

The Strength of Local Structures in Decentralized Network Formation

Jose M. Betancourt*

Yale University

November 2025

Abstract

I study dynamic network formation games in which agents meet stochastically and form links based on their valuation of the network. I show that these games can be represented in terms of the values agents assign to network sub-structures. Particularly, this characterizes potential games as those where all participants in a structure value it equally. When valuations are restricted to a finite set of repeated sub-structures, or *motifs*, the model exhibits *phase transitions*: small changes in motif values cause discontinuous shifts in network density.

JEL Classification: C73, D85, C72

Keywords: Random Networks, Graph Limits, Phase Transitions.

*Email: jose.betancourtvalencia@yale.edu.

I am grateful to Ben Golub, Tomás Rodríguez-Barraquer, Simon Boutin, Sanjeev Goyal, Roberto Corrao, Pau Milán, Sekhar Tatikonda, Elliot Lipnowski, Angelo Mele, Doron Ravid, Aniko Öry, Kevin Williams, Annie Chen, Kevin Yin, Laura Tenjo, and Eric Solomon for very informative discussions. I thank seminar participants at the Network Science and Economics Conference for helpful comments and suggestions. Refine.ink was used to proofread the paper for consistency and clarity.

1 Introduction

Economic outcomes often depend on who interacts with whom – whether through trade, information exchange, or collaboration. These structures often set the stage for how agents strategically interact with each other, but they are themselves evolving through time, responding to agents’ incentives to form or sever connections. Understanding the forces that drive these dynamics is a central part of describing any economic system with complex interactions. The main challenge with tackling this problem is the vast number of possible structures through which agents can interact. For example, there are around 10^7 possible networks with 5 agents, while the number of networks with 20 agents exceeds the number of atoms in the universe.

The process of forming a network often involves a combination of an element of randomness and strategic decisions. For example, one might randomly meet a friend of a friend at a social gathering, but the decision of whether or not to develop the relationship after this initial meeting is a conscious choice. Such interactions and choices happen constantly, making networks highly dynamic and fluid objects. A model that incorporates random meetings and strategic decisions gives rise to a time-dependent probability distribution over the space of networks. The goal of this paper is to study how the properties of this distribution depend on the incentives of agents to form the network.¹

To study the properties of dynamic network formation processes, I begin by characterizing some general properties of static network formation games. I show that any network formation game, where agents assign utilities to networks, has an equivalent representation where agents derive value from the sub-structures that are realized in the network. This means that the choice of forming/severing a link depends on the values of the structures that get created/destroyed in the process. This gives a simple economic interpretation to the values of structures: they capture the incentives of agents to deviate by creating or destroying these structures.

The interpretation that structure values capture the strategic incentives of players has an implication that greatly simplifies the analysis of the game. We say that an agent participates

¹Throughout the paper I focus on directed networks. Most of the results have counterparts in the corresponding undirected case, but require additional structure like transferable utilities.

in a structure if they have an outgoing link in it. I show that all agents that participate in a structure give it the same value if and only if the game is an exact potential game, meaning that there is a single function that captures the incentives of players to deviate. This is the first technical contribution of the paper. By focusing on potential games, we can characterize the game by specifying a single value to each structure.

Having established that static games have a structure value representation, I study the implications of this for a dynamic network formation game. Agents meet stochastically and decide whether to form or sever links based on deterministic utilities, a choice that is subject to idiosyncratic shocks. The second technical contribution of the paper is to show that this Markov process is reversible if and only if the static game with the deterministic utilities is a potential game. In this case, the stationary probability of the process can be computed explicitly: the probability of observing a network is proportional to the exponential of its potential. This is called a Gibbs measure, and its structure is ideal for studying how the model primitives affect the long-run properties of the network formation process.

To study the properties of large networks, it is useful to have a scheme that allows the size of the network to be scaled up. Motivated by the fact that potential games have a single value assigned to each structure, I study a class of potential games characterized by *motifs*. A motif is a fixed structure, such that every time the structure is realized in the network (allowing re-labeling of the nodes), agents that participate in it get the same value. By specifying a finite set of motifs, we can study the dynamics of the model for an arbitrary number of players. This captures scenarios in which incentives do not depend on the identity of the agents who interact, but only on the structure of interaction. An example of such a scenario is having a fixed cost of forming links and a payoff when this link gets reciprocated.

The next technical contribution of the paper is to show that the model with motif utilities becomes equivalent (in an appropriate graph limit sense) to an Erdős–Rényi model, where the parameter of the model is determined by an optimization process. The result follows from the literature on large deviations theory applied to large dense graphs ([Chatterjee and Varadhan, 2011](#); [Chatterjee and Diaconis, 2013](#)). This optimization problem corresponds to a trade-off between agents’ utilities and a quantity called *entropy*, which takes into account the exponentially increasing size of network space. Introducing multiple types of agents

generalizes this result, such that the process instead converges to a stochastic block model. In this case, the linking probabilities are determined by an analogous optimization process.

The nature of the optimization process allows for the emergence of discontinuities in the asymptotic properties of the process, which I refer to as *phase transitions*. This is the main conceptual result in the paper. Since the parameter in the equivalent Erdős–Rényi model is the maximizer of some objective function, even a continuous change in the values of motifs can lead to a discontinuous change in this parameter. At these transitions, the network discretely changes from a high-density phase to a low-density phase (and vice versa) as parameters are changed.

To illustrate how these phase transitions might arise naturally in an economic setting, I study a simple model of trade, where firms stochastically meet and decide whether to form trade relationships. Firms must incur some distance-dependent cost to state the intent to trade with another firm, and they obtain gains from trade when there is a mutual intent to form a partnership. If the gains from trade are changed continuously (with a tax, for example), a phase transition can be induced, breaking the trading network apart. More complex trading structures lead to diverse manifestations of phase transitions, illustrating the richness of this framework.

Together with the results on the static network formation game, the results on phase transitions shed light on the very complex process of network formation. The low-dimensional behavior of the asymptotics allows for potential applications of this framework to more complex phenomena where networked interactions are important, such as international trade or inter-bank lending. These results showcase novel phenomena that are relevant for the dynamics of and interventions on complex network structures.

1.1 Related Literature

This paper contributes to the broad literature on the analysis of endogenous network formation models. Early work focused on the properties of deterministic network formation games with finite players ([Jackson and Wolinsky, 1996](#); [Bala and Goyal, 2000](#)). The dynamics of network formation have been analyzed in similar models, motivated in their own right and as a selection device for the large number of equilibria that can arise in static models ([Bala](#)

and Goyal, 2000; Jackson and Watts, 2002; Currarini et al., 2009).² More recent work has analyzed the behavior of network formation games with large numbers of players, which can exhibit properties such as tipping points (Golub and Livne, 2010; Mele, 2017, 2022; Elliott et al., 2022).

Another branch of the literature on network formation includes actions beyond link formation/deletion, such as Hsieh et al. (2022); Sadler and Golub (2021). Badev (2021) considers simultaneous changes to the state of multiple network links. This paper differs from these in that more general utility functions for network formation are considered, but only in settings with single link evaluations and with no additional actions. For the case of forward-looking agents, Dutta et al. (2005) establishes equilibrium existence results in a more general setting, but lacks a tractable characterization of the resulting probabilities over the space of networks.

A special class of models that is of interest in econometric estimation are Exponential Random Graph Models (ERGMs)³. In these models, the probability associated with a network g is proportional to $\exp(\Phi(g))$, where $\Phi(g)$ is a linear combination of *sufficient statistics* of the network. The advantage of analyzing these models is twofold. First, it has been shown that they arise naturally from strategic network formation models under some regularity assumptions, which would otherwise be intractable (Butts, 2009; Mele, 2017; Christakis et al., 2020; Chandrasekhar and Jackson, 2025). The second advantage of ERGMs is that they provide a natural tool to analyze large graphs, provided that the sufficient statistics scale properly as the number of agents grows (Aristoff and Zhu, 2018; Mele, 2017, 2022). This paper contributes to the literature on ERGMs by characterizing a model of network formation that provides a microfoundation for distributions with arbitrary functions $\Phi(g)$, and hence any ERGM. The model can then be evaluated to see if the primitives satisfy desirable properties, which would yield criteria to evaluate the use of a given ERGM.

Finally, this paper contributes to the literature on graph limits by providing a characterization of the limiting properties of large networks in terms of model primitives. In the case of ERGMs, the model becomes intractable when this number is large but finite⁴. However,

²See Jackson (2005, 2011) for more complete surveys of network formation models and their properties.

³See Robins et al. (2007) for an overview of ERGMs

⁴Previous work on tackling the intractability of these models includes Boucher and Mourifié (2017), which

the distributions can be characterized in the limit of a large number of players. Results from the theory of graph limits (Chatterjee and Varadhan, 2011; Chatterjee and Diaconis, 2013; Mele, 2017) allow for a characterization of the limiting behavior of the process. This characterization makes it evident that the model can exhibit phase transitions⁵, a phenomenon that has been observed in the context of social network formation (Golub and Livne, 2010) and supply chain formation (Elliott et al., 2022).

2 A Static Network Formation Game

In this section I present a static model of network formation that will serve as a benchmark for the dynamic game developed in Section 3. I characterize the important properties of the structure of the static game that allow for a tractable study of the dynamic game.

2.1 Strategies and payoffs

I consider a game with $N \geq 2$ agents, where $\mathcal{J}_N := \{1, \dots, N\}$ denotes the set of players. In this game, agents choose who to connect to, and their payoffs are determined by the resulting structure of connections.

An agent's *strategy* is the subset of $\mathcal{J}_N$ that she chooses to connect to.⁶ I assume that agents cannot form connections with themselves, so agent i 's action set is $\mathcal{S}_{i,N} := 2^{\mathcal{J}_N \setminus \{i\}}$. A typical strategy is denoted with s_i .

A realization of all players' strategies can be characterized as a *network*. I use the notation ij to refer to the tuple (i, j) . A (directed) network g is the set of all dyads ij such that $j \in s_i$. Since I do not consider self-interactions, the set of possible dyads is $\mathcal{D}_N := \{ij \in \mathcal{J}_N^2 \mid i \neq j\}$. The set of all networks is, then, $\mathcal{G}_N := 2^{\mathcal{D}_N}$. For a given network g , g_{-i} denotes the subnetwork for which all links with i as the source are removed, such that

uses Markov Random Fields to reduce the complexity in estimation, and Graham (2017), which proposes consistent estimators of structural parameters that take into account degree heterogeneity. Mele (2017) and Mele (2022) use an exchange Markov chain Monte Carlo method to estimate parameters through maximum likelihood in ERGMs.

⁵Phase transitions are a central topic of analysis in statistical physics. See, for example, Holyst et al. (2000); Dorogovtsev et al. (2008); Squartini et al. (2015); Cimini et al. (2019) for applications of statistical physics to stochastic network analysis.

⁶I focus on pure strategies throughout the paper.

g_{-i} captures the strategies of all other players. Abusing notation, I often write the network as $g = (s_i, g_{-i})$ or $g = (s_1, \dots, s_N)$.

Since the strategies of all agents can be represented as a network, agents' payoffs are functions of the realization of the network. Specifically, let $U : \mathcal{J}_N \times \mathcal{G}_N \rightarrow \mathbb{R}$ be the payoff function, such that $U_i(g)$ is the payoff of agent i under network g . I also refer to $U_i(g)$ as the utility that agent i associates to g .

2.2 Potential games

The network formation game described above is fairly general, admitting $|\mathcal{G}_N| = 2^{N(N-1)}$ distinct action profiles. Because of its complexity, studying properties of the general game (such as its equilibria) becomes intractable for large N . In this section I characterize a special subset of games, called *potential games* (Monderer and Shapley, 1996), which have a structure that allows for a more tractable study of their properties. I also state necessary and sufficient conditions for the network formation game to be a potential game.

Definition 1. The network formation game $(\mathcal{J}_N, \mathcal{G}_N, U)$ is a potential game if there exists a function $\Phi : \mathcal{G}_N \rightarrow \mathbb{R}$, called the *potential*, such that

$$U_i(s_i, g_{-i}) - U_i(s'_i, g_{-i}) = \Phi(s_i, g_{-i}) - \Phi(s'_i, g_{-i}) \quad (1)$$

for all $i \in \mathcal{J}_N$, $g \in \mathcal{G}_N$ and $s_i, s'_i \in \mathcal{S}_{i,N}$.⁷

Potential games impose additional structure on the network formation game. This allows us to aggregate the incentives for unilateral deviation into a single function, as opposed to study each individual's payoff function. The existence of a potential yields some nice properties of the Nash equilibria of the game.

Remark 1. If the network formation game is a potential game, then all Nash equilibria are local maxima of the potential. That is, a network $g = (s_1, \dots, s_N)$ is a Nash equilibrium if and only if $\Phi(s_i, g_{-i}) \geq \Phi(s'_i, g_{-i})$ for all i and all s'_i . Additionally, since $\mathcal{G}_N$ is finite, there

⁷More specifically, this is a *cardinal* potential game. In an *ordinal* potential game only the signs of these differences would have to match.

exists a global maximum of Φ , so a Nash equilibrium in pure strategies always exists for potential games.

Given the added simplicity in studying properties of potential games, it is useful to have conditions under which an arbitrary network formation game is a potential game. This is difficult for general games, but the structure of a network formation game allows for a simplified characterization. In order to provide this characterization, it is useful to define a single-link “switching” function. Consider the function $\tau : \mathcal{D}_N \times \mathcal{G}_N \rightarrow \mathcal{G}_N$ defined by

$$\tau_{ij}(g) = \begin{cases} g \cup \{ij\} & \text{if } ij \notin g \\ g \setminus \{ij\} & \text{if } ij \in g. \end{cases} \quad (2)$$

Intuitively, the operation τ_{ij} creates the link ij if it is not present in a network and severs it if it is. These switching operations are useful to characterize single-link deviations in strategies, rather than an arbitrary change in the composition of an agent’s connections. With these single-link deviations, we can define the following condition on the payoff functions.

Definition 2. A set of payoff functions U is said to be *conservative* if the following condition on marginal utilities (MUs) holds:

$$\begin{aligned} & \underbrace{U_i(\tau_{ij}(g)) - U_i(g)}_{\text{MU of changing link } ij \text{ in network } g} + \underbrace{U_{i'}(\tau_{i'j'} \circ \tau_{ij}(g)) - U_{i'}(\tau_{ij}(g))}_{\text{MU of changing link } i'j' \text{ in network } \tau_{ij}(g)} \\ &= \underbrace{U_{i'}(\tau_{i'j'}(g)) - U_{i'}(g)}_{\text{MU of changing link } i'j' \text{ in network } g} + \underbrace{U_i(\tau_{ij} \circ \tau_{i'j'}(g)) - U_i(\tau_{i'j'}(g))}_{\text{MU of changing link } ij \text{ in network } \tau_{i'j'}(g)} \end{aligned} \quad (3)$$

for all $g \in \mathcal{G}_N$ and all $ij, i'j' \in \mathcal{D}_N$. That is, the sum of the marginal utilities of changing the state of links ij and $i'j'$ is independent of the order in which the changes are made.

Utility functions satisfying the property in Definition 2 are called conservative in analogy to conservative vector fields.⁸ The condition in Equation (3) can be thought of as the condition for the Poincaré Lemma in Footnote 8, which allows for the characterization of

⁸A C^1 vector field $\mathbf{F} : U \rightarrow \mathbb{R}^n$, with $U \subset \mathbb{R}^n$ open and simply connected, is said to be conservative if there exists a function ϕ such that $\mathbf{F}(\mathbf{x}) = \nabla\phi(\mathbf{x})$. According to the Poincaré Lemma, if $\forall i, j \in \{1, \dots, n\}$, $\frac{\partial F_i(\mathbf{x})}{\partial x_j} = \frac{\partial F_j(\mathbf{x})}{\partial x_i}$, then $\mathbf{F}$ is conservative (Warner, 1983).

a vector field as the gradient of a scalar field. That is, the information contained in a vector-valued function can be extracted from a scalar-valued function, effectively reducing the dimensionality of the problem. Proposition 1 connects this intuition to our network formation game.

Proposition 1. *The network formation game is a potential game if and only if the payoff functions are conservative.*

The equivalence between conservativeness and the existence of a potential is analogous to Theorem 2.8 in [Monderer and Shapley \(1996\)](#), but it is greatly simplified by using the network structure of the game. Since conservativeness is a necessary and sufficient condition for the game to be a potential game, it suffices to check conservativeness to evaluate whether a potential exists. We can go even further, characterizing the entire class of potential games. In order to do this, I use an alternate representation of the incentive structure of the game.

2.3 The value of group structures

To begin the characterization of the class of potential games, I use the following technical lemma.

Lemma 1. *For a network formation game $(\mathcal{J}_N, \mathcal{G}_N, U)$, there exists a unique function $V : \mathcal{J}_N \times \mathcal{G}_N \rightarrow \mathbb{R}$ such that*

$$U_i(g) = \sum_{g' \subseteq g} V_i(g') \quad (4)$$

for all $i \in \mathcal{J}_N$ and $g \in \mathcal{G}_N$. Additionally, this relation can be inverted as follows:

$$V_i(g) = \sum_{g' \subseteq g} (-1)^{|g \setminus g'|} U_i(g'). \quad (5)$$

I refer to the quantity $V_i(g)$ as the *value* that i assigns to the *structure* g . Despite their similarity to the payoff functions U , these objects are conceptually very different. To see this, let us consider the change in utility from adding a single link: $U_i(\tau_{ij}(g)) - U_i(g)$, with $ij \notin g$. This marginal utility only depends on the value $V_i(g')$ of structures g' that contain

ij and other links already present in g . A similar intuition holds for an arbitrary change in strategy, where the utility of deviating depends only on the value of the structures that are created or destroyed.

The representation of marginal utilities as the sum of values of structures that are created or destroyed gives a very nice representation of conservativeness. Let $\mathcal{J}_{\text{src}}(g) := \{i \mid \exists j : ij \in g\}$ be the set of *source nodes* of g . I also say that agent i *participates* in the structure g if $i \in \mathcal{J}_{\text{src}}(g)$. The following Proposition gives necessary and sufficient conditions for the game to be a potential game in terms of the values of structures.

Proposition 2. *The network formation game $(\mathcal{J}_N, \mathcal{G}_N, U)$ is a potential game if and only if*

$$V_i(g) = V_j(g) =: V_0(g) \quad \forall i, j \in \mathcal{J}_{\text{src}}(g). \quad (6)$$

Furthermore, the potential is given by

$$\Phi(g) = \sum_{g' \subseteq g} V_0(g'). \quad (7)$$

Proposition 2 states that the network formation game is a potential game if and only if all structures give the same value to all agents who are source nodes in the structure. This means that the path-independence property in Proposition 1 takes the form of a symmetry in the values of structures. But, what is special about the source nodes of the structures? Why doesn't this condition apply to all agents? The intuition behind this lies in a property called *choice-equivalence*.

Definition 3. Two sets of utility functions U and $\tilde{U}$ are *choice-equivalent* if

$$U_i(s_i, g_{-i}) - U_i(s'_i, g_{-i}) = \tilde{U}_i(s_i, g_{-i}) - \tilde{U}_i(s'_i, g_{-i}) \quad (8)$$

for all $g \in \mathcal{G}_N$, $i \in \mathcal{J}_N$ and $s_i, s'_i \in \mathcal{S}_{i,N}$.

If two sets of utility functions are choice-equivalent, then agents have the same incentives to change strategies under both of them. This means that they share important properties, such as their Nash equilibria and their potential (if the game is a potential game). Choice-

equivalence manifests itself in a clear way when utilities are represented in terms of structure values, as shown in the following Lemma.

Lemma 2. *Two sets of utility functions U and $\tilde{U}$, with corresponding structure values V and $\tilde{V}$ are choice-equivalent if and only if*

$$V_i(g) = \tilde{V}_i(g) \quad \forall i \in \mathcal{J}_{src}(g) \quad (9)$$

for all $g \in \mathcal{G}_N$.

Lemma 2 sheds light on what is special about source nodes. Take some structure g and some $i \notin \mathcal{J}_{src}(g)$. Then the value $V_i(g)$ can be changed arbitrarily and the resulting utilities will be choice-equivalent to the original ones. This means that for a given agent i , the only structures that matter for choice-equivalence (and for the game to be a potential game) are those for which it is a source node.

Example 1. To build an intuition for the results above, I now present a simple model of trade, which I generalize in Section 5. Consider N firms which can choose to form trade links with each other. Forming a trade link costs $c > 0$, and if two firms form links with each other, both of them get some value $v > 0$. Explicitly, firm i 's utility function is

$$U_i(g) = v \underbrace{\sum_{j \in \mathcal{J}_N} \mathbb{1}\{ij \in g, ji \in g\}}_{\text{reciprocated links}} - c \underbrace{\sum_{j \in \mathcal{J}_N} \mathbb{1}\{ij \in g\}}_{\text{outgoing links}}. \quad (10)$$

In this example, there are two types of structures to which agents assign value: outgoing links and reciprocated pairs of links. Since both firms get the same value from trade, the conditions in Proposition 2 hold, and the potential is

$$\Phi(g) = v \frac{1}{2} \underbrace{\sum_{i,j \in \mathcal{J}_N} \mathbb{1}\{ij \in g, ji \in g\}}_{\text{total reciprocated pairs}} - c \underbrace{\sum_{i,j \in \mathcal{J}_N} \mathbb{1}\{ij \in g\}}_{\text{total outgoing links}} \quad (11)$$

The factor of $1/2$ ensures that we are counting the number of appearances of a reciprocated pair, instead of all the links that are being reciprocated. $\diamond$

With this analysis, we can fully specify all utility functions for which the game is a potential game.

Remark 2. Suppose the network formation game $(\mathcal{J}_N, \mathcal{G}_N, U)$ is a potential game with potential Φ . Define the structure values

$$V_0(g) := \sum_{g' \subseteq g} (-1)^{|g \setminus g'|} \Phi(g'). \quad (12)$$

Then U is choice-equivalent to the utility functions

$$\tilde{U}_i(g) := \sum_{g' \subseteq g} V_0(g') \mathbb{1}\{i \in \mathcal{J}_{\text{src}}(g')\}. \quad (13)$$

Therefore, all utilities corresponding to potential games can be characterized, up to choice-equivalence, in terms of the potential.

Having characterized the structure of potential games in the static network formation game, I now formulate a model of dynamic network formation. The structure of the static game will be useful to characterize the properties of the dynamic game.

3 Dynamic Network Formation

In this section I study a dynamic process of network formation. In this model, agents meet stochastically and make choices on whether to change the state of their connections. Throughout this section, the set of agents is $\mathcal{J}_N$ and the space of networks is $\mathcal{G}_N$.

3.1 Stochastic meeting of agents

The network formation process occurs in continuous time. The state of the network at time t is $G_t \in \mathcal{G}_N$. Agents meet stochastically and, conditional on meeting, decide whether they want to change the state of the network. If the network state is g , agent i meets agent j at an exogenous Poisson rate $\lambda_{ij}(g)$, that can potentially depend on g . I impose the following restriction on the meeting rates:

Assumption 1. *The meeting rates satisfy $\lambda_{ij}(g) = \lambda_{ij}(\tau_{ij}(g))$ for all dyads ij and all networks g .*

Assumption 1 means that the meeting rate for the dyad ij is independent of whether it is present in the network, but can depend on the rest of the network. This is useful to obtain a tractable characterization of the stochastic dynamics.

