ESTIMATING AND UNDERSTANDING EXPONENTIAL RANDOM GRAPH MODELS

SOURAV CHATTERJEE AND PERSI DIACONIS

Abstract. We introduce a new method for estimating the parameters of exponential random graph models. The method is based on a large-deviations approximation to the normalizing constant shown to be consistent using theory developed by Chatterjee and Varadhan [15]. The theory explains a host of difficulties encountered by applied workers: many distinct models have essentially the same MLE, rendering the problems "practically" ill-posed. We give the first rigorous proofs of "degeneracy" observed in these models. Here, almost all graphs have essentially no edges or are essentially complete. We supplement recent work of Bhamidi, Bresler and Sly [6] showing that for many models, the extra sufficient statistics are useless: most realizations look like the results of a simple Erdős–Rényi model. We also find classes of models where the limiting graphs differ from Erdős–Rényi graphs and begin to make the link to models where the natural parameters alternate in sign.

1. Introduction

Graph and network data are increasingly common and a host of statistical methods have emerged in recent years. Entry to this large literature may be had from the research papers and surveys in Fienberg [21, 22]. One mainstay of the emerging theory are the exponential families

$$p_\beta(G) = \exp\left(\sum_{i=1}^{k} \beta_i T_i(G) + \psi(\beta)\right) \tag{1.1}$$

where $\beta = (\beta_1, \ldots, \beta_k)$ is a vector of real parameters, $T_1, T_2, \ldots, T_k$ are functions on the space of graphs (e.g., the number of edges, triangles, stars, cycles, …), and ψ is a normalizing constant. In this paper, T_1 is usually taken to be the number of edges (or a constant multiple of it).

We review the literature of these models in Section 2.1. Estimating the parameters in these models has proved to be a challenging task. First, the normalizing constant $\psi(\beta)$ is unknown. Second, very different values of β can give rise to essentially the same distribution on graphs.

Key words and phrases. Random graph, Erdős-Rényi, graph limit, Exponential Random Graphs, parameter estimation.

Sourav Chatterjee's research was partially supported by NSF grant DMS 1005312 and a Sloan Research Fellowship.

Persi Diaconis' research was partially supported by NSF grant DMS 0804324.

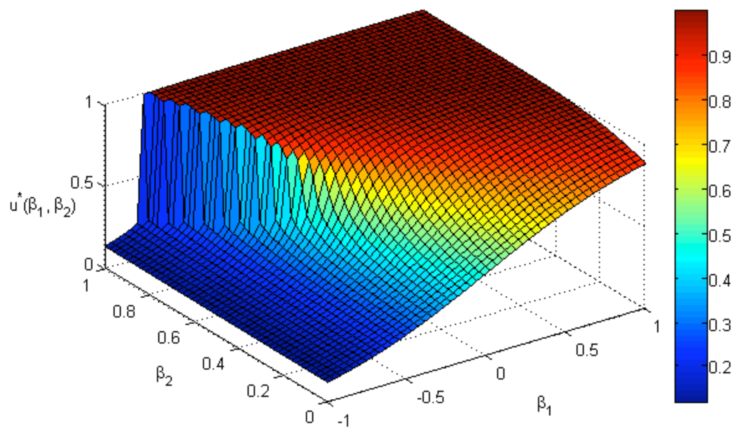


FIGURE 1. The plot of u^* against (β_1, β_2). There is a discontinuity on the left where u^* jumps from near 0 to near 1; this corresponds to a phase transition. (Picture by Sukhada Fadnavis.)

Here is an example: consider the model on simple graphs with n vertices,

$$p_{\beta_1,\beta_2}(G) = \exp\left(2\beta_1 E + \frac{6\beta_2}{n}\Delta - n^2\psi_n(\beta_1, \beta_2)\right) \tag{1.2}$$

where E, Δ denote the number of edges and triangles in the graph G. The normalization of the model ensures non-trivial large n limits. Without scaling, for large n, almost all graphs are empty or full. This model is studied by Strauss [52], Park and Newman [45, 46], Häggstrom and Jonasson [29], and many others.

Theorems 3.1 and 4.1 will show that for n large,

$$\psi_n(\beta_1, \beta_2) \simeq \sup_{0\le u\le 1}\left(\beta_1 u + \beta_2 u^3 - \frac{1}{2}u\log u - \frac{1}{2}(1-u)\log(1-u)\right). \tag{1.3}$$

The maximizing value of the right-hand side is denoted $u^*(\beta_1, \beta_2)$. A plot of this function appears in Figure 1. Theorem 4.2 shows that for any β_1 and $\beta_2 > 0$, with high probability, a pick from p_{β_1,β_2} is essentially the same as an Erdős–Rényi graph generated by including edges independently with probability $u^*(\beta_1, \beta_2)$. This phenomenon has previously been identified by Bhamidi et al. [6] and is discussed further in Section 2.1. Figure 2 shows the contour lines for Figure 1. All the (β_1, β_2) values on the same contour line lead to the same Erdős–Rényi model in the limit. Simulations show that the asymptotic results are valid for n as small as 30. Other methods for estimating normalizing constants are reviewed in Section 2.2.

