Sparse Maximum-Entropy Random Graphs with a Given Power-Law Degree Distribution

Pim van der Hoorn[1] · **Gabor Lippner**[2] ·
Dmitri Krioukov[3]

Abstract Even though power-law or close-to-power-law degree distributions are ubiquitously observed in a great variety of large real networks, the mathematically satisfactory treatment of random power-law graphs satisfying basic statistical requirements of realism is still lacking. These requirements are: sparsity, exchangeability, projectivity, and unbiasedness. The last requirement states that entropy of the graph ensemble must be maximized under the degree distribution constraints. Here we prove that the hypersoft configuration model, belonging to the class of random graphs with latent hyperparameters, also known as inhomogeneous random graphs or W-random graphs, is an ensemble of random power-law graphs that are sparse, unbiased, and either exchangeable or projective. The proof of their unbiasedness relies on generalized graphons, and on mapping the problem of maximization of the normalized Gibbs entropy of a random graph ensemble, to the graphon entropy maximization problem, showing that the two entropies converge to each other in the large-graph limit.

Keywords Sparse random graphs · Power-law degree distributions · Maximum-entropy graphs

Mathematics Subject Classification 05C80 · 05C82 · 54C70

✉ Pim van der Hoorn
w.vanderhoorn@northeastern.edu

1. Department of Physics, Northeastern University, 360 Huntington Ave, MS 210-177, Boston, MA 02115, USA

2. Department of Mathematics, Northeastern University, Boston, USA

3. Departments of Physics, Mathematics, and Electrical & Computer Engineering, Northeastern University, Boston, USA

Ⓐ Springer

1 Introduction

Random graphs have been used extensively to model a variety of real networks. Many of these networks, ranging from the Internet and social networks to the brain and the universe, have broad degree distributions, often following closely power laws [1–3], that the simplest random graph model, the Erdős–Rényi random graphs [4–6] with Poisson degree distributions, does not reproduce. To resolve this disconnect, several alternative models have been proposed and studied. The first one is the configuration model (CM), random graphs with a given degree sequence [7,8]. This model is a microcanonical ensemble of random graphs. Every graph in the ensemble has the same fixed degree sequence, e.g., the one observed in a snapshot of a real network, and every such graph is equiprobable in the ensemble. The ensemble thus maximizes Gibbs entropy subject to the constraint that the degree sequence is fixed. Yet given a real network snapshot, one cannot usually trust its degree sequence as some "ultimate truth" for a variety of reasons, including measurement imperfections, inaccuracies, and incompleteness, noise and stochasticity, and most importantly, the fact that most real networks are dynamic both at short and long time scales, growing often by orders of magnitude over years [1–3,9,10].

These factors partly motivated the development of the soft configuration model (SCM), random graphs with a given expected degree sequence, first considered in [11,12], and later corrected in [13–16], where it was shown that this correction yields a canonical ensemble of random graphs that maximize Gibbs entropy under the constraint that the expected degree sequence is fixed. In statistics, canonical ensembles of random graphs are known as exponential random graphs (ERGs) [17]. In [18,19] it was shown that the sparse CM and SCM are not equivalent, but they are equivalent in the case of dense graphs [20]. Yet the SCM still treats a given degree sequence as a fixed constraint, albeit not as a sharp but soft constraint. This constraint is in stark contrast with reality of many growing real networks, in which the degree of all nodes constantly change, yet the shape of the degree distribution and the average degree do not change, staying essentially constant in networks that grow in size even by orders of magnitude [1–3,9,10]. These observations motivated the development of the hypersoft configuration model [21–24].

1.1 Hypersoft Configuration Model (HSCM)

In the HSCM neither degrees nor even their expected values are fixed. Instead the fixed properties are the degree distribution and average degree. The HSCM with a given average degree and power-law degree distribution is defined by the exponential measure μ on the real line $\mathbb{R}$

$$\mu = e^{\alpha x}, \quad x \in \mathbb{R},$$

where $\alpha > 1$ is a constant, and by the Fermi–Dirac graphon $W : \mathbb{R}^2 \to [0, 1]$

$$W(x, y) = \frac{1}{e^{x+y} + 1}. \tag{1}$$

The volume-form measure μ establishes then probability measures

$$\mu_n = \frac{\mu|_{A_n}}{\mu(A_n)} = \alpha \, e^{\alpha(x - R_n)}$$

on intervals $A_n = (-\infty, R_n]$, where

$$R_n = \frac{1}{2} \log \frac{n}{\beta^2 \nu}, \quad \beta = 1 - \frac{1}{\alpha}, \tag{2}$$

and $\nu > 0$ is another constant. The constants $\alpha > 1$ and $\nu > 0$ are the only two parameters of the model. The HSCM random graphs of size n are defined by (W, A_n, μ_n) via sampling n i.i.d. points $x_1, \ldots, x_n$ on A_n according to measure μ_n, and then connecting pairs of points i and j at sampled locations x_i and x_j by an edge with probability $W(x_i, x_j)$.

An alternative equivalent definition is obtained by mapping (W, A_n, μ_n) to $(W_{I,n}, I, \mu_I)$, where $I = [0, 1]$, $\mu_I = 1$, and

$$W_{I,n}(x, y) = \frac{1}{\frac{n}{\beta^2 \nu} (xy)^{\frac{1}{\alpha}} + 1}, \quad x, y \in I, \tag{3}$$

In this definition, x_is are n i.i.d. random variables uniformly distributed on the unit interval $[0, 1]$, and vertices i and j are connected with probability $W_{I,n}(x_i, x_j)$.

Yet another equivalent definition, perhaps the most familiar and most frequently used one, is given by $(W_{P,n}, P, \mu_P)$, where interval $P = [\beta\nu, \infty)$, and measure μ_P on P is the Pareto distribution

$$\mu_P = \alpha (\beta\nu)^\alpha x^{-\gamma}, \quad x \in P,$$

$$W_{P,n}(x, y) = \frac{1}{\frac{\nu n}{xy} + 1}, \quad x, y \in P.$$

In this Pareto representation, the expected degree of a vertex at coordinate x is proportional to x [15,16].

Compared to the SCM where only edges are random variables while the expected degrees are fixed, the HSCM introduces another source of randomness—and hence entropy—coming from expected degrees that are also random variables. One obtains a particular realization of an SCM from the HSCM by sampling x_is from their fixed distribution and then freezing them. Therefore the HSCM is a probabilistic mixture of canonical ensembles, SCM ERGs, so that one may call the HSCM a *hypercanonical ensemble* given that latent variables x in the HSCM are called *hyperparameters* in statistics [25].

1.2 Properties of the HSCM

We prove in Theorem 2 that the distribution of degrees $\mathcal{D}$ in the power-law HSCM ensemble defined above converges—henceforth *convergence* always means the $n \to \infty$ limit, unless mentioned otherwise—to

$$\mathbb{P}(\mathcal{D} = k) = \alpha(\beta\nu)^\alpha \frac{\Gamma(k - \alpha, \beta\nu)}{k!} = \alpha(\beta\nu)^\alpha \frac{\Gamma(k - \alpha)}{\Gamma(k + 1)} (1 - P(k - \alpha, \beta\nu)), \tag{4}$$
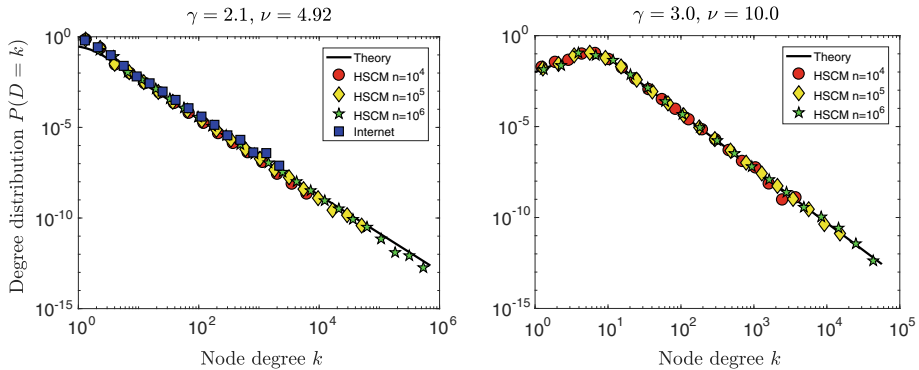
where $\Gamma(a, x)$ is the upper incomplete Gamma function, and $P(a, x)$ is the regularized lower incomplete Gamma function. Since $P(a, x) \sim (ex/a)^a$ for $a \gg ex$, while $\Gamma(k - \alpha)/\Gamma(k + 1) \sim k^{-(\alpha+1)}$ for $k \gg \alpha$, we get

$$\mathbb{P}(\mathcal{D} = k) \sim \alpha(\beta\nu)^\alpha k^{-\gamma}, \quad \gamma = \alpha + 1. \tag{5}$$

We also prove in Theorem 3 that the expected average degree in the ensemble converges to

$$\mathbb{E}[\mathcal{D}] = \nu. \tag{6}$$

That is, the degree distribution in the ensemble has a power tail with exponent γ, while the expected average degree is fixed to constant ν that does not depend on n, Fig. 1.

Fig. 1 Degree distribution in the HSCM, theory vs. simulations, and in the Internet. The theory curve in the left panel is Eq. (4) with $\alpha = 1.1$ ($\gamma = 2.1$) and $\nu = 4.92$. The simulation data shown by symbols is averaged over 100 random graphs for each graph size n. All the graphs are generated according to the HSCM with the same $\alpha = 1.1$ and $\nu = 4.92$. The average degrees, averaged over 100 random graphs, in the graphs of size 10^4, 10^5, and 10^6, are 1.73, 2.16, and 2.51, respectively. The Internet data comes from CAIDA's Archipelago measurements of the Internet topology at the Autonomous System level [26]. The number of nodes and the average degree in the Internet graph are 23, 752 and 4.92. The right panel shows the theoretical degree distribution curve in Eq. (4) with $\alpha = 2.0$ and $\nu = 10.0$ versus simulations of 100 random HSCM graphs of different sizes with the same α and ν. The average degrees in the graphs of size 10^4, 10^5, and 10^6, are 9.96, 9.98, and 10.0, respectively.

1.3 Unbiasedness and the Maximum-Entropy Requirement

While the average degree in the HSCM converges to a constant, and the degree distribution converges to a power law, the power-law HSCM is certainly just one of an infinite number of other models that possess these two properties. One example is random hyperbolic graphs [27] that also have constant average degree and power-law degree distribution, but have larger numbers of triangles, non-zero clustering in the limit. Is the HSCM an unbiased model of random power-law graphs with constant average degree? That is, are the HSCM random graphs characterized by only these two properties and no others? Or colloquially, is the HSCM the model of "maximally random" power-law graphs with a constant average degree. This question can be formally answered by checking whether the HSCM satisfies the maximum-entropy requirement.

A discrete distribution p_i, $i = 1, 2, \ldots$, is said to satisfy the maximum-entropy requirement subject to constraints $\sum_i p_i f_{ir} = \bar{f}_r$, $r = 1, 2, \ldots$, where f_rs are some real functions of states i, and $\bar{f}_r$s are a collection of real numbers, if the Gibbs/Shannon entropy of the distribution $S = -\sum_i p_i \log p_i$ is maximized subject to these constraints [28]. This entropy-maximizing distribution is known to be always unique, belonging to the exponential family of distributions, and it can be derived from the basic consistency axioms: uniqueness and invariance with respect to a change of coordinates, system independence and subset independence [29–31]. Since entropy is the unique measure of information satisfying the basic requirements of continuity, monotonicity, and system/subset independence [32], the maximum-entropy requirement formalizes the notion of encoding into the probability distribution p_i describing a stochastic system, all the available information about the system given to us in the form of the constraints above, and *not* encoding any other information not given to us. Since the entropy-maximizing distribution is unique, any other distribution necessarily but possibly implicitly introduces biases by encoding some additional ad-hoc information, and constrain-

ing some other system properties, concerning which we are not given any information, to some ad-hoc values. Clearly, such uncontrolled information injection into a model of a system may affect the predictions one may wish to make about the system using the model, and indeed it is known that given all the available information about a system, the predictive power of a model that describes the system is maximized by the maximum-entropy model [29–31]. Perhaps the best illustration of this predictive power is the predictive power of equilibrium statistical mechanics, which can be formulated almost fully in terms of the maximum-entropy principle [28].

To illustrate the maximum-entropy requirement in application to random graphs, suppose we are to define a random graph ensemble, and the *only* available information is that these random graphs must have n nodes and m edges. From the purely probabilistic perspective, any random graph ensemble satisfying these constraints—random m-stars or m-cycles, for instance, if $m < n$—would be an equally good one. Yet there is only one unique ensemble that satisfies not only these constraints but also the maximum-entropy requirement. This ensemble is $\mathcal{G}_{n,m}$ because in $\mathcal{G}_{n,m}$ *any* graph with n nodes and m edges is equally likely, so that the probability distribution on the set of all graphs with n nodes and m edges is uniform, and without further constraints, the uniform distribution is the maximum-entropy distribution on the state space, which in this case is $i = 1, \ldots, \binom{\binom{n}{2}}{m}$, the number of such graphs. Random m-stars or m-cycles, while satisfying the constraints, inject into the model, in this case explicitly, additional information about the graph structure that was not given to us. Clearly, predictions based on random m-stars versus $\mathcal{G}_{n,m}$ may be very different, as the first model trivially predicts that m-stars occur with probability 1, while they appear with a nearly zero probability in $\mathcal{G}_{n,m}$ if n and m are large.

A slightly less trivial example is the SCM. In this case the given information is that the expected degrees of nodes $i = 1, \ldots, n$ must be $k_i \in \mathbb{R}_+$, and the state space is all the $2^{\binom{n}{2}}$ graphs on n nodes. As shown in [13–16], the unique entropy-maximizing ensemble satisfying these constraints is given by random graphs in which nodes i and j are connected with probabilities $p_{ij} = 1/\left(kn/\kappa_i\kappa_j + 1\right)$, where $k = \sum_i k_i$, and κ_is are the unique solution of the system of n equations $\sum_j p_{ij} = k_i$. The popular Chung–Lu model [11,12] is different in that the connection probability there is $p_{ij}^{CL} = \min\left(k_ik_j/kn, 1\right)$, which can be thought of as a classical-limit approximation to the entropy-maximizing Fermi–Dirac p_{ij} above. While the CL ensemble also satisfies the desired constraints $\sum_j p_{ij}^{CL} = k_i$ (albeit only for sequences k_i s.t. $k_ik_j/kn \leq 1$), it does not satisfy the maximum-entropy requirement, so that it injects, in this case implicitly, some additional information into the ensemble, constraining some undesired properties of graphs in the ensemble to some ad-hoc values. Since the undesired information injection is implicit in this case, it may be quite difficult to detect and quantify all the biases introduced into the ensemble.

1.4 Main Results

The main result of this paper is the proof in Theorem 1 that the HSCM is unbiased, that is, that the HSCM random graphs maximize the Gibbs entropy of random graphs whose degree distribution and average degree converge to (4,6).

The first difficulty that we face in proving this result is how to properly formulate the entropy-maximization problem under these constraints. Indeed, we are to show that the probability distributions P_n that the HSCM defines on the set of n-sized graphs G_n maximizes the graph entropy

$$\mathcal{S}[P'_n] = -\sum_{G_n} P'_n(G_n) \log P'_n(G_n)$$

across all the distributions P'_n that define random graph ensembles with the degree distributions and average degrees converging to (4,6). These constraints are quite different than the SCM constraints, for example, because for any fixed n, we do not have a fixed set of constraints or sufficient statistics. Instead of introducing such sufficient statistics, e.g., expected degrees converging to a desired Pareto distribution, and proceeding from there, we show in Sect. 2 that the problem of graph entropy maximization under these constraints is equivalent to a graphon entropy maximization problem [33], i.e., to the problem of finding a graphon W that maximizes graphon entropy

$$\sigma_n[W'] = \iint_{A_n^2} H\left(W'(x,y)\right) d\mu_n(x) d\mu_n(y),$$

where $H(p) = -p \log p - (1-p) \log(1-p)$ is the entropy of a Bernoulli random variable with success probability p, across all the graphons W' that satisfy the constraint

$$(n-1) \int_{A_n} W'(x,y) d\mu_n(y) = \kappa_n(x),$$

where $\kappa_n(x) \approx \sqrt{\nu n} \, e^{-x}$ is the expected degree of a node at coordinate x in the power-law HSCM. We then prove in Proposition 1 that the unique solution to this graphon entropy maximization problem is given by the Fermi–Dirac graphon W in (1). The fact that the Fermi–Dirac graphon is the unique solution to this graphon entropy maximization problem is a reflection of the basic fact in statistical physics that the grand canonical ensemble of Fermi particles, which are edges of energy $x + y$ in our case, is the unique maximum-entropy ensemble with fixed expected values of energy and number of particles [34], in which the probability to find a particle at a state with energy $x + y$ is given by (1).