Conditional on meeting j , agent i makes a decision to keep or change the state of their relationship through a discrete choice rule, parametrized by a set of utility functions $U : \mathcal{J}_N \times \mathcal{G}_N \rightarrow \mathbb{R}$. Agent i has a baseline utility $U_i(g)$ associated with network g , and there are idiosyncratic shocks to this utility every time a choice is made. After meeting at time t , agent i chooses to change the state of link ij if and only if

$$(1 - \sigma)U_i(\tau_{ij}(G_t)) + \sigma\varepsilon_t^0 \geq (1 - \sigma)U_i(G_t) + \sigma\varepsilon_t^1. \quad (14)$$

where the shocks ε_t^k are i.i.d. (across k and across time) drawn from some distribution F_0 and $\sigma \in (0, 1)$ modulates the importance of noise in the choice process. As $\sigma \rightarrow 0$, the process becomes deterministic and the highest payoff option is chosen. As $\sigma \rightarrow 1$, the process becomes fully random and the state of each link is changed with probability 1/2. Let F_1 be the CDF of $\varepsilon_t^1 - \varepsilon_t^0$, such that the probability of the network changing from g to $\tau_{ij}(g)$, denoted by $p_{ij}(g)$, is given by

$$p_{ij}(g) = 1 - F_1 \left[\left(\frac{1 - \sigma}{\sigma} \right) (U_i(\tau_{ij}(g)) - U_i(g)) \right]. \quad (15)$$

With this characterization, we can now analyze how the dynamics of the model depend on the structure of the utility functions.

3.2 Dynamics and convergence

I assume at time $t = 0$ there is some initial probability distribution $\pi_0 \in \Delta(\mathcal{G}_N)$ over the space of networks. Let $\pi_t(g)$ be the probability of having $G_t = g$. The measure π_t evolves

according to the Kolmogorov forward equation:

$$\dot{\pi}_t(g) = \sum_{ij \in \mathcal{D}_N} \lambda_{ij}(g) [p_{ij}(\tau_{ij}(g))\pi_t(\tau_{ij}(g)) - p_{ij}(g)\pi_t(g)]. \quad (16)$$

Note that, as long as the distribution of utility shocks F_0 has full support, the switching probabilities $p_{ij}(g)$ are non-degenerate. Therefore, in this case, the network formation process corresponds to an aperiodic irreducible Markov process on $\mathcal{G}_N$. It follows from standard results in stochastic processes (e.g. Theorem 1.19 in [Durrett \(2016\)](#)) that there is a unique distribution π such that $\pi_t(g) \rightarrow \pi(g)$ for all $g \in \mathcal{G}_N$, regardless of the initial conditions. I now include additional structure in the stochasticity of choices, which allows tractability of the analysis to follow.

Assumption 2. *The shocks ε_t^k follow a type 1 extreme value (T1EV) distribution.*

Assumption 2 ensures that shocks have full support, which guarantees convergence of the probability distribution. Additionally, it implies that the switching probabilities $p_{ij}(g)$ are a logistic function of the marginal utility of switching $U_i(\tau_{ij}(g)) - U_i(g)$. Equivalently, they satisfy

$$\log \left(\frac{p_{ij}(g)}{p_{ij}(\tau_{ij}(g))} \right) = \left(\frac{1 - \sigma}{\sigma} \right) (U_i(\tau_{ij}(g)) - U_i(g)). \quad (17)$$

We see that there is an intuitive connection between the properties of the Markov chain (through the transition probabilities) and the utility functions of the agents. Given this connection, if the static game with utility functions U has some nice structure, we would expect the dynamic game to inherit some of this structure. This is formalized in the following Proposition for the case of potential games.

Proposition 3. *Consider the static game $(\mathcal{J}_N, \mathcal{G}_N, U)$. This game is a potential game if and only if the Markov chain of the dynamic process is reversible. Furthermore, if the potential of the static game is Φ , then the stationary distribution of the process is given by the Gibbs*

measure

$$\pi(g) = \frac{\exp\left[\left(\frac{1-\sigma}{\sigma}\right) \Phi(g)\right]}{\sum_{g' \in \mathcal{G}_N} \exp\left[\left(\frac{1-\sigma}{\sigma}\right) \Phi(g')\right]}. \quad (18)$$

Proposition 3 states that the same structure that allows for a tractable study of incentives in potential games allows us to fully characterize the stationary distribution in the dynamic game. This is because reversibility is mathematically equivalent to conservativeness.⁹ Reversibility allows us to explicitly construct the stationary distribution in terms of a function that encodes the transition rates of the stochastic process.¹⁰ Interestingly, it is precisely the potential function of the static game that determines the stationary distribution of the network formation process. Throughout the rest of the paper I exploit this form of the stationary distribution to study the properties of the dynamic game.

This characterization of the stationary distribution is related to the family of exponential random graph models (ERGMs). These assign a probability to a network of the form

$$\pi_{\text{ERGM}}(g) = \frac{\exp(\boldsymbol{\beta} \cdot \mathbf{S}(g))}{\sum_{g' \in \mathcal{G}} \exp(\boldsymbol{\beta} \cdot \mathbf{S}(g'))}, \quad (19)$$

where $\mathbf{S}(g)$ is a vector of sufficient statistics of the network and $\boldsymbol{\beta}$ is a vector of model parameters. Therefore, we see that if the potential of our network formation game is of the form $\Phi(g) = \boldsymbol{\beta} \cdot \mathbf{S}(g)$, the stationary distribution will be the distribution of an ERGM with sufficient statistics $\mathbf{S}$. This means that this myopic network formation model serves as a microfoundation for ERGMs, conditional on finding suitable utility functions¹¹. One property of ERGMs that is shared by the myopic network formation game is that the denominator in the Gibbs measure involves a sum over all possible networks, making its calculation intractable. This intractability causes problems for estimation of ERGMs (see [Chandrasekhar](#)

⁹Reversibility of a Markov jump process on a finite space $\mathcal{X}$ is usually stated in terms of the detailed balance condition: $W(i, j)\pi(i) = W(j, i)\pi(j)$ for all $i, j \in \mathcal{X}$, where $W(\cdot)$ are the transition rates and $\pi(\cdot)$ are the stationary probabilities. An equivalent characterization solely in terms of rates is Kolmogorov's criterion, which is analogous to conservativeness ([Norris, 1997](#)).

¹⁰In the theory of Markov processes, this function is also called a potential, or energy, function. For example, similar dynamics can be used to sample the Boltzmann distribution of the Ising model ([Glauber, 1963](#)).

¹¹[Mele \(2017\)](#) provides a microfoundation for a class of ERGMs with this model. [Chandrasekhar and Jackson \(2025\)](#) generalize this analysis to a wider class of utility functions.

and Jackson (2025) for a detailed description of the problems with estimating ERGMs and potential solutions). However, as I discuss in Section 4, for a certain class of models where the number of agents can be systematically increased, characterizing the large N behavior of this denominator allows for even further dimensionality reduction of the model.

3.3 Equilibrium selection

It is interesting to see how properties of the static game manifest in the dynamics of network formation. For the case of choice-equivalent utilities, we have the following:

Remark 3. If two sets of utility functions U and $\tilde{U}$ are choice-equivalent, then the switching probabilities $p_{ij}(g)$ are the same under U and $\tilde{U}$ for all dyads ij and networks g .

This means that the dynamics under two sets of choice-equivalent utility functions are *identical*. This helps to strengthen the intuition for why they must have the same potential, since they must have the same stationary distribution.

We can also analyze what networks get “selected” by the dynamic process. As a benchmark, we can analyze the case of $\sigma = 0$, for which the dynamics have no idiosyncratic shocks. As shown in Proposition 2 of Mele (2017), Nash equilibria are absorbing states of these dynamics, and the network converges to one of the Nash equilibria with probability 1. Therefore, the process with no shocks selects the Nash equilibria of the game.

Our characterization of the dynamics allows us to go beyond the no-shocks case for potential games. Previous work has used dynamic models to select equilibria which are robust to specific kinds of perturbations.¹² In our case, we can check which equilibria are present for small but non-zero σ .

Remark 4. As $\sigma \rightarrow 0^+$, the stationary distribution of the process converges to the uniform distribution across the potential-maximizing Nash equilibria (PMNE) of the game. A consequence of this is that if $g \notin \arg\max_{g'} \Phi(g')$, then $\pi(g) \rightarrow 0$.

This behavior of the stationary distribution *refines* the set of Nash equilibria of the game. This is because the set of PMNE of the game are robust (in the sense of having a

¹²See, for example, Bala and Goyal (2000) and Jackson and Watts (2002).

non-vanishing stationary probability) to small shocks in the choice process. Additionally, we can get some intuition for the properties of the dynamic process by analyzing the PMNE of the game.

Example 1 (Continued). Let us analyze the properties of the Nash equilibria and PMNE in our simple static trade example. For a given set of incoming links, an agent will want to either reciprocate all of them or not form any outgoing links, based on the value of v relative to c . This means that the Nash equilibria of the game are:

- the empty network if $v < c$,
- all networks g where $ij \in g \iff ji \in g$ if $v \geq c$.

To obtain the set of PMNE, we can use the potential computed in Equation (11). The PMNE, then, are

- the empty network if $v < 2c$,
- the complete network if $v > 2c$,
- all networks g where $ij \in g \iff ji \in g$ if $v = 2c$.

We can see that the set of PMNE drastically changes as v changes from being less than $2c$ to being larger. This intuition behind the drastic change in the PMNE is related to the drastic changes we observe for large networks in the next section. $\diamond$

4 Large Dense Networks

The results in the previous sections gave conditions under which our network formation process can be characterized in terms of a potential, and established some properties of the structure of such processes. This allowed us to identify the most probable networks with PMNE of the deterministic game. In this section, I study processes generated by a particular class of potential games as N grows large, revealing interesting typical behaviors that arise from the interplay of incentives and the exponentially increasing size of network space.

4.1 Network motifs

As illustrated by Proposition 2, conservativeness of utilities is tightly related to the value of social structures. We can study the large- N behavior of the model for games where there are a few structure *types* from which people derive utility. For example, in Example 1, firms derive value from outgoing links or reciprocated pairs. We can generalize this by fixing a few structure types and their values, and having agents derive this value every time they participate in one of these structures. Following the graph theory literature, I call these structures *motifs*. This characterization is particularly amenable to our analysis, since it creates a clean connection between the large N asymptotics of the networks and the model primitives.

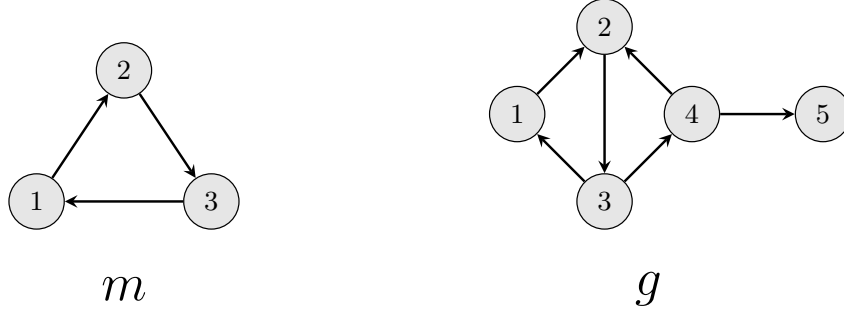
Let us begin by fixing the number of agents N . Formally, a motif m is a network over a set $\mathcal{J}_m = \{1, \dots, n_m\}$ of n_m nodes that has $e_m := |m|$ edges. We say that there is an instance of motif m in network g if there exists an injective map $\varphi : \mathcal{J}_m \rightarrow \mathcal{J}_N$ such that $\varphi(m) \subseteq g$, where $\varphi(m) := \{\varphi(i)\varphi(j) : ij \in m\}$ is the subnetwork induced by φ .

I will focus on dense networks, where the density of the network remains bounded away from 0 almost surely as $N \rightarrow \infty$.¹³ In order to obtain the correct scaling, I assume that agents derive value a_m/N^{n_m-2} from each realization of motif m that they are a source node of. If the value scaled faster, we would get a degenerate behavior in the asymptotics: either the network density would converge to 0 if $a_m < 0$ or to 1 if $a_m > 0$. If the value scaled slower, then the motif would not be relevant enough in influencing the behavior of agents, and the network density would converge to 1/2, as if the linking choices were uniformly random.

To obtain the total value that an agent i derives from a motif m , we need to calculate how many instances of the motif she participates in. For this, it is useful to count these instances and relate this count to the case of the complete network. There are N^{n_m} maps $\varphi : \mathcal{J}_m \rightarrow \mathcal{J}_N$, and the fraction of these that are injective converges to 1 as $N \rightarrow \infty$. Now consider the complete network. Every injective map results in a realization of m so, due to symmetry, agent i will participate in a fraction $|\mathcal{J}_{\text{src}}(m)|/N$ of the realizations of m in the

¹³This is because of the large body of work on the asymptotics of dense networks (Lovász, 2012). It is definitely of interest to study how similar models behave for sparse networks.

Figure 1: Motif counting



Notes. This figure illustrates the process of counting the realizations of motif m in network g . There are 3 injective mappings φ from the nodes of m to $\{1, 2, 3\}$ and 3 others to $\{2, 3, 4\}$ such that $\varphi(m) \subseteq g$. The degeneracy of the motif is $\delta_m = 3$ and it has $|\mathcal{J}_{\text{src}}(m)| = 3$ source nodes. Therefore, the participation density for nodes 1 and 4 is $b_{i,5}(m, g) = 3/3 \cdot 5^2 = 0.04$, for nodes 2 and 3 it is $b_{i,5}(m, g) = 6/3 \cdot 5^2 = 0.08$, and for node 5 it is $b_{5,5}(m, g) = 0$. The subgraph density of m in g is $b_5(m, g) = 0.048$.

complete network. With this, we can define a participation density of i in m .

Definition 4. The *participation density* of individual i in motif m given network g is

$$b_{i,N}(m, g) := \frac{N}{|\mathcal{J}_{\text{src}}(m)|} \frac{1}{N^{n_m}} \sum_{\substack{\varphi: \mathcal{J}_m \rightarrow \mathcal{J}_N \\ \varphi \text{ injective}}} \underbrace{\mathbb{1}\{\varphi(m) \subseteq g\}}_{\text{motif is present in network}} \underbrace{\mathbb{1}\{i \in \mathcal{J}_{\text{src}}(\varphi(m))\}}_{i \text{ is a source node of the motif}}. \quad (20)$$

Similarly, the *subgraph density* of motif m in network g is

$$b_N(m, g) := \frac{1}{N^{n_m}} \sum_{\substack{\varphi: \mathcal{J}_m \rightarrow \mathcal{J}_N \\ \varphi \text{ injective}}} \underbrace{\mathbb{1}\{\varphi(m) \subseteq g\}}_{\text{motif is present in network}} = \frac{1}{N} \sum_{i \in \mathcal{J}_N} b_{i,N}(m, g). \quad (21)$$

These definitions are useful since $b_{i,N}, b_N \in [0, 1)$ and $b_{i,N}(m, g) = b_N(m, g) = 0$ for the empty network and $b_{i,N}(m, g), b_N(m, g) \rightarrow 1$ as $N \rightarrow \infty$ for the complete network. This allows us to characterize the network in terms of these participation densities, which will directly enter the utilities.

To finish computing the number of realizations of m that i participates in, we need to account for the overcounting in how we compute the number of realizations. There can be multiple maps that give the same realization of the motif, as illustrated in Figure 1. Let δ_m , the *degeneracy* of motif m , be the number of injective maps $\varphi' : \mathcal{J}_m \rightarrow \mathcal{J}_m$ such that

$\varphi'(m) = m$. Having corrected for the counting, we have that the number of realizations of m that i participates in is $N^{n_m-1}|\mathcal{J}_{\text{src}}(m)|b_{i,N}(m, g)/\delta_m$. Therefore, the utility that agent i derives from motif m is

$$U_{i,N}^m(g; a_m) = N|\mathcal{J}_{\text{src}}(m)|\frac{a_m}{\delta_m}b_{i,N}(m, g). \quad (22)$$

If we consider a (finite) set of motifs $\mathcal{M}$, with values $\mathbf{a} := (a_m)_{m \in \mathcal{M}}$, the total utility she derives from these motifs is

$$U_{i,N}^{\mathcal{M}}(g; \mathbf{a}) = N \sum_{m \in \mathcal{M}} |\mathcal{J}_{\text{src}}(m)|\frac{a_m}{\delta_m}b_{i,N}(m, g). \quad (23)$$

To obtain the potential, we need the number of realizations of the motif in the network, which we can compute analogously with the subgraph density. Therefore, the potential for the game with N players is

$$\Phi_N^{\mathcal{M}}(g; \mathbf{a}) = N^2 \sum_{m \in \mathcal{M}} \frac{a_m}{\delta_m}b_N(m, g). \quad (24)$$

Having characterized a scheme under which we can study the model with a large number of players, we can now analyze the properties of the resulting networks.

4.2 Phase transitions

An astonishing result in the theory of graph limits is the possibility of discontinuous changes in the aggregate properties of the model caused by continuous changes in parameters. These types of phenomena fall under the category of *phase transitions* in the physics and applied mathematics literature, and the properties of systems near critical points (the parameter values where phase transitions occur) have been studied extensively¹⁴. In the case of our model, we will see that varying the values of motifs can induce a phase transition, causing a discontinuous change in the typical density of the networks.

The largest difficulty in analyzing the behavior of this model is characterizing the de-

¹⁴For a detailed overview of phase transitions in physics, see [Goldenfeld \(2018\)](#). For an application of the theory of phase transitions to ERGMs, see [Aristoff and Zhu \(2018\)](#).

nominator in Equation (18), which is intractable. This is because the number of terms in this sum is $2^{N(N-1)}$. Despite the difficulty in calculating this sum, this object encodes useful information about the system statistics.¹⁵ I now define the *scaled normalization constant*.

Definition 5. The scaled normalization constant of the system for motifs $\mathcal{M}$ with values $\mathbf{a}$ is

$$\zeta_N^{\mathcal{M}}(\mathbf{a}, \sigma) := \frac{1}{N^2} \log \left(\sum_{g \in \mathcal{G}_N} \exp \left[\left(\frac{1-\sigma}{\sigma} \right) \Phi_N^{\mathcal{M}}(g; \mathbf{a}) \right] \right). \quad (25)$$

The limiting scaled normalization constant is

$$\zeta^{\mathcal{M}}(\mathbf{a}, \sigma) := \lim_{N \rightarrow \infty} \zeta_N^{\mathcal{M}}(\mathbf{a}, \sigma). \quad (26)$$

To understand the $1/N^2$ re-scaling intuitively, note that to obtain dense networks, the participation densities must be typically non-vanishing. This means that the potential must be of order N^2 . Additionally, note that the size of network space scales as 2^{N^2} , so we expect the logarithm to scale as N^2 . Together, these arguments make us expect $\zeta_N^{\mathcal{M}}(\mathbf{a}, \sigma)$ to be of order 1.

We can now state the main result for the analysis of the motif model. An interesting aspect of the model is that the scaled partition function of the system is not only useful for calculating statistics of the limiting distribution, but rather is the central object that characterizes the typical behavior of the system as it navigates the space of networks.

Proposition 4. Suppose $a_m > 0$ for all motifs m with $e_m > 1$. Define the *entropy-adjusted potential (EAP)* as

$$\Gamma^{\mathcal{M}}(\rho; \mathbf{a}, \sigma) := \left(\frac{1-\sigma}{\sigma} \right) \underbrace{\sum_{m \in \mathcal{M}} \frac{a_m}{\delta_m} \rho^{e_m}}_{\text{motif values}} + \underbrace{H(\rho)}_{\text{entropy}} \quad (27)$$

where $H(p) := -p \log(p) - (1-p) \log(1-p)$ is the entropy of a Bernoulli random variable

¹⁵In the physics literature, this denominator is called the *partition function* (Huang, 1987).

with parameter p . Then the limiting normalization constant exists and is given by

$$\zeta^{\mathcal{M}}(\mathbf{a}, \sigma) = \max_{\rho \in [0,1]} \Gamma^{\mathcal{M}}(\rho; \mathbf{a}, \sigma). \quad (28)$$

Additionally, if the maximizer $\rho_{\mathcal{M}}^*(\mathbf{a}, \sigma)$ is unique, then as $N \rightarrow \infty$ the networks generated by the model become indistinguishable from those generated by an Erdős–Rényi random graph model with parameter $\rho_{\mathcal{M}}^*(\mathbf{a}, \sigma)$.¹⁶

Proposition 4 is similar to Theorem 2 in Mele (2017), and builds on it by extending the result to a microfounded model with arbitrary structures using the characterization in Proposition 2. This result essentially reduces the dimensionality of the complex network formation process to a single one-dimensional optimization process, where the maximizer ρ^* can be interpreted as the typical density of the resulting networks. The nature of this optimization process is interesting, since no single agent is solving the problem in Equation (28). To understand it, it is useful to analyze where the entropy term comes from. This term arises from the fact that the number of networks over N nodes with density ρ , to leading exponential order, is approximately $e^{N^2 H(\rho)}$. Therefore, as agents stochastically try to optimize their utility by building motifs, they get “stuck” exploring an exponentially growing region of network space, leading to a trade-off between utility and entropy.

It is important to note that this analysis is restricted to positive motif values for motifs that involve more than one link, since the characterization in terms of typical densities can break down when these values are negative, as shown in Mele (2017). This is of particular interest when analyzing the problem of estimation of model parameters, but I will restrict my attention to positive values, since this is still a rich enough case where there are non-trivial economic forces at play. One particularly interesting phenomenon that arises in the large N limit is the emergence of phase transitions. We can illustrate this with our simple model of trade.

Example 1 (Continued). We can easily see how our simple trade model can be formulated in terms of motifs. In this case, our set of motifs is $\mathcal{M} = \{m_1, m_2\}$, where $m_1 := \{12\}$ is the “outgoing links” motif and $m_2 := \{12, 21\}$ is the “reciprocated links” motif. The values

¹⁶See Appendix B for a detailed treatment of graph limits

assigned to these motifs are $\mathbf{a} = (-c, v)$ and their degeneracies are $\delta_{m_1} = 1$ and $\delta_{m_2} = 2$. Note that both motifs have $n_m = 2$, so their value is constant as N grows.

The entropy-adjusted potential Γ is shown in panel (a) of Figure 2 for a fixed value of c . Note that Γ is a continuous function of all its parameters. For low and high values of v , there is a unique local maximum of Γ . However, for v near $2c$, Γ has two local maxima, and their values coincide at $v = 2c$. This means that there is a discontinuous jump in the density that maximizes Γ at $v = 2c$. This happens despite Γ being continuous because, while the location of local maxima does change continuously, there can be discrete jumps between two different local maxima.

As shown in panel (b) of Figure 2, this discontinuous jump occurs for a wide range of values of v , and it defines a clear change between a phase of low-density and high-density networks. Thus, reducing the value of mutual links across the line of phase transitions can lead to the discontinuous collapse of the networks in the model. $\diamond$

The discrete jumps in the behavior of the typical density in the example have a striking resemblance to the change in the structure of the PMNE of the static game. If we compare the average density of the PMNE networks, we have a density of 0 for the empty network and 1 for the full network. This means that the average density of the PMNE networks jumps from 0 to 1 as v changes from being lower than $2c$ to being higher. This jump seems to remain for non-zero σ . With this intuition, we can define what a phase transition is, and why there is a connection to the behavior of the set of PMNE.