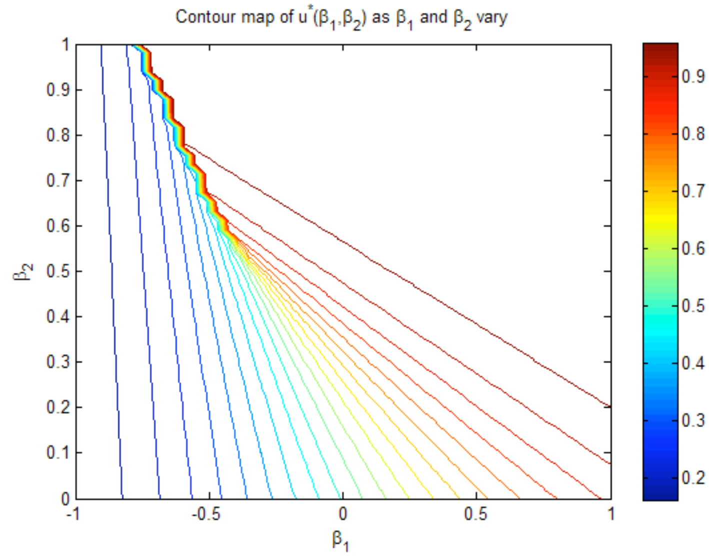


FIGURE 2. Contour lines for Figure 1. All pairs (β_1, β_2) on the same contour line correspond to the same value of u^* and hence those models will correspond to the same Erdős–Rényi model in the limit. The phase transition region is seen in the upper left-hand corner where all contour lines converge. (Picture by Sukhada Fadnavis.)

Our development uses the emerging tools of graph limits as developed by Lovász and coworkers. We give an overview in Section 2.3. Briefly, a sequence of graphs G_n converges to a limit if the proportion of edges, triangles, and other small subgraphs in G_n converges. There is a limiting object and the space of all these limiting objects serves as a useful compactification of the set of all graphs. Our theory works for functions $T_i(G)$ which are continuous in this topology. In their study of the large deviations of Erdős–Rényi random graphs, Chatterjee and Varadhan [15] derived the associated rate functions in the language of graph limit theory. Their work is crucial in the present development and is reviewed in Section 2.4.

Our main results are in Section 3 through Section 6. Working with general exponential models, Section 3 proves an extension of the approximation (1.3) for ψ_n (Theorem 3.1) and shows that, in the limit, almost all graphs from the model (1.1) are close to graphs where a certain functional is maximized. As will emerge, sometimes this maximum is taken on at a unique Erdős–Rényi model. Section 4 studies the problem for the model (1.1) when $\beta_2, \ldots, \beta_k$ are positive (β_1 may have any sign). It is shown that the large-deviations approximation for ψ_n can be easily calculated as a one-dimensional maximization (Theorem 4.1). Further, amplifying the results of Bhamidi et al. [6], it is shown that in these cases, almost all realizations of the model (1.1) are close to an Erdős–Rényi graph (or perhaps a finite mixture of Erdős–Rényi graphs) (Theorem 4.2). These mixture cases actually occur for natural parameter values. This explains a further difficulty found by applied workers who attempt to estimate parameters by using Monte Carlo to match observed counts of small subgraphs. Section 5 also gives a careful account of the phase transitions and near-degeneracies observed in the edge-triangle model (1.3).

Sections 6, 7 and 8 investigate cases where β_i is allowed to be negative. While the general case remains open (and appears complicated), in Section 6 it is shown that Theorems 4.1 and 4.2 hold as stated if $(\beta_i)_{2\le i\le k}$ are sufficiently small in magnitude. This requires a careful study of associated Euler–Lagrange equations. Section 7 shows how the results change for the model containing edges and triangles when β_2 is negative. For sufficiently large negative β_2, typical realizations look like a random bipartite graph. This is very different from the Erdős–Rényi model. The result generalizes to other models via an interesting analogy with the Erdős–Stone theorem from extremal graph theory. Finally, in Section 8 we discuss a model that exhibits transitivity, an important requirement for social networks.

2. Background

This section gives needed background and notation in three areas. Exponential graph models (Section 2.1), graph limits (Section 2.3), and large deviations (Section 2.4). Some new material is presented as well, e.g., the analysis of Monte Carlo maximum likelihood in Section 2.2.