Yet the solutions to the graph and graphon entropy maximization problems yield equivalent random graph ensembles only if the rescaled graph entropy $\mathcal{S}^*[P_n] = \mathcal{S}[P_n]/\binom{n}{2}$ converges to the graphon entropy $\sigma_n[W]$. Here we face another difficulty that since our ensembles are sparse, both $\mathcal{S}^*[P_n]$ and $\sigma_n[W]$ converge to zero, so that we are actually to prove that the two entropies converge to each other faster than either of them converges to zero. To this end we prove in Theorems 4 and 5 that both the graphon and graph entropies converges to zero as $\sigma_n[W], \mathcal{S}^*[P_n] \sim \nu \log(n)/n$. The key result then, also in Theorem 5, is the proof that if divided by the scaling factor of $\log(n)/n$, the difference between the graphon and graph entropies vanishes in the limit, $\lim_{n\to\infty}(n/\log n)|\mathcal{S}^*[P_n] - \sigma_n[W]| \to 0$, meaning that the two entropies do indeed converge to each other faster than to zero.

The combination of graphon (1) being the entropy maximizer, and the convergence of the rescaled graph entropy to the entropy of this graphon, implies the main result in Theorem 1 that the power-law HSCM is a graph entropy maximizer subject to the degree distribution and average degree constraints (4,6).

1.5 Exchangeability and Projectivity

In addition to the natural, dictated by real-world network data, requirements of constant, i.e., independent of graphs size n, average degree and power-law degree distribution, as well as the maximum-entropy requirement, dictated by the basic statistical considerations, a reasonable model of real networks must also satisfy two more requirements: exchangeability and projectivity.

Exchangeability takes care of the fact that node labels in random graph models are usually meaningless. Even though node labels in real networks often have some network-specific meaning, such as autonomous system numbers in the Internet [9], node labels in random graph models can be, and usually are, random integer indices $i = 1, 2, \ldots$. A random graph model is exchangeable if for any permutation σ of node indices i, the probabilities of any two graphs G and G_σ given by adjacency matrices $G_{i,j}$ and $G_{\sigma(i),\sigma(j)}$ are the same, $P(G) = P(G_\sigma)$ [35,36].

A random graph model is projective if there exists a map $\pi_{n \mapsto n'}$ from graphs of size n to graphs of size $n' < n$ such that the probability of graphs in the model satisfies $P(G_{n'}) = P(\pi_{n \mapsto n'}(G_n))$ [37,38]. If this condition is satisfied, then it is easy to see that the same model admits a dual formulation as an equilibrium model of graphs of a fixed size, or as a growing graph model [39]. If this requirement is not satisfied, then as soon as one node is added to a graph, e.g., due to the growth of a real network that this graph represents, then the resulting bigger graph is effectively sampled from a different distribution corresponding to the model with different parameters, necessarily affecting the structure of its existing subgraphs, a clearly unrealistic scenario. As the simplest examples, $\mathcal{G}_{n,p}$ is projective (the map $\pi_{n \mapsto n'}$ simply selects any subset of n nodes consisting of n' nodes), but $\mathcal{G}_{n,k/n}$ with constant k is not. In the first case, one can realize $\mathcal{G}_{n,p}$ by growing graphs adding nodes one at a time, and connecting each new node to all existing nodes with probability p, while in the second case such growth is impossible since the existing edges in the growing graphs must be removed with probability $1/n$ for the resulting graphs to be samples from $\mathcal{G}_{n,k/n}$ for each n.

The HSCM random graphs are manifestly exchangeable as any graphon-based ensemble [40,41]. Here we note that the fact that these graphs are both sparse and exchangeable is *not* by any means in conflict with the Aldous–Hoover theorem [35,42] that states that the limit graphon, mapped to the unit square, of any exchangeable sparse graph family is necessarily zero. Indeed, if mapped to unit square, the limit HSCM graphon $W_{I,n}$ (3) is zero as well. We also note that the convergence of $W_{I,n}$ to zero does not mean that the ensemble converges to infinite empty graphs. In fact, the expected degree distribution and average degree in the ensemble converge to (4,6) in the limit, as stated above.

If $\alpha = 2$, the HSCM ensemble is also projective, but only with a specific labeling of nodes breaking exchangeability. This can be seen by observing that the density of points on intervals A_n, $\delta_n = n/\mu(A_n) = \alpha \left(\beta^2 \nu\right)^{\alpha/2} n^{1-\alpha/2}$, and consequently on the whole real line $\mathbb{R}$ in the limit, is constant $\delta = \nu/2$ if $\alpha = 2$. In this case, the HSCM can be equivalently defined as a model of growing labeled graphs as follows: for $n = 1, 2, \ldots$, the location x_n of new node n belongs to A_n's increment, $x_n \in B_n = A_n \setminus A_{n-1}$ ($A_0 = \emptyset$), and sampled from μ restricted to this increment, i.e., from the probability measure $\tilde{\mu}_n = \mu|_{B_n}/\mu(B_n) = \alpha\, e^{\alpha(x - R_n + R_{n-1})}$. Having its location sampled, new node n then connects to all existing nodes $i = 1 \ldots n-1$ with probability given by (1).

This growing model is equivalent to the original equilibrium HSCM definition in Sect. 1.1 only asymptotically. However, the exact equivalence, for each n, between the equilibrium HSCM with ordered x_is, $x_i < x_{i+1}$, and its growing counterpart can be also achieved by ensuring that the joint distribution of x_is is exactly the same in both cases, using basic properties of Poisson point processes [39]. Specifically, the equilibrium definition in Sect. 1.1 must be adjusted by making the right boundary R_n of interval A_n not a fixed function of n (2), but a random variable $R_n = (1/2) \log(2V_n)$, where V_n is a random variable sampled from the Gamma distribution with shape n and rate $\delta = \nu/2$. Node n is then placed at random coordinate $x_n = R_n$, while the coordinates of the rest of $n-1$ nodes are sampled from probability measure μ_n—measure μ restricted to the random interval $A_n = (-\infty, R_n]$—and

then labeled in the increasing order of their coordinates. The growing model definition must be also adjusted: the coordinate x_{n+1} of the $n + 1$'th node is determined by $v_{n+1} = v_n + V$, where $v_0 = 0$, $v_i = (1/2)e^{2x_i}$, and V is a random variable sampled from the exponential distribution with rate $\delta = \nu/2$. One can show that coordinates x_i, both for finite and infinite n, in both the equilibrium and growing HSCM models defined this way, are equivalent realizations of the same Poisson point process on $\mathbb{R}$ with measure μ and rate δ, converging to the binomial sampling with R_n fixed to (2) [39].

The projective map $\pi_{n \mapsto n'}$ in the projectivity definition above, simply maps graphs G_n to their subgraphs induced by nodes $i = 1 \ldots n'$. We note that even though the growing HSCM is not exchangeable since it relies on labeling of nodes in the increasing order of their coordinates, it is nevertheless equivalent to the equilibrium HSCM with this ordered labeling, because the joint distribution of node coordinates, and the linking probability as a function of these coordinates are the same in both the equilibrium and growing HSCM definitions [39]. This observation suggests that there might exist a less trivial projective map such that the HSCM is both projective and exchangeable at the same time.

1.6 Other Remarks

We note that thanks to its projectiveness, the power-law HSCM was shown in [24] to be equivalent to a soft version of preferential attachment, a model of growing graphs in which new nodes connect to existing nodes with probabilities proportional to the expected degrees of existing nodes. It is well-known that similar to the HSCM, the degree distribution and average degree in graphs grown according to preferential attachment do not essentially change either as graphs grow [43,44]. If $\alpha = 2$, then the equivalence between the HSCM and soft preferential attachment is exact. If $\alpha \neq 2$, then the HSCM, even with ordered labeling, is not equivalent to soft preferential attachment, but it is equivalent to its adjusted version with a certain rate of (dis)appearance of edges between existing vertices [24].

We also note that the HSCM is the zero-clustering limit [45] of random hyperbolic graphs [27], where the $\alpha = 2$ case corresponds to the uniform density of points in the hyperbolic space $\mathbb{H}^d$, and where x_is are the radial coordinates of nodes i in the spherical coordinate system of the hyperboloid model of $\mathbb{H}^d$. These coordinates can certainly not be negative, but the expected fraction of nodes with negative coordinates in the HSCM is negligible: $\mu_n(\mathbb{R}_-) = (\beta^2 \nu/n)^{\alpha/2} \to 0$. In the zero-clustering limit, the angular coordinates of nodes in $\mathbb{H}^d$ are ignored in the hyperbolic graphon [27], which becomes equivalent to (1).

As a final introductory remark we note that among the rigorous approaches to sparse exchangeable graphs, the HSCM definition is perhaps closest to graphon processes and graphexes in [46–48]. In particular in [48], where the focus is on graph convergence to well-defined limits, two ensembles are considered. One ensemble, also appearing in [47], is defined by any graphon $W : \mathbb{R}_+^2 \to [0, 1]$ and any measure μ on $\mathbb{R}_+$ ($\mathbb{R}_+$ can be replaced with any measure space). Graphs of a certain expected size, which is a growing function of time $t > 0$, are defined by sampling points as the Poisson point process on $\mathbb{R}_+$ with intensity $t\mu$, then connecting pairs of points with the probability given by W, and finally removing isolated vertices. The other ensemble is even more similar to the HSCM. It is still defined by W and μ on $\mathbb{R}_+$, but the location of vertex n on $\mathbb{R}_+$ is sampled from $\mu_n = \mu|_{A_n}/\mu(A_n)$, where A_ns are finite-size intervals growing with n whose infinite union covers the whole $\mathbb{R}_+$. The latter ensemble is not exchangeable, but both ensembles are shown to converge to properly stretched graphons defined by W, yet only if the expected average degree grows to infinity in the limit. The HSCM definition is different—in particular all n vertices of n-sized

graphs are sampled from the same μ_n—ensuring exchangeability, and allowing for an explicit control of the degree distribution and average degree, which can be constant, but making the problem of graph convergence difficult. We do not further discuss graph convergence here, leaving it, as well as the generalization of the results to arbitrary degree distributions, for future publications.

1.7 Paper Organization

In the next section we first review in more detail the necessary background information and provide all the required definitions. In Sect. 3 we formally state all the results in the paper, while Sect. 4 contains all the proofs of these results.

2 Background Information and Definitions

2.1 Graph Ensembles and Their Entropy

A graph ensemble is a set of graphs $\mathcal{G}$ with probability measure P on $\mathcal{G}$. The Gibbs entropy of the ensemble is

$$S[P] = - \sum_{G \in \mathcal{G}} P(G) \log P(G) \tag{7}$$

Note that this is just the entropy of the random variable G with respect to the probability measure P. When G_n is a graph of size n, sampled from $\mathcal{G}$ according to measure P, we write $S[G_n]$ instead of $S[P]$. Given a set of constraints, e.g., in the form of graph properties fixed to given values, the maximum-entropy ensemble is given by P^* that maximizes $S[P]$ across all measures P that satisfy the constraints. These constraints can be either sharp (microcanonical) or soft (canonical), satisfied either exactly or on average, respectively. The simplest example of the constrained graph property is the number of edges, fixed to m, in graphs of size n. The corresponding microcanonical and canonical maximum-entropy ensembles are $\mathcal{G}_{n,m}$ and $\mathcal{G}_{n,p}$ with $p = m/\binom{n}{2}$, respectively. The P^* is respectively the uniform and exponential Boltzmann distribution $P(G) = e^{-H(G)}/Z$ with Hamiltonian $H(G) = \lambda m(G)$, where $m(G)$ is the number of edges in graph G, and the Lagrange multiplier λ is given by $p = 1/(e^\lambda + 1)$ [13].

When the constraints are given by the degrees of nodes, instead of the number of edges, we have the following characterization of the microcanonical and canonical ensemble.

2.1.1 Maximum-Entropy Graphs with a Given Degree Sequence (CM)

Given a degree sequence $\mathbf{d}_n = d_1 \ldots d_n$, the microcanonical ensemble of graphs that have this degree sequence is the configuration model (CM) [7,8]. The entropy-maximizing P^* is uniform on the set of all graphs that have degree sequence $\mathbf{d}_n$.

2.1.2 Maximum-Entropy Graphs with a Given Expected Degree Sequence (SCM)

If the sharp CM constraints are relaxed to soft constraints, the result is the canonical ensemble of the soft configuration model (SCM) [14,15]. Given an expected degree sequence $\mathbf{k}_n$, which in contrast to CM's $\mathbf{d}_n$, does not have to be a graphical sequence of non-negative integers,

but can be any sequence of non-negative real numbers, the SCM is defined by connecting nodes i and j with probabilities

$$p_{ij} = \frac{1}{e^{\lambda_i + \lambda_j} + 1}, \quad \text{where Lagrange multipliers } \lambda_i \text{ are the solution of} \tag{8}$$

$$k_i = \sum_{i < j} p_{ij}, \quad i = 1 \ldots n. \tag{9}$$

The entropy-maximizing P^* is the Boltzmann distribution with Hamiltonian $H(G) = \sum_i \lambda_i d_i(G)$, where $d_i(G)$ is the degree of node i in graph G [14,15].

2.2 Sparse Graphs with a Given Degree Distribution

Let $p(k), k = 0, 1, 2, \ldots$, be a probability density function with finite mean. Denote by $\mathcal{D}$ the corresponding random variable, and consider a sequence of graph ensembles that maximize Gibbs entropy under the constraint that for all k

$$\lim_{n \to \infty} \mathbb{P}(\mathcal{D}_n = k) = \mathbb{P}(\mathcal{D} = k), \tag{C1}$$

where $\mathcal{D}_n$ is the degree of a uniformly chosen node in the ensemble of graphs of size n. In other words, this is a maximum-entropy ensemble of graphs whose degree distribution converges to $p(k)$.

In addition to the degree distribution constraint, we also want our graphs to be sparse. The most common definition of *sparseness* seems to be that the number of edges is $o(n^2)$, so that the expected average degree can be unbounded. In contrast, here we use the term *sparse* to mean that the expected degree converges to the expected value of $\mathcal{D}$:

$$\lim_{n \to \infty} \mathbb{E}[\mathcal{D}_n] = \mathbb{E}[\mathcal{D}] := \nu. \tag{C2}$$

Constraint (C2) implies that the number of edges is $O(n)$. We note that, in general, (C2) does not follow from (C1), since convergence in distribution does not necessarily imply convergence in expectation.

We also note that constraints (C1,C2) are neither sharp nor soft, since they deal only with the $n \to \infty$ limits of the degree distribution and expected degree. We call these constraints *hypersoft*, since as we will see below, random graphs satisfying these constraints can be realized as random graphs with Lagrange multipliers that are not parameters but *hyperparameters* in the statistics terminology [25].

2.3 Maximum-Entropy Graphs with Hypersoft Constraints

Similar to the case of random graphs with a given (expected) degree sequence, we are to determine the distribution P that satisfies (C1) and (C2), and maximizes the Gibbs entropy. However, this task poses the question of what it means to maximize entropy under these limit constraints. In particular, unlike an ensemble of graphs with a given (expected) degree sequence, we are no longer dealing with a set of graphs of fixed size, but with a sequence of graphs $(G_n)_{n \geq 1}$ of varying sizes. To answer this question, and to give a proper definition of entropy maximization under hypersoft constraints, we consider graphon-based ensembles of graphs.

2.3.1 Graphon-Based Graph Ensembles

In the simplest case, a graphon is a symmetric integrable function $W : [0, 1]^2 \to [0, 1]$. Graphons, or more precisely graphon equivalence classes, consisting of all functions W_σ such that $W(x, y) = W_\sigma(\sigma(x), \sigma(y))$ under all measure preserving transformations $\sigma : [0, 1] \to [0, 1]$, are well-defined limits of dense graph families [41,49]. One can think of the interval $[0, 1]$ as the continuum limit of node indices i, and of W as the limit of graphs' adjacency matrices. Equivalently, $W(x, y)$ is the probability that there exists the edge between "nodes" x and y. Graphons are an application to graphs of a class of earlier results on exchangeable arrays in statistics [35,50], and are better known as the *connection probability* in random graphs with latent parameters in sociology [51–54] and network science [21,22], also known in graph theory as *inhomogeneous random graphs* [55]. Here we use the term *graphon* to refer to any symmetric function $W : A \times A \to [0, 1]$, for some $A \subseteq \mathbb{R}$.