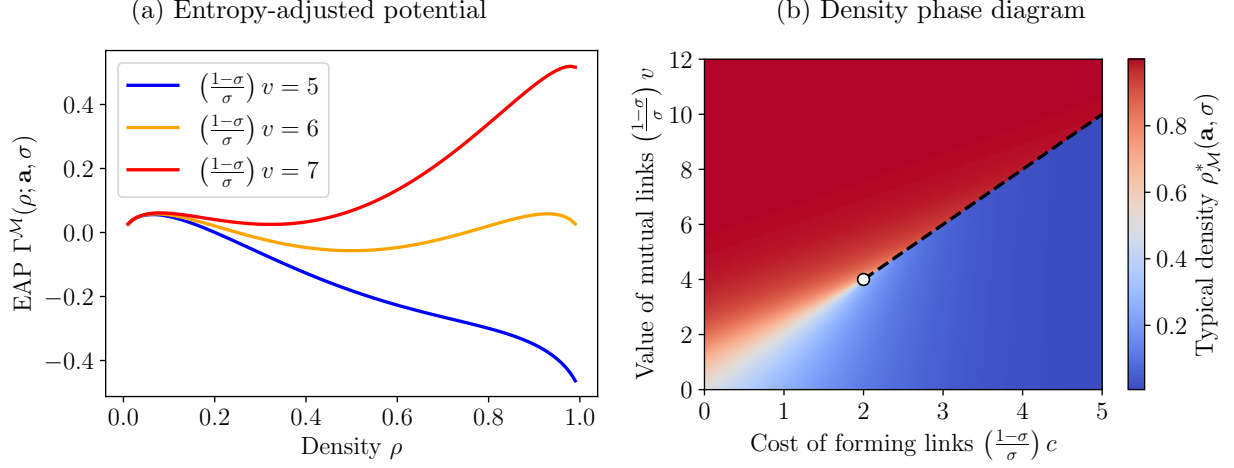
Definition 6. The system is said to undergo a *phase transition* at parameters $(\mathbf{a}, \sigma)$ if $\rho_{\mathcal{M}}^*(\mathbf{a}, \sigma)$ is discontinuous in $\mathbf{a}$.¹⁷ A phase transition at parameters $(\mathbf{a}, \sigma)$ is said to be *incentive-driven* if $\lim_{\sigma \rightarrow 0^+} \rho_{\mathcal{M}}^*(\mathbf{a}, \sigma)$ is discontinuous in $\mathbf{a}$.¹⁸ If a phase transition is not incentive-driven, it is said to be *entropy-driven*.

The intuition behind phase transitions being incentive-driven is that there is a stark change in the structure of the PMNE that survives having non-zero noise. If noise becomes

¹⁷In the physics literature, there are other kinds of phase transitions that involve non-analyticities, not only discontinuities.

¹⁸This is related to the concept of zero-temperature phase transitions in statistical physics. See [Goldenfeld \(2018\)](#) for a detailed discussion.

Figure 2: Phase transition induced by motif values



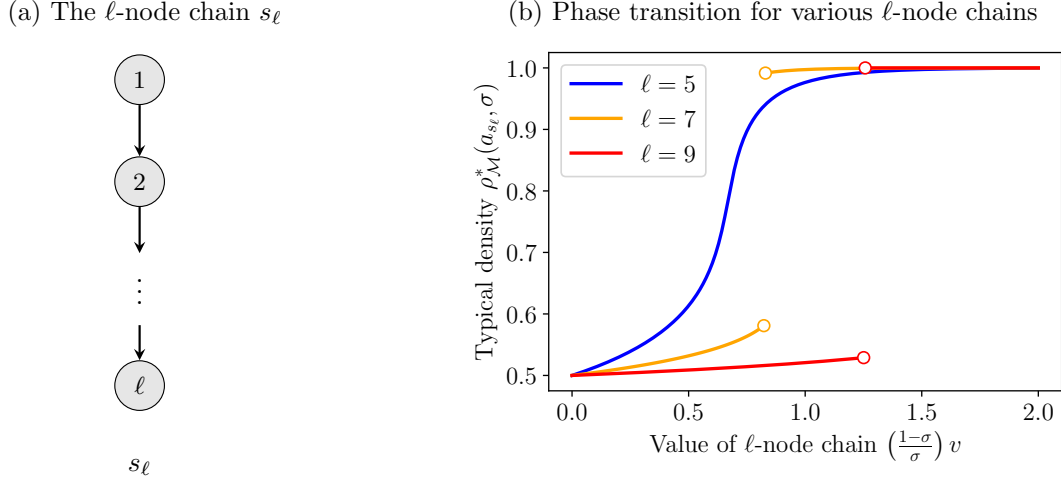
Notes. This figure illustrates the mechanisms behind phase transitions with two motif values. Panel (a) shows the behavior of the entropy-adjusted potential $\Gamma^M(\rho; \mathbf{a}, \sigma)$ as the value of mutual links v is varied, while the value of single connections is kept constant at $(\frac{1-\sigma}{\sigma})c = 3$. Panel (b) shows the value of the typical density for different values of cost c and the value of mutual-links v .

too strong, however, the phase transition can disappear, since incentives are not strong enough to drive the transition. This can be seen in the lower-left corner of Figure 2, panel (b). We can also find examples of entropy-driven phase transitions, where phase transitions appear when the motifs are sufficiently complex.

Example 2. To illustrate how complexity in the motifs can drive phase transitions, consider a very simple supply chain formation game. To produce a product, ℓ firms must come together to form a supply chain. We say a supply chain is formed if a firm connects to another, which connects to another, etc. for a total of $\ell - 1$ connections. There are no costs to forming links, just a positive payoff from forming the supply chain (to be specified below). Therefore, since there are no incentives to sever links, the unique Nash equilibrium (and hence PMNE) of the deterministic game is the complete network. This means that this model cannot have an incentive-driven phase transition.

This game corresponds to a motif model with a single motif: an ℓ -node directed chain. Formally, the ℓ -node chain is the network $s_\ell = \{12, 23, \dots, (\ell - 1)\ell\}$. The degeneracy of this motif is $\delta_{s_\ell} = 1$. If there are N players, the payoff that each firm gets for each chain it participates in is $v/N^{\ell-2}$ for some $v > 0$. Hence, v is the value of the supply chain motif.

Figure 3: Phase transition in the ℓ -node chain model



Notes. This figure shows the behavior of the typical density under the ℓ -node chain model. Panel (a) shows an illustration of the ℓ -node chain motif. Panel (b) shows how the typical density depends on the value of the chain v for different chain lengths ℓ .

The typical density of the network as a function of supply chain value is shown in Figure 3. The 5-node model shows a continuous change from the $\rho = 1/2$ density obtained from shock-dominated choices to the $\rho = 1$ density from the PMNE. In contrast, the 7-node and 9-node models show a large discontinuity in the typical density as the value of the chain increases. This can be interpreted as a transition from a shock-dominated phase to a utility-dominated phase. $\diamond$

An important ingredient for phase transitions to appear is that agents must derive utility from structures that are “sufficiently complex” to generate these effects. For example, if the only motif we considered were a single link, the typical density would be a continuous function of the utility of this motif. A similar complexity requirement was found in the analysis of supply networks in Elliott et al. (2022), where there is a discontinuity in the probability of successful production as a function of how likely individual production ties are successful. They find that the complexity of production, defined as the number of different inputs required to produce some good, is fundamental in generating this discontinuity. It seems, therefore, that the feedback generated by having multiple relationships is a recurring feature of models that generate these types of phenomena.

4.3 Heterogeneous agents

Our previous analysis showed that the network formation model can generate non-trivial phenomena even when all agents have the same utility. In this section, I expand on this by analyzing the network formation game for heterogeneous agents. Heterogeneity is a natural property of real network formation processes. For example, individuals have the tendency to form disproportionately more connections with people similar to themselves than to others. This is a phenomenon known as homophily (McPherson et al., 2001; Pin and Rogers, 2016). Another setting where heterogeneity is important is when networks are spatially embedded, such as in trade routes, since distance between agents becomes an important factor when determining whether or not to form relationships. Here, I study how heterogeneity in preferences interacts with network effects to determine the aggregate properties of large networks.

To characterize agents' heterogeneity, I classify them into a finite number of types. Let Θ be the set of agents' types, with $|\Theta| = L < \infty$. For a given number of agents N , their types are given by $\theta_1^N, \dots, \theta_N^N \in \Theta$. Let $\hat{\mathbf{w}}^N := (\hat{w}_\theta^N)_{\theta \in \Theta}$, where $\hat{w}_\theta^N$ is the fraction of the N agents that are of type θ . As $N \rightarrow \infty$, let $\hat{\mathbf{w}}^N \rightarrow \mathbf{w}$ for some $\mathbf{w} \in \Delta(\Theta)$ with full support over Θ .

In addition to deriving value from network motifs, which only depend on network structure and not on agents' types, I now allow agents' utilities to depend on the distribution of types of their neighbors. This allows us to systematically include heterogeneity in the agents, while also maintaining group-formation incentives. Define the unnormalized empirical distribution of agent i 's neighbors in network g as

$$\hat{z}_{i,\theta}^N(g) = \sum_{j=1}^N \mathbb{1}\{\theta_j^N = \theta\} \mathbb{1}\{ij \in g\} \quad \forall \theta \in \Theta. \quad (29)$$

Agents' utility functions now include a component $u : \Theta \times \mathbb{R}_+^\Theta \rightarrow \mathbb{R}$, which depends on their own type and the distribution of their neighbors' types. With this, the utility of agent i in network g is determined by the motif incentives, given by Equation (23), and the value of

their neighborhood:

$$U_{i,N}(g) = \underbrace{U_{i,N}^{\mathcal{M}}(g; \mathbf{a})}_{\text{motif values}} + \underbrace{u_{\theta_i^N}[\hat{\mathbf{z}}_i^N(g)]}_{\text{neighborhood value}}. \quad (30)$$

To characterize the limiting distribution of networks, I make some regularity assumptions on u .

Assumption 3. *The neighborhood utility function $u : \Theta \times \mathbb{R}_+^\Theta \rightarrow \mathbb{R}$ has the following properties:*

- *Homogeneity: for all $\theta \in \Theta$, $\mathbf{z} \in \mathbb{R}_+^\Theta$, and $c > 0$, u satisfies $u_\theta[c\mathbf{z}] = cu_\theta[\mathbf{z}]$.*
- *Concave decomposition: for all $\theta \in \Theta$, the function u_θ can be decomposed into*

$$u_\theta(\mathbf{z}) = \sum_{\theta' \in \Theta} c_{\theta\theta'} z_{\theta'} + \tilde{u}_\theta(\mathbf{z}), \quad (31)$$

where $c_{\theta\theta'} \in \mathbb{R}$ and $\tilde{u}_\theta$ is continuous, concave and monotonically increasing in all its arguments.

These assumptions capture economic environments where agents' payoffs depend smoothly on the composition of their neighbors. In spatially embedded networks, types may represent locations, so we might want to consider distance-dependent costs. In production or trade networks, types can represent sectors, and the value of a firm's neighborhood reflects how production scales with inputs (in a CES aggregator, for example), net of costs. Homogeneity ensures that payoffs scale naturally with network size, while concavity and monotonicity guarantee diminishing returns to additional similar neighbors, yielding stable and economically interpretable asymptotic properties.

In the case of homogeneous agents, the distribution resulting from the network formation process is concentrated around networks with some typical density. To extend this result to heterogeneous agents, we need a more flexible characterization of the limiting object. It turns out to be sufficient to characterize the typical density of links between types θ and θ' . Formally, let $\mathcal{K}_\Theta$ be the set of functions $\psi : \Theta^2 \rightarrow [0, 1]$, which I denote as *kernels*.

Intuitively, if the process converges to a density kernel ψ^* , then $\psi_{\theta, \theta'}^*$ is the fraction of agents of type θ' that agents of type θ will typically form a connection with.

Another generalization that must be made to account for agent heterogeneity is of the method of calculating motif densities. Intuitively, the density of motif m in an Erdős–Rényi network with density ρ is ρ^{e_m} . This allows for a clean characterization of the limiting typical networks, as seen in Proposition 4. In order to generalize this counting to networks with density kernel ψ , the density must now depend on the kernel and the fraction of agents of a given type. Formally, for a kernel $\psi \in \mathcal{K}_\Theta$ we define the density of motif m as¹⁹

$$b[m, \psi; \mathbf{w}] := \sum_{\boldsymbol{\theta} \in \Theta^{n_m}} \left(\prod_{i \in \mathcal{I}_m} w_{\theta_i} \right) \left(\prod_{ij \in m} \psi_{\theta_i \theta_j} \right). \quad (32)$$

Note that this definition of motif density is equal to ρ^{e_m} for the kernel satisfying $\psi_{\theta\theta'} = \rho$ for all θ, θ' , which matches our intuition. With this definition, we can study how group formation incentives interact with neighborhood values.

Theorem 1. *Suppose that $a_m > 0$ for all motifs with $e_m > 1$ and the neighborhood utility function satisfies Assumption 3. Define the entropy-adjusted potential of this model as*

$$\Gamma_\Theta^{\mathcal{M}}(\psi; \mathbf{a}, \mathbf{w}, \sigma) := \left(\frac{1-\sigma}{\sigma} \right) \underbrace{\sum_{m \in \mathcal{M}} \frac{a_m}{\delta_m} b[m, \psi; \mathbf{w}]}_{\text{motif values}} + \sum_{\theta \in \Theta} w_\theta \left[\underbrace{\sum_{\theta' \in \Theta} w_{\theta'} H(\psi_{\theta\theta'})}_{\text{entropy}} + \left(\frac{1-\sigma}{\sigma} \right) \underbrace{u_\theta[(w_{\theta'} \psi_{\theta\theta'})_{\theta' \in \Theta}]}_{\text{neighborhood value}} \right] \quad (33)$$

Then the limiting scaled normalization constant is given by

$$\zeta_\Theta^{\mathcal{M}}(\mathbf{a}, \mathbf{w}, \sigma) = \max_{\psi \in \mathcal{K}_\Theta} \Gamma_\Theta^{\mathcal{M}}(\psi; \mathbf{a}, \mathbf{w}, \sigma). \quad (34)$$

Furthermore, if the maximizer $\psi^{\mathcal{M}, \Theta}(\mathbf{a}, \mathbf{w}, \sigma)$ is unique, then as $N \rightarrow \infty$ the networks generated by the model become indistinguishable from those generated by a directed stochastic block model with edge probabilities between types θ and θ' given by the kernel $\psi_{\theta, \theta'}^{\mathcal{M}, \Theta}(\mathbf{a}, \mathbf{w})$.

Theorem 1 provides a similar reduction in dimensionality as Proposition 4, where the

¹⁹For a formal justification for the introduction of this limiting object, see Appendix B.

complex model of network formation can be understood in terms of a limiting kernel. The same forces drive the result, in the sense that agents must explore an exponentially growing region of network space, so the entropic term appears in the optimization problem. However, now there is heterogeneity in agents' incentives, leading agents of different types to explore in different directions, which results in a potentially non-uniform kernel. In addition to the neighborhood incentives, there is still a global feedback in the form of the motif values. This model provides an additional dimension to the complex interplay between incentives and network formation frictions, in that network structure can now affect the composition of who agents interact with.

5 A Simple Model of Trade

In order to build economic intuition on the forces driving the phenomena in Section 4, I expand the model of trade from Example 1. In this setting, assigning values to motifs and neighborhoods arises naturally from the incentives of firms.

5.1 Technology and incentives

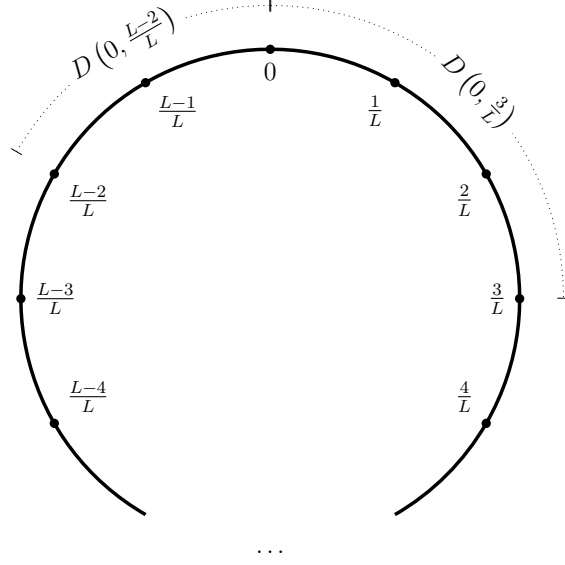
Consider a set of firms that are spatially embedded in a circular city with unit circumference. There are L equally spaced-out locations where firms can be located. A firm's type is determined by its location, such that $\Theta = \{0, \frac{1}{L}, \dots, \frac{L-1}{L}\}$. The main quantity that determines the incentives of a firm to trade with another is the distance to the other firm, defined as

$$D(\theta, \theta') = \min\{|\theta - \theta'|, 1 - |\theta - \theta'|\}. \quad (35)$$

This definition of distance is illustrated in Figure 4. As $N \rightarrow \infty$, the fractions of firms of different types converges to $\mathbf{w}$.

Firms play a myopic network formation game, where the presence of a link ij signifies that firm i wishes to establish a trade route with firm j . To establish this link, firm i faces a fixed cost of $\gamma D(\theta_i, \theta_j)$, which is proportional to the distance between the firms. If both firms wish to establish a trade route, this route becomes active and generates a revenue of

Figure 4: Distance in the trade model



Notes. This figure illustrates the distance between firms of different types in the trade model in Section 5. The set of types is $\Theta = \{0, \frac{1}{L}, \dots, \frac{L-1}{L}\}$ and the distance is given by $D(\theta, \theta') = \min\{|\theta - \theta'|, 1 - |\theta - \theta'|\}$. In this representation, the city has a circumference of 1.

v. Hence, for a fixed N , the utility of agent i in network g is

$$U_{i,N}(g; \gamma, v) = v \underbrace{\sum_{j \in \mathcal{J}_N} \mathbb{1}\{ij \in g, ji \in g\}}_{\text{reciprocated links}} - \gamma \underbrace{\sum_{j \in \mathcal{J}_N} D(\theta_i, \theta_j) \mathbb{1}\{ij \in g\}}_{\text{outgoing links}}. \quad (36)$$

This utility can be decomposed into two parts: the value of participating in a mutual-links motif and the cost of establishing trade intentions with a neighborhood.

In terms of the notation in section 4, firms assign value to a single motif, such that $\mathcal{M} = \{m_2\}$, where $m_2 = \{12, 21\}$ is the “mutual links” motif. The degeneracy of this motif is $\delta_{m_2} = 2$ and its value is v . The cost of establishing trade intentions can be written as a neighborhood distribution utility, given by

$$u_\theta[\mathbf{z}] = -\gamma \sum_{\theta' \in \Theta} z_{\theta'} D(\theta, \theta'). \quad (37)$$

Using this formulation, we can use our previous results on the limiting properties of the

network formation game to study the effect of trading incentives on aggregate outcomes.

5.2 Spatial structure of trade

Characterizing the network formation game in terms of motif and neighborhood values allows us to formulate the problem of finding the typical density kernel $\psi^{\mathcal{M},\Theta}(v, \mathbf{w}, \sigma)$ that characterizes the limiting properties of the system. Following Theorem 1, the limiting scaled normalization constant of this model is given by

$$\zeta_{\Theta}^{\mathcal{M}}(v, \mathbf{w}, \sigma) = \max_{\psi \in \mathcal{K}_{\Theta}} \left[\sum_{\theta, \theta' \in \Theta^2} w_{\theta} w_{\theta'} \left(\left(\frac{1-\sigma}{\sigma} \right) \left[\frac{v}{2} \psi_{\theta\theta'} \psi_{\theta'\theta} - \gamma \psi_{\theta\theta'} D(\theta, \theta') \right] + H(\psi_{\theta\theta'}) \right) \right], \quad (38)$$

and the typical density kernel $\psi^{\mathcal{M},\Theta}(v, \mathbf{w}, \sigma)$ is the maximizer. Interestingly, the solution to this problem can be characterized in terms of our solution to the problem with only motif values.

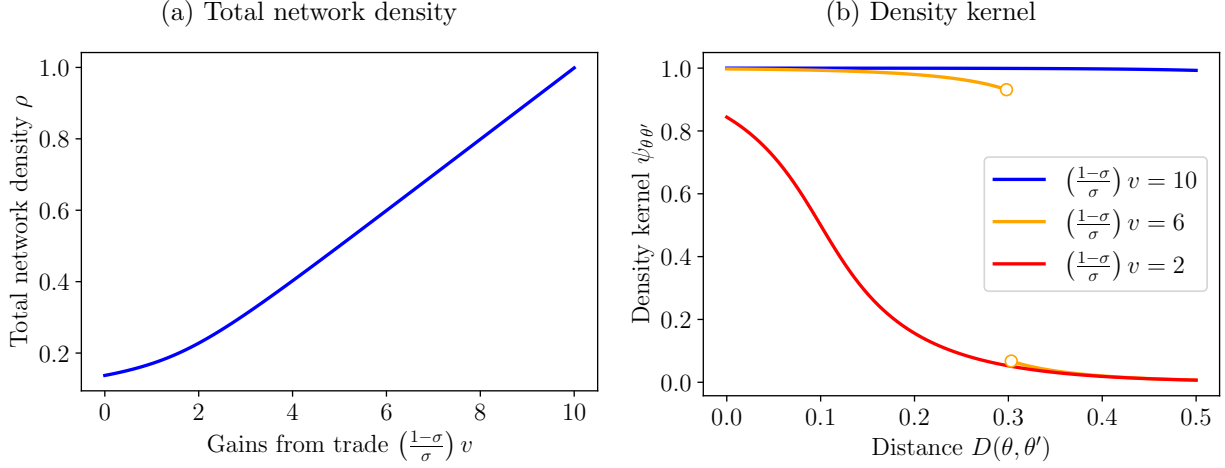
Lemma 3. *Let $\rho_{\mathcal{M}_0}^*(\mathbf{a}, \sigma)$ be the typical density of the model with motifs $\mathcal{M}_0 = \{m_1, m_2\}$, where $m_1 = \{12\}$ and $m_2 = \{12, 21\}$. Suppose the maximizer $\psi^{\mathcal{M},\Theta}(v, \mathbf{w}, \sigma)$ in the problem above is unique. Then $\psi_{\theta\theta'}^{\mathcal{M},\Theta}(v, \mathbf{w}, \sigma)$ is given by $\rho_{\mathcal{M}_0}^*(-\gamma D(\theta, \theta'), v, \sigma)$.*

Intuitively, this result says that in this model there is effectively a network formation process with motif incentives for each pair (θ, θ') , and these do not affect each other for different pairs. This means that our results for phase transitions in the homogeneous agent model hold locally for this model. Therefore, for a fixed θ , $\psi_{\theta\theta'}^{\mathcal{M},\Theta}(v, \mathbf{w})$ can exhibit a large jump as θ' is varied.

This characterization allows us to think about the effect of varying the value of gains from trade v . This is relevant for policy interventions, such as taxes on trade. In the homogeneous agent model, such a change can trigger a phase transition. However, the heterogeneity in the types of firms can make it such that these effects are greatly mitigated in the total network density.

To build intuition on this result, consider the case where the limiting distribution $\mathbf{w}$ is the uniform distribution over Θ . The total network density is $\rho = \sum_{(\theta, \theta') \in \Theta^2} w_{\theta} w_{\theta'} \psi_{\theta\theta'}$. With

Figure 5: Effect of motif values on network density



Notes. This figure shows the behavior of the typical density and the density kernel as the value of mutual links is changed. There are $L = 100$ types and the cost per unit distance is $\gamma = 10$. Panel (a) shows the total network density as a function of the value of gains from trade. Panel (b) shows the density kernel as a function of distance for different values of v .

this, we can study the differential effects of changing v on the total network density and on the density kernel, as shown in Figure 5. In this case, increasing the gains from trade increases the total density of the network in almost imperceptible jumps. In contrast, the density kernel for any two given types can exhibit a stark discontinuous transition. We see that understanding the phase transition structure in the motif-only model gives us insight into the mechanisms of these “local” phase transitions as distance changes. Particularly, this model gives rise to clear high-density neighborhoods where firms are trading, and a discontinuous drop in trade with firms outside this neighborhood.