2.1. Exponential random graphs.

Let $\mathcal{G}_n$ be the space of all simple graphs on n labled vertices ("simple" means undirected, with no loops or multiple edges). Thus $\mathcal{G}_n$ contains $2^{\binom{n}{2}}$ elements. A variety of models in active use can be presented in exponential form

$$p_\beta(G) = \exp\left(\sum_{i=1}^k \beta_i T_i(G) - \psi(\beta)\right) \tag{2.1}$$

where $\beta = (\beta_1, \dots, \beta_k)$ is a vector of real parameters, $T_1, T_2, \dots, T_k$ are real-valued functions on $\mathcal{G}_n$, and $\psi(\beta)$ is a normalizing constant. Usually, T_i are taken to be counts of various subgraphs, e.g., $T_1(G) = \#$ edges in G, $T_2(G) = \#$ triangles in G, ... The main results of Section 3 work for more general "continuous functions" on graph space, such as the degree sequence or the eigenvalues of the adjacency matrix. This allows models with sufficient statistics of the form $\sum_{i=1}^n \beta_i d_i(G)$ with $d_i(G)$ the degree of vertex i. See, e.g., [14].

These exponential models were used by Holland and Leinhardt [32] in the directed case. Frank and Strauss [24] developed them, showing that if T_i are chosen as edges, triangles, and stars of various sizes, the resulting random graph edges form a Markov random field. A general development is in Wasserman and Faust [54]. Newer developments are summarized in Snijders et al. [51]. Finally, Rinaldo et al. [47] develop the geometric theory for this class of models with extensive further references.

A major problem in this field is the evaluation of the constant $\psi(\beta)$ which is crucial for carrying out maximum likelihood and Bayesian inference. As far as we know, there is no feasible analytic method for approximating ψ when n is large. Physicists have tried the technique of mean-field approximations; see Park and Newman [45, 46] for the case where T_1 is the number of edges and T_2 is the number of two-stars or the number of triangles. Mean-field approximations have no rigorous foundation, however, and are known to be unreliable in related models such as spin glasses [53]. For exponential graph models, Chatterjee and Dey [13] prove that they work for some restricted ranges of $\{\beta_i\}$: values where the graphs are shown to be essentially Erdős–Rényi graphs (see Theorem 4.2 below and [6]).

A host of techniques for approximating the normalizing constant using various Monte Carlo schemes have been proposed. As explained in Section 2.2, these include the MCMLE procedure of Geyer and Thompson [28] (see example below). The bridge sampling approach of Gelman and Meng [27] also builds on techniques suggested by physicists to estimate free energy ($\psi(\beta)$ in our context). The equi-energy sampler of Kou et al. [36] can also be harnessed to estimate ψ.

Alas, at present writing these procedures do not seem very useful. Snijders [50] and Handcock [31] demonstrate this empirically with further discussion in [51]. One theoretical explanation for the poor performance of these techniques comes from the work of Bhamidi et al. [6]. Most of the algorithms above require a sample from the model (2.1). This is most often done by using a local Markov chain based on adding or deleting edges (via Metropolis or Glauber dynamics). These authors show that if the parameters are non-negative, then for large n,

- either the p_β model is essentially the same as an Erdős–Rényi model (in which case the Markov chain mixes in $n^2 \log n$ steps);
- or the Markov chain takes exponential time to mix.

Thus, in cases where the model is not essentially trivial, the Markov chains required to carry MCMLE procedures cannot be usefully run to stationarity.

Two other approaches to estimation are worth mentioning. The pseudo-likelihood approach of Besag [5] is widely used because of its ease of implementation. Its properties are at best poorly understood: it does not directly maximize the likelihood and in empirical comparisions (see, e.g., [17]), has appreciably larger variability than the MLE. Comets and Janžura [16] prove consistency and asymptotic normality of the maximum pseudo-likelihood estimator in certain Markov random field models. Chatterjee [12] shows that it is consistent for estimating the temperature parameter of the Sherrington-Kirkpatrick model of spin glasses. The second approach is Snijders' [50] suggestion to use the Robbins–Monroe optimization procedure to compute solutions to the moment equations $E_\beta(T(G)) = T(G^*)$ where G^* is the observed graph. While promising, the approach requires generating points from p_β for arbitrary β. The only way to do this at present is by MCMC and the results of [6] suggest this may be impractical.

Practical Remark. One use for the normalizing constant is to enable maximum likelihood estimates of the β parameter in the model (1.1). This requires evaluating $\psi(\beta)$ on a fine grid in β space and then carrying out the maximization by classical methods (e.g., a grid search). Iterative refinement may be used when honing in at the maximum. The theory developed below allows for refining the estimate of $\psi(\beta)$ along the following lines. Consider the situation of Section 4 below where $\beta_2, \ldots, \beta_k$ are positive. Theorem 4.2 shows that the exponential model is close to an Erdős–Rényi graph with parameter u^* determined by an equation similar to (1.3). Let $q(G|\beta) = \exp(\sum_{i=1}^k \beta_i T_i(G))$ be the unnormalized density. Generate independent, identically distributed random graphs G_i from the Erdős–Rényi model $p_{u^*}(G)$. The estimator

$$\frac{1}{N}\sum \frac{q_\beta(G_i)}{p_{u^*}(G_i)}$$

is unbiased for $\exp\psi(\beta)$. Many similar variations can be concocted by combining present theory with the host of algorithms reviewed by Gelman and Meng [27, Sect. 3.4].