Let μ be a probability measure on $A \subseteq \mathbb{R}$ and W a graphon. Then the standard graphon-based ensemble $\mathcal{G}_{W,\mu,A}$ known as W-random graphs [40] is the ensemble of random n-sized graphs G_n defined by first i.i.d. sampling n node coordinates $\mathbf{x}_n = x_1, \ldots, x_n \in A$ according to μ, and then connecting every node pair (i, j), independently, with probability $W(x_i, x_j)$.

To be able to satisfy the hypersoft constraints we generalize this ensemble as follows. Let μ be a measure on $\mathbb{R}$, W a graphon, and let $\mathbf{A} = (A_1, A_2, \ldots)$, $A_n \subseteq A_{n+1}$, be an infinite sequence of growing subsets of $\mathbb{R}$ such that $\bigcup_{n \geq 1} A_n = \mathbb{R}$ and $\mu(A_n) < \infty$. We then define the graphon ensemble $\mathcal{G}_{W,\mu,\mathbf{A}} = (\mathcal{G}_{W,\mu_n,n})_{n \geq 1}$ to be random graphs $(G_n)_{n \geq 1}$ defined by the graphon W and measures

$$\mu_n = \frac{\mu|_{A_n}}{\mu(A_n)},$$

which are the probability measures on A_n associated with μ. To sample a graph G_n from this ensemble, n i.i.d. points $x_i \in A_n$ are first sampled according to measure μ_n, and then pairs of nodes i and j are connected by an edge with probability $W(x_i, x_j)$.

We remark that if $A_n = A = [0, 1]$ and μ is the uniform measure, we are in the most classical settings of W-random graphs [40]. In the case of an arbitrary $\mathbf{A}$ and measure μ, our random graph ensemble is similar to a model considered recently in [48], Sect. 2.4. There, a growing graph construction is considered where G_n is created from G_{n-1} by sampling the coordinate x_n of node n according to measure μ_n, and then connecting it to all existing nodes i, independently with probability $W(x_n, x_i)$. The main difference between such sampling and the one above is that in the former case, the coordinates x_i of different nodes $i = 1 \ldots n$ are sampled from different measures μ_i, thus breaking exchangeability, while in the latter case, all x_is are sampled from the same measure μ_n.

2.3.2 Bernoulli and Graphon Entropies

Given the coordinates $\mathbf{x}_n$, edges in our random graphs are independent Bernoulli random variables, albeit with different success probabilities $W(x_i, x_j)$. The conditional Bernoulli entropy of random graphs with fixed coordinates $\mathbf{x}_n$ is thus

$$\mathcal{S}[G_n|\mathbf{x}_n] = \sum_{i<j} H\left(W(x_i, x_j)\right),$$

where H is the Bernoulli entropy,

$$H(p) = -p \log p - (1 - p) \log(1 - p), \quad 0 \leq p \leq 1. \tag{10}$$

The *graphon entropy* [41,50,56–59] of W with respect to μ on A is defined as

$$\sigma[W, \mu, A] = \iint_{A^2} H\left(W(x, y)\right) d\mu(x) d\mu(y), \tag{11}$$

which we write as $\sigma[W, \mu]$, when μ has support A and no confusion arises. In addition, if G_n is a graph in the graphon ensemble $\mathcal{G}_{W,\mu_n,n}$, we write $\sigma[G_n]$ for $\sigma[W, \mu_n, A_n]$.

Since for any two discrete random variables X and Y, the expectation of the conditional entropy of X given Y is a lower bound for X's entropy, $\mathcal{S}[X] \geq \mathbb{E}[\mathcal{S}[X|Y]]$, the graphon entropy definition implies

$$\mathcal{S}[G_n] \geq \mathbb{E}[\mathcal{S}[G_n|\mathbf{x}_n]] = \binom{n}{2}\mathbb{E}[H(W(x, y))] = \binom{n}{2}\sigma[G_n], \tag{12}$$

The graphon entropy is thus a lower bound for the rescaled Gibbs entropy defined as

$$\mathcal{S}^*[G_n] = \frac{\mathcal{S}[G_n]}{\binom{n}{2}}. \tag{13}$$

Before we give our definition of the maximum-entropy ensemble of sparse graphs, it is instructive to consider the case of dense graphs.

2.3.3 Dense Maximum-Entropy Graphs with a Given Degree Distribution

Consider a sequence $(\mathbf{d}_n)_{n\geq 1}$ of degree sequences $\mathbf{d}_n = (d_1 \ldots d_n)_n$, $(d_i)_n \geq (d_{i+1})_n$, for which there exist constants $0 < c_1 \leq c_2$ such that $c_1 n \leq (d_i)_n \leq c_2 n$ for any $(d_i)_n \in \mathbf{d}_n$. Now let $(\mathcal{G}_n)_{n\geq 1}$ be a sequence of microcanonical ensembles of random graphs, that is, CMs, defined by $\mathbf{d}_n$. If there exists a function $f : [0, 1] \to (0, 1)$ such that for any $c \in (0, 1)$

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\mathcal{D}_n}{n} \leq c\right) = \mathbb{P}\left(f(U) \leq c\right), \tag{14}$$

where $\mathcal{D}_n$ is the degree of a random node in random graph $G_n \in \mathcal{G}_n$ or equivalently, a uniformly sampled $(d_i)_n \in \mathbf{d}_n$, and U is a uniform random variable on $[0, 1]$, then it was proven in [20] that the limit of the CM sequence $(\mathcal{G}_n)_{n\geq 1}$ is given by the graphon

$$W(x, y) = \frac{1}{e^{g(x)+g(y)} + 1}, \tag{15}$$

where $g(x)$ is such that

$$f(x) = \int_0^1 W(x, y) \, dy = \int_A W_A(x, y) \, d\mu(y) = \int_A \frac{1}{e^{x+y} + 1} \, d\mu(y), \tag{16}$$

where A is the image of g and $\mu = g^{-1}$ (functions f and g are continuous and strictly increasing almost everywhere on $[0, 1]$ [20]).

Some important observations are in order here. First, we note that (14) is very similar to (C1), with the exception that (14) implies that the degrees of all nodes are $O(n)$, so that the graphs are dense. In particular, in our graphs we have that $\mathcal{D}_n = o(n)$ so that $\mathcal{D}_n/n \to 0$.

Second, consider the problem of maximizing the graphon entropy under the constraint given by (16). We will show in Proposition 1 that the solution to this problem is given by (15), where g is defined by (16). Hence the graphon (15) obtained in [20] maximizes the graphon entropy under the constraint (16) imposed by the limit f of the sequence of rescaled degree sequences $(\mathbf{d}_n/n)_{n\geq 1}$.

Third, Theorem D.5 in [41] states that in dense W-random graph ensembles $\mathcal{G}_{W,\mu,A}$

$$\lim_{n \to \infty} \mathcal{S}^*[G_n] = \sigma[W, \mu, A],\tag{17}$$

meaning that the rescaled Gibbs entropy of W-random graphs converges to the graphon entropy of W.

Given f, this result suggests to consider the family of W-random graph ensembles $\tilde{G}_n$ defined by the Fermi–Dirac graphon (15) with g given by (16), which we call the dense hypersoft configuration model (HSCM). The distribution of rescaled degrees $\mathcal{D}_n/n$ in these HSCM graphs converges to f, cf. (14), while the limit of these graphs is also W since the limit of any dense W-random graphs is W [60]. That is, the limit of HSCM ensembles $\tilde{G}_n$ and the limit of CM ensembles $\mathcal{G}_n$ with any $(\mathbf{d}_n/n)_{n \geq 1}$ converging to f, are the same Fermi–Dirac graphon W (15). Since the two ensembles have the same graphon limit, their rescaled Gibbs entropies converge to the same value, equal, thanks to (17), to the graphon entropy, even though for any finite n the two ensembles are quite different.

Fourth, if we replace the sequence of degree sequences $(\mathbf{d}_n)_{n \geq 1}$ with a sequence of expected degree sequences $(\mathbf{k}_n)_{n \geq 1}$ converging upon rescaling to f, and then replace the sequence of CMs with the corresponding sequence of SCMs, then the limit of this SCM sequence is the same graphon W (15) [20], so that the rescaled SCM Gibbs entropy also converges to the same graphon entropy, squeezed, for any finite n, between the CM entropy [7,61,62] and the HSCM entropy (12). In other words, dense CM, SCM, and HSCM are all equivalent in the limit, versus the sparse case where the equivalence is broken [18,19] since the graphon is zero in the limit.

The key point here however is that the rescaled degree distribution in the dense HSCM converges to a well-defined limit, i.e., satisfies the hypersoft constraints (14), and that the HSCM Gibbs entropy converges to the graphon entropy. Therefore if we *define* a maximum-entropy ensemble under given hypersoft constraints f to be an ensemble that: (1) satisfies these constraints, i.e., has a degree distribution converging to f in the limit, (2) maximizes graphon entropy under these constraints given by (16), and (3) has rescaled Gibbs entropy converging to the graphon entropy, then the dense HSCM ensemble is trivially such an ensemble. In addition, this ensemble is the unique maximum-entropy hypersoft ensemble in the dense case [20].

These observations instruct us to extend this definition of maximum-entropy hypersoft dense graphs to sparse graphs, where we naturally replace the dense hypersoft constraints (14) with sparse hypersoft constraints (C1,C2). However, things become immediately less trivial in this case. In particular, we face the difficulty that since the limit graphon of any sparse exchangeable graph ensemble is zero according to the Aldous–Hoover theorem [35,42], the entropy of this graphon is zero as well. Since this entropy is zero, $|\mathcal{S}^*[G_n] - \sigma[G_n]| \to 0$ in (17) does not necessarily imply that the rescaled Gibbs entropy converges to the graphon entropy. We address this difficulty next.

2.3.4 Rescaled Graphon Entropy of Sparse Graphs

Consider again the generalized graphon ensemble $\mathcal{G}_{W,\mu_n,n}$ of random graphs defined in Sect. 2.3.1, and their graphon entropy defined as

$$\sigma[G_n] = \iint_{A_n^2} H\left(W(x, y)\right) d\mu_n(x) d\mu_n(y).\tag{18}$$

If $W \neq \{0, 1\}$ everywhere, then for any finite n, $\sigma[G_n]$ is positive, but if the ensemble is sparse, then $\lim_{n \to \infty} \sigma[G_n] = 0$.

To address this problem we rescale the graphon entropy $\sigma[G_n]$ such that upon rescaling it converges to a positive constant. That is, let a_n be a sequence such that

$$\lim_{n \to \infty} a_n \sigma[G_n] = a \in (0, \infty). \tag{19}$$

This rescaling does not affect the graphon entropy maximization problem, because maximizing $\sigma[G_n]$ for every n as a functional of W under a given constraint is equivalent to maximizing $a_n \sigma[G_n]$ under the same constraint.

Upon this rescaling, we see that the rescaled Gibbs entropy converges to the graphon entropy, generalizing (17), if

$$\lim_{n \to \infty} a_n \left| \mathcal{S}^*[G_n] - \sigma[G_n] \right| = 0, \tag{20}$$

in which case $a_n \mathcal{S}^*[G_n]$ converges to a. This condition implies that the rescaled Gibbs entropy $\mathcal{S}^*[G_n]$ converges to the graphon entropy $\sigma[G_n]$ faster than either of them converge to zero.

2.3.5 Sparse Maximum-Entropy Graphs with a Given Degree Distribution

With the graphon rescaling in the previous section, we can now *define* a graphon ensemble $\mathcal{G}_{W,\mu_n,n}$ to be a maximum-entropy ensemble under the sparse hypersoft constraints (C1,C2) if:

(1) the degree distribution and expected degree in $\mathcal{G}_{W,\mu_n,n}$ converge to (C1,C2);
(2) graphon entropy $\sigma[G_n]$ of $\mathcal{G}_{W,\mu_n,n}$ is maximized for every n under the constraint imposed by (C1,C2); and
(3) (20) holds, with a_n given by (19).

Our main result (Theorem 1) is that the sparse power-law hypersoft configuration model, defined next, is a maximum-entropy model under hypersoft constraints.

2.3.6 Sparse Power-Law Hypersoft Configuration Model (Sparse HSCM)

The sparse power-law HSCM(α, ν) is defined as the graphon ensemble $\mathcal{G}_{W,\mu_n,n} = \mathcal{G}_{W,\mu,\mathbf{A}}$, Sect. 2.3.1, with

$$W(x, y) = \frac{1}{e^{x+y} + 1}, \tag{21}$$

$$\mu = e^{\alpha x}, \quad \alpha > 1, \tag{22}$$

$$A_n = (-\infty, R_n], \tag{23}$$

$$R_n = \frac{1}{2} \log \frac{n}{\nu \beta^2}, \quad \nu > 0, \quad \beta = 1 - \frac{1}{\alpha}, \tag{24}$$

$$\mu_n = \frac{\mu|_{A_n}}{\mu(A_n)} = \alpha \, e^{\alpha(x - R_n)}. \tag{25}$$

The dense power-law HSCM is recovered from the above definition by setting $\nu = \tilde{\nu}n$, where $\tilde{\nu}$ is a constant, in which case $R_n = R = -(1/2) \log(\tilde{\nu}\beta^2)$, $A_n = A = (-\infty, R]$, and $\mu_n = \mu = \alpha \, e^{\alpha(x-R)}$.

3 Results

In this section we formally state our results, and provide brief overviews of their proofs appearing in subsequent sections. The main result is Theorem 1, stating that the HSCM(α, ν) defined in Sect. 2.3.6 is a maximum-entropy model under hypersoft power-law degree distribution constraints, according to the definition in Sect. 2.3.5. This result follows from Theorems 2–5 and Proposition 1. Theorems 2,3 establish the limits of the degree distribution and expected average degree in the HSCM(α, ν). Proposition 1 states that HSCM's graphon uniquely maximizes the graphon entropy under the constraints imposed by the degree distribution. Theorem 4 establishes proper graphon rescaling and the limit of the rescaled graphon. Finally, the most critical and involved Theorem 5 proves that the rescaled Gibbs entropy of the HSCM converges to its rescaled graphon entropy.

3.1 Main Result

Let Y be a Pareto random variable with shape $\alpha > 1$, scale $\nu\beta > 0$, $\beta = 1 - 1/\alpha$, so that Y's probability density function is

$$P_Y(y) = \alpha (\nu\beta)^\alpha y^{-\gamma}, \quad y \geq \nu\beta, \quad \gamma = \alpha + 1, \quad \text{and} \tag{26}$$

$$\mathbb{P}(Y > y) = \begin{cases} (\nu\beta)^\alpha y^{-\alpha} & \text{if } y \geq \nu\beta \\ 1 & \text{otherwise.} \end{cases} \tag{27}$$

Let $\mathcal{D}$ be a discrete random variable with probability density function

$$\mathbb{P}(\mathcal{D} = k) = \mathbb{E}\left[\frac{Y^k}{k!} e^{-Y}\right], \quad k = 0, 1, 2, \ldots, \tag{28}$$

which is the mixed Poisson distribution with mixing parameter Y [63]. Then it follows that

$$\mathbb{E}[\mathcal{D}] = \mathbb{E}[Y] = \nu, \tag{29}$$

and since Y is a power law with exponent γ, the tail of $\mathcal{D}$'s distribution is also a power law with the same exponent [63]. In particular, $\mathbb{P}(\mathcal{D} = k)$ is given by (4). Therefore if $\mathcal{D}$ is the degree of a random node in a random graph ensemble, then graphs in this ensemble are sparse and have a power-law degree distribution.

Our main result is:

Theorem 1 *For any $\alpha > 1$ and $\nu > 0$, HSCM(α, ν) is a maximum entropy ensemble of random graphs under the hypersoft constraints (C1,C2) with $\mathbb{P}(\mathcal{D} = k)$ and ν defined by (26–29).*

3.2 The Limit of the Degree Distribution in the HSCM

The degree $\mathcal{D}_n$ of a random node i in a random HSCM(α, ν) graph of size n, conditioned on the node coordinates $\mathbf{x}_n = x_1 \ldots x_n$, is the sum of $n - 1$ independent Bernoulli random variables with success probabilities $W(x_i, x_j)$, $j \neq i$. The distribution of this sum can be approximated by the mixed Poisson distribution with the mixing parameter $\sum_{j\neq i} W(x_i, x_j)$. Therefore after first integrating over x_j with $j \neq i$ and then over x_i, the distribution of $\mathcal{D}_n$ is approximately the mixed Poisson distribution

$$\mathbb{P}(\mathcal{D}_n = k) = \mathbb{E}\left[\frac{(\kappa_n(X))^k}{k!} e^{-\kappa_n(X)}\right], \quad \text{so that } \mathbb{E}[\mathcal{D}_n] = \mathbb{E}[\kappa_n(X)],$$

where the random variable X has density μ_n (25), and the mixing parameter $\kappa_n(x)$ is the expected degree of a node at coordinate x:

$$\kappa_n(x) = (n-1)w_n(x), \tag{30}$$

$$w_n(x) = \int_{A_n} W(x, y)\, d\mu_n(y), \tag{31}$$

where A_n is given by (23).