6 Conclusion

I have studied a dynamic model of network formation in which agents stochastically meet and myopically decide whether to form or sever links. The paper makes two main contributions: it establishes an alternate formulation of the game in terms of structure values and characterizes the asymptotic behavior of the process as the number of players grows large. When all players in a structure give it the same value, the static game is a potential game and the stationary distribution of the dynamic model can be explicitly computed. This allows us to

study the effect of changing the value of these local structures on the aggregate properties of the network. In large populations, the process can be asymptotically approximated by an Erdős–Rényi process. The typical density of this process can be discontinuous as a function of the values of structures, a phenomenon known as a phase transition.

This work is a stepping stone towards improving our understanding of the economic forces behind the formation of large networks. It gives a clear relation between the incentives to form links and the frictions agents encounter when forming the network, materialized in an entropy term. Despite the simplicity of the motif model, it captures important spillover effects present in real network formation processes. Mainly, forming a structure not only affects current payoffs, but amplifies the probability of forming future links by allowing the formation of even more structures. This self-reinforcing behavior is the driving force behind the phase transitions we observe.

These results open many avenues for further exploration. While a myopic network formation model is a convenient starting point to study aggregate properties, it is natural to ask how the results change when agents are forward-looking. Care is needed, though, since the space of networks grows super-exponentially, so assuming agents can perfectly form expectations over future network realizations might be inappropriate. Another direction to explore is how to extend these results beyond potential games. Potential games have the very desirable property of being equivalent to reversibility (Proposition 3), which greatly simplifies the computation of the stationary distribution. However, the large changes at the phase transitions suggest that continuous deviations from potential games (in an appropriate sense) might preserve these stark jumps. I hope this work inspires further exploration in these directions.

References

- Aristoff, David and Lingjiong Zhu**, “On the phase transition curve in a directed exponential random graph model,” *Advances in Applied Probability*, March 2018, 50 (01), 272–301.
- Badev, Anton**, “Nash Equilibria on (Un)Stable Networks,” *Econometrica*, 2021, 89 (3), 1179–1206.

- Bala, Venkatesh and Sanjeev Goyal**, “A Noncooperative Model of Network Formation,” *Econometrica*, September 2000, *68* (5), 1181–1229.
- Baranyai, Zolt**, “On the factorization of the complete uniform hypergraphs,” *Infinite and finite sets*, 1974.
- Boeckner, Derek**, “Directed graph limits and directed threshold graphs.” Ph.D., The University of Nebraska - Lincoln, United States – Nebraska 2013. ISBN: 978-1-303-04202-7 Publication Title: ProQuest Dissertations and Theses 3559217.
- Boucher, Vincent and Ismael Mourifié**, “My friend far, far away: a random field approach to exponential random graph models,” *The Econometrics Journal*, October 2017, *20* (3), S14–S46.
- Butts, Carter**, “Using potential games to parameterize ERG models,” *University of California, Irvine working paper*, 2009, *1*.
- Chandrasekhar, Arun G and Matthew O Jackson**, “A Network Formation Model Based on Subgraphs,” *Review of Economic Studies*, November 2025, *92* (6), 3741–3787.
- Chatterjee, Sourav and Persi Diaconis**, “Estimating and understanding exponential random graph models,” *The Annals of Statistics*, October 2013, *41* (5).
- **and S.R.S. Varadhan**, “The large deviation principle for the Erdős-Rényi random graph,” *European Journal of Combinatorics*, October 2011, *32* (7), 1000–1017.
- Christakis, Nicholas, James Fowler, Guido W. Imbens, and Karthik Kalyanaraman**, “An empirical model for strategic network formation,” in “The Econometric Analysis of Network Data,” Elsevier, 2020, pp. 123–148.
- Cimini, Giulio, Tiziano Squartini, Fabio Saracco, Diego Garlaschelli, Andrea Gabrielli, and Guido Caldarelli**, “The statistical physics of real-world networks,” *Nature Reviews Physics*, January 2019, *1* (1), 58–71.
- Currarini, Sergio, Matthew O. Jackson, and Paolo Pin**, “An Economic Model of Friendship: Homophily, Minorities, and Segregation,” *Econometrica*, 2009, *77* (4), 1003–1045.
- Dembo, Amir and Ofer Zeitouni**, *Large Deviations Techniques and Applications*, Vol. 38 of *Stochastic Modelling and Applied Probability*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.
- Diao, Peter, Dominique Guillot, Apoorva Khare, and Bala Rajaratnam**, “Model-free consistency of graph partitioning,” 2016. Version Number: 1.
- Dorogovtsev, S. N., A. V. Goltsev, and J. F. F. Mendes**, “Critical phenomena in complex networks,” *Reviews of Modern Physics*, October 2008, *80* (4), 1275–1335.
- Durrett, Richard**, *Essentials of Stochastic Processes* Springer Texts in Statistics, Cham: Springer International Publishing, 2016.

- Dutta, Bhaskar, Sayantan Ghosal, and Debraj Ray**, “Farsighted network formation,” *Journal of Economic Theory*, June 2005, 122 (2), 143–164.
- Elliott, Matthew, Benjamin Golub, and Matthew V. Leduc**, “Supply Network Formation and Fragility,” *American Economic Review*, August 2022, 112 (8), 2701–2747.
- Glauber, Roy J.**, “Time-Dependent Statistics of the Ising Model,” *Journal of Mathematical Physics*, February 1963, 4 (2), 294–307.
- Goldenfeld, Nigel**, *Lectures on Phase Transitions and the Renormalization Group*, 1 ed., CRC Press, March 2018.
- Golub, Benjamin and Yair Livne**, “Strategic random networks and tipping points in network formation,” *Unpublished paper, Department of Economics, Harvard University.*, 2010.
- Graham, Bryan S.**, “An Econometric Model of Network Formation With Degree Heterogeneity,” *Econometrica*, 2017, 85 (4), 1033–1063.
- Hołyst, Janusz A., Krzysztof Kacperski, and Frank Schweitzer**, “Phase transitions in social impact models of opinion formation,” *Physica A: Statistical Mechanics and its Applications*, September 2000, 285 (1-2), 199–210.
- Hsieh, Chih-Sheng, Michael D. König, and Xiaodong Liu**, “A Structural Model for the Coevolution of Networks and Behavior,” *The Review of Economics and Statistics*, March 2022, 104 (2), 355–367.
- Huang, Kerson**, *Statistical mechanics*, 2nd ed ed., New York: Wiley, 1987.
- Jackson, Matthew O.**, “A Survey of Network Formation Models: Stability and Efficiency,” in Gabrielle Demange and Myrna Wooders, eds., *Group Formation in Economics*, 1 ed., Cambridge University Press, January 2005, pp. 11–57.
- , *Social and economic networks*, Princeton, N.J. Woodstock: Princeton University Press, 2011.
- **and Alison Watts**, “The Evolution of Social and Economic Networks,” *Journal of Economic Theory*, October 2002, 106 (2), 265–295.
- **and Asher Wolinsky**, “A Strategic Model of Social and Economic Networks,” *Journal of Economic Theory*, October 1996, 71 (1), 44–74.
- Lovász, László**, *Large networks and graph limits* number volume 60. In ‘American Mathematical Society colloquium publications.’, Providence, Rhode Island: American Mathematical Society, 2012.
- Luenberger, David G.**, *Optimization by vector space methods* Series in decision and control, nachdr. ed., New York, NY Chichester Weinheim: Wiley, 1998.

- McPherson, Miller, Lynn Smith-Lovin, and James M Cook**, “Birds of a Feather: Homophily in Social Networks,” *Annual Review of Sociology*, August 2001, *27* (1), 415–444.
- Mele, Angelo**, “A Structural Model of Dense Network Formation,” *Econometrica*, 2017, *85* (3), 825–850.
- , “A Structural Model of Homophily and Clustering in Social Networks,” *Journal of Business & Economic Statistics*, July 2022, *40* (3), 1377–1389.
- Monderer, Dov and Lloyd S. Shapley**, “Potential Games,” *Games and Economic Behavior*, May 1996, *14* (1), 124–143.
- Norris, J. R.**, “Continuous-time Markov chains II,” in “Markov Chains” Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1997, pp. 108–127.
- Pin, Paolo and Brian W. Rogers**, “Stochastic Network Formation and Homophily,” in Yann Bramoullé, Andrea Galeotti, and Brian W. Rogers, eds., *The Oxford Handbook of the Economics of Networks*, Oxford University Press, April 2016, pp. 137–166.
- Robins, Garry, Pip Pattison, Yuval Kalish, and Dean Lusher**, “An introduction to exponential random graph (p^*) models for social networks,” *Social Networks*, May 2007, *29* (2), 173–191.
- Sadler, Evan and Benjamin Golub**, “Games on Endogenous Networks,” 2021. Version Number: 8.
- Squartini, Tiziano, Joey De Mol, Frank Den Hollander, and Diego Garlaschelli**, “Breaking of Ensemble Equivalence in Networks,” *Physical Review Letters*, December 2015, *115* (26), 268701.
- Stanley, Richard P.**, *Enumerative combinatorics: volume 1* number 49. In ‘Cambridge studies in advanced mathematics.’, second edition ed., Cambridge New York: Cambridge University Press, 2012.
- Varadhan, S.R.S.**, “Large Deviations,” 2010.
- Warner, Frank W.**, *Foundations of Differentiable Manifolds and Lie Groups*, Vol. 94 of *Graduate Texts in Mathematics*, New York, NY: Springer New York, 1983.

Appendix

A Proofs

A.1 Proofs of Lemmas

A.1.1 Proof of Lemma 1

This is an application of the Möbius inversion formula for functions of subsets. See Proposition 3.7.2 in [Stanley \(2012\)](#) for a proof for a more general setting and Example 3.8.3 for an explicit statement of the result.

A.1.2 Proof of Lemma 2

First, note that choice-equivalence for utility functions U and $\tilde{U}$ holds if and only if

$$U_i(\tau_{ij}(g)) - U_i(g) = \tilde{U}_i(\tau_{ij}(g)) - \tilde{U}_i(g) \quad (\text{A.1})$$

for all g, i and ij . The “only if” direction is immediate, since a one-link change is a particular case of a deviation. For the “if” direction, notice that any general deviation can be achieved by changing single links, and adding the marginal utility of single-link deviations gives a telescopic sum that results in the marginal utility of the entire deviation.

I now show that the condition (A.1) holds if and only if condition (9) holds. For the rest of the proof, i is fixed. First, fix some network g . We have that

$$U_i(g \cup \{ij\}) - U_i(g \setminus \{ij\}) = \sum_{g' \subseteq g \cup \{ij\}} V_i(g') \mathbb{1}\{ij \in g'\}, \quad (\text{A.2})$$

and a similar expression involving $\tilde{U}_i$. This means that, regardless of whether $ij \in g$, the marginal utility of changing the link ij only depends on the value of structures that include the link ij .²⁰ In particular, this expression only involves structures g' for which $i \in \mathcal{J}_{\text{src}}(g')$. The “if” direction, then, is immediate. If all structures g' for which $i \in \mathcal{J}_{\text{src}}(g')$ satisfy

²⁰If $ij \notin g$, then $U_i(\tau_{ij}(g)) - U_i(g)$ is simply the negative of this expression.

$V_i(g') = \tilde{V}_i(g')$, then condition (A.1) is satisfied.

For the “only if” direction, we proceed by induction on the size of the structure.

Base case: Fix g such that $i \in \mathcal{J}_{\text{src}}(g)$ and $|g| = 1$. We can write $g = \{ij\}$. Note that

$$\begin{aligned} U_i(\tau_{ij}(g)) - U_i(g) &= U_i(\emptyset) - U_i(\{ij\}) = -V_i(\{ij\}), \\ \tilde{U}_i(\tau_{ij}(g)) - \tilde{U}_i(g) &= \tilde{U}_i(\emptyset) - \tilde{U}_i(\{ij\}) = -\tilde{V}_i(\{ij\}). \end{aligned} \quad (\text{A.3})$$

Since condition (A.1) holds, we must have $V_i(\{ij\}) = \tilde{V}_i(\{ij\})$. Therefore, all structures with $i \in \mathcal{J}_{\text{src}}(g)$ and $|g| = 1$ satisfy $V_i(g) = \tilde{V}_i(g)$.

Induction hypothesis: For a given $n \geq 1$, assume that all g such that $i \in \mathcal{J}_{\text{src}}(g)$ and $|g| \leq n$ satisfy $V_i(g) = \tilde{V}_i(g)$.

Inductive step: Fix g such that $i \in \mathcal{J}_{\text{src}}(g)$ and $|g| = n + 1$. Take some j such that $ij \in g$. We have that

$$\begin{aligned} U_i(\tau_{ij}(g)) - U_i(g) &= - \sum_{g' \subseteq g} V_i(g') \mathbb{1}\{ij \in g'\}, \\ \tilde{U}_i(\tau_{ij}(g)) - \tilde{U}_i(g) &= - \sum_{g' \subseteq g} \tilde{V}_i(g') \mathbb{1}\{ij \in g'\}. \end{aligned} \quad (\text{A.4})$$

Note that all structures $g' \subset g$ satisfy $|g'| \leq n$. If $ij \in g'$, then $i \in \mathcal{J}_{\text{src}}(g')$. Therefore, subtracting the two equations above, only the term for the structure $g' = g$ survives (note that $ij \in g$):

$$[U_i(\tau_{ij}(g)) - U_i(g)] - [\tilde{U}_i(\tau_{ij}(g)) - \tilde{U}_i(g)] = -[V_i(g) - \tilde{V}_i(g)].$$

Using condition (A.1), we obtain $V_i(g) = \tilde{V}_i(g)$. This completes the proof.

A.1.3 Proof of Lemma 3

The divergence in $H'(u)$ near 0 and 1 ensures the solution to the problem in Equation (38) is interior. For $\theta' \neq \theta$, the first-order condition for $\psi_{\theta\theta'}$ is

$$w_\theta w_{\theta'} \left\{ \left(\frac{1-\sigma}{\sigma} \right) [v\psi_{\theta'\theta}^* - \gamma D(\theta, \theta')] + H'(\psi_{\theta\theta'}^*) \right\} = 0. \quad (\text{A.5})$$

Now, suppose the solution is such that $\psi_{\theta\theta'}^* \neq \psi_{\theta'\theta}^*$. Since $D(\theta, \theta') = D(\theta', \theta)$, making the change $\psi_{\theta\theta'} \leftrightarrow \psi_{\theta'\theta}$ would yield the same value of the objective function. However, this violates the uniqueness of the solution. Therefore, if the solution is unique, we must have $\psi_{\theta\theta'}^* = \psi_{\theta'\theta}^*$.

The argument above implies that $\psi_{\theta\theta'}^*$ satisfies

$$\left(\frac{1-\sigma}{\sigma} \right) (v\psi_{\theta\theta'}^* - \gamma D(\theta, \theta')) + H'(\psi_{\theta\theta'}^*) = 0. \quad (\text{A.6})$$

This is precisely the first-order condition of the problem in Equation (28) with motif values $\mathbf{a} = (-\gamma D(\theta, \theta'), v)$.

A.2 Proof of Proposition 1

First, note that the game is a potential game with potential Φ if and only if

$$U_i(\tau_{ij}(g)) - U_i(g) = \Phi(\tau_{ij}(g)) - \Phi(g). \quad (\text{A.7})$$

The argument for this is analogous to the argument at the beginning of the proof of Lemma 2. Intuitively, any deviation can be constructed from one-link deviations, and the marginal utility of an arbitrary deviation is the sum of the marginal utilities of one-link deviations.

To prove the Proposition, it is useful to establish the following technical lemma:

Lemma 4. *Consider a function $\phi : \mathcal{D}_N \times \mathcal{G}_N \rightarrow \mathbb{R}$ with the property that $\phi_{ij}(g) = -\phi_{ij}(\tau_{ij}(g))$. The function satisfies*

$$\phi_{ij}(g) + \phi_{i'j'}(\tau_{ij}(g)) = \phi_{i'j'}(g) + \phi_{ij}(\tau_{i'j'}(g)) \quad (\text{A.8})$$

for all $ij, i'j' \in \mathcal{D}_N$ and $g \in \mathcal{G}_N$ if and only if there exists a function $\Psi : \mathcal{G}_N \rightarrow \mathbb{R}$ such that

$$\phi_{ij}(g) = \Psi(\tau_{ij}(g)) - \Psi(g) \quad (\text{A.9})$$

for all $ij \in \mathcal{D}_N$ and $g \in \mathcal{G}_N$. Furthermore, for a given ϕ , this function Ψ is unique up to an additive constant.

Proof. The “if” direction can be immediately obtained by directly checking Equation (A.8) using the expression for ϕ in terms of Ψ .

The “only if” direction can be proven by construction. Fix a network $g \in \mathcal{G}_N$ and let $n = |g|$ be its size. Let $\gamma = (\gamma_1, \dots, \gamma_n)$ be an ordering of the links in g , meaning that $\gamma_\ell \in g$ for $\ell = 1, \dots, n$ and $\gamma_\ell \neq \gamma_{\ell'}$ if $\ell \neq \ell'$. From this ordering, we can construct a sequence of networks recursively by letting $g_0^\gamma := \emptyset$ and setting $g_\ell^\gamma := \tau_{\gamma_\ell}(g_{\ell-1}^\gamma)$ for $\ell > 0$. Let us define the following quantity:

$$\Psi^\gamma(g) := \sum_{\ell=1}^n \phi_{\gamma_\ell}(g_{\ell-1}^\gamma). \quad (\text{A.10})$$

I now prove that this value does not depend on the specific ordering of the links chosen.

First, let us consider an ordering γ^0 that differs from γ only at positions ℓ_0 and $\ell_0 + 1$, where $1 \leq \ell_0 < n$. Since both γ and γ^0 are orderings, this means that γ^0 is γ with the entries ℓ_0 and $\ell_0 + 1$ swapped. From this, we must have $g_{\ell_0-1}^{\gamma^0} = g_{\ell_0-1}^\gamma$, and $g_{\ell_0+2}^{\gamma^0} = g_{\ell_0+2}^\gamma$ if $\ell_0 < n - 1$. Therefore, computing $\Psi^\gamma(g) - \Psi^{\gamma^0}(g)$ yields

$$\Psi^\gamma(g) - \Psi^{\gamma^0}(g) = \phi_{\gamma_{\ell_0}}(g_{\ell_0-1}^\gamma) + \phi_{\gamma_{\ell_0+1}}(g_{\ell_0}^\gamma) - \phi_{\gamma_{\ell_0}^0}(g_{\ell_0-1}^{\gamma^0}) - \phi_{\gamma_{\ell_0+1}^0}(g_{\ell_0}^{\gamma^0}). \quad (\text{A.11})$$

Note that we can write $g_{\ell_0}^{\gamma^0} = \tau_{\gamma_{\ell_0+1}}(g_{\ell_0-1}^\gamma)$. Therefore, this difference can be written as

$$\Psi^\gamma(g) - \Psi^{\gamma^0}(g) = \phi_{\gamma_{\ell_0}}(g_{\ell_0-1}^\gamma) + \phi_{\gamma_{\ell_0+1}}(\tau_{\gamma_{\ell_0}}(g_{\ell_0-1}^\gamma)) - \phi_{\gamma_{\ell_0+1}}(g_{\ell_0-1}^\gamma) - \phi_{\gamma_{\ell_0}}(\tau_{\gamma_{\ell_0+1}}(g_{\ell_0-1}^\gamma)). \quad (\text{A.12})$$

From the condition in Equation (A.8), this must vanish, such that $\Psi^\gamma(g) = \Psi^{\gamma^0}(g)$. Any ordering of the links in g can be obtained starting from γ through a sequence of “swaps”

like the one we just analyzed. Therefore, the quantity $\Psi^\gamma(g)$ will be the same regardless of the ordering γ chosen. Thus, we can define a quantity that depends solely on the network $\Psi(g) := \Psi^\gamma(g)$ for any ordering γ .

Now we want to prove that our construction of Ψ satisfies Equation (A.9). Fix a network g and a link $ij \in \mathcal{D}_N$. Without loss of generality, suppose that $ij \notin g$ ²¹. Let γ^1 be some ordering of the links in g and γ^2 be the ordering of the links in $\tau_{ij}(g)$ such that $\gamma_\ell^2 = \gamma_\ell^1$ for $1 \leq \ell \leq |g|$ and $\gamma_{|g|+1}^2 = ij$. We can write our Ψ functions using these orderings, such that

$$\Psi(\tau_{ij}(g)) - \Psi(g) = \sum_{\ell=1}^{|g|+1} \phi_{\gamma_\ell^2}(g_{\ell-1}^{\gamma^2}) - \sum_{\ell=1}^{|g|} \phi_{\gamma_\ell^1}(g_{\ell-1}^{\gamma^1}) = \phi_{ij}(g), \quad (\text{A.13})$$

as desired.

Finally, to prove uniqueness up to an additive constant, suppose that there are two functions Ψ_1 and Ψ_2 satisfying Equation (A.9) for all $g \in \mathcal{G}_N$ and $ij \in \mathcal{D}_N$. Let $K_0 := \Psi_1(\emptyset) - \Psi_2(\emptyset)$, and suppose there exists a g_0 such that $\Psi_1(g_0) - \Psi_2(g_0) \neq K_0$. From Equation (A.9), we have that, for all $g \in \mathcal{G}_N$ and $ij \in \mathcal{D}_N$

$$\begin{aligned} \Psi_1(g) - \Psi_2(g) &= [\Psi_1(g) + \phi_{ij}(g)] - [\Psi_2(g) + \phi_{ij}(g)] \\ &= \Psi_1(\tau_{ij}(g)) - \Psi_2(\tau_{ij}(g)). \end{aligned} \quad (\text{A.14})$$

Since this procedure can be repeated with any sequence of dyads to reach any network, we conclude that $\Psi_1(g) - \Psi_2(g) = K_0$ for all networks, which contradicts our assumption. $\square$

We can now prove Proposition 1. To do this, define $\phi_{ij}(g) := U_i(\tau_{ij}(g)) - U_i(g)$. Conservativeness, then, is equivalent to Equation (A.8). Lemma 4, then, establishes that conservativeness is equivalent to the existence of a potential, and that this potential is unique up to an additive constant.

²¹Since $\phi_{ij}(g) = -\phi_{ij}(\tau_{ij}(g))$, it suffices to prove that Equation (A.9) holds for g to prove that it holds for $\tau_{ij}(g)$, so we can choose to prove it for the smaller network.

A.3 Proof of Proposition 2

Let V be the structure values associated to the utility functions U . Define the utilities $\tilde{U}$ by

$$\tilde{U}_i(g) := \sum_{g' \subseteq g} V_i(g') \mathbb{1}\{i \in \mathcal{J}_{\text{src}}(g')\}, \quad (\text{A.15})$$

with corresponding values

$$\tilde{V}_i(g) := V_i(g) \mathbb{1}\{i \in \mathcal{J}_{\text{src}}(g)\}. \quad (\text{A.16})$$

By Lemma 2, these utilities are choice-equivalent to U . Therefore, to prove Proposition 2 it is sufficient to prove that $\tilde{V}$ satisfies condition (6) if and only if the game is a potential game.