2.2. **A simple example.** In this section we treat the simplest exponential graph model, the Erdős–Rényi model. Here the relevant Markov chains for carrying out the Monte Carlo estimates of normalizing constants described at the end of Section 2.1 can be explicitly diagonalized and estimates for the variance of various estimators are available in closed form. The main findings are these: for graphs with n vertices,

- the Metropolis algorithm for sampling from p_β converges in order $n^2 \log n$ steps;
- the variance of MCMLE estimates of the normalizing constant is exponential in n^2, rendering them impractical.

The model to be studied is

$$p_\beta(G) = z(\beta)^{-1} e^{\beta E(G)} \tag{2.2}$$

for $-\infty < \beta < \infty$ a fixed parameter, $z(\beta)$ the normalizing constant, and $E(G)$ the number of edges in G. This is just the Erdős–Rényi model with edges included independently with parameter $p = e^\beta/(1+e^\beta)$. Here, the normalizing constant is

$$z(\beta)^{-1} = (1+e^\beta)^{\binom{n}{2}}$$

and $\beta \geq 0$ corresponds to $p \geq 1/2$. *We suppose throughout this section that $\beta \geq 0$.*

A natural Markov chain for generating from p_β is the Metropolis algorithm:

$$\begin{aligned}
&\bullet \text{ From } G \text{ pick } (i,j),\ 1 \le i < j \le n, \text{ uniformly.}\\
&\bullet \text{ If } (i,j) \text{ is not in } G, \text{ add this edge.}\\
&\bullet \text{ If } (i,j) \text{ is in } G, \text{ delete it with probability } e^{-\beta}\\
&\quad \text{and leave it with probability } 1-e^{-\beta}.
\end{aligned} \tag{2.3}$$

Call the transition matrix of this Markov chain $K(G,G')$. The following theorem gives an explicit spectral decomposition of K. It is useful to identify a graph G with the binary indicator of its edges, a vector $x_G \in C_2^m$ with $m = \binom{n}{2}$.

THEOREM 2.1. *For the Metropolis Markov chain K of (2.3), with $m = \binom{n}{2}$,*

(1) *K is reversible with stationary distribution $p_\beta(G)$ of (2.2).*

(2) *For each $\xi \in C_2^m$ there is an eigenvalue β_ξ with eigenfunction $\psi_\xi(x)$ given by*

$$\psi_\xi(x) = (-1)^{\xi\cdot x} e^{\frac{\beta}{2}(|\xi|-2\xi\cdot x)}, \qquad \beta_\xi = 1 - \frac{|\xi|(1+e^{-\beta})}{m}.$$

Here $|\xi|$ is the number of ones in ξ and $\xi\cdot x$ is the usual inner product. The eigenfunctions are orthornormal in $L^2(p_\beta)$.

(3) *The $L^2(p_\beta)$ or chi-square distance from stationarity, starting at $G \leftrightarrow x_G$ is*

$$\chi^2_G(\ell) = \sum_{G'} \frac{(K^\ell(G,G') - p_\beta(G'))^2}{p_\beta(G')} = \sum_{\xi\neq 0} \psi_\xi^2(x_G)\beta_\xi^{2\ell}.$$

(4) *For $0 \le \beta \le 1$, as n tends to infinity, $\ell^* = \frac{m(\log m + c)}{2(1+e^{-\beta})}$ steps are necessary and sufficient to drive $x^2_G(\ell)$ to zero:*

$$\lim_{n\to\infty} \chi^2_\emptyset(\ell^*) = e^{e^{\beta-c}} - 1, \qquad \lim_{n\to\infty} \chi^2_{K_n}(\ell^*) = e^{e^{-\beta-c}} - 1;$$

$$\chi^2_\emptyset(\ell^*) \ge e^\beta m \left(1 - \frac{(1+e^{-\beta})}{m}\right)^{2\ell^*},$$

$$\chi^2_{K_n}(\ell^*) \ge e^{-\beta} m \left(1 - \frac{(1+e^{-\beta})}{m}\right)^{2\ell^*}.$$

Proof. For (1), the Metropolis algorithm is reversible by construction [30]. For (2), the Metropolis chain is a product chain on the product space C_2^m with component chain

$$\begin{pmatrix} 0 & 1 \\ e^{-\beta} & 1-e^{-\beta} \end{pmatrix}$$

with stationary distribution

$$\pi(0) = \frac{1}{1+e^\beta},\ \pi(1) = \frac{e^\beta}{1+e^\beta}.$$

This two-state chain has (right) eigenfunctions/eigenvalues (orthonormal in $L^2(\pi)$)

$$\psi_0(0) = \psi_0(1) = 1, \qquad \psi_1(0) = e^{\beta/2}, \qquad \psi_1(1) = -e^{-\beta/2},$$
$$\beta_0 = 1, \qquad \beta_1 = -e^{-\beta}.$$