In the $n \to \infty$ limit, the distribution of the expected degree $\kappa_n(X)$ of X converges to the Pareto distribution with shape α and scale $\nu\beta$. To prove this, we use the observation that the mass of measure μ_n is concentrated towards the right end of the interval $A_n = (-\infty, R_n]$, where $R_n \gg 1$ for large n. Therefore, not only the contributions coming from negative x, y are negligible, but we can also approximate the Fermi–Dirac graphon $W(x, y)$ in (21) with its classical limit approximation

$$\widehat{W}(x, y) = e^{-(x+y)} \tag{32}$$

on $\mathbb{R}_+^2$. In addition, the expected degree function $w_n(x)$ can be approximated with $\widehat{w}_n(x)$ defined by

$$\widehat{w}_n(x) = \begin{cases} \omega_n e^{-x} & \text{if } 0 \le x \le R_n \\ 0 & \text{otherwise} \end{cases}, \text{ where} \tag{33}$$

$$\omega_n = \int_0^{R_n} e^{-x}\, d\mu_n(x) = \frac{1 - e^{-(\alpha-1)R_n}}{\beta e^{R_n}} = \left(\frac{\nu}{n}\right)^{\frac{1}{2}} + o\left(n^{-\frac{1}{2}}\right), \tag{34}$$

so that the expected degree of a node at coordinate x can be approximated by

$$\widehat{\kappa}_n(x) = n\widehat{w}_n(x) = e^{-x}\left((\nu n)^{\frac{1}{2}} + o\left(n^{\frac{1}{2}}\right)\right).$$

To see that $\widehat{\kappa}_n(X)$ converges to a Pareto random variable, note that since X has density μ_n, it follows that for all $t > \nu\beta$

$$\mathbb{P}\left(\widehat{\kappa}_n(X) > t\right) = \mathbb{P}\left(X < \log\frac{n\omega_n}{t}\right) = e^{-\alpha R_n}\left(\frac{n\omega_n}{t}\right)^\alpha$$
$$= (\nu\beta)^\alpha\, t^{-\alpha}\, (1 + o(1)) = \mathbb{P}(Y > t)\,(1 + o(1)),$$

where Y is a Pareto-distributed random variable (27). We therefore have the following result, the full proof of which can be found in Sect. 4.2.2:

Theorem 2 (HSCM(α, ν) satisfies (C1)) *Let $\alpha > 1$, $\nu > 0$, and $\mathcal{D}_n$ be the degree of a uniformly chosen vertex in the HSCM(α, ν) graphs of size n. Then, for each $k = 0, 1, 2, \ldots$,*

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{D}_n = k) = \mathbb{P}(\mathcal{D} = k),$$

where $\mathbb{P}(\mathcal{D} = k)$ is given by (28).

3.3 The Limit of the Expected Average Degree in the HSCM

The expected degree of a random node in n-sized HSCM graphs is, for any fixed i,

$$\mathbb{E}[\mathcal{D}_n] = \sum_{j \ne i} \mathbb{E}\left[W(X_i, X_j)\right] = (n-1)\mathbb{E}[W(X, Y)], \tag{35}$$

where X and Y are independent random variables with distribution μ_n. Approximating $W(x, y)$ with $\widehat{W}(x, y)$ on $\mathbb{R}_+^2$, and using $e^{2R_n} = n/\nu\beta^2$, we have

$$n \iint_0^{R_n} \widehat{W}(x, y)\, d\mu_n(x)\, d\mu_n(y) = n \left(\int_0^{R_n} e^{-x}\, d\mu_n(x) \right)^2 = \frac{n}{\beta^2 e^{2R_n}} \left(1 - e^{-(\alpha-1)R_n} \right)^2$$

$$= \nu + o(1).$$

All other contributions are shown to also vanish in the $n \to \infty$ limit in Sect. 4.2.3, where the following theorem is proved:

Theorem 3 (HSCM satisfies (C2)) *Let $\alpha > 1$, $\nu > 0$, and $\mathcal{D}_n$ be the degree of a uniformly chosen vertex in the HSCM(α, ν) graphs of size n. Then*

$$\lim_{n\to\infty} \mathbb{E}[\mathcal{D}_n] = \nu.$$

3.4 HSCM Maximizes Graphon Entropy

Let $A \subseteq \mathbb{R}$ be some interval, μ a measure on A with $\mu(A) < \infty$, and suppose some μ-integrable function $w : A \to \mathbb{R}$ is given. Consider the graphon entropy maximization problem under the constraint

$$w(x) = \int_A W(x, y)\, d\mu(y). \tag{36}$$

That is, the problem is to find a symmetric function W^* that maximizes graphon entropy $\sigma[W, \mu, A]$ in (11) and satisfies the constraint above, for fixed A and μ.

We note that this problem is a "continuous version" of the Gibbs entropy maximization problem in the SCM ensemble in Sect. 2.1.2. The following proposition, which we prove in Sect. 4.3.1, states that the solution to this problem is a "continuous version" of the SCM solution (8):

Proposition 1 (HSCM Maximizes Graphon Entropy) *Suppose there exists a solution W^* to the graphon entropy maximization problem defined above. Then this solution has the following form:*

$$W^*(x, y) = \frac{1}{e^{\lambda(x)+\lambda(y)} + 1}, \tag{37}$$

where function λ is μ-almost-everywhere uniquely defined on A by (36).

This proposition proves that for each n, the HSCM(α, ν)'s Fermi–Dirac graphon (21) maximizes the graphon entropy under constraint (31), because $\mathbf{A}$ and μ in the HSCM(α, ν) are chosen such that $\lambda(x) = x$. This is always possible as soon as $\lambda(x)$ is invertible, cf. Sect. 2.3.3. For each n, interval A_n (23) and measure μ_n (25) in the HSCM(α, ν) can be mapped to [0, 1] and 1, respectively, in which case $\lambda_n(x) = R_n + (1/\alpha) \log x$, leading to (3). In other words, node coordinates x in the original HSCM(α, ν) definition in Sect. 2.3.6, and their coordinates $\tilde{x}$ in its equivalent definition with $A_n = [0, 1]$ and $\mu_n = 1$ are related by

$$\tilde{x} = e^{\alpha(x-R_n)}.$$

3.5 Graphon Entropy Scaling and Convergence

To derive the rate of convergence of HSCM's graphon entropy to zero, it suffices to consider the classical limit approximation $\widehat{W}$ (32) to W (21) on $\mathbb{R}_+^2$. Its Bernoulli entropy (10) is

$$H\left(\widehat{W}(x, y)\right) = (x + y)\widehat{W}(x, y) - (1 - \widehat{W}(x, y))\log(1 - \widehat{W}(x, y)).$$

Since the most of the mass of μ_n is concentrated near R_n and $\widehat{W}(R_n, R_n) \to 0$, the second term is negligible. Integrating the first term over $[0, R_n]$, we get

$$\iint_0^{R_n} (x + y)\widehat{W}(x, y)\, d\mu_n(x)\, d\mu_n(y)$$

$$= \alpha^2 e^{-2\alpha R_n} \iint_0^{R_n} (x + y)e^{(\alpha-1)(x+y)}\, dx\, dy$$

$$= \frac{2R_n}{\beta^2 e^{2R_n}} + O(n^{-1}) = \frac{\nu}{n}\log\frac{n}{\nu\beta^2} + O(n^{-1}) = \nu\frac{\log n}{n} + O(n^{-1}), \qquad (38)$$

from which we obtain the proper scaling as $\log(n)/n$. All further details behind the proof of the following theorem are in Sect. 4.3.2.

Theorem 4 (Graphon Entropy Convergence) *Let $\sigma[G_n]$ be the graphon entropy* (18) *in the* HSCM(α, ν) *ensemble with any $\alpha > 1$ and $\nu > 0$. Then, as $n \to \infty$,*

$$\left|\frac{n\,\sigma[G_n]}{\log n} - \nu\right| = O\left(1/\log n\right).$$

This theorem implies that $\sigma[G_n]$ goes to zero as $\log(n)/n$, while $n\,\sigma[G_n]/\log n$ goes to ν as $1/\log n$.

3.6 Gibbs Entropy Scaling and Convergence

The last part of Theorem 1 is to prove that the rescaled Gibbs entropy (13) of the HSCM converges to its graphon entropy (18) faster than the latter converges to zero.

The graphon entropy is a trivial lower bound for the rescaled Gibbs entropy, Sect. 2.3.2, so the problem is to find an appropriate upper bound for the latter converging to the graphon entropy. To identify such an upper bound, we rely on an argument similar to [41]. Specifically, we first partition A_n into m intervals I_t that induce a partition of A_n^2 into rectangles $I_{st} = I_s \times I_t$, $s, t = 1 \ldots m$. We then approximate the graphon by its average value on each rectangle. Such approximation brings in an error term on each rectangle. We then show that the Gibbs entropy is upper-bounded by the entropy of the averaged graphon, plus the sum of entropies of indicator random variables M_i which take value t if the coordinate x_i of node i happens to fall within interval I_t. The smaller the number of intervals m, the smaller the total entropy of these random variables M_i, but the larger the sum of the error terms coming from graphon averaging, because rectangles I_{st} are large. The smaller they are, the smaller the total error term, but the larger the total entropy of the M_i's. The crux of the proof is to find a "sweet spot"—the right number of intervals guaranteeing the proper balance between these two types of contributions to the upper bound, which we want to be tighter than the rate of the convergence of the graphon entropy to zero.

This program is executed in Sect. 4.4, where we prove the following theorem:

Theorem 5 (Gibbs Entropy Convergence) *Let* $\sigma[G_n]$ *be the graphon entropy* (18) *and* $\mathcal{S}^*[G_n]$ *be the rescaled Gibbs entropy* (13) *in the* HSCM(α, ν) *ensemble with any* $\alpha > 1$ *and* $\nu > 0$. *Then*

$$\lim_{n \to \infty} \frac{n}{\log n} \left| \mathcal{S}^*[G_n] - \sigma[G_n] \right| = 0, \text{ and}$$

$$\lim_{n \to \infty} \frac{2 \, \mathcal{S}[G_n]}{n \log n} = \nu.$$

We remark that this theorem implies that

$$\mathcal{S}[G_n] \sim \frac{\nu}{2} n \log n,$$

which is the leading term of the (S)CM Gibbs entropy obtained in [23]. It is also instructive to compare this scaling of Gibbs entropy with its scaling in dense ensembles with $\lim_{n \to \infty} \sigma[G_n] = \sigma[W, \mu, A] = \sigma \in (0, \infty)$ [41]:

$$\mathcal{S}[G_n] \sim \frac{\sigma}{2} n^2.$$

Finally, it is worth mentioning that even though we use the Fermi–Dirac graphon W (21) to define our W-random graphs, the same convergence results could be obtained for W-random graphs defined by any other graphon W' such that

$$\lim_{n \to \infty} n \, \mathbb{E} \left[\left| W(X, Y) - W'(X, Y) \right| \right] = 0,$$

with X and Y having density μ_n. In fact, to establish the required limits, we use the classical limit approximation graphon $\widehat{W}$ instead of W. Therefore there exists a vast equivalence class of W-random graphs defined by graphons W' that all have the same limit degree distribution (C1) and average degree (C2), and whose rescaled Gibbs entropy converges to the graphon entropy of the Fermi–Dirac W. However, it follows from Proposition 1 that among all these ensembles, only the W-random graph ensemble defined by the Fermi–Dirac graphon (21) uniquely maximizes the graphon entropy (11) for each n, which, by our definition of maximum-entropy ensembles under hypersoft constraints, is a necessary condition for graph entropy maximization.

4 Proofs

In this section we provide the proofs of all the results stated in the previous section. In Sect. 4.1 we begin with some preliminary results on the accuracy of the approximation of the Fermi–Dirac graphon W (21) by the classical limit approximation $\widehat{W}$. In the same section we also establish results showing that the main contribution of the integration with respect to μ_n is on the positive part $[0, R_n]$ of the interval A_n as defined by (23). In particular, we show that all contributions coming from the negative part of this interval, i.e., $\mathbb{R}_-$, are $o(n^{-1})$, which means that for all our results the negative part of the support of our measure μ_n is negligible. We then proceed with proving Theorems 2 and 3 in Sect. 4.2. The proofs of Proposition 1 and Theorem 4 can be found in Sect. 4.3. Finally, the convergence of the rescaled Gibbs entropy to the graphon entropy (Theorem 5) is given in Sect. 4.4.

4.1 The Classical Limit Approximation of the Fermi–Dirac Graphon

We will use $e^{-(x+y)}$ as an approximation to the graphon W to compute all necessary limits. To be precise we define

$$\widehat{W}(x, y) = \min\{e^{-(x+y)}, 1\} \tag{39}$$

and show that differences between the integrals of W and $\widehat{W}$ converge to zero as n tends to infinity. Note that, instead of $\widehat{W}$, we could have also worked with the integral expressions involving W, which might have led to better bounds. However, these integrals tend to evaluate to combinations of hypergeometric functions, while the integrals of $\widehat{W}$ are much easier to evaluate and are sufficient for our purposes.

By the definition of $\widehat{W}(x, y)$ we need to consider separately, the intervals $(-\infty, 0]$ and $(0, R_n]$. Since graphons are symmetric functions, this leads to the following three different cases:

I) $-\infty < x, y \leq 0$ II) $-\infty < y \leq R_n$ and $0 < x \leq R_n$,   III) $0 < x, y \leq R_n$.

For case I) we note that $W, \widehat{W} \leq 1$ and

$$\iint_{-\infty}^{0} d\mu_n(y) \, d\mu_n(x) = O\left(n^{-\alpha}\right). \tag{40}$$

With this we obtain the following result, which shows that for both W and $\widehat{W}$, only the integration over $(0, R_n]^2$, i.e. case III, matters.

Lemma 1

$$\iint_{-\infty}^{R_n} \widehat{W}(x, y) \, d\mu_n(y) \, d\mu_n(x) - \iint_{0}^{R_n} \widehat{W}(x, y) \, d\mu_n(y) \, d\mu_n(x) = O\left(n^{-\frac{\alpha+1}{2}}\right)$$

and the same result holds if we replace $\widehat{W}$ with W.

Proof First note that

$$\int_{-\infty}^{-R_n} \int_{-\infty}^{R_n} \widehat{W}(x, y) \, d\mu_n(y) \, d\mu_n(x) \leq \int_{-\infty}^{-R_n} d\mu_n(y) = O\left(n^{-\alpha}\right). \tag{41}$$

We show that

$$\int_{-R_n}^{0} \int_{0}^{R_n} \widehat{W}(x, y) \, d\mu_n(y) \, d\mu_n(x) = O\left(n^{-\frac{\alpha+1}{2}}\right), \tag{42}$$

which together with (40) and (41) implies the first result. The result for W follows by noting that $W \leq \widehat{W}$.

We split the integral (42) as follows

$$\int_{-R_n}^{0} \int_{0}^{R_n} \widehat{W}(x, y) \, d\mu_n(y) \, d\mu_n(x)$$

$$= \int_{-R_n}^{0} \int_{0}^{-x} d\mu_n(y) \, d\mu_n(x) + \int_{-R_n}^{0} \int_{-x}^{R_n} e^{-(x+y)} \, d\mu_n(y) \, d\mu_n(x).$$

For the first integral we compute

$$\int_{-R_n}^{0} \int_{0}^{-x} d\mu_n(y) \, d\mu_n(x) = \alpha e^{-2\alpha R_n} \int_{-R_n}^{0} \left(e^{-\alpha x} - 1\right) e^{\alpha x} \, dx$$

$$= e^{-2\alpha R_n} \left(\alpha R_n + e^{-\alpha R_n} - 1\right) = O\left(\log(n) n^{-\alpha}\right)$$

Finally, the second integral evaluates to

$$\frac{\alpha}{\beta} e^{-2\alpha R_n} \int_{-R_n}^0 \left(e^{(\alpha-1)R_n} - e^{-(\alpha-1)x} \right) e^{(\alpha-1)x} \, dx$$

$$= \frac{\alpha}{\beta} e^{-2\alpha R_n} \left(\frac{e^{(\alpha-1)R_n} - 1}{\alpha - 1} - R_n \right) = O\left(n^{-\frac{\alpha+1}{2}} \right),$$

from which the result follows since $\alpha > (\alpha + 1)/2$. $\qquad \square$

We can show a similar result for $H(W)$ and $H(\widehat{W})$.