First, suppose the game is a potential game with potential Φ . Fix a network $g = (s_i, g_{-i})$. Note that

$$\Phi(s_i, g_{-i}) - \Phi(\emptyset, g_{-i}) = \tilde{U}_i(s_i, g_{-i}) - \tilde{U}_i(\emptyset, g_{-i}) = \tilde{U}_i(s_i, g_{-i}). \quad (\text{A.17})$$

The last equality comes from the fact that $i \notin \mathcal{J}_{\text{src}}(\emptyset, g_{-i})$, so the payoff associated to it is 0. To simplify notation, denote with $\hat{g}_{-i}$ the network $(\emptyset, g_{-i})$. Using our inversion formula from Lemma 1 applied to the potential, define V_Φ as the unique values such that

$$\Phi(g) := \sum_{g' \subseteq g} V_\Phi(g'). \quad (\text{A.18})$$

Now, note that we can write the utility functions as

$$\tilde{U}_i(g) = \Phi(g) - \Phi(\hat{g}_{-i}) = \sum_{g' \subseteq g} V_\Phi(g') \mathbb{1}\{g' \not\subseteq \hat{g}_{-i}\}. \quad (\text{A.19})$$

Note that a structure $g' \subseteq g$ satisfies $g' \not\subseteq \hat{g}_{-i}$ if and only if $i \in \mathcal{J}_{\text{src}}(g')$. This means that

$$\tilde{U}_i(g) = \sum_{g' \subseteq g} V_\Phi(g') \mathbb{1}\{i \in \mathcal{J}_{\text{src}}(g')\}. \quad (\text{A.20})$$

Using Lemma 1, we know that the structure values for a set of utilities are unique. This implies that

$$V_\Phi(g) \mathbb{1}\{i \in \mathcal{J}_{\text{src}}(g)\} = \tilde{V}_i(g) = V_i(g) \mathbb{1}\{i \in \mathcal{J}_{\text{src}}(g)\}. \quad (\text{A.21})$$

Therefore, $V_i(g) = V_\Phi(g)$ for all $i \in \mathcal{J}_{\text{src}}(g)$.

Now suppose that $V_i(g) = V_0(g)$ for all $i \in \mathcal{J}_{\text{src}}(g)$ and let $V_0(\emptyset)$ be arbitrary. I will now prove that the game is a potential game. We can write the utilities $\tilde{U}$ as

$$\tilde{U}_i(g) = \sum_{g' \subseteq g} V_0(g') \mathbb{1}\{i \in \mathcal{J}_{\text{src}}(g')\}. \quad (\text{A.22})$$

We can prove that the game is a potential game by proposing a potential and checking that it captures the marginal utilities of deviating. Define the proposed potential as

$$\Phi_V(g) := \sum_{g' \subseteq g} V_0(g'). \quad (\text{A.23})$$

As shown in the proof of Proposition 1, it is sufficient to show that the potential captures the incentives of a single-link deviation. First, fix g such that $ij \in g$. Then we have

$$\tilde{U}_i(\tau_{ij}(g)) - \tilde{U}_i(g) = - \sum_{g' \subseteq g} V_0(g') \mathbb{1}\{i \in \mathcal{J}_{\text{src}}(g'), ij \in g'\}. \quad (\text{A.24})$$

Note that $ij \in g' \implies i \in \mathcal{J}_{\text{src}}(g')$. Therefore, we can write this marginal utility as

$$\tilde{U}_i(\tau_{ij}(g)) - \tilde{U}_i(g) = - \sum_{g' \subseteq g} V_0(g') \mathbb{1}\{ij \in g'\}. \quad (\text{A.25})$$

Similarly, we can compute the change in the potential from the deviation

$$\Phi_V(\tau_{ij}(g)) - \Phi_V(g) = - \sum_{g' \subseteq g} V_0(g') \mathbb{1}\{ij \in g'\} = \tilde{U}_i(\tau_{ij}(g)) - \tilde{U}_i(g). \quad (\text{A.26})$$

If $ij \notin g$, define $g' := \tau_{ij}(g)$ and apply the steps above to reach the same result. Therefore, the game is a potential game with potential Φ_V .

A.4 Proof of Proposition 3

First, let us prove that the existence of the potential implies reversibility of the process. By definition, the potential satisfies

$$\Phi(\tau_{ij}(g)) - \Phi(g) = U_i(\tau_{ij}(g)) - U_i(g) = \left(\frac{\sigma}{1 - \sigma} \right) \log \left(\frac{p_{ij}(g)}{p_{ij}(\tau_{ij}(g))} \right). \quad (\text{A.27})$$

The Markov process is reversible if there exists a distribution π such that the detailed balance condition is satisfied:

$$\lambda_{ij}(g)p_{ij}(g)\pi(g) = \lambda_{ij}(\tau_{ij}(g))p_{ij}(\tau_{ij}(g))\pi(\tau_{ij}(g)) \quad (\text{A.28})$$

for all $g \in \mathcal{G}_N$ and $ij \in \mathcal{D}_N$. This is sufficient since the only non-zero transition probabilities correspond to pairs of networks that differ by only one link. Since $\lambda_{ij}(\tau_{ij}(g)) = \lambda_{ij}(g)$, we have that

$$\begin{aligned} \lambda_{ij}(g)p_{ij}(g)e^{\left(\frac{1-\sigma}{\sigma}\right)\Phi(g)} &= \lambda_{ij}(\tau_{ij}(g))p_{ij}(\tau_{ij}(g)) \left[\frac{p_{ij}(g)}{p_{ij}(\tau_{ij}(g))} \right] e^{\left(\frac{1-\sigma}{\sigma}\right)\Phi(g)} \\ &= \lambda_{ij}(\tau_{ij}(g))p_{ij}(\tau_{ij}(g))e^{\left(\frac{1-\sigma}{\sigma}\right)[\Phi(\tau_{ij}(g)) - \Phi(g)]} e^{\left(\frac{1-\sigma}{\sigma}\right)\Phi(g)} \\ &= \lambda_{ij}(\tau_{ij}(g))p_{ij}(\tau_{ij}(g))e^{\left(\frac{1-\sigma}{\sigma}\right)\Phi(\tau_{ij}(g))}. \end{aligned} \quad (\text{A.29})$$

Define the following probability distribution over the space of networks:

$$\tilde{\pi}(g) = \frac{\exp \left[\left(\frac{1-\sigma}{\sigma} \right) \Phi(g) \right]}{\sum_{g' \in \mathcal{G}_N} \exp \left[\left(\frac{1-\sigma}{\sigma} \right) \Phi(g') \right]}. \quad (\text{A.30})$$

Dividing the previous expression by $\sum_{g' \in \mathcal{G}_N} \exp \left[\left(\frac{1-\sigma}{\sigma} \right) \Phi(g') \right]$, we have that

$$\lambda_{ij}(g)p_{ij}(g)\tilde{\pi}(g) = \lambda_{ij}(\tau_{ij}(g))p_{ij}(\tau_{ij}(g))\tilde{\pi}(\tau_{ij}(g)), \quad (\text{A.31})$$

so the process is reversible. In addition to Assumption 2 (from which we have irreducibility and aperiodicity), satisfying detailed balance implies that $\tilde{\pi}$ is the unique stationary distribution of the Markov chain.

Now we can prove that reversibility of the process implies existence of the potential.

Recall that under Assumption 2, there exists a unique non-degenerate stationary distribution π to which the system converges. If the process is reversible, it satisfies the detailed balance condition:

$$\lambda_{ij}(g)p_{ij}(g)\pi(g) = \lambda_{ij}(\tau_{ij}(g))p_{ij}(\tau_{ij}(g))\pi(\tau_{ij}(g)) \quad (\text{A.32})$$

Since the stationary distribution has full support and $\lambda_{ij}(g) = \lambda_{ij}(\tau_{ij}(g))$, we have that

$$\log(\pi(\tau_{ij}(g))) - \log(\pi(g)) = \log\left(\frac{p_{ij}(g)}{p_{ij}(\tau_{ij}(g))}\right). \quad (\text{A.33})$$

for all $g \in \mathcal{G}_N$ and $ij \in \mathcal{D}_N$. Since the stationary probabilities are a function of the network state only, we have that any function

$$\Phi(g) = C + \left(\frac{\sigma}{1-\sigma}\right) \log(\pi(g)), \quad (\text{A.34})$$

with $C \in \mathbb{R}$, will satisfy

$$\Phi(\tau_{ij}(g)) - \Phi(g) = \left(\frac{\sigma}{1-\sigma}\right) \log\left(\frac{p_{ij}(g)}{p_{ij}(\tau_{ij}(g))}\right) = U_i(\tau_{ij}(g)) - U_i(g), \quad (\text{A.35})$$

so the static game is a potential game.

Using the proof of the first implication, we have that the potential characterizes the stationary distribution of the chain. Namely, the stationary distribution is given by Equation (A.30).

A.5 Proof of Proposition 4

For definitions and important results, see Appendix B.

First, we want to find the set of graphons that solve the optimization problem in Theorem 3. Note that in our case, the function $\mathcal{T}$ corresponds to our scaled potential. These technically correspond to subgraph densities but, as pointed out in Appendix B, homomorphism densities and subgraph densities are equivalent for large networks. Specifically, in the proofs of Theorem 3 and 4, a subleading correction will be removed by the limsup and liminf.

Therefore, we have that for a given graphon $h \in \mathcal{W}$,

$$\mathcal{T}(h) = \sum_{m \in \mathcal{M}} \frac{a_m}{\delta_m} \tilde{b}(m, h), \quad (\text{A.36})$$

where $\tilde{b}(m, h)$ is the homomorphism density of motif m into graphon h . To solve this problem, I will first prove that this optimization problem is solved by a constant graphon.

Fix some graphon $h \in \mathcal{W}$ and a motif $m \in \mathcal{M}$. For our case of interest, Hölder's inequality states that for two functions $f, g : [0, 1]^n \rightarrow [0, 1]$ and for $p, q > 0$ such that $1/p + 1/q = 1$, we have that

$$\int_{[0,1]^n} f(\mathbf{x})g(\mathbf{x}) dx_1 \dots dx_n \leq \left(\int_{[0,1]^n} f(\mathbf{x})^p dx_1 \dots dx_n \right)^{1/p} \left(\int_{[0,1]^n} g(\mathbf{x})^q dx_1 \dots dx_n \right)^{1/q}, \quad (\text{A.37})$$

with equality if and only if there exist $c_1, c_2 \in \mathbb{R}$, not both zero, such that $c_1 f(\mathbf{x})^p = c_2 g(\mathbf{x})^q$ almost everywhere. Applying this repeatedly with appropriate choices of p and q , we have that for a collection of k functions $f_i : [0, 1]^n \rightarrow [0, 1]$, the following holds:

$$\int_{[0,1]^n} \left(\prod_{i=1}^k f_i(\mathbf{x}) \right) dx_1 \dots dx_n \leq \prod_{i=1}^k \left(\int_{[0,1]^n} f_i(\mathbf{x})^k dx_1 \dots dx_n \right)^{1/k}. \quad (\text{A.38})$$

Applying this result to the homomorphism density of m in h , we obtain

$$\begin{aligned} \tilde{b}(m, h) &= \int_{[0,1]^{n_m}} \left(\prod_{ij \in m} h(x_i, x_j) \right) dx_1 \dots dx_{n_m} \\ &\leq \prod_{ij \in m} \left(\int_{[0,1]^{n_m}} h(x_i, x_j)^{e_m} dx_1 \dots dx_{n_m} \right)^{1/e_m} \\ &= \int_{[0,1]^2} h(x, y)^{e_m} dx dy, \end{aligned} \quad (\text{A.39})$$

where $e_m = |m|$ is the number of edges in the motif m . Additionally, note that the above is always an equality if $e_m = 1$.

Recall that we assumed $a_m > 0$ for all m with $e_m > 1$. We can apply the previous result

to the motifs with $e_m > 1$ to bound our function $\mathcal{T}$ for an arbitrary graphon h :

$$\mathcal{T}(h) \leq \sum_{m \in \mathcal{M}} \frac{a_m}{\delta_m} \int_{[0,1]^2} h(x, y)^{e_m} dx dy. \quad (\text{A.40})$$

Define the function $R : [0, 1] \rightarrow \mathbb{R}$ as

$$R(s) := \sum_{m \in \mathcal{M}} \frac{a_m}{\delta_m} s^{e_m} + H(s). \quad (\text{A.41})$$

Let $\mathcal{H}(h) := \int_{[0,1]^2} H(h(x, y)) dx dy$. From the bound above, we see that

$$\mathcal{T}(h) + \mathcal{H}(h) \leq \int_{[0,1]^2} R(h(x, y)) dx dy. \quad (\text{A.42})$$

Let $R^* := \sup_s R(s)$. Using the extreme value theorem, we know this supremum is actually a maximum, and is achieved at some s^* . Therefore, we have that

$$\mathcal{T}(h) + \mathcal{H}(h) \leq \int_{[0,1]^2} R(s^*) dx dy. \quad (\text{A.43})$$

Now, the graphon h^* given by $h^*(x, y) = s^*$ saturates Hölder's inequality and, therefore, satisfies

$$\mathcal{T}(h^*) + \mathcal{H}(h^*) = \int_{[0,1]^2} R(s^*) dx dy, \quad (\text{A.44})$$

so it solves the optimization problem in Theorem 3.

Uniqueness of the solution follows the same argument as Theorem 4.1 in [Chatterjee and Diaconis \(2013\)](#). Therefore, convergence to the Erdős–Rényi model immediately follows from Theorem 4.

A.6 Proof of Theorem 1

To prove Theorem 1, we want to find the colored graphon that solves the optimization problem in Theorem 7 in Appendix B. Intuitively, we want to prove that this problem is solved by a graphon that is piece-wise constant.

Our potential is now composed of two terms $\mathcal{T}_1$ and $\mathcal{T}_2$, corresponding to the motif and neighbor utilities, respectively. These can be written as

$$\begin{aligned}\mathcal{T}_1(h, c) &= \sum_{m \in \mathcal{M}} \frac{a_m}{\delta_m} \tilde{b}(m, h), \\ \mathcal{T}_2(h, c) &= \sum_{\theta \in \Theta} \int_{[0,1]} \mathbb{1}\{c(x) = \theta\} u_\theta \left(\left[\int_{[0,1]} h(x, y) \mathbb{1}\{c(y) = \theta'\} dy \right]_{\theta' \in \Theta} \right) dx. \end{aligned} \quad (\text{A.45})$$

To simplify notation, it is convenient to define “subgraphons” for every pair of types. Without loss of generality, we can consider the coloring to be ordered²², such that for some ordering of types $\theta^1, \dots, \theta^L$, the coloring is

$$c(x) = \sum_{\theta^i \in \Theta} \theta^i \mathbb{1} \left\{ x \geq \sum_{j < i} w_{\theta^j}, x < \sum_{j \leq i} w_{\theta^j} \right\}. \quad (\text{A.46})$$

The subgraphon associated to types θ^i and θ^j is, then, $h_{\theta^i \theta^j} : [0, 1]^2 \rightarrow [0, 1]$ defined by

$$h_{\theta^i \theta^j}(x, y) := h \left(\sum_{k < i} w_{\theta^k} + w_{\theta^i} x, \sum_{k < j} w_{\theta^k} + w_{\theta^j} y \right). \quad (\text{A.47})$$

This will allow us to consider the variational problem on the whole graphon as a variational problem on the subgraphons $h_{\theta\theta'}$.

To begin, we fix a colored graphon (h, c) and find a bound on $\mathcal{T}_1(h)$ in a similar manner to before. For a given motif m , we can break up the integration domain into “boxes” where all the types are fixed. Specifically, we can write

$$\begin{aligned}\tilde{b}(m, h) &= \sum_{\theta \in \Theta^{n_m}} \int_{[0,1]^{n_m}} \mathbb{1}\{c(x_k) = \theta_k \forall k\} \left(\prod_{ij \in m} h(x_i, x_j) \right) dx_1 \dots dx_{n_m} \\ &= \sum_{\theta \in \Theta^{n_m}} \left(\prod_{i \in \mathcal{N}_m} w_{\theta_i} \right) \int_{[0,1]^{n_m}} \left(\prod_{ij \in m} h_{\theta_i \theta_j}(x_i, x_j) \right) dx_1 \dots dx_{n_m}. \end{aligned} \quad (\text{A.48})$$

²²This is because $\mathcal{T}$ is the same for all measurable re-labelings of the nodes.

We will now use the following result, obtained by iterating Hölder's inequality:

$$\int_{[0,1]^{n_m}} \left(\prod_{ij \in m} f_{ij}(x_i, x_j) \right) dx_1 \dots dx_{n_m} \leq \prod_{ij \in m} \|f_{ij}\|_{e_m}. \quad (\text{A.49})$$

Applying this to the homomorphism density, we obtain

$$\tilde{b}(m, h) \leq \sum_{\theta \in \Theta^{n_m}} \left(\prod_{i \in \mathcal{N}_m} w_{\theta_i} \right) \prod_{ij \in m} \|h_{\theta_i \theta_j}\|_{e_m}. \quad (\text{A.50})$$

We can use this to bound $\mathcal{T}_1$. Note that we can assume without loss of generality that $e_m > 1$ for all motifs m , since any motifs with $e_m = 1$ can be absorbed into the linear part of the neighborhood utility. Recall that we assumed that $a_m > 0$ for all motifs with $e_m > 1$. Then, we have that $a_m > 0$ for all m , so we obtain the bound

$$\mathcal{T}_1(h, c) \leq \sum_{m \in \mathcal{M}} \frac{a_m}{\delta_m} \sum_{\theta \in \Theta^{n_m}} \left(\prod_{i \in \mathcal{N}_m} w_{\theta_i} \right) \prod_{ij \in m} \|h_{\theta_i \theta_j}\|_{e_m}. \quad (\text{A.51})$$

Note that this bound has different norms based on the size of the motif. To obtain a bound that uses a single norm for all graphons, we can use the fact that for a space with $\|1\|_1 = 1$, for $1 \leq p \leq q < \infty$, the corresponding norms satisfy $\|h\|_p \leq \|h\|_q$, with equality iff h is constant almost everywhere. Define $e^* := \max_{m \in \mathcal{M}} e_m$. Then we have the following bound:

$$\mathcal{T}_1(h, c) \leq \sum_{m \in \mathcal{M}} \frac{a_m}{\delta_m} \sum_{\theta \in \Theta^{n_m}} \left(\prod_{i \in \mathcal{N}_m} w_{\theta_i} \right) \prod_{ij \in m} \|h_{\theta_i \theta_j}\|_{e^*}. \quad (\text{A.52})$$

To find a bound on $\mathcal{T}_2$, I will use the following technical Lemma.

Lemma 5. *For any $n \geq 1$, $h \in \mathcal{W}$ and $\kappa \in \mathbb{R}$,*

$$\int_{[0,1]^2} [H(h(x, y)) + \kappa h(x, y)] dx dy \leq H(\|h\|_n) + \kappa \|h\|_n. \quad (\text{A.53})$$

Proof. Note that the function $G(u) := H(u) + \kappa u$ has a unique maximizer, which I call $\rho^*(\kappa)$.

I consider the cases $\|h\|_n \leq \rho^*(\kappa)$ and $\|h\|_n > \rho^*(\kappa)$ separately.

First, consider the case $\|h\|_n \leq \rho^*(\kappa)$. Since G is a concave function, we can use Jensen's

inequality to obtain

$$\int_{[0,1]^2} G(h(x,y)) dx dy \leq G(\|h\|_1). \quad (\text{A.54})$$

Since $G(u)$ is non-decreasing for $u \leq \rho^*(\kappa)$ and $\|h\|_1 \leq \|h\|_n$, we obtain the bound for this case.

For the case $\|h\|_n > \rho^*(\kappa)$, I show that the optimization problem

$$\max_{h' \in \mathcal{W}} \int [H(h'(x,y)) + \kappa h'(x,y)] dx dy \quad \text{s.t.} \quad \|h'\|_n \geq \rho, \quad (\text{A.55})$$

for $\rho := \|h\|_n > \rho^*(\kappa)$, is solved by the constant graphon $h'(x,y) = \rho$. Note that G satisfies

$$\lim_{u \rightarrow 0} G'(u) = +\infty, \quad \lim_{u \rightarrow 1} G'(u) = -\infty. \quad (\text{A.56})$$

This means that the constraints $h'(x,y) \in [0,1]$ will not be binding. Using Theorem 9.4.1 in [Luenberger \(1998\)](#), we have that the solution h_n must be a stationary point of the Lagrangian

$$\mathcal{L}(h') := \int [H(h'(x,y)) + \kappa h'(x,y) + \mu(\rho^n - h'(x,y)^n)] dx dy, \quad (\text{A.57})$$

for some Lagrange multiplier μ . This means that it must satisfy

$$H'(h_n(x,y)) + \kappa - n\mu h_n(x,y)^{n-1} = 0, \quad (\text{A.58})$$

with $\mu \geq 0$ and the complementary slackness condition

$$\mu \left[\rho^n - \int_{[0,1]^2} h_n(x,y)^n dx dy \right] = 0. \quad (\text{A.59})$$

Note that we must have $\mu > 0$, since $H'(h_n) + \kappa = 0$ would imply $h_n = \rho^*(\kappa) \implies \|h_n\|_n = \rho^*(\kappa)$, which violates the constraint. With $\mu > 0$, the function $H'(u) + \kappa - n\mu u^{n-1}$ is strictly decreasing, and diverges as it approaches 0 and 1, so it must have a unique root. This means that there is a unique solution to the problem where the constraint binds, corresponding to

$h_n = \rho$. Since our original graphon is in the feasible set of this problem, it is bounded by the value of the functional at the constant graphon. This yields the bound for the second case. $\square$

Let us write $\mathcal{T}_2$ in terms of the decomposition of u :

$$\mathcal{T}_2(h, c) = \sum_{\theta \in \Theta} w_\theta \int_{[0,1]} \left[\sum_{\theta' \in \Theta} c_{\theta\theta'} w_{\theta'} \int_{[0,1]} h_{\theta\theta'}(x, y) dy + \tilde{u}_\theta \left(\left[w_{\theta'} \int_{[0,1]} h_{\theta\theta'}(x, y) dy \right]_{\theta' \in \Theta} \right) \right] dx. \quad (\text{A.60})$$

In addition to this, we can also write the entropy term in terms of the subgraphons:

$$\mathcal{H}[h] = \sum_{\theta, \theta' \in \Theta} w_\theta w_{\theta'} \int_{[0,1]^2} H(h_{\theta\theta'}(x, y)) dx dy. \quad (\text{A.61})$$

Note that the Lemma 5 allows us to jointly bound $\mathcal{T}_2 + \mathcal{H}$. Specifically, we obtain

$$\begin{aligned} \mathcal{T}_2(h, c) + \mathcal{H}(h) &\leq \sum_{\theta, \theta' \in \Theta} w_\theta w_{\theta'} [H(\|h_{\theta\theta'}\|_{e^*}) + c_{\theta\theta'} \|h_{\theta\theta'}\|_{e^*}] \\ &\quad + \sum_{\theta \in \Theta} w_\theta \int_{[0,1]} \tilde{u}_\theta \left(\left[w_{\theta'} \int_{[0,1]} h_{\theta\theta'}(x, y) dy \right]_{\theta' \in \Theta} \right) dx. \end{aligned} \quad (\text{A.62})$$

Using concavity of the $\tilde{u}_\theta$ functions, we can apply Jensen's inequality to obtain

$$\int_{[0,1]} \tilde{u}_\theta \left(\left[w_{\theta'} \int_{[0,1]} h_{\theta\theta'}(x, y) dy \right]_{\theta' \in \Theta} \right) \leq \tilde{u}_\theta \left(\left[w_{\theta'} \int_{[0,1]^2} h_{\theta\theta'}(x, y) dx dy \right]_{\theta' \in \Theta} \right) = \tilde{u}_\theta ([w_{\theta'} \|h_{\theta\theta'}\|_1]_{\theta' \in \Theta}). \quad (\text{A.63})$$

Since $\|h_{\theta\theta'}\|_1 \leq \|h_{\theta\theta'}\|_{e^*}$, using monotonicity of $\tilde{u}_\theta$ yields

$$\int_{[0,1]} \tilde{u}_\theta \left(\left[w_{\theta'} \int_{[0,1]} h_{\theta\theta'}(x, y) dy \right]_{\theta' \in \Theta} \right) \leq \tilde{u}_\theta ([w_{\theta'} \|h_{\theta\theta'}\|_{e^*}]_{\theta' \in \Theta}). \quad (\text{A.64})$$

Together with the previous bound, this yields

$$\mathcal{T}_2(h, c) + \mathcal{H}(h) \leq \sum_{\theta \in \Theta} w_\theta \left[\sum_{\theta' \in \Theta} w_{\theta'} [H(\|h_{\theta\theta'}\|_{e^*}) + u_\theta([w_{\theta'} \|h_{\theta\theta'}\|_{e^*}]_{\theta' \in \Theta})] \right]. \quad (\text{A.65})$$

As in the main text, let $\mathcal{K}_\Theta$ be the set of functions $\psi : \Theta^2 \rightarrow [0, 1]$. For $\psi \in \mathcal{K}_\Theta$, define the function

$$Q(\psi) := \sum_{m \in \mathcal{M}} \frac{a_m}{\delta_m} \sum_{\theta \in \Theta^{n_m}} \left(\prod_{i \in \mathcal{N}_m} w_{\theta_i} \right) \prod_{ij \in m} \psi_{\theta_i \theta_j} + \sum_{\theta \in \Theta} w_\theta \left[\sum_{\theta' \in \Theta} w_{\theta'} [H(\psi_{\theta\theta'}) + u_\theta([w_{\theta'} \psi_{\theta\theta'}]_{\theta' \in \Theta})] \right]. \quad (\text{A.66})$$

Then the bounds above imply that, for an arbitrary colored graphon (h, c) ,

$$\mathcal{T}(h, c) + \mathcal{H}(h) \leq Q([\|h_{\theta\theta'}\|_{e^*}]_{\theta, \theta' \in \Theta}). \quad (\text{A.67})$$

Define ψ^* to be

$$\psi^* \in \operatorname{argmax}_{\psi \in \mathcal{K}_\Theta} Q(\psi). \quad (\text{A.68})$$

This optimality implies that any graphon satisfies the following bound:

$$\mathcal{T}(h, c) + \mathcal{H}(h) \leq Q(\psi^*). \quad (\text{A.69})$$

Let h^* be the piecewise constant graphon given by $h_{\theta\theta'}^*(x, y) = \psi_{\theta\theta'}^*$ for all θ, θ', x, y . Note that this graphon satisfies $\|h_{\theta\theta'}^*\|_n = \psi_{\theta\theta'}^*$ for all $n \geq 1$. In addition, all the inequalities used to construct the bounds above are equalities for this case, so we conclude that

$$\mathcal{T}(h^*, c) + \mathcal{H}(h^*) = Q(\psi^*). \quad (\text{A.70})$$

Therefore, the graphon h^* solves the variational problem in Theorem 7.