By elementary computations [19, Sect. 6], the product chain has eigenfunctions the product of these component eigenfunctions/values yielding (2). Formula (3) follows from elementary spectral theory (see, e.g., [48]). For (4), starting from the empty graph $G = \emptyset$ corresponds to $x_\emptyset = 0$ and then

$$x_\emptyset^2(\ell) = \sum_{j=1}^{m} e^{\beta j} \binom{m}{j} \left(1 - \frac{j(1+e^{-\beta})}{m}\right)^{2\ell}.$$

Similarly, starting at the complete graph $K_n \leftrightarrow x_{K_n} = (1, \ldots, 1)$ and

$$x_{K_n}^2(\ell) = \sum_{j=1}^{m} e^{-\beta j} \binom{m}{j} \left(1 - \frac{j(1+e^{-\beta})}{m}\right)^{2\ell}.$$

Now the stated results follow from elementary calculus (for upper bounds) and just using the first term in the sums above (for the lower bound). $\square$

Note that the right-hand sides of the limits in (4) tend to zero as c tends to ∞. Thus there is a cutoff in convergence at ℓ^*. More crudely, for the simple model (2.2), order $m \log m$ steps are necessary and sufficient for convergence for all values of β and all starting states. This remains true for total variation. More complicated models can have more complicated mixing behavior [6]. The calculations for the Metropolis algorithm can be simply adapted for Glauber dynamics with very similar conclusions.

In applications, Markov chains such as the Metropolis algorithm are used to estimate normalizing constants or their ratios. Consider an exponential graph model p_β (as in (2.1)) on $\mathcal{G}_n$ with normalizing constant $z(\beta)$. Several estimates of $z(\beta)$ are discussed in Section 2.1. These include:

Importance sampling. Generate $G_1, G_2, \ldots, G_N$ from a Markov chain with known stationary distribution $Q(G)$ and use

$$\hat{z}_I = \frac{1}{N} \sum_{j=1}^{N} \frac{\exp\left\{\sum_{i=1}^{k} \beta_i T_i(G_j)\right\}}{Q(G)}. \tag{2.4}$$

This is an unbiased estimate of $z(\beta)$. This requires knowing Q. (For example, an Erdős–Rényi model may be used.) If Q is only known up to a normalizing constant, say $Q = z\bar{Q}$, then

$$\frac{\sum_{j=1}^{N} \exp\{\sum_{i=1}^{k} \beta_i T_i(G_j)\}/\bar{Q}(G_j)}{\sum_{j=1}^{N} 1/\bar{Q}(G_j)}$$

may be used.

MCMLE. Generate $G_1, G_2, \ldots, G_N$ with stationary distribution p_{β^0} and use

$$\hat{z}_M = \frac{1}{N} \sum_{j=1}^{N} \exp\left\{(\beta_i - \beta_i^0) T_i(G_j)\right\}. \tag{2.5}$$

This is an unbiased estimate of $z(\beta)/z(\beta^0)$.

Acceptance ratio. Generate $G_1, \ldots, G_{N_1}$ with stationary distribution p_{β^0} and $G_1', \ldots, G_{N_2}'$ with stationary distribution p_β and use

$$\hat{z}_A = \frac{\frac{1}{N_1} \sum_{j=1}^{N_1} \exp\left\{\sum_j \beta_i T_i(G_j)\right\} \alpha(G_j)}{\frac{1}{N_2} \sum_{j=1}^{N_2} \exp\left\{\sum_j \beta_i^0 T(G_j')\right\} \alpha(G_j)}. \tag{2.6}$$

Here α can be any function on graph space. The numerator is an unbiased estimator of $c/z(\beta^0)$. The denominator is an unbiased estimator of $c/z(\beta)$ with $c = \sum_{G \in \mathcal{G}_n} \exp\{\sum_{i=1}^{k} (\beta_i + \beta_i^0) T_i(G)\} \alpha(G)$. Thus the ratio estimates $z(\beta)/z(\beta_0)$. Common choices of $\alpha(G)$ are the constant function, or $\alpha(G) = \exp\{\frac{1}{2} \sum (\beta_i - \beta_i^0) T_i(G)\}$. See [27] for history and efforts to optimize α.