Lemma 2

$$\iint_{-\infty}^{R_n} H\left(\widehat{W}(x, y)\right) d\mu_n(y) \, d\mu_n(x) - \iint_0^{R_n} H\left(\widehat{W}(x, y)\right) d\mu_n(y) \, d\mu_n(x) = \epsilon_n,$$

where

$$\epsilon_n = \begin{cases} O\left(\log(n)n^{-\alpha}\right) & \text{if } 1 < \alpha < 2, \\ O\left(\log(n)^2 n^{-2}\right) & \text{if } \alpha = 2, \\ O\left(n^{-\frac{2+\alpha}{2}}\right) & \text{if } \alpha > 2, \end{cases}$$

Moreover, the same result holds if we replace $\widehat{W}$ with W.

Proof We will first prove the result for $\widehat{W}$. For this we split the interval A_n into three parts $(-\infty, -R_n], [-R_n, 0]$ and $(0, R_n]$ and show that the integrals on all ranges other than $[0, R_n]^2$ are bounded by a term that scales as ϵ_n.

Since $H(p) \leq \log(2)$ for all $0 \leq p \leq 1$ it follows from (40) and (41) that, for all $\alpha > 1$,

$$\iint_{-\infty}^0 H\left(\widehat{W}(x, y)\right) d\mu_n(y) \, d\mu_n(x) = O\left(n^{-\alpha}\right) = o(\epsilon_n),$$

$$\int_{-\infty}^{R_n} \int_{-\infty}^{-R_n} H\left(\widehat{W}(x, y)\right) d\mu_n(y) \, d\mu_n(x) = O\left(n^{-\alpha}\right) = o(\epsilon_n).$$

Hence, using the symmetry of $\widehat{W}$, we only need to consider the integration over $(0, R_n] \times (-R_n, 0]$.

First we compute

$$H\left(\widehat{W}(x, y)\right) = e^{-(x+y)}(x + y) - (1 - e^{-(x+y)}) \log\left(1 - e^{-(x+y)}\right)$$

and observe that

$$-(1 - e^{-z}) \log\left(1 - e^{-z}\right) \leq e^{-2z} \quad \text{for all large enough } z. \tag{43}$$

Now let $\delta > 0$ be such that (43) holds for all $z \geq \delta$. Then $\delta < R_n$ for sufficiently large n and we split the integration as follows

$$\int_0^{R_n} \int_{-R_n}^0 H\left(\widehat{W}(x, y)\right) d\mu_n(y) \, d\mu_n(x)$$

$$= \int_0^\delta \int_{-R_n}^0 H\left(\widehat{W}(x, y)\right) d\mu_n(y) \, d\mu_n(x) + \int_\delta^{R_n} \int_{-R_n}^0 H\left(\widehat{W}(x, y)\right) d\mu_n(y) \, d\mu_n(x).$$

The first integral is $O(n^{-\alpha})$. For the second we note that $x + y > \delta$ for all $y > \delta - x$ and hence

$$\int_{\delta}^{R_n} \int_{-R_n}^{0} H\left(\widehat{W}(x, y)\right) d\mu_n(y) \, d\mu_n(x)$$
$$\leq \int_{\delta}^{R_n} \int_{-R_n}^{\delta-x} \log(2) \, d\mu_n(y) \, d\mu_n(x) + \int_{\delta}^{R_n} \int_{\delta-x}^{0} e^{-2(x+y)} \, d\mu_n(y) \, d\mu_n(x).$$

For the second integral we obtain,

$$\int_{\delta}^{R_n} \int_{\delta-x}^{0} e^{-2(x+y)} \, d\mu_n(y) \, d\mu_n(x) = \begin{cases} O\left(\log(n)n^{-\alpha}\right) & \text{if } 1 < \alpha < 2, \\ O\left(\log(n)^2 n^{-2}\right) & \text{if } \alpha = 2, \\ O\left(n^{-\frac{2+\alpha}{2}}\right) & \text{if } \alpha > 2, \end{cases}$$

while for the first integral we have

$$\int_{\delta}^{R_n} \int_{-R_n}^{\delta-x} \log(2) \, d\mu_n(y) d\mu_n(x) \leq \alpha e^{-2\alpha R_n} \int_{\delta}^{R_n} e^{\alpha\delta} \, dx = O\left(\log(n)n^{-\alpha}\right).$$

Therefore we conclude that

$$\int_{0}^{R_n} \int_{-\infty}^{0} H\left(\widehat{W}(x, y)\right) d\mu_n(y) d\mu_n(x) = \begin{cases} O\left(\log(n)n^{-\alpha}\right) & \text{if } 1 < \alpha < 2, \\ O\left(\log(n)n^{-2}\right) & \text{if } \alpha = 2, \\ O\left(n^{-\frac{4+\alpha}{2}}\right) & \text{if } \alpha > 2, \end{cases}$$

which yields the result for $\widehat{W}$.

For W we first compute that

$$H(W(x, y)) = \log\left(1 + e^{x+y}\right) - \frac{(x + y)e^{x+y}}{1 + e^{x+y}}$$
$$= \log\left(1 + e^{x+y}\right) - (x + y) + (x + y)\left(1 - \frac{e^{x+y}}{1 + e^{x+y}}\right)$$
$$= \log\left(1 + e^{-(x+y)}\right) + \frac{(x + y)}{1 + e^{x+y}} \leq \log\left(1 + e^{-(x+y)}\right) + (x + y)e^{-(x+y)}.$$

Comparing this upper bound to $H(\widehat{W}(x, y))$ and noting that $\log\left(1 + e^{-z}\right) \leq e^{-2z}$ for large enough z, the result follows from the computation done for $\widehat{W}$. □

With these two lemmas we now establish two important results on the approximations of W. The first shows that if X and Y are independent with distribution μ_n, then $\widehat{W}(X, Y)$ converges in expectation to $W(X, Y)$, faster than n^{-1}.

Proposition 2 *Let X, Y be independent with density μ_n and $\alpha > 1$. Then, as $n \to \infty$,*

$$\mathbb{E}\left[\left|W(X, Y) - \widehat{W}(X, Y)\right|\right] = O\left(n^{-\frac{\alpha+1}{2}}\right)$$

Proof Since

$$\mathbb{E}\left[\left|W(X, Y) - \widehat{W}(X, Y)\right|\right] = \iint_{-\infty}^{R_n} \left|W(x, y) - \widehat{W}(x, y)\right| d\mu_n(y) \, d\mu_n(x),$$

and $\left|W(x, y) - \widehat{W}(x, y)\right| \leq 1$ it follows from Lemma 1 that it is enough to consider the integral

$$\iint_0^{R_n} \left|W(x, y) - \widehat{W}(x, y)\right| d\mu_n(y) d\mu_n(x).$$

For this we note that

$$\left|W(x, y) - \widehat{W}(x, y)\right| \leq e^{-2(x+y)}.$$

Hence we obtain

$$\iint_0^{R_n} \left|W(x, y) - \widehat{W}(x, y)\right| d\mu_n(y) d\mu_n(x)$$

$$\leq \left(\int_0^{R_n} e^{-2x} d\mu_n(x)\right)^2 = \begin{cases} O\left(n^{-\alpha}\right) & \text{if } 1 < \alpha < 2 \\ O\left(\log(n)^2 n^{-2}\right) & \text{if } \alpha = 2 \\ O\left(n^{-2}\right) & \text{if } \alpha > 2. \end{cases}$$

Since all these terms are $o(n^{-(\alpha+1)/2})$, the result follows. □

Next we show that also $H(\widehat{W}(X, Y))$ converges in expectation to $H(W(X, Y))$, faster than n^{-1}.

Proposition 3 *Let X, Y be independent with density μ_n and $\alpha > 1$. Then, as $n \to \infty$,*

$$\mathbb{E}\left[\left|H\left(W(X, Y)\right) - H\left(\widehat{W}(X, Y)\right)\right|\right] = \begin{cases} O\left(\log(n) n^{-\alpha}\right) & \text{if } 1 < \alpha < 2 \\ O\left(\log(n)^3 n^{-2}\right) & \text{if } \alpha = 2 \\ O\left(\log(n) n^{-2}\right) & \text{if } \alpha > 2. \end{cases}$$

Proof Similar to the previous proof we now use Lemma 2 to show that it is enough to consider the integral

$$\iint_0^{R_n} \left|H\left(W(x, y)\right) - H\left(\widehat{W}(x, y)\right)\right| d\mu_n(y) d\mu_n(x).$$

Define

$$\delta_W(x, y) = \frac{1}{e^{2(x+y)} + e^{x+y}},$$

and note that $\widehat{W}(x, y) = W(x, y) + \delta_W(x, y)$. Now fix x, y. Then, by the Taylor-Lagrange theorem, there exists a $W(x, y) \leq c_W(x, y) \leq \widehat{W}(x, y)$ such that

$$\left|H\left(W(x, y)\right) - H\left(\widehat{W}(x, y)\right)\right| = \left|H'\left(c_W(x, y)\right)\right| \delta_W(x, y)$$
$$\leq \left(\left|H'\left(W(x, y)\right)\right| + \left|H'\left(\widehat{W}(x, y)\right)\right|\right) \delta_W(x, y).$$

Next we compute that

$$\left|H'\left(W(x, y)\right)\right| = (x + y), \quad \text{and} \quad \left|H'\left(\widehat{W}(x, y)\right)\right| = \log\left(e^{x+y} - 1\right),$$

where $\log(e^{x+y} - 1) \leq (x + y)$ for all $x + y \geq 1$. We now split the integral and bound it as follows

$$\iint_0^{R_n} \left|H\left(W(x, y)\right) - H\left(\widehat{W}(x, y)\right)\right| d\mu_n(y) d\mu_n(x)$$

$$= \int_0^1 \int_0^{1-x} \left| H\left(W(x,y)\right) - H\left(\widehat{W}(x,y)\right) \right| d\mu_n(y)\, d\mu_n(x)$$

$$+ \int_0^1 \int_{1-x}^{R_n} \left| H\left(W(x,y)\right) - H\left(\widehat{W}(x,y)\right) \right| \delta_W(x,y)\, d\mu_n(y)\, d\mu_n(x)$$

$$+ \int_1^{R_n} \int_0^{R_n} \left| H\left(W(x,y)\right) - H\left(\widehat{W}(x,y)\right) \right| \delta_W(x,y)\, d\mu_n(y)\, d\mu_n(x)$$

$$\leq \log(2) \int_0^1 \int_0^{1-x} d\mu_n(y)\, d\mu_n(x)$$

$$+ 2 \int_0^1 \int_{1-x}^{R_n} (x+y)\delta_W(x,y)\, d\mu_n(y)\, d\mu_n(x)$$

$$+ 2 \int_1^{R_n} \int_0^{R_n} (x+y)\delta_W(x,y)\, d\mu_n(y)\, d\mu_n(x)$$

$$\leq \log(2) \int_0^1 \int_0^{1-x} d\mu_n(y)\, d\mu_n(x) + 4 \iint_0^{R_n} (x+y)\delta_W(x,y)\, d\mu_n(y)\, d\mu_n(x).$$

The first integral is $O\left(\log(n)n^{-\alpha}\right)$, while for the second we have

$$4 \iint_0^{R_n} (x+y)\delta_W(x,y)\, d\mu_n(y)\, d\mu_n(x)$$

$$\leq 8R_n \left(\int_0^{R_n} e^{-2x}\, d\mu_n(x) \right)^2 = \begin{cases} O\left(\log(n)n^{-\alpha}\right) & \text{if } 1 < \alpha < 2 \\ O\left(\log(n)^3 n^{-2}\right) & \text{if } \alpha = 2 \\ O\left(\log(n)n^{-2}\right) & \text{if } \alpha > 2. \end{cases}$$

Comparing these scaling to the ones from Lemma 2 we see that the former are dominating, which finishes the proof. □

4.2 Proofs for Node Degrees in the `HSCM`

In this section we give the proofs of Theorems 2 and 3. Denote by D_i the degree of node i and recall that $\mathcal{D}_n$ is the degree of a node, sampled uniformly at random. Since the node labels are interchangeable we can, without loss of generality, consider D_1 for $\mathcal{D}_n$.

For Theorem 3 we use (35). We show that if X and Y are independent with distribution μ_n, then $\mathbb{E}\left[\widehat{W}(X,Y)\right] \to \nu$. The final result will then follow from Proposition 2.

The proof of Theorem 2 is more involved. Given the coordinates $X_1, \ldots X_n$ we have $\mathcal{D}_n = \sum_{j=2}^n W(X_1, X_j)$. We follow the strategy from [64, Theorem 6.7] and construct a coupling between $\mathcal{D}_n$ and a mixed Poisson random variable P_n, with mixing parameter $n\widehat{w}_n(X)$, where $\widehat{w}_n(x)$ is given by (33) and X has distribution μ_n.

In general, a coupling between two random variables X and Y consists of a pair of new random variables $(\widehat{X}, \widehat{Y})$, with some joint probability distribution, such that $\widehat{X}$ and $\widehat{Y}$ have the same marginal probabilities as, respectively X and Y. The advantage of a coupling is that we can tune the joint distribution to fit our needs. For our proof we construct a coupling $(\widehat{\mathcal{D}}_n, \widehat{P}_n)$, such that

$$\lim_{n \to \infty} \mathbb{P}\left(\widehat{\mathcal{D}}_n \neq \widehat{P}_n\right) = 0.$$

Hence, since $\widehat{\mathcal{D}}_n$ and $\widehat{P}_n$ have the same distribution as, respectively $\mathcal{D}_n$ and P_n we have

$$\mathbb{P}\left(\mathcal{D}_n = k\right) = \mathbb{P}\left(P_n = k\right) + \mathbb{P}\left(\widehat{\mathcal{D}}_n = k, \widehat{\mathcal{D}}_n \neq \widehat{P}_n\right),$$

from which it follows that $|\mathbb{P}(\mathcal{D}_n = k) - \mathbb{P}(P_n = k)| \to 0$. Finally, we show that

$$\lim_{n\to\infty} \mathbb{P}(n\widehat{w}_n(X) > k) = \mathbb{P}(Y > k),$$

where Y is a Pareto random variable with shape $\nu\beta$ and scale α, i.e. it has probability density $P_Y(y)$ given by (26). This implies that the mixed Poisson random variable P_n with mixing parameter $n\widehat{w}_n(X)$ converges to a mixed Poisson $\mathcal{D}$ with mixing parameter Y, which proves the result.

Before we give the proofs of the two theorems we first establish some technical results needed to construct the coupling, required for Theorem 2, the proof of which is given in Sect. 4.2.2.

4.2.1 Technical Results on Poisson Couplings and Concentrations

We first establish a general result for couplings between mixed Poisson random variables, where the mixing parameters converge in expectation.

Lemma 3 *Let X_n and Y_n be random variables such that,*

$$\lim_{n\to\infty} \mathbb{E}[|X_n - Y_n|] = 0.$$

Then, if P_n and Q_n are mixed Poisson random variables with, respectively, parameters X_n and Y_n,

$$\mathbb{P}(P_n \neq Q_n) = O(\mathbb{E}[|X_n - Y_n|]),$$

and in particular

$$\lim_{n\to\infty} \mathbb{P}(P_n \neq Q_n) = 0.$$

Proof Let $a_n = \mathbb{E}[|X_n - Y_n|]$ and define the event $A_n = \{|X_n - Y_n| \leq \sqrt{a_n}\}$. Then, since

$$\lim_{n\to\infty} \mathbb{P}(|X_n - Y_n| > \sqrt{a_n}) \leq \lim_{n\to\infty} \frac{\mathbb{E}[|X_n - Y_n|]}{\sqrt{a_n}} = 0,$$

it is enough to show that

$$\lim_{n\to\infty} \mathbb{P}(P_n \neq Q_n, A_n) = 0.$$

Take $\hat{P}_n$ to be mixed Poisson with parameter $X_n + \sqrt{a_n}$. In addition, let V_n and Z_n be mixed Poisson with parameter $\min\{X_n + \sqrt{a_n} - Y_n, 0\}$ and $\sqrt{a_n}$, respectively. Then, since on A_n we have $X_n + \sqrt{a_n} > Y_n$ we get, using Markov's inequality,

$$\mathbb{P}\left(P_n \neq Q_n, A_n\right) \leq \mathbb{P}\left(P_n \neq Q_n, P_n = \hat{P}_n, A_n\right) + \mathbb{P}\left(\hat{P}_n \neq P_n, A_n\right)$$

$$= \mathbb{P}\left(\hat{P}_n \neq Q_n, A_n\right) + \mathbb{P}\left(Z_n \neq 0, A_n\right)$$

$$= \mathbb{P}\left(V_n \neq 0, A_n\right) + \mathbb{P}\left(Z_n \neq 0, A_n\right)$$

$$\leq \mathbb{P}\left(V_n \geq 1, A_n\right) + \mathbb{P}\left(Z_n \geq 1, A_n\right) \leq \mathbb{E}\left[|X_n + \sqrt{a_n} - Y_n|\right] + \sqrt{a_n}$$

$$\leq \mathbb{E}\left[|X_n - Y_n|\right] + 2\sqrt{a_n} = O(\sqrt{a_n}).$$

Since $a_n \to 0$ by assumption, this finishes the proof. □

Next we show that $\widehat{W}(X, Y)$ converges in expectation to $\widehat{w}_n(X)$ when X, Y have distribution μ_n. We also establish an upper bound on the rate of convergence, showing that it converges faster than n^{-1}.