To show that only graphons that are constant almost everywhere solve the problem, assume for the sake of contradiction that there is a solution $\hat{h}$ that is not constant almost

everywhere. Then the inequalities above are strict, and we would obtain a strict improvement by “flattening” the solution in a way that preserves the norms of the subgraphons $\|\hat{h}_{\theta\theta'}\|_{e^*}$.

The convergence to the directed stochastic block model when the optimizer is unique is a direct consequence of Theorem 8.

B Graph Limits

In this section, I present the relevant definitions and results for the theory of large dense graphs. The results in Sections B.1 and B.2 are presented without proof. These are taken from Appendix D in Mele (2017), which is an excellent introduction to the topic, and builds on the results in Chatterjee and Varadhan (2011) and Chatterjee and Diaconis (2013). The reader is invited to read this, and the references therein, for a detailed discussion on the topic²³. Section B.3 extends these results to colored graphs, allowing us to characterize the limiting behavior of the model with heterogeneous agents. In order to make a clearer connection to the literature on graph limits, the notation in this section will differ from the rest of the paper.

B.1 Graphons

This section presents the relevant definitions for the theory of graph limits, following the outline presented in Chatterjee and Diaconis (2013) adapted to directed graphs. For this appendix, a directed graph is an ordered pair $G = (V, A)$, where V is a set of vertices and A is a set of arcs (directed edges). Consider a sequence G_N of simple directed graphs whose number of nodes tends to infinity. For every fixed simple graph H , let $|\text{hom}(H, G)|$ denote the number of homomorphisms of H into G , which is the number of edge-preserving maps from $V(H)$ into $V(G)$. That is, a map $\varphi : V(H) \rightarrow V(G)$ is a homomorphism if $(i, j) \in A(H) \implies (\varphi(i), \varphi(j)) \in A(G)$. Normalizing by the possible number of maps from $V(H)$ to $V(G)$ yields the *homomorphism density*

$$t(H, G) := \frac{|\text{hom}(H, G)|}{|V(G)|^{|V(H)|}}, \quad (\text{B.1})$$

which corresponds to the probability of a uniformly random mapping $V(H) \rightarrow V(G)$ being a homomorphism.

The importance of analyzing homomorphism densities is twofold. First, they provide a way to “probe” large graphs in order to understand their properties. Second, they’re relevant

²³Also see Lovász (2012) for a broader introduction to graph limits.

in our context since motif utilities can be approximated as homomorphism densities for large graphs, as is discussed below.

Motif utilities do not depend explicitly on the number of homomorphisms of a motif into the network, but rather the number of times it appears as a subgraph. This can be captured using the concept of *subgraph densities*. Let $|\text{sub}(H, G)|$ be the number of *injective* maps from $V(H)$ to $V(G)$ that are homomorphisms. Clearly $|\text{sub}(H, G)| \leq |\text{hom}(H, G)|$. The normalized motif densities in our model are given by subgraph densities, defined by

$$s(H, G) := \frac{|\text{sub}(H, G)|}{|V(G)|^{|V(H)|}}. \quad (\text{B.2})$$

It can be shown that

$$t(H, G) - \frac{1}{|V(G)|} \binom{|V(H)|}{2} \leq s(H, G) \leq t(H, G). \quad (\text{B.3})$$

Therefore, for a given H , characterizing homomorphism densities for large graphs is equivalent to characterizing subgraph densities.

A possible notion of convergence for the sequence $\{G_N\}$ is that the densities $t(H, G_N)$ converge to some value for every finite graph H . The work of Lovász and coauthors (see [Lovász \(2012\)](#) for an overview) established the existence of an object that characterizes this convergence for undirected graphs. That is, there exists an object from which the limiting homomorphism densities $t(H, \cdot)$ can be obtained. An analogous result for directed graphs was established in [Boeckner \(2013\)](#). This “limiting object” is a function $h \in \mathcal{W}$, where $\mathcal{W}$ is the set of measurable functions $[0, 1]^2 \rightarrow [0, 1]$. These objects are called “graphons”. Conversely, every function in $\mathcal{W}$ arises as the limit of an appropriate sequence of directed graphs. If H is a simple directed graph with $V(H) = \{1, \dots, k\}$, the homomorphism density of H into h is defined as

$$t(H, h) := \int_{[0, 1]^k} \left(\prod_{(i, j) \in A(H)} h(x_i, x_j) \right) dx_1 \dots dx_k. \quad (\text{B.4})$$

A sequence of graphs $\{G_N\}$ is said to converge to h if

$$\lim_{N \rightarrow \infty} t(H, G_N) = t(H, h) \quad (\text{B.5})$$

for all finite simple directed graphs H .

The intuition is that as $N \rightarrow \infty$, the interval $[0, 1]$ represents a “continuum” of vertices and $h(x, y)$ is the probability that there is an arc going from x to y .²⁴ For the case of directed Erdős–Rényi graphs $G(N, p)$ with fixed p , the limit graph is represented by the graphon that is equal to p for all $(x, y) \in [0, 1]^2$. For a fixed $p \in (0, 1)$, this corresponds to a dense random graph model.

Finite simple graphs have a canonical representation as graphons. For a given graph G over N nodes, its associated graphon is given by

$$h^G(x, y) := \mathbb{1}\{(\lceil Nx \rceil, \lceil Ny \rceil) \in A(G)\}. \quad (\text{B.6})$$

The notion of convergence in $\mathcal{W}$ in terms of homomorphism densities can be metrized using the *cut distance*. For two graphons h_1 and h_2 , it is defined as

$$d_{\square}(h_1, h_2) := \sup_{S, T \subseteq [0, 1]} \left| \int_{S \times T} [h_1(x, y) - h_2(x, y)] dx dy \right|. \quad (\text{B.7})$$

For our purposes, it will be useful to work in a different space that has some useful topological properties. Let Σ be the set of measure-preserving bijections $\sigma : [0, 1] \rightarrow [0, 1]$. Define an equivalence relation on $\mathcal{W}$ by setting $h_1 \sim h_2$ if $h_1(x, y) = h_2^{\sigma}(x, y) := h_2(\sigma x, \sigma y)$ for some $\sigma \in \Sigma$. Let $\tilde{h}$ be the closure of the orbit $\{h^{\sigma}\}$ in $(\mathcal{W}, d_{\square})$. Let $\tilde{\mathcal{W}} := \mathcal{W} / \sim$ and let τ be the map $\tau h \mapsto \tilde{h}$. The distance $\delta_{\square}$ on the space $\tilde{\mathcal{W}}$ is defined as

$$\delta_{\square}(\tilde{h}_1, \tilde{h}_2) = \inf_{\sigma} d_{\square}(h_1, h_2^{\sigma}), \quad (\text{B.8})$$

such that $(\tilde{\mathcal{W}}, \delta_{\square})$ is a metric space. For any finite directed graph G , let G be its associated graphon and let $\tilde{G} := \tilde{h}^G \in \tilde{\mathcal{W}}$ be its corresponding orbit.

²⁴This intuition is more explicit in the case of W random graphs.

An important aspect of homomorphism densities is that they are continuous functions in this space. This is useful in the analysis of the limiting behavior of ERGMs, as discussed below.

B.2 Results on graph limits

A central tool used in characterizing the limiting behavior of ERGMs is the theory of large deviations for random graph models. Fix some $p \in (0, 1)$. Let $\tilde{\mathbb{P}}_{N,p}$ be the measure induced on $\tilde{\mathcal{W}}$ by the directed Erdős–Rényi model with parameter p . Additionally, define the function $I_p : [0, 1] \rightarrow \mathbb{R}$ as

$$I_p(u) := u \log \left(\frac{u}{p} \right) + (1 - u) \log \left(\frac{1 - u}{1 - p} \right). \quad (\text{B.9})$$

The domain of this function can be extended to $\tilde{\mathcal{W}}$ by defining

$$\mathcal{I}_p(\tilde{h}) := \int_{[0,1]^2} I_p(h(x, y)) \, dx \, dy \quad (\text{B.10})$$

for any $h \in \tau^{-1}(\tilde{h})$. Building off the results in [Chatterjee and Varadhan \(2011\)](#), we have the following result for directed graphs.

Theorem 2 (Theorem 8 in [Mele \(2017\)](#)). *For each fixed $p \in (0, 1)$, the sequence $\tilde{\mathbb{P}}_{N,p}$ satisfies a large deviation principle on the space $(\tilde{\mathcal{W}}, \delta_\square)$ at speed N^2 with rate function $\mathcal{I}_p$. Explicitly, for any closed set $\tilde{F} \subseteq \tilde{\mathcal{W}}$,*

$$\limsup_{N \rightarrow \infty} \frac{1}{N^2} \log(\tilde{\mathbb{P}}_{N,p}(\tilde{F})) \leq - \inf_{\tilde{h} \in \tilde{F}} \mathcal{I}_p(\tilde{h}) \quad (\text{B.11})$$

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

$$\liminf_{N \rightarrow \infty} \frac{1}{N^2} \log(\tilde{\mathbb{P}}_{N,p}(\tilde{U})) \geq - \inf_{\tilde{h} \in \tilde{U}} \mathcal{I}_p(\tilde{h}). \quad (\text{B.12})$$

Now suppose we have a sequence of measures π_N on $\mathcal{G}_N$ given by

$$\pi_N(G) = \exp\{N^2(\mathcal{T}(\tilde{G}) - \psi_N)\}, \quad (\text{B.13})$$

where ψ_N is a normalization constant given by

$$\psi_N := \frac{1}{N^2} \log \left(\sum_{G \in \mathcal{G}_N} e^{N^2 \mathcal{T}(\tilde{G})} \right). \quad (\text{B.14})$$

In the case of [Mele \(2017\)](#) and our motif utilities, the function $\mathcal{T}$ is a linear combination of homomorphism densities. Define the *entropy functional* $\mathcal{H} : \tilde{\mathcal{W}} \rightarrow \mathbb{R}$

$$\mathcal{H}(\tilde{h}) = \int_{[0,1]^2} H(h(x, y)) \, dx \, dy \quad (\text{B.15})$$

for some $h \in \tau^{-1}(\tilde{h})$ and H is given by

$$H(u) := -u \log(u) - (1 - u) \log(1 - u). \quad (\text{B.16})$$

Adapting the results in [Chatterjee and Diaconis \(2013\)](#), [Mele \(2017\)](#) finds the following result on the normalization constant of these models.

Theorem 3 (Theorem 10 in [Mele \(2017\)](#)). *If $\mathcal{T} : \tilde{\mathcal{W}} \rightarrow \mathbb{R}$ is a bounded continuous function, then*

$$\psi := \lim_{N \rightarrow \infty} \psi_N = \sup_{\tilde{h} \in \tilde{\mathcal{W}}} \left[\mathcal{T}(\tilde{h}) + \mathcal{H}(\tilde{h}) \right]. \quad (\text{B.17})$$

Using the properties of the space $\tilde{\mathcal{W}}$, an even stronger result on convergence can be obtained. This allows us to characterize the limiting behavior of the statistical properties of ERGMs.

Theorem 4 (Theorem 18 in [Mele \(2017\)](#)). *Let $\tilde{M}^*$ be the set of maximizers of the variational problem (B.17). 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 ,

$$\mathbb{P}\{\delta_{\square}(\tilde{G}_N, \tilde{M}^*) > \eta\} \leq Ce^{-N^2\kappa},$$

where $\mathbb{P}$ denotes the probability measure implied by the model.

Building on these results, we can characterize the limiting behavior of models with heterogeneity in the nodes.

B.3 Colored graphons

In order to account for node heterogeneity, I use the framework of *colored graphs*, particularly as laid out in [Diao et al. \(2016\)](#). Let Θ be a set of colors (or types in our case). A directed colored graph Q is a tuple (V, A, C) , where $C : V \rightarrow \Theta$ represents the coloring of the vertices of Q . Finally, let $\mathcal{Q}_N^\Theta$ be the set of colored graphs with color set Θ .

Similar to the case of uncolored graphs, we can consider a way to “probe” colored graphs using the homomorphism densities of other colored graphs into them. Let R and Q be directed colored graphs. We define $|\text{hom}_\Theta(R, Q)|$ to be the number of homomorphisms of $(V(R), A(R))$ into $(V(Q), A(Q))$ that preserve the coloring of the vertices. The *colored homomorphism density* of R into Q , then, is

$$t_\Theta(R, Q) := \frac{|\text{hom}_\Theta(R, Q)|}{|V(Q)|^{|V(R)|}}. \quad (\text{B.18})$$

As in the uncolored case, we can characterize the convergence of a sequence of colored graphs by studying their colored homomorphism densities. In order to do this, we need to define appropriate limiting objects.

A colored directed graphon is a pair $q := (h[q], c[q])$, where $h[q] \in \mathcal{W}$ is a directed graphon and $c[q]$ is a measurable function $c[q] : [0, 1] \rightarrow \Theta$, which I denote the *coloring*. The space of colored directed graphons is denoted with $\mathcal{W}_\Theta$. The colored homomorphism density of R

into q , assuming $V(R) = \{1, \dots, k\}$, is defined as

$$t_\Theta(R, q) := \int_{[0,1]^k} \left(\prod_{(i,j) \in A(R)} h[q](x_i, x_j) \right) \left(\prod_{i=1}^k \mathbb{1}\{c[q](x_i) = C(R)_i\} \right) dx_1 \dots dx_k. \quad (\text{B.19})$$

As in the uncolored case, there is a canonical representation of colored graphs as colored graphons. For a given colored graph Q , its associated graphon $q^Q = (h[q^Q], c[q^Q])$ is given by

$$h[q^Q] = h^{(V(Q), A(Q))}, \quad c[q^Q](x) = C(Q)_{\lceil V(Q)|x \rceil}. \quad (\text{B.20})$$

Note that for the uncolored graph $G = (V(Q), A(Q))$, we have $h[q^Q] = h^G$.

We can also construct a topology in the space of colored graphons. In order to do this, we need a new notion of distance. The *colored cut distance* between two graphons is given by

$$d_\Theta^\Theta(q_1, q_2) := d_\square(h[q_1], h[q_2]) + d_\Theta(c[q_1], c[q_2]), \quad (\text{B.21})$$

where

$$d_\Theta(c, c') := \frac{1}{2} \sum_{\theta \in \Theta} \int_{[0,1]} \mathbb{1}\{x \in c^{-1}(\theta) \Delta c'^{-1}(\theta)\} dx \quad (\text{B.22})$$

is a distance on the space of colorings.²⁵ The equivalence class in this space is defined similarly to the uncolored case, but we have to take into account that the bijections σ also act on the colorings. Define the equivalence relation $\sim_\Theta$ by setting $q_1 \sim_\Theta q_2$ if $h[q_1](x, y) = h[q_2^\sigma](x, y) = h[q_2](\sigma x, \sigma y)$ and $c[q_1](x) = c[q_2^\sigma](x) = c[q_2](\sigma x)$ for some $\sigma \in \Sigma$. Define $\tilde{q}$ to be the equivalence class of q , and let $\tilde{\mathcal{W}}_\Theta := \mathcal{W}_\Theta / \sim_\Theta$. Additionally, let τ_Θ be the map such that $\tau_\Theta(q) = \tilde{q}$. For any finite colored graph Q , let $\tilde{Q} := \tilde{q}^Q \in \tilde{\mathcal{W}}_\Theta$ be its corresponding orbit.

²⁵Equivalently, it is the Lebesgue measure of the set where c and c' do not agree.

The natural distance on this space is given by

$$\delta_{\square}^{\Theta}(\tilde{q}_1, \tilde{q}_2) := \inf_{\sigma} d_{\square}^{\Theta}(q_1, q_2^{\sigma}). \quad (\text{B.23})$$

It is also convenient to define an analogous equivalence on the space of colorings. Let $\mathcal{C}$ be the set of measurable functions $c : [0, 1] \rightarrow \Theta$ and define $\tilde{\mathcal{C}}$ as the corresponding space of orbits. Define the coloring distance in this space as

$$\delta_{\Theta}(\tilde{c}_1, \tilde{c}_2) = \inf_{\sigma} d_{\Theta}(c_1, c_2^{\sigma}). \quad (\text{B.24})$$

Note that every coloring orbit $\tilde{c}$ is uniquely identified (up to null sets) with a color vector $\mathbf{w}(\tilde{c}) := (\lambda(c^{-1}(\theta))_{\theta \in \Theta} \in \Delta(\Theta)$, where λ is the Lebesgue measure, for any representative $c \in \tilde{c}$. Therefore, we can also uniquely define a coloring orbit $\tilde{c}_{\mathbf{w}}$ for a given color vector $\mathbf{w}$. This equivalence with color vectors allows us to prove the following property of the space of colorings:

Lemma 6. *The space $(\tilde{\mathcal{C}}_{\Theta}, \delta_{\Theta})$ is compact.*

Proof. To prove this, I show that $(\tilde{\mathcal{C}}_{\Theta}, \delta_{\Theta})$ is isometric, up to a factor,9 with $\Delta(\Theta)$ equipped with the L^1 norm. This is a closed and bounded subset of $\mathbb{R}^{|\Theta|}$, hence compact by Heine-Borel.

Consider two colorings $\tilde{c}$ and $\tilde{c}'$. The sets that are assigned to θ cannot overlap over more than $\min\{w_{\theta}(\tilde{c}), w_{\theta}(\tilde{c}')\}$. This implies that $\delta_{\Theta}(\tilde{c}, \tilde{c}') \geq \|\mathbf{w}(\tilde{c}) - \mathbf{w}(\tilde{c}')\|_1$. Additionally, note that this bound can be achieved by ordering both colorings by intervals of length $\min\{w_{\theta}(\tilde{c}), w_{\theta}(\tilde{c}')\}$, and having the remaining intervals be mismatched. Hence,

$$\delta_{\Theta}(\tilde{c}, \tilde{c}') = \frac{1}{2} \|\mathbf{w}(\tilde{c}) - \mathbf{w}(\tilde{c}')\|_1. \quad (\text{B.25})$$

This proves the isometry (with a factor of 1/2), which implies compactness of $(\tilde{\mathcal{C}}_{\Theta}, \delta_{\Theta})$. $\square$

The space $(\tilde{\mathcal{W}}_{\Theta}, \delta_{\square}^{\Theta})$ has similar topological properties to $(\tilde{\mathcal{W}}, \delta_{\square})$. There is one particular result that is of interest to us, since it allows us to characterize the limiting behavior of the measures induced by random graph models.

Theorem 5. *The space $(\tilde{\mathcal{W}}_\Theta, \delta_\square^\Theta)$ is compact.*

Proof. This proof follows similarly to the proof of Theorem 3.7 in [Diao et al. \(2016\)](#) for the undirected case. The strategy is to prove that any sequence of colored graphons $q_1, q_2, \dots$ has a convergent subsequence. Since $(\tilde{\mathcal{W}}_\Theta, \delta_\square^\Theta)$ is a metric space, this would imply it is compact.

First, note that there is a sequence of measure-preserving bijections $\sigma_n : [0, 1] \rightarrow [0, 1]$ such that the colored graphons $q_n^{\sigma_n}$ have colorings $c[q_n^{\sigma_n}]$ that are intervals that follow some fixed order of Θ . Therefore, without loss of generality, we can consider that the colored graphons q_n already have colorings arranged by intervals. Since the vectors $\mathbf{w}(\tilde{c}[q_n])$ lie on the simplex (which is compact), there exists a subsequence n_k that converges to some color vector $\mathbf{w}_0$. This implies that the limit

$$\lim_{k \rightarrow \infty} c[q_{n_k}](x) =: c_0(x) \quad (\text{B.26})$$

exists almost everywhere, and $\lambda(c[q_{n_k}]^{-1}(\theta)\Delta c_0^{-1}(\theta)) \rightarrow 0$ for all $\theta \in \Theta$.

The proof now follows the proof of Lemma 5 in [Mele \(2017\)](#), which adapts the proof of Theorem 9.23 in [Lovász \(2012\)](#) to directed graphons. Since we know we can take a subsequence whose coloring converges from any sequence of colored graphons, suppose the colorings $c[q_n]$ converge. Now apply the argument from the proof of Lemma 5 in [Mele \(2017\)](#) to the graphons $h[q_n]$. Note that the partitions $\mathcal{P}_{n,1}$ can be chosen to respect the partition defined by $c[q_n]^{-1}(\theta)$ for $\theta \in \Theta$. Since for $k > 1$ the partitions $\mathcal{P}_{n,k}$ are refinements of this one, they also respect the coloring partition. The procedure in that proof then gives us a subsequence q_{n_l} such that $d_\square(h[q_{n_l}], h_0) \rightarrow 0$ for some graphon h_0 .²⁶ Now, define the graphon $q_0 := (h_0, c_0)$. We have that

$$\delta_\square^\Theta(q_n, q_0) \leq d_\square(h[q_n], h_0) + d_\Theta(c[q_n], c_0). \quad (\text{B.27})$$

Under the subsequence $(n_l)_l$, both terms on the right vanish, meaning $q_{n_l} \rightarrow q_0$ in the metric space $(\tilde{\mathcal{W}}_\Theta, \delta_\square^\Theta)$. □

²⁶The proof of Lemma 5 in [Mele \(2017\)](#) technically yields a sequence that converges under $\delta_\square$. However, it invokes Lemmas 3.1.20 and 3.1.21 in [Boeckner \(2013\)](#), which state properties of convergence under $d_\square$. Using these results with $d_\square$ and carrying out the argument in the proof of Lemma 5 in [Mele \(2017\)](#) yields $d_\square(h[q_{n_l}], h_0) \rightarrow 0$.