All of these estimators involve things like $\mathbb{E}_\beta(f(G))$ with $f(G)$ an exponentially large function. In the remainder of this section we investigate the variance of these estimates in the Erdős–Rényi case. To ease notation, suppose that all Markov chains start in stationarity. Let $K(G, G')$ be a reversible Markov chain on $\mathcal{G}_n$ with stationary distribution $P(G)$. Suppose that K has eigenvalues β_ξ and eigenfunctions ψ_ξ for $\xi \in C_2^m$. Let f be a function on $\mathcal{G}_n$. Expand $f(G) = \sum_\xi \hat{f}(\xi) \psi_\xi(x_G)$, with

$$\hat{f}(\xi) = \sum_G f(G) \psi_\xi(x_G) p_\beta(G).$$

Let $G_1, G_2, \ldots, G_N$ be a stationary realization from K. Proposition 2.1 in [4] shows that the estimator $\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} f(G_i)$ is unbiased with variance

$$\operatorname{var}(\hat{\mu}) = \frac{1}{N^2} \sum_{\xi \neq 0} \left|\hat{f}(\xi)\right|^2 W_N(\xi) \tag{2.7}$$

where

$$W_N(\xi) = \frac{N + 2\beta_\xi - N\beta_\xi^2 + 2\beta_\xi^{N+1}}{(1 - \beta_\xi)^2}.$$

For large N, the asymptotic variance is

$$\begin{aligned} \sigma_\infty^2(\hat{\mu}) := \lim_{N \to \infty} N \operatorname{var}(\hat{\mu}) &= \sum_{\xi \neq 0} \left|\hat{f}(\xi)\right|^2 \frac{1 + \beta_\xi}{1 - \beta_\xi} \\ &\leq \frac{2}{1 - \beta_1} \|f\|_{2,0}^2 := \bar{\sigma}_\infty^2. \end{aligned} \tag{2.8}$$

Here β_1 is the second eigenvalue and

$$\|f\|_{2,0}^2 = \sum_{\xi \neq 0} |\hat{f}(\xi)|^2 = \sum_{G \in \mathcal{G}_n} f^2(G) p_\beta(G) - \Big(\sum_{G \in \mathcal{G}_n} f(G) p_\beta(G) \Big)^2.$$

For the Erdős–Rényi model (2.2) with the Markov chain (2.3), all the quantities needed above are available in closed form:

LEMMA 2.2. *With notation as in Theorem* 2.1*, let* $f(G) = e^{aE(G)}$*. Then*

$$\hat{f}(\xi) = (1 - e^a)^{|\xi|}(1 + e^{a+\beta})^{m-|\xi|}.$$

As an example, we compute the usual bound for the asymptotic variance of the MCMLE estimate (2.5). More precise calculations based on (2.7) do not change the basic message; the standard deviation is exponentially larger than the mean.

PROPOSITION 2.3. *For* $\beta \geq 0$ *and* $\beta_0 \geq 0$ *in the Erdős–Rényi model* (2.2)*, the MCMLE estimate for the ratio of normalizing constants* (2.5) *is unbiased with mean*

$$\mu = \left(\frac{1 + e^{\beta_0}}{1 + e^{\beta}} \right)^m.$$

The second eigenvalue is $\beta_1 = 1 - (1 + e^{-\beta})/m$*. The variance bound is*

$$\bar{\sigma}_\infty^2 = \frac{2}{1 - \beta_1} \|f\|_{2,0}^2 \quad \textit{with} \quad \|f\|_{2,0}^2 = \mu^2 \left[\left(1 + \left(\frac{1 - e^{\beta - \beta_0}}{1 + e^{\beta}} \right)^2 \right)^m - 1 \right].$$

It follows that, if $\beta_0 \neq \beta$*,* $\bar{\sigma}_\infty^2/\mu^2$ *tends to* ∞ *exponentially fast as* n *tends to infinity.*

For example if $\beta_0 = 2$ and $\beta = 1$ then $\mu \doteq (2.2562)^m$ and $\bar{\sigma}_\infty^2/\mu^2 \doteq \frac{m}{2.7358}[(1.042)^m - 1]$. If $n = 30$, $\bar{\sigma}_\infty/\mu \doteq 95,431$. If $n = 100$, the ratio is huge.

2.3. **Graph limits.** In a sequence of papers [9, 10, 11, 25, 37, 38, 39, 40, 41, 42, 43], Laszlo Lovász and coauthors V.T. Sós, B. Szegedy, C. Borgs, J. Chayes, K. Vesztergombi, A. Schrijver, and M. Freedman have developed a beautiful, unifying theory of graph limits. (See also the related work of Austin [2] and Diaconis and Janson [18] which traces this back to work of Aldous [1], Hoover [33] and Kallenberg [35].) This sheds light on topics such as graph homomorphisms, Szemerédi's regularity lemma, quasi-random graphs, graph testing and extremal graph theory, and has even found applications in statistics and related areas (see e.g., [14]). Their theory has been developed for dense graphs (number of edges comparable to the square of number of vertices) but parallel theories for sparse graphs are beginning to emerge [7].