Lemma 4 *Let X, Y be independent random variables with density μ_n. Then, as $n \to \infty$,*

$$\mathbb{E}\left[|\widehat{W}(X, Y) - \widehat{w}_n(X)|\right] = O\left(\log(n)n^{-\alpha} + n^{-\frac{1+\alpha}{2}}\right)$$

Proof Recall that

$$\widehat{w}_n(x) = \begin{cases} \omega_n e^{-x} & \text{if } 0 \leq x \leq R_n \\ 0 & \text{otherwise} \end{cases}, \quad \text{where}$$

$$\omega_n = \int_0^{R_n} e^{-x} \, d\mu_n(x) = \frac{1 - e^{-(\alpha-1)R_n}}{\beta e^{R_n}}.$$

Hence, it follows that

$$\mathbb{E}\left[|\widehat{W}(X, Y) - \widehat{w}_n(X)|\right] = \int_{-\infty}^{0} \int_{-\infty}^{R_n} \widehat{W}(x, y) \, d\mu_n(y) \, d\mu_n(x)$$

$$+ \int_0^{R_n} \int_{-\infty}^{R_n} |\widehat{W}(x, y) - \widehat{w}_n(x)| \, d\mu_n(y) \, d\mu_n(x),$$

where the first integral is $O(n^{-(\alpha+1)/2})$ by Lemma 1.

To deal with the second integral we first compute

$$\int_0^{R_n} \int_{-\infty}^{-x} d\mu_n(y) \, d\mu_n(x) = \alpha R_n e^{-2\alpha R_n} = O\left(\log(n)n^{-\alpha}\right).$$

Therefore we have

$$\int_0^{R_n} \int_{-\infty}^{R_n} |\widehat{W}(x, y) - \widehat{w}_n(x)| \, d\mu_n(y) d\mu_n(x)$$

$$= \int_0^{R_n} \int_{-\infty}^{-x} d\mu_n(y) d\mu_n(x) + \int_0^{R_n} \int_{-x}^{R_n} \left|e^{-(x+y)} - \omega_n e^{-x}\right| d\mu_n(y) d\mu_n(x)$$

$$\leq O\left(\log(n)n^{-\alpha}\right) + \int_0^{R_n} \int_0^{R_n} e^{-x} \left|e^{-y} - \omega_n\right| d\mu_n(y) d\mu_n(x)$$

$$= O\left(\log(n)n^{-\alpha}\right) + \omega_n \int_0^{R_n} \left|e^{-y} - \omega_n\right| d\mu_n(y)$$

We proceed to compute the last integral and show that it is $O(n^{-(\alpha+1)/2})$. For this we note that since

$$e^{-y} \leq \omega_n \iff y \leq -\log(\omega_n),$$

we have

$$\int_0^{R_n} \left| e^{-y} - \omega_n \right| d\mu_n(y) = \int_0^{-\log(\omega_n)} (\omega_n - e^{-y}) d\mu_n(y) + \int_{-\log(\omega_n)}^{R_n} (e^{-y} - \omega_n) d\mu_n(y)$$

For the first integral we compute

$$\int_0^{-\log(\omega_n)} (\omega_n - e^{-y}) d\mu_n(y) = \alpha e^{-\alpha R_n} \left(\frac{\omega_n^{1-\alpha}}{\alpha} - \frac{\omega_n^{1-\alpha}}{\alpha-1} - \frac{\omega_n}{\alpha} + \frac{1}{\alpha-1} \right)$$

$$= \frac{e^{-\alpha R_n}}{\beta} - \frac{\omega_n e^{-\alpha R_n}}{\alpha-1} - \omega_n e^{-\alpha R_n} = \frac{e^{-\alpha R_n}}{\beta} - \frac{\omega_n e^{-\alpha R_n}}{\beta}.$$

Similar calculations yield

$$\int_{-\log(\omega_n)}^{R_n} (e^{-y} - \omega_n) d\mu_n(y) = \frac{e^{-R_n}}{\beta} - \omega_n - \frac{\omega_n^{1-\alpha} e^{-\alpha R_n}}{\alpha-1},$$

and hence,

$$\int_0^{R_n} \left| e^{-y} - \omega_n \right| d\mu_n(y) = \frac{1}{\beta} \left(e^{-R_n} + e^{-\alpha R_n} \right) - \omega_n - \frac{\omega_n e^{-\alpha R_n}}{\beta} - \frac{\omega_n^{1-\alpha} e^{-\alpha R_n}}{\alpha-1}$$

$$\leq \frac{1}{\beta} \left(e^{-R_n} + e^{-\alpha R_n} \right) - \omega_n.$$

We now use this last upper bound, together with

$$\omega_n = \frac{1 - e^{-(\alpha-1)R_n}}{\beta e^{R_n}} = \frac{1}{\beta} \left(e^{-R_n} - e^{-\alpha R_n} \right),$$

to obtain

$$\omega_n \int_0^{R_n} \left| e^{-y} - \omega_n \right| d\mu_n(y) \leq \frac{\omega_n}{\beta} \left(e^{-R_n} + e^{-\alpha R_n} \right) - \omega_n^2$$

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

$$= \frac{2}{\beta^2} e^{-(\alpha+1)R_n} = O\left(n^{-\frac{1+\alpha}{2}} \right),$$

from which the result follows. □

4.2.2 Proof of Theorem 2

We start with constructing the coupling $(\widehat{\mathcal{D}}_n, \widehat{P}_n)$, between $\mathcal{D}_n$ and P_n such that

$$\lim_{n \to \infty} \mathbb{P}\left(\widehat{\mathcal{D}}_n \neq \widehat{P}_n \right) = 0. \tag{44}$$

First, let I_j be the indicator that the edge $(1, j)$ is present in G_n. Let $\mathbf{X}_n = X_1, \ldots, X_n$ denote the coordinates of the nodes. Then, conditioned on these, we have that I_j are independent Bernoulli random variables with probability $W(X_1, X_j)$, while $D_1 = \sum_{j=2}^n I_j$. Now, let Q_n be a mixed Poisson with parameter $\sum_{j=2}^n W(X_1, X_j)$. Then, see for instance [64, Theorem 2.10], there exists a coupling $(\widehat{D}_1, \widehat{Q}_n)$ of D_1 and Q_n, such that

$$\mathbb{P}\left(\widehat{D}_1 \neq \widehat{Q}_n \middle| \mathbf{X}_n \right) \leq \sum_{j=2}^n W(X_1, X_j)^2.$$

Therefore, we get that

$$\mathbb{P}\left(\widehat{D}_1 \neq \widehat{Q}_n\right) \leq (n-1) \iint_{-\infty}^{R_n} W(x,y)^2 \, d\mu_n(y) \, d\mu_n(x)$$

$$\leq n \left(\int_{-\infty}^{R_n} e^{-2x} \, d\mu_n(x)\right)^2 = \begin{cases} O\left(n^{-(\alpha-1)}\right) & \text{if } 1 < \alpha < 2 \\ O\left(\log(n)n^{-1}\right) & \text{if } \alpha = 2 \\ O\left(n^{-1}\right) & \text{if } \alpha > 2 \end{cases}.$$

Next, since X_1 and X_j are independent for all $2 \leq j \leq n$ we use Proposition 2 together with Lemma 4 to obtain

$$\mathbb{E}\left[\left|\sum_{j=2}^n W(X_1, X_j) - n\widehat{w}_n(X_1)\right|\right]$$

$$\leq (n-1)\mathbb{E}\left[|W(X_1, X_2) - \widehat{w}_n(X_1)|\right] + \mathbb{E}\left[\widehat{w}_n(X_1)\right]$$

$$\leq (n-1)\mathbb{E}\left[\left|W(X_1, X_2) - \widehat{W}(X_1, X_2)\right|\right] + (n-1)\mathbb{E}\left[\left|\widehat{W}(X_1, X_2) - \widehat{w}_n(X_1)\right|\right]$$

$$+ \mathbb{E}\left[\widehat{w}_n(X_1)\right]$$

$$= O\left(n^{-\frac{\alpha-1}{2}}\right).$$

from which it follows that

$$\lim_{n\to\infty} \mathbb{E}\left[\left|\sum_{j=2}^n W(X_1, X_j) - n\widehat{w}_n(X_1)\right|\right] = 0. \tag{45}$$

Now let $\widehat{P}_n$ be a mixed Poisson random variable with mixing parameter $n\widehat{w}_n(X_1)$. Then by (45) and Lemma 3

$$\lim_{n\to\infty} \mathbb{P}\left(\widehat{P}_n \neq \widehat{Q}_n\right) = 0$$

and (44) follows from

$$\mathbb{P}\left(\widehat{\mathcal{D}}_n \neq \widehat{P}_n\right) \leq \mathbb{P}\left(\widehat{\mathcal{D}}_n \neq \widehat{Q}_n\right) + \mathbb{P}\left(\widehat{P}_n \neq \widehat{Q}_n\right).$$

As a result we have that

$$\lim_{n\to\infty} |\mathbb{P}\left(\mathcal{D}_n = k\right) - \mathbb{P}\left(P_n = k\right)| = \lim_{n\to\infty} \left|\mathbb{P}\left(\widehat{\mathcal{D}}_n = k\right) - \mathbb{P}\left(\widehat{P}_n = k\right)\right|$$

$$= \lim_{n\to\infty} \mathbb{P}\left(\widehat{\mathcal{D}}_n = k, \widehat{\mathcal{D}}_n \neq \widehat{P}_n\right) \leq \lim_{n\to\infty} \mathbb{P}\left(\widehat{\mathcal{D}}_n \neq \widehat{P}_n\right) = 0, \tag{46}$$

We will now prove that if X has density μ_n, then for any $t > 0$

$$\lim_{n\to\infty} |\mathbb{P}\left(n\widehat{w}_n(X) > t\right) - \mathbb{P}\left(Y > t\right)| = 0. \tag{47}$$

A sequence $\{P_n\}_{n\geq 1}$ of mixed Poisson random variables, with mixing parameters Z_n converges to a mixed Poisson random variable P with mixing parameter Z, if Z_n converge to Z in distribution, see for instance [63]. Therefore, since $\mathcal{D}$ is mixed Poisson with mixing parameter Y, (47) implies

$$\lim_{n\to\infty} \mathbb{P}\left(P_n = k\right) = \mathbb{P}\left(\mathcal{D} = k\right)$$

which, combined with (46), yields

$$\lim_{n \to \infty} \mathbb{P}\left(\mathcal{D}_n = k\right) = \mathbb{P}\left(\mathcal{D} = k\right).$$

To establish (47), first define

$$\epsilon_n = e^{-(\alpha-1)R_n} - e^{-2\alpha R_n}$$

so that

$$n\widehat{w}_n(x) = \frac{n}{\beta e^{R_n}}\left(1 - e^{-(\alpha-1)R_n}\right) = \sqrt{\nu n}\, e^{-x}\left(1 - \epsilon_n\right).$$

Next, for all $0 \le x \le R_n$ we have that

$$\nu\beta\left(1 - \epsilon_n\right) \le n\widehat{w}_n(x) \le \sqrt{\nu n}\left(1 - \epsilon_n\right)$$

and hence

$$\mathbb{P}\left(n\widehat{w}_n(X) > t\right) = \begin{cases} \mathbb{P}\left(X < \log\left(\frac{n\omega_n}{t}\right)\right) & \text{if } \nu\beta\left(1 - \epsilon_n\right) \le t \le \sqrt{\nu n}\left(1 - \epsilon_n\right) \\ 1 & \text{if } t < \nu\beta\left(1 - \epsilon_n\right) \\ 0 & \text{else.} \end{cases}$$

Moreover, for any $\nu\beta\left(1 - \epsilon_n\right) \le t \le \sqrt{\nu n}\left(1 - \epsilon_n\right)$

$$\mathbb{P}\left(n\widehat{w}_n(X) > t\right) = \mathbb{P}\left(X < \log\left(\frac{n\omega_n}{t}\right)\right) = \int_{-\infty}^{\log\left(\frac{n\omega_n}{t}\right)} d\mu_n(x)$$

$$= e^{-\alpha R_n}\left(\frac{n\omega_n}{t}\right)^\alpha = \left(n\omega_n e^{-R_n}\right)^\alpha t^{-\alpha} = \left(\frac{\nu\beta}{t}\right)^\alpha\left(1 - \epsilon_n\right)^\alpha. \quad (48)$$

Now, fix $t < \nu\beta$. Then, for large enough n it holds that $t < \nu\beta(1 - \epsilon_n)$ in which case (47) holds trivially, since both probabilities are 1. Hence we can assume, without loss of generality that $t \ge \nu\beta$. Then for n large enough such $t \le \sqrt{\nu n}\left(1 - \epsilon_n\right)$, it follows from (48) and (27) that

$$|\mathbb{P}\left(n\widehat{w}_n(X) > t\right) - \mathbb{P}\left(Y > t\right)| = \left|\left(\frac{\nu\beta}{t}\right)^\alpha\left(1 - \epsilon_n\right)^\alpha - \left(\frac{\nu\beta}{t}\right)^\alpha\right|$$

$$= \left(\frac{\nu\beta}{t}\right)^\alpha\left(1 - \left(1 - \epsilon_n\right)^\alpha\right) \le 1 - \left(1 - \epsilon_n\right)^\alpha,$$

which converges to zero as $n \to \infty$ and hence proves (47).

4.2.3 Proof of Theorem 3

First using Lemma 1, we compute that

$$(n-1)\mathbb{E}\left[\widehat{W}(X, Y)\right] = (n-1)\iint_{-\infty}^{R_n} \widehat{W}(x, y)\, d\mu_n(y)\, d\mu_n(x)$$

$$= (n-1)\iint_0^{R_n} e^{-(x+y)}\, d\mu_n(y)\, d\mu_n(x) + O\left(n^{-\frac{\alpha-1}{2}}\right)$$

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

$$= \nu \left(1 - \frac{1}{\beta^2 n}\right) \left(1 - e^{-\alpha R_n}\right)^2 + O\left(n^{-\frac{\alpha-1}{2}}\right) = \nu + O\left(n^{-\frac{\alpha-1}{2}}\right),$$

where the last line follows since $-1 < -\alpha/2 < -(\alpha-1)/2$.

Next recall (35)

$$\mathbb{E}[\mathcal{D}_n] = \sum_{j=2}^{n} \mathbb{E}\left[W(X_1, X_j)\right] = (n-1)\mathbb{E}\left[W(X, Y)\right].$$

Therefore, using Lemma 4 we have

$$|\mathbb{E}[\mathcal{D}_n] - \nu| = |(n-1)\mathbb{E}[W(X, Y)] - \nu|$$
$$\leq (n-1)\mathbb{E}\left[\left|W(X, Y) - \widehat{W}(X, Y)\right|\right] + (n-1)\left|\mathbb{E}\left[\widehat{W}(X, Y)\right] - \nu\right|$$
$$\leq O\left(\log(n)n^{-(\alpha-1)}\right) + O\left(n^{-\frac{\alpha-1}{2}}\right),$$

which yields the result.

4.3 Proofs for Graphon Entropy

Here we derive the properties of the graphon entropy σ of our graphon W given by (21). We first give the proof of Proposition 1, and then that of Theorem 4.

4.3.1 Proof of Proposition 1

Recall that, given a measure μ on some interval $A \subseteq \mathbb{R}$ with $\mu(A) < \infty$ and a μ-integrable function $w(x)$, we consider the problem of maximizing $\sigma[W, \mu]$ under the constraint (36),

$$w(x) = \int_A W(x, y) \, d\mu(y).$$

In particular we need to show that if a solution exists, it is given by (37). Therefore, suppose there exists at least one graphon W which satisfies the constraint. For this we use the technique of Lagrange multipliers from variation calculus [65,66].

