phase*.

To be precise, let $\widetilde{M}^* \subset \widetilde{\mathcal{W}}$ be the set of maximizers of the variational problem and let G_n be a graph on n vertices drawn from the ERGM model implied by function $\mathcal{T}$. The next theorem shows that as n grows large, the network $\widetilde{G}_n$ must be close to the set $\widetilde{M}^*$. If the set consists of a single graph, then this is equivalent to a weak law of large numbers for G_n.

THEOREM 18 *Let $\widetilde{M}^*$ be the set of maximizers of the variational problem (67). Let G_n be a graph on n vertices drawn from the model implied by function $\mathcal{T}$. Then for any $\eta > 0$ there exist $C, \kappa > 0$ such that for any n*

$$\mathbb{P}(\delta_\square(\widetilde{G}_n, \widetilde{M}^*) > \eta) \leq Ce^{-n^2\kappa}$$

where $\mathbb{P}$ denotes the probability measure implied bu the model.

Proof. The proof is identical to the proof of Theorem 3.2 in Diaconis and Chatterjee (2011) ■

For the model we analyze in this paper, the result specializes to the following theorem.

THEOREM 19 *Consider the model above in (68) and assume $\theta_2 > 0$. Let G_n be the directed graph implied by the model.*

1. *If the maximization problem in Theorem 11 has a unique solution μ^*, then $G_n \to G_d(n, \mu^*)$ in probability as $n \to \infty$.*

2. *If the maximization problem in Theorem 11 has two solutions $\mu_1^* < \frac{1}{2} < \mu_2^*$, then G_n is drawn from a mixture of directed Erdos-Renyi graphs $G_d(n, \mu_1^*)$ and $G_d(n, \mu_2^*)$, as $n \to \infty$.*

Proof. It is an application of Theorem 18. ■

The previous results consider the limit as $n \to \infty$. However, for fixed n, the speed of convergence of the model to the stationary distribution π_n can be studied using the previous results. The model evolves according to a Glauber dynamics: essentially it behaves like a random Gibbs sampler.

In particular, when the maximization problem in Theorem 11 has a unique solution, the Markov chain of networks converges in an order $n^2 \log n$ steps. However, when the maximization problem in Theorem 11 has two solutions $\mu_1^* < \frac{1}{2} < \mu_2^*$, the convergence is exponentially slow, i.e. there exists a constant $C > 0$ such that the number of steps needed to reach stationarity are $O(e^{Cn})$. This is true for any local chain, i.e. a chain that updates $o(n)$ links per iteration.

The main convergence result that is proven in Bhamidi et al. (2011) is extended to our directed network formation model in the following proposition.

PROPOSITION 4 ***(Convergence rates)*** *Assume $\beta, \gamma > 0$ in any of the models in Theorem 12.*

1. *If the variational problem has a unique solution, we say that the parameters belong to the* high temperature *region. The chain of networks generated by the model mixes in order $n^2 \log n$ steps.*

2. *If the variational problem has two local maxima, we say that the parameters belong to the* low temperature *region. The chain of networks generated by the model mixes in order e^{n^2} steps. This holds for any local dynamics, i.e. a dynamics that updates an $o(n)$ number of links per period.*

Proof. See Bhamidi et al. (2011), Thm. 5 and 6 ■

The main reason for the slow convergence in the bi-modal regime is that a local chain makes small steps. The solution to this problem is to allow the sampler to perform larger steps. However, large steps are not sufficient. Indeed, we need to be able to make large steps of order n: in other words we need a large step whose size is a function of n.

The result of asymptotically independent edges (Theorem 7 in Bhamidi et al. (2011)) is proven above in our Theorem 19.