Lovász and coauthors define the limit of a sequence of dense graphs as follows. We quote the definition verbatim from [40] (see also [10, 11, 18]). Let G_n be a sequence of simple graphs whose number of nodes tends to infinity. For every fixed simple graph H, let $|\hom(H,G)|$ denote the number of homomorphisms of H into G (i.e., edge-preserving maps $V(H) \to V(G)$, where $V(H)$ and $V(G)$ are the vertex sets). This number is normalized to get the homomorphism density

$$t(H,G) := \frac{|\hom(H,G)|}{|V(G)|^{|V(H)|}}. \tag{2.9}$$

This gives the probability that a random mapping $V(H) \to V(G)$ is a homomorphism.

Note that $|\hom(H,G)|$ is not the count of the number of copies of H in G, but is a constant multiple of that if H is a complete graph. For example, if H is a triangle, $|\hom(H,G)|$ is the number of triangles in G multiplied by six. On the other hand if H is, say, a 2-star (i.e. a triangle with one edge missing) and G is a triangle, then the number of copies of H in G is zero, while $|\hom(H,G)| = 3^3 = 27$.

Suppose that the graphs G_n become more and more similar in the sense that $t(H,G_n)$ tends to a limit $t(H)$ for every H. One way to define a limit of the sequence $\{G_n\}$ is to define an appropriate limit object from which the values $t(H)$ can be read off.

The main result of [40] (following the earlier equivalent work of Aldous [1] and Hoover [33]) is that indeed there is a natural "limit object" in the form of a function $h \in \mathcal{W}$, where $\mathcal{W}$ is the space of all measurable functions from $[0,1]^2$ into $[0,1]$ that satisfy $h(x,y) = h(y,x)$ for all x,y.

Conversely, every such function arises as the limit of an appropriate graph sequence. This limit object determines all the limits of subgraph densities: if H is a simple graph with $V(H) = [k] = \{1,\ldots,k\}$, let

$$t(H,h) = \int_{[0,1]^k} \prod_{(i,j)\in E(H)} h(x_i,x_j)\, dx_1 \ldots dx_k. \tag{2.10}$$

Here $E(H)$ denotes the edge set of H. A sequence of graphs $\{G_n\}_{n\geq 1}$ is said to converge to h if for every finite simple graph H,

$$\lim_{n\to\infty} t(H,G_n) = t(H,h). \tag{2.11}$$

Intuitively, the interval $[0,1]$ represents a 'continuum' of vertices, and $h(x,y)$ denotes the probability of putting an edge between x and y. For example, for the Erdős–Rényi graph $G(n,p)$, if p is fixed and $n \to \infty$, then the limit graph is represented by the function that is identically equal to p on $[0,1]^2$.

These limit objects, i.e., elements of $\mathcal{W}$, are called "graph limits" or "graphons" in [10, 11, 40]. A finite simple graph G on $\{1,\ldots,n\}$ can also be represented as a graph limit f^G is a natural way, by defining

$$f^G(x,y) = \begin{cases} 1 & \text{if } (\lceil nx\rceil, \lceil ny\rceil) \text{ is an edge in } G \\ 0 & \text{otherwise.} \end{cases} \tag{2.12}$$

The definition makes sense because $t(H,f^G) = t(H,G)$ for every simple graph H and therefore the constant sequence $\{G,G,\ldots\}$ converges to the graph limit f^G. Note that this allows *all* simple graphs, irrespective of the number of vertices, to be represented as elements of a single abstract space, namely $\mathcal{W}$.

With the above representation, it turns out that the notion of convergence in terms of subgraph densities outlined above can be captured by an explicit metric on $\mathcal{W}$, the so-called *cut distance* (originally defined for finite graphs by Frieze and Kannan [26]). Start with the space $\mathcal{W}$ of measurable functions $f(x,y)$ on $[0,1]^2$ that satisfy $0 \le f(x,y) \le 1$ and $f(x,y) = f(y,x)$. Define the cut distance

$$d_\Box(f,g) := \sup_{S,T\subseteq[0,1]} \left| \int_{S\times T} [f(x,y) - g(x,y)]\, dxdy \right|. \tag{2.13}$$

Introduce in $\mathcal{W}$ an equivalence relation: Let Σ be the space of measure preserving bijections $\sigma : [0,1] \to [0,1]$. Say that $f(x,y) \sim g(x,y)$ if $f(x,y) = g_\sigma(x,y) := g(\sigma x, \sigma y)$ for some $\sigma \in \Sigma$. Denote by $\widetilde{g}$ the closure in $(\mathcal{W}, d_\Box)$ of the orbit $\{g_\sigma\}$. The quotient space is denoted by $\widetilde{\mathcal{W}}$ and τ denotes the natural map $g \to \widetilde{g}$. Since $d_\Box$ is invariant under σ one can define on $\widetilde{\mathcal{W}}$, the natural distance $\delta_\Box$ by

$$\delta_\Box(\widetilde{f},\widetilde{g}) := \inf_\sigma d_\Box(f, g_\sigma) = \inf_\sigma d_\Box(f_\sigma, g) = \inf_{\sigma_1,\sigma_2} d_\Box(f_{\sigma_1}, g_{\sigma_2})$$

making $(\widetilde{\mathcal{W}}, \delta_\Box)$ into a metric space. To any finite graph G, we associate f^G as in (2.12) and its orbit $\widetilde{G} = \tau f^G = \widetilde{f}^G \in \widetilde{\mathcal{W}}$.