Having this topological property, we can now state the result that allows for the generalization of the graph convergence results to the colored case. Consider a sequence of empirical fractions of colors $\mathbf{w}^N$ that converges to some $\mathbf{w}$. Consider the measures $\tilde{\mathbb{P}}_{N,p}^\Theta$ on $\tilde{\mathcal{W}}_\Theta$ generated by the following procedure:

1. Take the set of N nodes $\{1, \dots, N\}$ and deterministically fix a coloring such that the coloring fractions are $\mathbf{w}_N$.
2. Draw a directed Erdős–Rényi random graph on the nodes with parameter p .

Note that the specific coloring of the nodes does not matter, since one could relabel the nodes and run the same procedure without affecting the measures. The next theorem characterizes the limiting behavior of these measures, which later allows us to characterize the asymptotic behavior of the Gibbs measures for Theorem 1.

Theorem 6. *For each fixed $p \in (0, 1)$, the sequence $\tilde{\mathbb{P}}_{N,p}^\Theta$ satisfies a large deviation principle in the space $(\tilde{\mathcal{W}}_\Theta, \delta_\square^\Theta)$ at speed N^2 with rate function*

$$\mathcal{J}_{\mathbf{w},p}^\Theta(\tilde{q}) = \begin{cases} \mathcal{I}_p(\tilde{h}[\tilde{q}]) & \text{if } \tilde{c}[\tilde{q}] = \tilde{\mathbf{c}}_{\mathbf{w}}, \\ \infty & \text{otherwise,} \end{cases} \quad (\text{B.28})$$

where $\mathcal{I}_p$ is defined in Equation (B.10). Explicitly, for any closed set $\tilde{F} \subseteq \tilde{\mathcal{W}}_\Theta$,

$$\limsup_{N \rightarrow \infty} \frac{1}{N^2} \log(\tilde{\mathbb{P}}_{N,p}^\Theta(\tilde{F})) \leq - \inf_{\tilde{q} \in \tilde{F}} \mathcal{J}_{\mathbf{w},p}^\Theta(\tilde{q}) \quad (\text{B.29})$$

and for any open set $\tilde{U} \subseteq \tilde{\mathcal{W}}_\Theta$,

$$\liminf_{N \rightarrow \infty} \frac{1}{N^2} \log(\tilde{\mathbb{P}}_{N,p}^\Theta(\tilde{U})) \geq - \inf_{\tilde{q} \in \tilde{U}} \mathcal{J}_{\mathbf{w},p}^\Theta(\tilde{q}). \quad (\text{B.30})$$

Proof. In order to prove the LDP result on the space $\tilde{\mathcal{W}}_\Theta$, I will use the fact that an LDP is satisfied for the uncolored graphons. Given the structure of the process that generates the random colored graphons, enough structure is preserved (as needed in Lemma 7) to conclude that these also satisfy an LDP.

First, I will prove that an LDP is satisfied in the product space $\mathcal{W} \times \mathcal{C}_\Theta$, with the product topology. Define the function $\Psi : \tilde{\mathcal{W}}_\Theta \rightarrow \tilde{\mathcal{W}} \times \tilde{\mathcal{C}}_\Theta$ by

$$\Psi(\tilde{q}) := (\tilde{h}[\tilde{q}], \tilde{c}[\tilde{q}]). \quad (\text{B.31})$$

This function “separates” the orbit of a colored graphon into its uncolored graphon and its coloring. Clearly, many colored graphons can have the same components $(\tilde{h}, \tilde{c})$, so the function Ψ is not injective.

Note that the product topology can be induced by the metric

$$\delta_\otimes^\Theta((\tilde{h}_1, \tilde{c}_1), (\tilde{h}_2, \tilde{c}_2)) := \delta_\square(\tilde{h}_1, \tilde{h}_2) + \delta_\Theta(\tilde{c}_1, \tilde{c}_2). \quad (\text{B.32})$$

It is easy to see that Ψ is continuous using this metric. Indeed, by definition of $\delta_\square^\Theta$,

$$\begin{aligned} \delta_\square^\Theta(\tilde{q}_1, \tilde{q}_2) &= \inf_{\sigma} [d_\square(h[q_1], h[q_2^\sigma]) + d_\Theta(c[q_1], c[q_2^\sigma])] \\ &\geq \delta_\otimes^\Theta((\tilde{h}[\tilde{q}_1], \tilde{c}[\tilde{q}_1]), (\tilde{h}[\tilde{q}_2], \tilde{c}[\tilde{q}_2])) \\ &= \delta_\otimes^\Theta(\Psi(\tilde{q}_1), \Psi(\tilde{q}_2)), \end{aligned} \quad (\text{B.33})$$

the mapping is Lipschitz and, therefore, continuous. It is also surjective, since for any orbits $\tilde{h}$ and $\tilde{c}$, a colored graphon can be constructed from elements of these orbits, and has a corresponding orbit in $\tilde{\mathcal{W}}_\Theta$. Also note that since $\tilde{\mathcal{W}}$ and $\tilde{\mathcal{C}}_\Theta$ are compact metric spaces, so is $\tilde{\mathcal{W}} \times \tilde{\mathcal{C}}_\Theta$.

Let us show that the pushforward measures $\tilde{\mathbb{P}}_{N,p}^\Theta \circ \Psi^{-1}$ satisfy an LDP. Note that these are simply the product measure of the Erdős–Rényi measure on $\tilde{\mathcal{W}}$ and the Dirac measure centered on the correct empirical coloring on $\tilde{\mathcal{C}}_\Theta$:

$$\tilde{\mathbb{P}}_{N,p}^\Theta \circ \Psi^{-1} = \tilde{\mathbb{P}}_{N,p} \otimes \delta_{\tilde{\mathbf{c}}_{\mathbf{w}_N}}, \quad (\text{B.34})$$

since the uncolored graphons are generated independent of the colorings. The complex structure that comes from coupling the graphon to the coloring is lost once they are decoupled using Ψ . Because the empirical distributions converge to $\mathbf{w}$, it follows that the measures

$\delta_{\tilde{c}_{\mathbf{w}N}}$ satisfy an LDP at *any* speed with rate function

$$\mathcal{I}_{\mathbf{w}}^{\Theta}(\tilde{c}) = \begin{cases} 0 & \text{if } \tilde{c} = \tilde{c}_{\mathbf{w}}, \\ \infty & \text{otherwise.} \end{cases} \quad (\text{B.35})$$

In particular, this implies that these measures satisfy a large deviations principle at speed N^2 . Additionally, note that $\mathcal{I}_{\mathbf{w}}^{\Theta}$ and $\mathcal{I}_p$ are good rate functions, since $\tilde{\mathcal{C}}_{\Theta}$ and $\tilde{\mathcal{W}}$ are compact metric spaces. Using Lemma 8 and Theorem 2, we have that the measures $\tilde{\mathbb{P}}_{N,p}^{\Theta} \circ \Psi^{-1}$ satisfy a large deviations principle on the product space $\tilde{\mathcal{W}} \times \tilde{\mathcal{C}}_{\Theta}$ at speed N^2 with rate function

$$\mathcal{I}_{\mathbf{w},p}^{\Theta,\text{prod}}(\tilde{h}, \tilde{c}) = \begin{cases} \mathcal{I}_p(\tilde{h}) & \text{if } \tilde{c} = \tilde{c}_{\mathbf{w}}, \\ \infty & \text{otherwise.} \end{cases} \quad (\text{B.36})$$

Having shown that the pushforward measures satisfy an LDP in $\tilde{\mathcal{W}} \times \tilde{\mathcal{C}}_{\Theta}$, we need to show that the conditions in Lemma 7 are satisfied to obtain an LDP in $\tilde{\mathcal{W}}_{\Theta}$.

First, note that the support of $\tilde{\mathbb{P}}_{N,p}^{\Theta} \otimes \delta_{\tilde{c}_{\mathbf{w}N}}$ is composed of the orbits $(\tilde{G}, \tilde{c}_{\mathbf{w}N})$, where $G \in \mathcal{G}_N$ are directed graphs over N vertices, which is a discrete set. The fibers $\Psi^{-1}((\tilde{G}, \tilde{c}_{\mathbf{w}N}))$ are not discrete, but they contain the set of colored directed graphs with color fractions $\mathbf{w}_N$ and graphon orbit $\tilde{G}$. We will choose this to be our set Υ_N (in the notation of Lemma 7):

$$\Upsilon_N((\tilde{G}, \tilde{c}_{\mathbf{w}N})) := \Psi^{-1}((\tilde{G}, \tilde{c}_{\mathbf{w}N})) \cap \tilde{\mathcal{Q}}_N^{\Theta}, \quad (\text{B.37})$$

where $\tilde{\mathcal{Q}}_N^{\Theta} \subset \tilde{\mathcal{W}}_{\Theta}$ is the set of colored graphon orbits generated from directed colored graphons over N nodes with colorings over Θ .

I now relate the probability associated to a colored graphon to the probability of its corresponding uncolored graphon. Consider colored graphs over N , let τ be a permutation of $\{1, \dots, N\}$ and let C_0^N be the (fixed) coloring used to generate $\tilde{\mathbb{P}}_{N,p}^{\Theta}$. For a finite graph G with N nodes, the probability associated to $\tilde{G}$ is

$$\tilde{\mathbb{P}}_{N,p}(\{\tilde{G}\}) = p^{|A(G)|} (1-p)^{N(N-1)-|A(G)|} \left(\frac{N!}{\sum_{\tau} \mathbb{1}_{\{\tau A(G) = A(G)\}}} \right). \quad (\text{B.38})$$

The first factor is the probability of drawing a specific representative directed graph. The second counts the number of directed graphs with orbit $\tilde{G}$: it counts all re-labelings of the nodes and takes care of degeneracy by counting the automorphisms of G . Similarly, for a colored graph Q over N nodes (with the correct coloring), its associated probability is

$$\tilde{\mathbb{P}}_{N,p}^\Theta(\{\tilde{Q}\}) = p^{|A(Q)|}(1-p)^{N(N-1)-|A(Q)|} \left(\frac{\prod_{\theta \in \Theta} (Nw_\theta^N)!}{\sum_\tau \mathbb{1}\{\tau A(Q) = A(Q), \tau C(Q) = C_0^N\}} \right). \quad (\text{B.39})$$

Now the re-labelings happen within each color, and the degeneracy must also take into account permutations that respect the coloring. Since the matching of the coloring is an additional restriction, we can see that for the restricted graph $G = (V(Q), A(Q))$:

$$\tilde{\mathbb{P}}_{N,p}^\Theta(\{\tilde{Q}\}) \geq \frac{(\prod_{\theta \in \Theta} (Nw_\theta^N)!)}{N!} \tilde{\mathbb{P}}_{N,p}(\{\tilde{G}\}). \quad (\text{B.40})$$

Define the weights $\alpha_N(\tilde{Q}, (\tilde{G}, \tilde{c}_{\mathbf{w}^N}))$ by

$$\alpha_N(\tilde{Q}, (\tilde{G}, \tilde{c}_{\mathbf{w}^N})) = \tilde{\mathbb{P}}_{N,p}^\Theta(\{\tilde{Q}\}|\{\tilde{G}\}) \quad \text{for } \tilde{Q} \in \Upsilon_N((\tilde{G}, \tilde{c}_{\mathbf{w}^N})). \quad (\text{B.41})$$

From the argument above, we have the following bound on these weights:

$$\alpha_N(\tilde{Q}, (\tilde{G}, \tilde{c}_{\mathbf{w}^N})) \geq \frac{(\prod_{\theta \in \Theta} (Nw_\theta^N)!)}{N!}. \quad (\text{B.42})$$

It is useful to note the Stirling bounds on the factorial:

$$\sqrt{2\pi n} n^{n+\frac{1}{2}} e^{-n} \leq n! \leq e n^{n+\frac{1}{2}} e^{-n}. \quad (\text{B.43})$$

With this, we conclude that

$$\liminf_{N \rightarrow \infty} \inf_{G \in \mathcal{G}_N} \min_{\tilde{Q} \in \Upsilon_N((\tilde{G}, \tilde{c}_{\mathbf{w}^N}))} \frac{1}{N^2} \log(\alpha_N(\tilde{Q}, (\tilde{G}, \tilde{c}_{\mathbf{w}^N}))) = 0. \quad (\text{B.44})$$

Hence, the subexponential weight floor condition in Lemma 7 is satisfied.

The uniform fiber approximation condition is satisfied by the result in Lemma 10, with

$$\epsilon_N = \frac{\kappa_1}{\sqrt{\log(N)}} + 2|\Theta|N^{-\kappa_2} \quad (\text{B.45})$$

for some constants κ_1 and κ_2 . Therefore, applying Lemma 7 we have that $\tilde{\mathbb{P}}_{N,p}^\Theta$ satisfies a large deviations principle on $\tilde{\mathcal{W}}_\Theta$ at speed N^2 with rate function $\mathcal{J}_{\mathbf{w},p}^\Theta$. This completes the proof. $\square$

Using the previous result, we can state the convergence theorems that yield Theorem 1 in the paper.

Theorem 7. *Consider a sequence of colorings $C_0^1, C_0^2, \dots$ such that their empirical color fractions $\mathbf{w}^N$ converge to some $\mathbf{w}$. Additionally, consider an Exponential Random Graph model where the measure is given by*

$$\pi_N(G) = \exp \left\{ N^2 [\mathcal{U}(\tilde{Q}(G)) - \psi_N] \right\}, \quad (\text{B.46})$$

where $Q(G) = (G, C_0^N)$ is the colored directed graph induced by G and the coloring C_0^N . If $\mathcal{U}$ is a continuous bounded function on $\tilde{\mathcal{W}}_\Theta$, then the normalization constant ψ_N given by

$$\psi_N = \frac{1}{N^2} \log \left(\sum_{G \in \mathcal{G}_N} e^{N^2 \mathcal{U}(\tilde{Q}(G))} \right) \quad (\text{B.47})$$

converges to

$$\psi := \sup_{\substack{\tilde{q} \in \tilde{\mathcal{W}}_\Theta \\ \tilde{c}[\tilde{q}] = \tilde{c}_{\mathbf{w}}}} (\mathcal{U}(\tilde{q}) + \mathcal{H}(\tilde{h}[\tilde{q}])). \quad (\text{B.48})$$

Proof. This proof follows the same procedure as the proof of Theorem 3.1 in Chatterjee and Diaconis (2013) for undirected graphons and Theorem 10 of Mele (2017) for directed graphons. Let $\tilde{B} \subseteq \tilde{\mathcal{W}}_\Theta$ be a Borel set. For each N , let $B_N \subseteq \mathcal{W}_\Theta$ be the (finite) set

$$B_N := \{q \in \mathcal{W}_\Theta : \tilde{q} \in \tilde{B} \text{ and } q = q^Q \text{ for some } Q \text{ with } C(Q) = C_0^N \text{ and } (V(Q), A(Q)) \in \mathcal{G}_N\}. \quad (\text{B.49})$$

That is, for all $q \in B_N$, there is a colored directed graph Q that has the correct coloring C_0^N and corresponds to a directed graph over N nodes. Recall that $\tilde{\mathbb{P}}_{N,p}^\Theta$ corresponds to the measure induced on $\tilde{\mathcal{W}}_\Theta$ by fixing the correct coloring C_0^N and sampling the edges using a directed Erdős–Rényi model. Therefore, we have that

$$|B_N| = 2^{N(N-1)} \tilde{\mathbb{P}}_{N,1/2}^\Theta(\tilde{B}). \quad (\text{B.50})$$

Using the result from Theorem 6, we have that for all closed $\tilde{F} \subseteq \tilde{\mathcal{W}}_\Theta$,

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{1}{N^2} \log(\tilde{\mathbb{P}}_{N,p}^\Theta(\tilde{F})) &= \limsup_{N \rightarrow \infty} \frac{1}{N^2} [\log(|F_N|) - N(N-1) \log(2)] \\ &= \limsup_{N \rightarrow \infty} \frac{1}{N^2} \log(|F_N|) - \log(2) \\ &\leq - \inf_{\tilde{q} \in \tilde{F}} \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q}). \end{aligned} \quad (\text{B.51})$$

This implies that

$$\limsup_{N \rightarrow \infty} \frac{1}{N^2} \log(|F_N|) \leq \log(2) - \inf_{\tilde{q} \in \tilde{F}} \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q}). \quad (\text{B.52})$$

Similarly, for open sets $\tilde{U} \subseteq \tilde{\mathcal{W}}_\Theta$ we have

$$\liminf_{N \rightarrow \infty} \frac{1}{N^2} \log(|U_N|) \geq \log(2) - \inf_{\tilde{q} \in \tilde{U}} \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q}). \quad (\text{B.53})$$

Fix $\epsilon > 0$. Since $\mathcal{U}$ is bounded, there is a finite set R such that the intervals $\{(a, a + \epsilon) : a \in R\}$ cover the range of $\mathcal{U}$. For each $a \in R$, let $\tilde{F}^a := \mathcal{U}^{-1}([a, a + \epsilon])$. By continuity of $\mathcal{U}$, the $\tilde{F}^a$ are closed. Now, note that

$$e^{N^2 \psi_N} = \sum_{G \in \mathcal{G}_N} e^{N^2 \mathcal{U}(\tilde{Q}(G))} \leq \sum_{a \in R} e^{N^2(a+\epsilon)} |F_N^a|. \quad (\text{B.54})$$

The last inequality comes from counting the directed graphs that induce a colored graphon orbit in $\tilde{F}_N^a$, which yields $|F^a|$, and bounding $\mathcal{U}$ in this interval. We can further bound this

by

$$e^{N^2\psi_N} \leq |R| \sup_{a \in R} e^{N^2(a+\epsilon)} |F_N^a|. \quad (\text{B.55})$$

Using Eq. (B.52), this yields

$$\limsup_{N \rightarrow \infty} \psi_N \leq \sup_{a \in R} \left(a + \epsilon + \log(2) - \inf_{\tilde{q} \in \tilde{F}^a} \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q}) \right). \quad (\text{B.56})$$

Since each $\tilde{q} \in \tilde{F}^a$ satisfies $\mathcal{U}(\tilde{q}) \geq a$, we have

$$\sup_{\tilde{q} \in \tilde{F}^a} (\mathcal{U}(\tilde{q}) + \log(2) - \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q})) \geq \sup_{\tilde{q} \in \tilde{F}^a} (a + \log(2) - \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q})) = a + \log(2) - \inf_{\tilde{q} \in \tilde{F}^a} \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q}). \quad (\text{B.57})$$

Substituting above,

$$\begin{aligned} \limsup_{N \rightarrow \infty} \psi_N &\leq \epsilon + \sup_{a \in R} \left(\sup_{\tilde{q} \in \tilde{F}^a} (\mathcal{U}(\tilde{q}) + \log(2) - \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q})) \right) \\ &= \epsilon + \sup_{\tilde{q} \in \tilde{\mathcal{W}}_\Theta} (\mathcal{U}(\tilde{q}) + \log(2) - \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q})). \end{aligned} \quad (\text{B.58})$$

Our rate function selects colored graphon orbits that have the correct coloring fraction, so we can write

$$\limsup_{N \rightarrow \infty} \psi_N \leq \epsilon + \sup_{\substack{\tilde{q} \in \tilde{\mathcal{W}}_\Theta \\ \tilde{c}[\tilde{q}] = \tilde{c}_{\mathbf{w}}}} (\mathcal{U}(\tilde{q}) + \log(2) - \tilde{\mathcal{I}}_{1/2}(\tilde{h}[\tilde{q}])) = \epsilon + \sup_{\substack{\tilde{q} \in \tilde{\mathcal{W}}_\Theta \\ \tilde{c}[\tilde{q}] = \tilde{c}_{\mathbf{w}}}} (\mathcal{U}(\tilde{q}) + \tilde{\mathcal{H}}(\tilde{h}[\tilde{q}])). \quad (\text{B.59})$$

where $\mathcal{H}$ is as defined in Equation (B.15).

Similarly, for each a , let $\tilde{U}^a := \mathcal{U}^{-1}((a, a + \epsilon))$, which is open by continuity of $\mathcal{U}$. Note that

$$e^{N^2\psi_N} \geq \sup_{a \in R} e^{N^2a} |U_N^a|. \quad (\text{B.60})$$

By Eq. (B.53),

$$\liminf_{N \rightarrow \infty} \psi_N \geq \sup_{a \in R} \left(a + \log(2) - \inf_{\tilde{q} \in \tilde{U}^a} \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q}) \right). \quad (\text{B.61})$$

Each $\tilde{q} \in \tilde{U}^a$ satisfies $\mathcal{U}(\tilde{q}) \leq a + \epsilon$. Therefore,

$$\begin{aligned} \sup_{\tilde{q} \in \tilde{U}^a} (\mathcal{U}(\tilde{q}) + \log(2) - \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q})) &\leq \sup_{\tilde{q} \in \tilde{U}^a} (a + \epsilon + \log(2) - \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q})) \\ &= a + \epsilon + \log(2) - \inf_{\tilde{q} \in \tilde{U}^a} \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q}). \end{aligned} \quad (\text{B.62})$$

Substituting in our previous bound yields

$$\begin{aligned} \liminf_{N \rightarrow \infty} \psi_N &\geq -\epsilon + \sup_{a \in R} \left(\sup_{\tilde{q} \in \tilde{U}^a} (\mathcal{U}(\tilde{q}) + \log(2) - \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q})) \right) \\ &= -\epsilon + \sup_{\tilde{q} \in \mathcal{V}_\Theta} (\mathcal{U}(\tilde{q}) + \log(2) - \mathcal{J}_{\mathbf{w},1/2}^\Theta(\tilde{q})) \\ &= -\epsilon + \sup_{\substack{\tilde{q} \in \mathcal{V}_\Theta \\ \tilde{c}[\tilde{q}] = \tilde{c}_{\mathbf{w}}}} (\mathcal{U}(\tilde{q}) + \tilde{\mathcal{H}}(\tilde{h}[\tilde{q}])). \end{aligned} \quad (\text{B.63})$$

Since ϵ is arbitrary, we have $\psi_N \rightarrow \psi$, which completes the proof. $\square$

Theorem 8. *Let $\tilde{M}^*$ be the set of maximizers of the variational problem (B.48). Let Q_N be a colored graph on N vertices drawn from the model implied by function $\mathcal{U}$. Then, for any $\eta > 0$, there exist $C, \kappa > 0$ such that, for any N ,*

$$\mathbb{P}\{\delta_\square^\Theta(\tilde{Q}_N, \tilde{M}^*) > \eta\} \leq Ce^{-N^2\kappa},$$

where $\mathbb{P}$ denotes the probability measure implied by the model.

Proof. This proof follows the same procedure as the proof of Theorem 3.2 in Chatterjee and Diaconis (2013). $\square$

B.4 Technical results

This section contains results that are used in the proofs of the results in the previous section.