To set up the framework, let $\mathcal{W}$ denote the space of symmetric functions $W : A \times A \to [0, 1]$ which satisfy

$$\iint_{A^2} W(x, y) \, d\mu(y) \, d\mu(x) < \infty.$$

Observe that $\mathcal{W}$ is a convex subset of the Banach space $\mathcal{W}_{\mathbb{R}}$, of all symmetric, μ-integrable, functions $W : A \times A \to \mathbb{R}$ and that the function w is an element of $L^1(-\infty, R)$ with respect to the measure μ, which is also Banach. Denote the latter space by $L^1_{A,\mu}$. We slightly abuse notation and write σ for the functional

$$W \mapsto \iint_{A^2} H(W(x, y)) \, d\mu(y) \, d\mu(x) = \sigma[W, \mu],$$

and define the functional $\mathcal{F} : \mathcal{W}_{\mathbb{R}} \to L^1_{A,\mu}$ by

$$\mathcal{F}(W)(x) = w(x) - \int_A W(x, y) \, d\mu(y).$$

Then we need to solve the following Euler–Lagrange equation for some Lagrange multiplier functional $\lambda : L^1_{A,\mu} \to \mathbb{R}$,

$$\frac{\partial}{\partial W} (\sigma(W) - \Lambda \circ \mathcal{F}(W)) = 0,$$

with respect to the Fréchet derivative. By Riesz Representation Theorem, we have that for any functional $\Lambda : L^1_{A,\mu} \to \mathbb{R}$, there exists a $\lambda \in L^\infty$, uniquely defined on A, such that

$$\Lambda(f) = \int_A \lambda(x) f(x) \, d\mu(x).$$

Hence our Euler–Lagrange equation becomes

$$\frac{\partial}{\partial W} \left(\sigma(W) - \int_A \lambda(x) \left(w(x) - \int_A W(x, y) \, d\mu(y) \right) d\mu(x) \right) = 0. \qquad (49)$$

In addition, since W is symmetric we have

$$\iint_{A^2} \lambda(x) W(x, y) \, d\mu(y) d\mu(x) = \frac{1}{2} \iint_{A^2} (\lambda(x) + \lambda(y)) W(x, y) \, d\mu(y) \, d\mu(x)$$

and hence, by absorbing the factor $1/2$ into λ we can rewrite our Euler–Lagrange equation as

$$\frac{\partial}{\partial W} \left(\sigma(W) - \int_A \lambda(x) w(x) \, d\mu(x) + \iint_{A^2} (\lambda(x) + \lambda(y)) W(x, y) \, d\mu(y) \, d\mu(x) \right) = 0. \qquad (50)$$

For the two derivatives we have

$$\frac{\partial \sigma(W)}{\partial W} = \log(1 - W(x, y)) - \log(W(x, y)) = -\log \left(\frac{W(x, y)}{1 - W(x, y)} \right),$$

$$\frac{\partial}{\partial W} \left(\int_A \lambda(x) w(x) \, d\mu(x) + \iint_{A^2} (\lambda(x) + \lambda(y)) W(x, y) \, d\mu(y) \, d\mu(x) \right) = \lambda(x) + \lambda(y). \qquad (51)$$

Hence, we need to solve the equation

$$-\log \left(\frac{W(x, y)}{1 - W(x, y)} \right) = \lambda(x) + \lambda(y),$$

which gives (37).

There is, however, a small technicality related to the computation of the derivative (51). This is caused by the fact that H is only defined on $[0, 1]$. In particular, for $W_1, W_2 \in \mathcal{W}$, it could be that $W_1 + W_2 > 1$ on some subset $C \subseteq A \times A$ and hence $H(W_1 + W_2)$ is not well defined. To compute the Fréchet derivative we need $H(W_1 + \varepsilon W_2)$ to be well defined for any $W_1, W_2 \in \mathcal{W}$ and some, small enough, $\varepsilon > 0$. To this end we fix a $0 < \delta < 1$ and define

$$H_\delta(x) = H \left((1 - \delta)x + \frac{\delta}{2} \right),$$

which is just H, stretched out to the interval $(-\delta/2(1 - \delta), 1 + \delta/2(1 - \delta))$. Similarly we define σ_δ, using H_δ and consider the corresponding graphon entropy maximization problem.

Then we can compute $\partial \sigma_\delta / \partial W$, by taking a $W' \in \mathcal{W}$ and $\varepsilon > 0$ such that $W + \varepsilon W' < 1 + \delta/2(1 - \delta))$. Then $H_\delta(W + \varepsilon W')$ is well-defined and using the chain rule we obtain

$$\frac{\partial H_\delta(W + \varepsilon W')}{\partial \varepsilon} = -(1 - \delta)W' \log\left(\frac{(1 - \delta)(W + \varepsilon W') + \delta/2}{1 - (1 - \delta)(W + \varepsilon W') - \delta/2}\right),$$

from which it follows that

$$\frac{\partial \sigma_\delta(W)}{\partial W} = -(1 - \delta) \log\left(\frac{(1 - \delta)W + \delta/2}{1 - (1 - \delta)W - \delta/2}\right).$$

Therefore we have the following equation

$$-(1 - \delta) \log\left(\frac{(1 - \delta)(W + \delta/2)}{1 - (1 - \delta)W - \delta/2}\right) = \lambda(x) + \lambda(y),$$

which leads to a solution W_δ of the form

$$W_\delta(x, y) = \frac{1 - \frac{\delta}{2}\left(1 + e^{\frac{\lambda(x)+\lambda(y)}{(1-\delta)}}\right)}{(1 - \delta)\left(1 + e^{\frac{\lambda(x)+\lambda(y)}{(1-\delta)}}\right)}.$$

Since $\delta < 1$ it follows, using elementary algebra, that $W_\delta(x, y) \in [0, 1]$ for all $x, y \in [0, 1]$. From this we conclude that $W_\delta \in \mathcal{W}$. Moreover W_δ converges to (37) as $\delta \to 0$. Since $\sigma_\delta \to \sigma$ as $\delta \to 0$, we obtain the graphon that maximizes the entropy σ, where the function λ is determined by the constraint (36).

For uniqueness, suppose there exist two solutions, λ_1 and λ_2 in $L^1_{A,\mu}$, to (36) and for which the graphon entropy is maximized. Let $f(x) = \lambda_1(x) - \lambda_2(x) \in L^1_{A,\mu}$, so that $\lambda_1 = \lambda_2 + f$. Since λ_1 satisfies (49) it follows that, due to linearity of the derivative,

$$\frac{\partial}{\partial W} \int_A f(x) W(x, y) \, d\mu(y) \, d\mu(x) = 0.$$

Now since

$$\frac{\partial}{\partial W} \int_A f(x) W(x, y) \, d\mu(y) \, d\mu(x) = f(x),$$

it follows that $f = 0$, μ almost everywhere on A and hence $\lambda_1 = \lambda_2$, μ almost everywhere on A.

4.3.2 Proof of Theorem 4

First note that Proposition 3 implies that the difference in expectation between $H(W)$ and $H(\widehat{W})$ converges to zero faster than $\log(n)/n$, so that for the purpose of Theorem 4 we can approximate W with $\widehat{W}$. Hence we are left to show that the rescaled entropy $n\sigma[\widehat{W}, \mu_n]/\log(n)$ converges to ν. By Lemma 2 the integration over all regimes except $[0, R_n]^2$ goes to zero faster than $\log(n)/n$ and therefore we only need to consider integration over $[0, R_n]^2$. The main idea for the rest of the proof is that in this range

$$H\left(\widehat{W}(x, y)\right) \approx (x + y)\widehat{W}(x, y).$$

Let us first compute the integral over $[0, R_n]^2$ of the right hand side in the above equation.

$$\iint_0^{R_n} (x + y)\widehat{W}(x, y)\, d\mu_n(y)\, d\mu_n(x)$$

$$= \alpha^2 e^{-2\alpha R_n} \iint_0^{R_n} (x + y)e^{(\alpha-1)(x+y)}\, dy\, dx$$

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

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

$$= \frac{2R_n e^{-2R_n}}{\beta^2} + O(n^{-1}) = \nu n^{-1}\log(n) + O(n^{-1})$$

which implies that

$$\lim_{n\to\infty} \frac{n}{\log(n)} \iint_0^{R_n} (x + y)\widehat{W}(x, y)\, d\mu_n(y)\, d\mu_n(x) = \nu. \tag{52}$$

Next we show that

$$\frac{n}{\log(n)} \left| \iint_0^{R_n} \left(H\left(\widehat{W}(x, y)\right) - (x + y)\widehat{W}(x, y) \right) d\mu_n(y)\, d\mu_n(x) \right| = O\left(\log(n)^{-1}\right), \tag{53}$$

which, together with (52), gives

$$\lim_{n\to\infty} \frac{n}{\log(n)} \iint_0^{R_n} H\left(\widehat{W}(x, y)\right) d\mu_n(y)\, d\mu_n(x) = \nu. \tag{54}$$

We compute that

$$H\left(\widehat{W}(x, y)\right) = e^{-(x+y)}(x + y) - (1 - e^{-(x+y)})\log\left(\frac{e^{x+y} - 1}{e^{x+y}}\right)$$

$$= (x + y)\widehat{W}(x, y) - (1 - e^{-(x+y)})\log\left(1 - e^{-(x+y)}\right). \tag{55}$$

Note that $-(1 - e^{-z})\log\left(1 - e^{-z}\right) \le e^{-z}$ for all $z \ge 0$. Hence, it follows from (55) that, on $[0, R_n]^2$,

$$\left| H\left(\widehat{W}(x, y)\right) - (x + y)\widehat{W}(x, y) \right| = -\left(1 - e^{-(x+y)}\right)\log\left(1 - e^{-(x+y)}\right) \le e^{-(x+y)}$$

$$= \widehat{W}(x, y),$$

so that by Theorem 3,

$$\frac{n}{\log(n)} \left| \iint_0^{R_n} \left(H\left(\widehat{W}(x, y)\right) - (x + y)\widehat{W}(x, y) \right) d\mu_n(y)\, d\mu_n(x) \right|$$

$$\le \frac{n}{\log(n)} \iint_0^{R_n} \widehat{W}(x, y)\, d\mu_n(y)\, d\mu_n(x) = O\left(\log(n)^{-1}\right).$$

To conclude, we have

$$\sigma[\widehat{W}, \mu_n] = \iint_{-\infty}^{R_n} H\left(\widehat{W}(x, y)\right) d\mu_n(y)d\mu_n(x)$$

$$= \iint_0^{R_n} H\left(\widehat{W}(x, y)\right) d\mu_n(y)d\mu_n(x) + O\left(n^{-\alpha} + \log(n)n^{-\frac{\alpha+1}{2}}\right)$$

and hence, using (54),

$$\lim_{n\to\infty} \frac{n\sigma[\widehat{W}, \mu_n]}{\log(n)} = \nu.$$

4.4 Proof of Theorem 5

In this section we first formalize the strategy behind the proof of Theorem 5, briefly discussed in Sect. 3.6. This strategy relies on partitioning the interval A_n into non-overlapping subintervals. We then construct a specific partition satisfying certain requirements, and finish the proof of the theorem.

4.4.1 Averaging W by a Partition of A_n

We follow the strategy from [41]. First recall that for a graph G_n generated by the HSCM,

$$S[G_n] \geq \mathbb{E}\left[S[G_n|\mathbf{x}_n]\right] = \binom{n}{2}\mathbb{E}\left[H(W(x_1, x_2))\right] = \binom{n}{2}\sigma[G_n],$$

and hence

$$S^*[G_n] \geq \sigma[G_n],$$

where $S^*[G_n] = S[G_n]/\binom{n}{2}$ denotes the normalized Gibbs entropy.

Therefore, the key ingredient is to find a matching upper bound. For this we partition the range $A_n = (-\infty, R_n]$ of our probability measure $\mu_n(x)$ into intervals and approximate $W(x, y)$ by its average over the box in which x and y lie.

To be more precise, let m_n be any increasing sequence of positive integers, $\{\rho_t\}_{0\leq t\leq m_n}$ be such that

$$\rho_0 = -\infty, \quad \text{and} \quad \rho_{m_n} = R_n,$$

and consider the partition of $(-\infty, R_n]$ given by

$$I_t = (\rho_{t-1}, \rho_t] \quad \text{for } 1 \leq t \leq m_n.$$

Now define $J_n(x) = t \iff x \in I_t$, and let M_i be the random variable $J_n(X_i)$, where X_i has density function μ_n for any vertex i. The value of M_i equal to t indicates that vertex i happens to lie within interval I_t. Denoting the vector of these random variable by $\mathbf{M}_n = M_1, \ldots, M_n$, and their entropy by $S[\mathbf{M}_n]$, we have that

$$S[G_n] \leq S[\mathbf{M}_n] + S[G_n|\mathbf{M}_n]$$
$$\leq nS[M] + \binom{n}{2}\iint_{-\infty}^{R_n} H(\widetilde{W}(x, y)) d\mu_n(y) d\mu_n(x),$$

where $\widetilde{W}(x, y)$ is the average of W over the square $I_{J_n(x)} \times I_{J_n(y)}$. That is,

$$\widetilde{W}_n(x, y) = \frac{1}{\mu_{x,y}} \int_{I_{J_n(x)}} \int_{I_{J_n(y)}} W(u, v) d\mu_n(v) d\mu_n(u), \tag{56}$$

with

$$\mu_{x,y} = \int_{I_{J_n(x)}} \int_{I_{J_n(y)}} d\mu_n(v) \, d\mu_n(u),$$

the measure of the box to which (x, y) belongs.

The first step in our proof is to investigate how well $\widetilde{W}_n$ approximates W. More specifically, we want to understand how the difference $|\sigma[W, \mu_n] - \sigma[\widetilde{W}_n, \mu_n]|$ scales, depending on the specific partition. Note that for any partition ρ_t of A_n into m_n intervals we have that

$$\mathcal{S}[M] = -\sum_{t=1}^{m_n} \mathbb{P}\,(M = t) \log\left(\mathbb{P}\,(M = t)\right) \leq \log(m_n),$$

where the upper bound is achieved on the partition which is uniform according to measure μ_n. Since

$$\frac{n}{\log(n)} \frac{n\mathcal{S}[M]}{\binom{n}{2}} = \frac{2\mathcal{S}[M]}{\log(n)(1 - 1/n)},$$

it is enough to find a partition ρ_t, with $\log(m_n) = o(\log(n))$, such that

$$\lim_{n \to \infty} \frac{n}{\log(n)} |\sigma[W, \mu_n] - \sigma[\widetilde{W}_n, \mu_n]| = 0, \tag{57}$$

This then proves Theorem 5, since

$$\lim_{n \to \infty} \frac{n}{\log(n)} \left| \mathcal{S}^*[G_n] - \sigma[G_n] \right| = \lim_{n \to \infty} \frac{n}{\log(n)} \left(\mathcal{S}^*[G_n] - \sigma[G_n] \right)$$

$$\leq \lim_{n \to \infty} \left(\frac{n}{\log(n)} \frac{n\mathcal{S}[M]}{\binom{n}{2}} + \frac{n}{\log(n)} |\sigma[W, \mu_n] - \sigma[\widetilde{W}_n, \mu_n]| \right)$$

$$\leq \lim_{n \to \infty} \left(\frac{2\log(m_n)}{\log(n)(1 - 1/n)} + \frac{n}{\log(n)} |\sigma[W, \mu_n] - \sigma[\widetilde{W}_n, \mu_n]| \right) = 0.$$

4.4.2 Constructing the Partition

We will take $I_1 = (-\infty, -R_n]$ and partition the remaining interval $[-R_n, R_n]$ into $\log(n)^2$ equal parts. To this end, let n be sufficiently large so that $n \geq \nu\beta^2$, take $m_n = \lceil \log(n)^2 \rceil + 1$, and define the partition ρ_t by

$$\rho_0 = -\infty, \quad \rho_1 = -R_n, \quad \text{and} \quad \rho_t = \rho_{t-1} + \frac{2R_n}{m_n - 1} \quad \text{for all } t = 2, \ldots, m_n.$$

Note that $\log(m_n) = O(\log \log n) = o(\log(n))$ so that all that is left, is to prove (57). In addition,

$$\frac{n}{\log(n)} \int_{-\infty}^{-R_n} \int_{-\infty}^{R_n} H(\widetilde{W}(x, y)) \, d\mu_n(y) \, d\mu_n(x) = O\left(\log(n)^{-1} n^{1-\alpha}\right),$$

and the same holds if we replace $\widetilde{W}$ with W. Hence it follows that in order to establish (57) we only need to consider the integral over the square $[-R_n, R_n] \times [-R_n, R_n]$. That is, we need to show

$$\lim_{n \to \infty} \frac{n}{\log(n)} \left| \int\int_{-R_n}^{R_n} \left(H\,(W(x, y)) - H\left(\widetilde{W}(x, y)\right) \right) d\mu_n(y) \, d\mu_n(x) \right| = 0, \tag{58}$$

For this we compare $\sigma[W, \mu_n]$ and $\sigma[\widetilde{W}, \mu_n]$, based on the mean value theorem which states that

$$\text{for any } a \le b, \quad \exists c \in [a, b], \quad \text{such that } H(b) - H(a) = H'(c)(a - b).$$

Here $|H'(c)| = |\log(c) - \log(1 - c)|$, and due to the symmetry of H we get, for any $0 < a \le c \le b < 1$,

$$|H'(c)| \le \frac{\min\{H(a), H(b)\}}{\min\{a, b, 1 - a, 1 - b\}}. \tag{59}$$

Note that

$$0 < \min_{u \in I_{M_n(x)}, \, v \in I_{M_n(y)}} W(u, v) \le \widetilde{W}_n(x, y) \le \max_{u \in I_{M_n(x)}, \, v \in I_{M_n(y)}} W(u, v) < 1, \tag{60}$$

for all $x, y \in [-R_n, R_n]$. In addition, for all $x, y \in [-R_n, R_n]$ and $(u, v) \in I_{M_n(x)} \times I_{M_n(y)}$ we have $|x + y - u - v| \le 2R_n/m_n \le 2/\log(n)$ and thus $|1 - \exp(x + y - u - v)| \le 3|x + y - u - v|$ by the mean value theorem. Therefore we have

$$|W(x, y) - W(u, v)| = \frac{\left|e^{u+v} - e^{x+y}\right|}{(1 + e^{x+y})(1 + e^{u+v})} \le \min\left\{ \frac{|e^{u+v-x-y} - 1|}{1 + e^{u+v}}, \frac{|1 - e^{x+y-u-v}|}{1 + e^{u+v}} \right\}$$

$$\le 3|u + v - x - y| \min\{W(u, v), 1 - W(u, v)\},$$

By symmetry we obtain a similar upper bound with $W(x, y)$ instead of $W(u, v)$ and hence we conclude

$$|W(x, y) - W(u, v)| \le 3|u + v - x - y| \min\{W(x, y), 1 - W(x, y), W(u, v), 1 - W(u, v)\}. \tag{61}$$

Then, for any $x + y \le u + v$ there are c, d such that $x + y \le c + d \le u + v$, so by (59) and (61) we get

$$|H(W(x, y)) - H(W(u, v))| = |H'(W(c, d))(W(x, y) - W(u, v))|$$

$$\le \frac{\min\{H(W(x, y)), H(W(u, v))\} |W(x, y) - W(u, v)|}{\min\{W(x, y), 1 - W(x, y), W(u, v), 1 - W(u, v)\}}$$

$$\le 3|x + y - u - v| \cdot \min\{H(W(x, y)), H(W(u, v))\}. \tag{62}$$

Next, for the partition ρ_t we have

$$|I_t| = \frac{2R_n}{m_n - 1} \le \frac{2R_n}{\log(n)^2} = \frac{\log(n) - \log(\nu\beta^2)}{\log(n)^2} \le \frac{1}{\log(n)},$$

for $t \ge 2$. In addition, (60) implies that for $(x, y) \in I_t \times I_s$, with $t, s \ge 2$,

$$W(\rho_{t-1}, \rho_{s-1}) \le \widetilde{W}_n(x, y), \quad W(x, y) \le W(\rho_t, \rho_s),$$

and thus there is a pair $x_{s,t}, y_{s,t} \in I_s \times I_t$ such that $\widetilde{W}_n(x, y) = W(x_{s,t}, y_{s,t})$ on $I_s \times I_t$. Therefore, using (62), we get

$$|H(W(x, y)) - H(\widetilde{W}_n(x, y))| = |H(W(x, y)) - H(W(x_{s,t}, y_{s,t}))|$$

$$\le 3|x + y - x_{s,t} - y_{s,t}|H(W(x, y))$$

$$\le 3(|I_t| + |I_s|)H(W(x, y)) \le \frac{6}{\log(n)}H(W(x, y))$$

Finally, integrating the above equation over the whole square $[-R_n, R_n] \times [-R_n, R_n]$ we obtain

$$\left| \iint_{-R_n}^{R_n} \left(H\left(W(x, y) \right) - H\left(\widetilde{W}(x, y) \right) \right) d\mu_n(y) \, d\mu_n(x) \right| \leq \frac{6}{\log(n)} \sigma[W, \mu_n],$$

which proves (58) since

$$\lim_{n\to\infty} \frac{n}{\log(n)} |\sigma[W, \mu_n] - \sigma[\widetilde{W}_n, \mu_n]| \leq \lim_{n\to\infty} \frac{6n}{(\log n)^2} \sigma[W, \mu_n] = 0$$

thanks to Theorem 4.

References

1. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwanga, D.-U.: Complex networks: structure and dynamics. Phys. Rep. **424**, 175–308 (2006). doi:10.1016/j.physrep.2005.10.009
2. Newman, M.E.J.: Networks: An Introduction. Oxford University Press, Oxford (2010)
3. Barabási, A.-L.: Network Science. Cambridge University Press, Cambridge (2016)
4. Solomonoff, R., Rapoport, A.: Connectivity of random nets. B Math. Biophys. **13**(2), 107–117 (1951). doi:10.1007/BF02478357
5. Gilbert, E.N.: Random graphs. Ann. Math. Stat. **30**(4), 1141–1144 (1959). doi:10.1214/aoms/1177706098
6. Erdős, P., Rényi, A.: On random graphs. Publ. Math. **6**, 290–297 (1959)
7. Bender, E.A., Canfield, E.R.: The asymptotic number of labeled graphs with given degree sequences. J. Comb. Theory Ser. A **24**(3), 296–307 (1978). doi:10.1016/0097-3165(78)90059-6
8. Molloy, M., Reed, B.: A critical point for random graphs with a given degree sequence. Random Struct. Algorithm **6**, 161–179 (1995)
9. Dhamdhere, A., Dovrolis, C.: Twelve years in the evolution of the internet ecosystem. IEEE/ACM Trans. Netw. **19**(5), 1420–1433 (2011). doi:10.1109/TNET.2011.2119327
10. Newman, M.E.J.: Clustering and preferential attachment in growing networks. Phys. Rev. E **64**(2), 025102 (2001). doi:10.1103/PhysRevE.64.025102
11. Chung, F., Linyuan, L.: Connected components in random graphs with given expected degree sequences. Ann. Comb. **6**(2), 125–145 (2002). doi:10.1007/PL00012580
12. Chung, F., Linyuan, L.: The average distances in random graphs with given expected degrees. Proc. Natl. Acad. Sci. USA **99**(25), 15879–82 (2002). doi:10.1073/pnas.252631999
13. Park, J., Newman, M.E.J.: Statistical mechanics of networks. Phys. Rev. E **70**, 66117 (2004). doi:10.1103/PhysRevE.70.066117
14. Bianconi, G.: The entropy of randomized network ensembles. EPL **81**(2), 28005 (2008). doi:10.1209/0295-5075/81/28005
15. Garlaschelli, D., Loffredo, M.: Maximum likelihood: extracting unbiased information from complex networks. Phys. Rev. E **78**(1), 015101 (2008). doi:10.1103/PhysRevE.78.015101
16. Squartini, T., Garlaschelli, D.: Analytical maximum-likelihood method to detect patterns in real networks. New J. Phys. **13**(8), 083001 (2011). doi:10.1088/1367-2630/13/8/083001
17. Holland, P.W., Leinhardt, S.: An exponential family of probability distributions for directed graphs. J. Am. Stat. Assoc. **76**(373), 33–50 (1981). doi:10.1080/01621459.1981.10477598
18. Anand, K., Bianconi, G.: Entropy measures for networks: toward an information theory of complex topologies. Phys. Rev. E **80**, 045102(R) (2009). doi:10.1103/PhysRevE.80.045102
19. Squartini, T., de Mol, J., den Hollander, F., Garlaschelli, D.: Breaking of ensemble equivalence in networks. Phys. Rev. Lett. **115**(26), 268701 (2015). doi:10.1103/PhysRevLett.115.268701
20. Chatterjee, S., Diaconis, P., Sly, A.: Random graphs with a given degree sequence. Ann. Appl. Probab. **21**(4), 1400–1435 (2011). doi:10.1214/10-AAP728
21. Caldarelli, G., Capocci, A., Rios, P.D.L., Muñoz, M.A.: Scale-free networks from varying vertex intrinsic fitness. Phys. Rev. Lett. **89**(25), 258702 (2002). doi:10.1103/PhysRevLett.89.258702
22. Boguñá, M., Pastor-Satorras, R.: Class of correlated random networks with hidden variables. Phys. Rev. E **68**, 36112 (2003). doi:10.1103/PhysRevE.68.036112

23. Anand, K., Krioukov, D., Bianconi, G.: Entropy distribution and condensation in random networks with a given degree distribution. Phys. Rev. E **89**(6), 062807 (2014). doi:10.1103/PhysRevE.89.062807

24. Zuev, K., Papadopoulos, F., Krioukov, D.: Hamiltonian dynamics of preferential attachment. J. Phys. A **49**(10), 105001 (2016). doi:10.1088/1751-8113/49/10/105001

25. Evans, M., Rosenthal, J.S.: Probability and Statistics: The Science of Uncertainty. W.H. Freeman and Co, New York (2009)

26. Claffy, K., Hyun, Y., Keys, K., Fomenkov, M., Krioukov, D.: Internet mapping: from art to science. In: 2009 Cybersecurity Appl Technol Conf Homel Secur (2009). doi:10.1109/CATCH.2009.38

27. Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., Boguñá, M.: Hyperbolic geometry of complex networks. Phys. Rev. E **82**, 36106 (2010). doi:10.1103/PhysRevE.82.036106

28. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. **106**(4), 620–630 (1957). doi:10.1103/PhysRev.106.620

29. Shore, J., Johnson, R.: Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans. Inf. Theory **26**(1), 26–37 (1980). doi:10.1109/TIT.1980.1056144

30. Tikochinsky, Y., Tishby, N.Z., Levine, R.D.: Consistent inference of probabilities for reproducible experiments. Phys. Rev. Lett. **52**(16), 1357–1360 (1984). doi:10.1103/PhysRevLett.52.1357

31. Skilling, J.: The axioms of maximum entropy, In: Maximum-Entropy and Bayesian Methods in Science and Engineering, pp. 173–187. Springer, Dordrecht (1988). doi:10.1007/978-94-009-3049-0_8

32. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948). doi:10.1002/j.1538-7305.1948.tb01338.x

33. Krioukov, D.: Clustering implies geometry in networks. Phys. Rev. Lett. **116**(20), 208302 (2016). doi:10.1103/PhysRevLett.116.208302

34. Kapur, J.N.: Maximum-Entropy Models in Science and Engineering. Wiley, New Delhi (1989)

35. Aldous, D.J.: Representations for partially exchangeable arrays of random variables. J. Multivar. Anal. **11**(4), 581–598 (1981). doi:10.1016/0047-259X(81)90099-3

36. Diaconis, P., Janson, S.: Graph limits and exhcangeable random graphs. Rend di Matemtica **28**, 33–61 (2008)

37. Kallenberg, O.: Foundations of Modern Probability. Springer, New York (2002)

38. Shalizi, C.R., Rinaldo, A.: Consistency under sampling of exponential random graph models. Ann. Stat. **41**(2), 508–535 (2013). doi:10.1214/12-AOS1044

39. Krioukov, D., Ostilli, M.: Duality between equilibrium and growing networks. Phys. Rev. E **88**(2), 022808 (2013). doi:10.1103/PhysRevE.88.022808

40. Lovász, L., Szegedy, B.: Limits of dense graph sequences. J. Comb. Theory Ser. B **96**(6), 933–957 (2006). doi:10.1016/j.jctb.2006.05.002

41. Janson, S.: Graphons, cut norm and distance, couplings and rearrangements. NYJM Monogr, vol. 4 (2013)

42. Hoover, D.N.: Relations on probability spaces and arrays of random variables. Technical report, Institute for Adanced Study, Princeton (1979)

43. Dorogovtsev, S.N., Mendes, J.F.F., Samukhin, A.N.: Structure of growing networks with preferential linking. Phys. Rev. Lett. **85**(21), 4633–4636 (2000). doi:10.1103/PhysRevLett.85.4633

44. Krapivsky, P.L., Redner, S., Leyvraz, F.: Connectivity of growing random networks. Phys. Rev. Lett. **85**(21), 4629–4632 (2000). doi:10.1103/PhysRevLett.85.4629

45. Aldecoa, R., Orsini, C., Krioukov, D.: Hyperbolic graph generator. Comput. Phys. Commun. **196**, 492–496 (2015). doi:10.1016/j.cpc.2015.05.028

46. Caron, F., Fox, E.B.: Sparse graphs using exchangeable random measures. (2014). arXiv:1401.1137

47. Veitch, V., Roy, D.M.: The class of random graphs arising from exchangeable random measures. (2015). arXiv:1512.03099

48. Borgs, C., Chayes, J.T., Cohn, H., Holden, N.: Sparse exchangeable graphs and their limits via graphon processes. (2016). arXiv:1601.07134

49. Lovász, L.: Large Networks and Graph Limits. American Mathematical Society, Providence, RI (2012)

50. Aldous, D.J.: Exchangeability and related topics. In: Ecole d'Ete de Probabilites de Saint-Flour XIII, pp. 1–198. Springer, Berlin (1983). doi:10.1007/BFb0099421

51. McFarland, D.D., Brown, D.J.: Social distance as a metric: a systematic introduction to smallest space analysis. In: Bonds of Pluralism: The Form and Substance of Urban Social Networks, pp. 213–252. Wiley, New York (1973)

52. Faust, K.: Comparison of methods for positional analysis: structural and general equivalences. Soc. Netw. **10**(4), 313–341 (1988). doi:10.1016/0378-8733(88)90002-0

53. McPherson, J.M., Ranger-Moore, J.R.: Evolution on a dancing landscape: organizations and networks in dynamic Blau space. Soc. Forces **70**(1), 19–42 (1991). doi:10.1093/sf/70.1.19

54. Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. J. Am. Stat. Assoc. **97**(460), 1090–1098 (2002). doi:10.1198/016214502388618906

55. Bollobás, B., Janson, S., Riordan, O.: The phase transition in inhomogeneous random graphs. Random Struct. Algorithm **31**(1), 3–122 (2007). doi:10.1002/rsa.20168
56. Hatami, H., Janson, S., Szegedy, B.: Graph properties, graph limits and entropy. (2013). arXiv:1312.5626
57. Chatterjee, S., Varadhan, S.R.S.: The large deviation principle for the Erds-Rényi random graph. Eur. J. Comb. **32**(7), 1000–1017 (2011). doi:10.1016/j.ejc.2011.03.014
58. Chatterjee, S., Diaconis, P.: Estimating and understanding exponential random graph models. Ann. Stat. **41**(5), 2428–2461 (2013). doi:10.1214/13-AOS1155
59. Radin, C., Sadun, L.: Singularities in the entropy of asymptotically large simple graphs. J. Stat. Phys. **158**(4), 853–865 (2015). doi:10.1007/s10955-014-1151-3
60. Borgs, C., Chayes, J.T., Lovász, L., Sós, V.T., Vesztergombi, K.: Convergent sequences of dense graphs I: subgraph frequencies, metric properties and testing. Adv. Math. **219**(6), 1801–1851 (2008). doi:10.1016/j.aim.2008.07.008
61. Bianconi, G.: Entropy of network ensembles. Phys. Rev. E **79**, 36114 (2009)
62. Barvinok, A., Hartigan, J.A.: The number of graphs and a random graph with a given degree sequence. Random Struct. Algorithm **42**(3), 301–348 (2013). doi:10.1002/rsa.20409
63. Grandell, J.: Mixed Poisson Processes. Chapman & Hall/CRC, London (1997)
64. van der Hofstad, R.: Random Graphs and Complex Networks. Cambridge University Press, Cambridge, UK (2016)
65. Aubin, J.P.: Applied Functional Analysis. Wiley, New York (2000)
66. Gelfand, I.M., Fomin, S.V.: Calculus of Variations. Dover Publications, New York (2000)