Lemma 7. *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact metric spaces and $\Psi : \mathcal{X} \rightarrow \mathcal{Y}$ be a continuous surjection. Consider a sequence of probability measures $(\mu_N)_{N \in \mathbb{N}}$ on $\mathcal{X}$, and their pushforwards by Ψ : $\nu_N := \mu_N \circ \Psi^{-1}$. Suppose that for all N and $y \in \text{supp}(\nu_N)$, there exists a finite set $\Upsilon_N(y) \subset \Psi^{-1}(\{y\})$ such that:*

1. (Finite support) *There exist weights $\alpha_N(x, y) > 0$, $x \in \Upsilon_N(y)$, with $\sum_{x \in \Upsilon_N(y)} \alpha_N(x, y) = 1$ such that the Ψ -disintegration of μ_N is*

$$\mu_N(dx) = \int \mu_N^y(dx) \nu_N(dy), \quad \mu_N^y = \sum_{x \in \Upsilon_N(y)} \alpha_N(x, y) \delta_x \text{ for } y \in \text{supp}(\nu_N). \quad (\text{B.64})$$

2. (Subexponential weight floor) *The weights satisfy*

$$\liminf_{N \rightarrow \infty} \inf_{y \in \mathcal{Y}} \min_{x \in \Upsilon_N(y)} \frac{1}{a_N} \log(\alpha_N(x, y)) = 0, \quad (\text{B.65})$$

3. (Uniform fiber approximation) *There exists a sequence $(\epsilon_N)_{N \in \mathbb{N}}$, with $\epsilon_N \rightarrow 0$, such that $\Upsilon_N(y)$ is an ϵ_N net of $\Psi^{-1}(y)$ for all $y \in \text{supp}(\nu_N)$.*

Then, if the sequence $(\nu_N)_{N \in \mathbb{N}}$ satisfies a large deviation principle on $\mathcal{Y}$ at speed a_N with rate function I_Y , the sequence $(\mu_N)_{N \in \mathbb{N}}$ satisfies a large deviation principle on $\mathcal{X}$ at speed a_N with rate function $I_X := I_Y \circ \Psi$.

Proof. The strategy for this proof is to show that the measures μ_N satisfy the condition for Bryc's Theorem (Theorem 4.4.2 in [Dembo and Zeitouni \(2010\)](#)). For a topological space $\mathcal{Z}$, let $C_b(\mathcal{Z})$ denote the set of real-valued continuous bounded functions. We need to show that the sequence μ_N is exponentially tight, and that for all $\phi \in C_b(\mathcal{X})$, the following holds:

$$\lim_{N \rightarrow \infty} \frac{1}{a_N} \log \left(\int e^{-a_N \phi} d\mu_N \right) = - \inf_{x \in \mathcal{X}} \{ \phi(x) + I_X(x) \}. \quad (\text{B.66})$$

Note that the sequences μ_N and ν_N are automatically exponentially tight since the spaces are compact metric (this also automatically implies that all rate functions are good). Therefore, we just need to show that the limit above exists, and that it corresponds to the expression on the right. If this is the case, Bryc's theorem states that μ_N satisfies the LDP with speed a_N and rate function I_X .

First, fix a function $\phi \in C_b(\mathcal{X})$. Define the fiberwise infimum of ϕ :

$$\Phi(y) := \inf\{\phi(x) : \Psi(x) = y\}. \quad (\text{B.67})$$

Now, note that since Ψ is continuous surjective and $\mathcal{X}$ is compact, the correspondence $\Gamma(y) := \Psi^{-1}(y)$ is non-empty, compact-valued and has a closed graph. Therefore, since ϕ is continuous, Berge's maximum theorem states that $\Phi \in C_b(\mathcal{Y})$.

By definition, we have that $\phi(x) \geq \Phi(\Psi(x))$ for all $x \in \mathcal{X}$. Therefore,

$$\int e^{-a_N \phi(x)} d\mu_N(x) \leq \int e^{-a_N \Phi(\Psi(x))} d\mu_N(x) = \int e^{-a_N \Phi(y)} d\nu_N(y). \quad (\text{B.68})$$

Taking limits and applying Varadhan's Lemma (Theorem 4.3.1 in [Dembo and Zeitouni \(2010\)](#)) to the integral over $\mathcal{Y}$ yields

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{1}{a_N} \log \left(\int e^{-a_N \phi(x)} d\mu_N(x) \right) &\leq \limsup_{N \rightarrow \infty} \frac{1}{a_N} \log \left(\int e^{-a_N \Phi(y)} d\nu_N(y) \right) \\ &= - \inf_{y \in \mathcal{Y}} \{ \Phi(y) + I_Y(y) \} \\ &= - \inf_{x \in \mathcal{X}} \{ \Phi(\Psi(x)) + I_Y(\Psi(x)) \} \\ &= - \inf_{x \in \mathcal{X}} \{ \phi(x) + I_X(x) \}, \end{aligned} \quad (\text{B.69})$$

where the last equality used the infimum in the definition of Φ . This yields the limsup bound.

To obtain the liminf bound, we need the structure of the Ψ -disintegration of μ_N . Fix some $\eta > 0$. Then, for all y there exists some $x_0 \in \Psi^{-1}(y)$ such that $\phi(x_0) \leq \Phi(y) + \eta/2$. Since ϕ is continuous and $\mathcal{X}$ is compact, by the Heine-Cantor theorem, ϕ must be uniformly continuous. Together with the ϵ_N net condition, this implies that there is some N_η (independent of y) such that for all $N > N_\eta$ there is some $x_1^N \in \Upsilon_N(y)$ that satisfies $\phi(x_1^N) \leq \phi(x_0) + \eta/2$, which implies $\phi(x_1^N) \leq \Phi(y) + \eta$. Now, note that

$$\sum_{x \in \Upsilon_N(y)} \alpha_N(x, y) e^{-a_N \phi(x)} \geq \alpha_N(x_1^N, y) e^{-a_N \phi(x_1^N)} \geq \left[\min_{x \in \Upsilon_N(y)} \alpha_N(x, y) \right] e^{-a_N [\Phi(y) + \eta]}. \quad (\text{B.70})$$

Therefore, for N large enough we obtain

$$\begin{aligned}
\int e^{-a_N \phi(x)} d\mu_N(x) &= \int \left(\sum_{x \in \Upsilon_N(y)} \alpha_N(x, y) e^{-a_N \phi(x)} \right) d\nu_N(y) \\
&\geq \int \left[\min_{x \in \Upsilon_N(y)} \alpha_N(x, y) \right] e^{-a_N [\Phi(y) + \eta]} d\nu_N(y) \\
&\geq \left[\inf_{y' \in \mathcal{Y}} \min_{x \in \Upsilon_N(y')} \alpha_N(x, y') \right] \int e^{-a_N [\Phi(y) + \eta]} d\nu_N(y). \tag{B.71}
\end{aligned}$$

Taking logs and limits, and applying Varadhan's Lemma to $\Phi + \eta$ yields

$$\begin{aligned}
&\liminf_{N \rightarrow \infty} \frac{1}{a_N} \log \left(\int e^{-a_N \phi(x)} d\mu_N(x) \right) \\
&\geq \liminf_{N \rightarrow \infty} \inf_{y' \in \mathcal{Y}} \min_{x \in \Upsilon_N(y')} \frac{1}{a_N} \log(\alpha_N(x, y')) - \eta - \inf_{y \in \mathcal{Y}} \{\Phi(y) + I_Y(y)\} \tag{B.72}
\end{aligned}$$

Using the subexponential weight floor condition and noting that this bound holds for any η , we conclude that

$$\begin{aligned}
\liminf_{N \rightarrow \infty} \frac{1}{a_N} \log \left(\int e^{-a_N \phi(x)} d\mu_N(x) \right) &\geq - \inf_{y \in \mathcal{Y}} \{\Phi(y) + I_Y(y)\} \\
&= - \inf_{x \in \mathcal{X}} \{\phi(x) + I_X(x)\}. \tag{B.73}
\end{aligned}$$

Since the two bounds coincide, we have that the limit in Eq. (B.66) exists for all $\phi \in C_b(\mathcal{X})$, and the expression is satisfied with $I_X = I_Y \circ \Psi$. Therefore, the conditions in Bryc's Theorem hold. This completes the proof. $\square$

Lemma 8 (Theorem 2.6 in [Varadhan \(2010\)](#)). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. Let the sequences of measures $(\mu_N)_{N \in \mathbb{N}}$ on $\mathcal{X}$ and $(\nu_N)_{N \in \mathbb{N}}$ on $\mathcal{Y}$ satisfy large deviations principles at the same speed a_N with good rate functions I_X and I_Y , respectively. Then the sequence $(\mu_N \otimes \nu_N)_{N \in \mathbb{N}}$ satisfies a large deviations principle at speed a_N with rate function*

$$I_{X \times Y}(x, y) = I_X(x) + I_Y(y). \tag{B.74}$$

Definition 7. For a directed graph $G = (V, A)$ and for $X, Y \subseteq V$, let $a_G(X, Y)$ denote the

number of arcs that originate from a node in X and terminate in a node in Y . Let

$$d_G(X, Y) := \frac{a_G(X, Y)}{|X||Y|} \quad (\text{B.75})$$

denote the density of arcs from X to Y .

Definition 8. Consider a directed graph $G = (V, A)$ and a partition $\mathcal{P} = \{V_1, \dots, V_k\}$ of V . Define the weighted directed graph $G_{\mathcal{P}}$ on V as the complete directed graph whose arc (u, v) has weight $d_G(V_i, V_j)$ if $u \in V_i$ and $v \in V_j$.

Lemma 9. *There exists a constant C_0 such that for every $\eta \in (0, 1)$ and every directed graph $G = (V, A)$, V has a partition $\mathcal{P}$ into $k \leq \exp(C_0/\eta^2)$ classes such that*

$$d_{\square}(G, G_{\mathcal{P}}) \leq \eta. \quad (\text{B.76})$$

Proof. Fix $\eta \in (0, 1)$. Let $N = |V|$ and let B be the adjacency matrix of G for some ordering of $[N] := (1, \dots, N)$. Consider the bipartite *undirected* graph with adjacency matrix

$$\tilde{B} = \begin{pmatrix} 0 & B^{\top} \\ B & 0 \end{pmatrix}. \quad (\text{B.77})$$

Define the left and right sets $L := \{1, \dots, N\}$ and $R := \{N + 1, \dots, 2N\}$. A consequence of the Weak Regularity Lemma for undirected graphs (Lemma 9.3 in [Lovász \(2012\)](#)), there is a partition $\mathcal{Q}$ of $[2N]$ into $k \leq \exp(C_1/\eta^2)$ classes such that

$$d_{\square}(\tilde{B}, \tilde{B}_{\mathcal{Q}}) \leq \frac{\eta}{4}, \quad (\text{B.78})$$

where C_1 is independent of the graph and of N . Since refining a partition can only decrease the cut distance to the original graph, we can consider the refinement $\mathcal{Q} \vee \{L, R\}$. This gives partitions $\mathcal{Q}_1$ and $\mathcal{Q}_2$ of L and R with k_1 and k_2 classes, respectively, with $k_1, k_2 \leq k$. Note that $\mathcal{Q}_2$ induces a partition $\mathcal{Q}'_2 = (Q'_{2,i})_{1 \leq i \leq k_2}$ of L given by

$$Q'_{2,i} = \{2N + 1 - x : x \in Q_{2,i}\}. \quad (\text{B.79})$$

We can further consider the refinement $\mathcal{P} := \mathcal{Q}_1 \vee \mathcal{Q}_2$ of L , which satisfies $|\mathcal{P}| \leq k_1 k_2$. We can extend $\mathcal{P}$ to a partition $\mathcal{Q}'$ of $L \cup R$. Let $P(x)$ and $Q'(x)$ be the partition classes of $\mathcal{P}$ and $\mathcal{Q}'$ that contain element x . Then we can define $\mathcal{Q}'$ by

$$Q'(x) = \begin{cases} P(x) & \text{if } x \in L, \\ P(2N + 1 - x) & \text{if } x \in R. \end{cases} \quad (\text{B.80})$$

Note that $\mathcal{Q}'$ is a refinement of $\mathcal{Q}$. Therefore, we have that

$$d_{\square}(\tilde{B}, \tilde{B}_{\mathcal{Q}'}) \leq d_{\square}(\tilde{B}, \tilde{B}_{\mathcal{Q}}) \leq \frac{\eta}{4}. \quad (\text{B.81})$$

Returning to the directed graph G , the cut distance between G and $G_{\mathcal{P}}$ can be written as

$$d_{\square}(G, G_{\mathcal{P}}) = \frac{1}{N^2} \max_{S, T \subseteq L} \left| \sum_{i \in S, j \in T} (B_{ij} - (B_{\mathcal{P}})_{ij}) \right|. \quad (\text{B.82})$$

Note that any cut S, T can be implemented in the bipartite graph by choosing appropriate $S' \subseteq L$ and $T' \subseteq R$ (a cross-cut). Therefore,

$$d_{\square}(G, G_{\mathcal{P}}) \leq 4 \frac{1}{(2N)^2} \max_{S, T \subseteq L \cup R} \left| \sum_{i \in S, j \in T} (\tilde{B}_{ij} - (\tilde{B}_{\mathcal{Q}'})_{ij}) \right| = 4d_{\square}(\tilde{B}, \tilde{B}_{\mathcal{Q}'}). \quad (\text{B.83})$$

Therefore, for the partition $\mathcal{P}$, we have

$$d_{\square}(G, G_{\mathcal{P}}) \leq \eta. \quad (\text{B.84})$$

Additionally, we have that

$$|\mathcal{P}| \leq k_1 k_2 \leq k^2 \leq \exp(2C_1/\eta^2). \quad (\text{B.85})$$

This completes the proof. $\square$

Definition 9. For integer $N > 0$, for $i \in \{0, \dots, N-1\}$ denote with I_i^N the set $(i/N, (i +$

1)/N]. An N -permutation σ_τ is a measurable bijection determined by a permutation $\tau \in S_N$ such that

$$\sigma_\tau(x) = x + \frac{\tau(i) - i}{N} \quad (\text{B.86})$$

for $x \in I_i^N$. Let Σ_N denote the set of N -permutations.

Lemma 10. *There exist constants $\kappa_1, \kappa_2, N_0 > 0$ such that for all integer $N > N_0$, all colored graphons Q over N nodes, and all $\tilde{q} \in \tilde{\mathcal{W}}_\Theta$ such that $(\tilde{h}[\tilde{q}], \tilde{c}[\tilde{q}]) = (\tilde{h}^Q, \tilde{c}^Q)$, there exists an N -permutation $\hat{\sigma} \in \Sigma_N$ such that $q^{\hat{\sigma}} := (\hat{\sigma}h^Q, c^Q)$ satisfies*

$$\delta_\square^\Theta(\tilde{q}^{\hat{\sigma}}, \tilde{q}) \leq \frac{\kappa_1}{\sqrt{\log(N)}} + 2|\Theta|N^{-\kappa_2}. \quad (\text{B.87})$$

Proof. To simplify notation, denote $h_1 := h^Q$ and $c_1 := c^Q$. Without loss of generality, we can set the coloring c_1 such that the sets $c_1^{-1}(\theta)$ are pairwise disjoint intervals. This is because the condition $(\tilde{h}[\tilde{q}], \tilde{c}[\tilde{q}]) = (\tilde{h}^Q, \tilde{c}^Q)$ and the distance $\delta_\square^\Theta(\tilde{q}^{\sigma_0}, \tilde{q})$ only involve orbits. Denote with $\mathcal{K} = \{K_{\theta_1}, \dots, K_{\theta_{|\Theta|}}\}$ the partition of $[0, 1]$ into the intervals corresponding to this coloring.

Since the orbits of the graphon and the coloring match, there is a representative $q_2 = (h_2, c_2) \in \tilde{q}$ such that $c_2 = c_1$. Additionally, there exists a measure-preserving bijection σ_{12} such that $h_2 = \sigma_{12}h_1$. Now, note that for all $\sigma_0 \in \Sigma_N$,

$$\delta_\square^\Theta(\tilde{q}^{\sigma_0}, \tilde{q}) = \inf_{\sigma} \{d_\square(\sigma_0 h_1, \sigma h_2) + d_\Theta(c_1, \sigma c_2)\}, \quad (\text{B.88})$$

where the infimum is over measure-preserving bijections. I will bound this distance by searching over color-preserving bijections. That is, σ such that $\sigma(K_\theta) = K_\theta$ for all θ . We have that

$$\delta_\square^\Theta(\tilde{q}^{\sigma_0}, \tilde{q}) \leq \inf_{\sigma \text{ color-preserving}} \{d_\square(\sigma_0 h_1, \sigma h_2) + d_\Theta(c_1, \sigma c_2)\} = \inf_{\sigma \text{ color-preserving}} d_\square(\sigma_0 h_1, \sigma h_2). \quad (\text{B.89})$$

Fix $\eta > 0$. By Lemma 9, there exists a partition $\mathcal{P}$ of $\{1, \dots, N\}$ with $k \leq \exp(C_0/\eta^2)$

classes such that, for the directed graph $G(Q)$,

$$d_{\square}(G(Q), G(Q)_{\mathcal{P}}) \leq \eta. \quad (\text{B.90})$$

For the graphon h_1 , this implies that there is a partition $\mathcal{J}$ of $[0, 1]$, where each partition class is composed of unions of intervals I_i^N , such that the graphon u given by

$$u(x, y) = \frac{1}{|J_a||J_b|} \int_{J_a \times J_b} h_1(x', y') dx' dy' \quad \text{for } (x, y) \in J_a \times J_b \quad (\text{B.91})$$

satisfies

$$d_{\square}(h_1, u) \leq \eta. \quad (\text{B.92})$$

Since the cut distance is invariant under measure-preserving bijections, we have that for any σ_0, σ ,

$$d_{\square}(\sigma_0 h_1, \sigma_0 u), d_{\square}(\sigma \circ \sigma_{12} h_1, \sigma \circ \sigma_{12} u) \leq \eta. \quad (\text{B.93})$$

Therefore, by the triangle inequality, for any fixed color-preserving measure-preserving bijection σ

$$\delta_{\square}^{\Theta}(\tilde{q}^{\sigma_0}, \tilde{q}) \leq d_{\square}(\sigma_0 h_1, \sigma h_2) \leq 2\eta + d_{\square}(\sigma_0 u, \sigma \circ \sigma_{12} u). \quad (\text{B.94})$$

To control the distance $d_{\square}(\sigma_0 u, \sigma \circ \sigma_{12} u)$, let us first choose a color-preserving measure-preserving bijection σ such that the family $(\sigma \circ \sigma_{12}(J_a) \cap K_{\theta})_{1 \leq a \leq k}$ are pairwise disjoint intervals (modulo null sets) within K_{θ} . Let $\sigma_{\text{int}} := \sigma \circ \sigma_{12}$. Now I construct an N -permutation that approximates the transport properties of σ_{int} . Define the transport matrix

$$x_{a\theta} = \lambda(J_a \cap \sigma_{\text{int}}^{-1} K_{\theta}), \quad (\text{B.95})$$

where λ is the Lebesgue measure. Intuitively, this captures how much mass from J_a is sent

to K_θ under σ_{int} . Additionally, define $p_a := \lambda(J_a)$ and $w_\theta := \lambda(K_\theta)$. Note that

$$\sum_{\theta} x_{a\theta} = p_a, \quad \sum_a x_{a\theta} = w_\theta. \quad (\text{B.96})$$

Now, since Np_a and Nw_θ are integers for all a, θ (because they originate from a graphon over N vertices), by Lemma 3 in [Baranyai \(1974\)](#) there exists a matrix $(n_{a\theta})_{a,\theta}$ such that

$$n_{a\theta} \in \{\lfloor Nx_{a\theta} \rfloor, \lceil Nx_{a\theta} \rceil\}, \quad \sum_{\theta} n_{a\theta} = Np_a, \quad \sum_a n_{a\theta} = Nw_\theta. \quad (\text{B.97})$$

Define $y_{a\theta} = n_{a\theta}/N$. Note that $|y_{a\theta} - x_{a\theta}| \leq 1/N$. Now define an N -permutation $\hat{\sigma}$ that sends $n_{a\theta}$ intervals I_i^N from J_a to K_θ . This permutation exists by the feasibility result above. Under this construction,

$$\lambda(J_a \cap \hat{\sigma}^{-1}(K_\theta)) = y_{a\theta}. \quad (\text{B.98})$$

Furthermore, the sets $\hat{\sigma}(J_a) \cap K_\theta$ can be chosen to be intervals, since only their measure is constrained. Furthermore, these can follow the same order as the intervals $\sigma_{\text{int}}(J_a) \cap K_\theta$.

Now consider a single interval K_θ . Within this interval, the intervals $\hat{\sigma}(J_a) \cap K_\theta$ and $\sigma_{\text{int}}(J_a) \cap K_\theta$ follow the same order. Without loss of generality, assume they are ordered according to a . Define the set

$$E_\theta := \{x \in K_\theta : \hat{\sigma}^{-1}(x) \text{ belongs to a different class than } \sigma_{\text{int}}^{-1}(x)\}. \quad (\text{B.99})$$

Since the measure of E is determined by the discrepancies in the boundaries of the intervals,

we have that

$$\begin{aligned}
\lambda(E_\theta) &\leq \sum_{a=1}^k \left\{ \underbrace{\left| \sum_{b=1}^{a-1} (x_{b\theta} - y_{b\theta}) \right|}_{\text{left endpoints}} + \underbrace{\left| \sum_{b=1}^a (x_{b\theta} - y_{b\theta}) \right|}_{\text{right endpoints}} \right\} \\
&\leq \sum_{a=1}^k \left\{ \sum_{b=1}^{a-1} |x_{b\theta} - y_{b\theta}| + \sum_{b=1}^a |x_{b\theta} - y_{b\theta}| \right\} \\
&\leq \frac{1}{N} \sum_{a=1}^k (2a - 1) \\
&= \frac{k^2}{N}
\end{aligned} \tag{B.100}$$

Letting $E := \bigcup_{\theta \in \Theta} E_\theta$, we get that the total discrepancy in the classes of the preimages is

$$\lambda(E) = \sum_{\theta} \lambda(E_\theta) \leq \frac{|\Theta|k^2}{N}. \tag{B.101}$$

Since $\sigma_{\text{int}} u$ and $\hat{\sigma} u$ have the same value on the sets $\sigma_{\text{int}}^{-1}(J_a) \times \sigma_{\text{int}}^{-1}(J_b)$ and $\hat{\sigma}^{-1}(J_a) \times \hat{\sigma}^{-1}(J_b)$, we have that they must match on $E^c \times E^c$. Therefore, since $|u| \leq 1$, we obtain

$$\|\sigma_{\text{int}} u - \hat{\sigma} u\|_1 = \int_{[0,1]^2 \setminus E^c \times E^c} |\sigma_{\text{int}} u - \hat{\sigma} u| \leq 1 - (1 - \lambda(E))^2 \leq 2\lambda(E) \leq \frac{2|\Theta|k^2}{N}. \tag{B.102}$$

Since the L^1 norm upper-bounds the cut norm, we have that

$$d_{\square}(\sigma_{\text{int}} u, \hat{\sigma} u) \leq \frac{2|\Theta|k^2}{N}. \tag{B.103}$$

Going back to our distance of interest, we have that

$$\delta_{\square}^{\Theta}(\tilde{q}^{\hat{\sigma}}, \tilde{q}) \leq 2\eta + d_{\square}(\sigma_{\text{int}} u, \hat{\sigma} u) \leq 2\eta + \frac{2|\Theta|k^2}{N}. \tag{B.104}$$

Choose $\eta = C_1/\sqrt{\log(N)}$ for some $C_1 > 0$. This yields

$$\delta_{\square}^{\Theta}(\tilde{q}^{\hat{\sigma}}, \tilde{q}) \leq \frac{2C_1}{\sqrt{\log(N)}} + \frac{2|\Theta|N^{2C_0/C_1^2}}{N}. \tag{B.105}$$

Choosing $C_1 > \sqrt{2C_0}$, we obtain the desired bound.

□