The papers by Lovász and coauthors establish many important properties of the metric space $\widetilde{\mathcal{W}}$ and the associated notion of graph limits. For example, $\widetilde{\mathcal{W}}$ is compact. A pressing objective is to understand what functions from $\widetilde{\mathcal{W}}$ into $\mathbb{R}$ are continuous. Fortunately, it is an easy fact that the homomorphism density $t(H,\cdot)$ is continuous for any finite simple graph H [10, 11]. There are other, more complicated functions that are continuous (see, e.g., [3]).

2.4. **Large deviations for random graphs.** Let $G(n,p)$ be the random graph on n vertices where each edge is added independently with probability p. This model has been the subject of extensive investigations since the pioneering work of Erdős and Rényi [20], yielding a large body of literature (see [8, 34] for partial surveys).

Recently, Chatterjee and Varadhan [15] formulated a large deviation principle for the Erdős–Rényi graph, in the same way as Sanov's theorem [49] gives a large deviation principle for an i.i.d. sample. The formulation and proof of this result makes extensive use of the properties of the topology described in Section 2.3.

Let $I_p : [0,1] \to \mathbb{R}$ be the function

$$I_p(u) := \frac{1}{2} u \log \frac{u}{p} + \frac{1}{2}(1-u) \log \frac{1-u}{1-p}. \tag{2.14}$$

The domain of the function I_p can be extended to $\mathcal{W}$ as

$$I_p(h) := \int_0^1 \int_0^1 I_p(h(x,y))dxdy. \tag{2.15}$$

The function I_p can be defined on $\widetilde{\mathcal{W}}$ by declaring $I_p(\widetilde{h}) := I_p(h)$ where h is any representative element of the equivalence class $\widetilde{h}$. Of course, this raises the question whether I_p is well defined on $\widetilde{\mathcal{W}}$. It was proved in [15] that the function I_p is indeed well defined on $\widetilde{\mathcal{W}}$ and is lower semicontinuous under the cut metric $\delta_\Box$.

The random graph $G(n,p)$ induces probability distributions $\mathbb{P}_{n,p}$ on the space $\mathcal{W}$ through the map $G \to f^G$ and $\widetilde{\mathbb{P}}_{n,p}$ on $\widetilde{\mathcal{W}}$ through the map $G \to f^G \to \widetilde{f}^G = \widetilde{G}$. The large deviation principle for $\widetilde{\mathbb{P}}_{n,p}$ on $(\widetilde{\mathcal{W}}, \delta_\Box)$ is the main result of [15].

Theorem 2.4 (Chatterjee and Varadhan [15]). *For each fixed $p \in (0,1)$, the sequence $\widetilde{\mathbb{P}}_{n,p}$ obeys a large deviation principle in the space $\widetilde{\mathcal{W}}$ (equipped with the cut metric) with rate function I_p defined by* (2.15). *Explicitly, this means that for any closed set $\widetilde{F} \subseteq \widetilde{\mathcal{W}}$,*

$$\limsup_{n\to\infty} \frac{1}{n^2} \log \widetilde{\mathbb{P}}_{n,p}(\widetilde{F}) \leq - \inf_{\widetilde{h}\in\widetilde{F}} I_p(\widetilde{h}). \tag{2.16}$$

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

$$\liminf_{n\to\infty} \frac{1}{n^2} \log \widetilde{\mathbb{P}}_{n,p}(\widetilde{U}) \geq - \inf_{\widetilde{h}\in\widetilde{U}} I_p(\widetilde{h}). \tag{2.17}$$

3. Exponential random graphs

Let $T : \widetilde{\mathcal{W}} \to \mathbb{R}$ be a bounded continuous function on the metric space $(\widetilde{\mathcal{W}}, \delta_\Box)$. Fix n and let $\mathcal{G}_n$ denote the set of simple graphs on n vertices. Then T induces a probability mass function p_n on $\mathcal{G}_n$ defined as:

$$p_n(G) := e^{n^2(T(\widetilde{G})-\psi_n)}.$$

Here $\widetilde{G}$ is the image of G in the quotient space $\widetilde{\mathcal{W}}$ as defined in Section 2.2 and ψ_n is a constant such that the total mass of p_n is 1. Explicitly,

$$\psi_n = \frac{1}{n^2} \log \sum_{G\in\mathcal{G}_n} e^{n^2 T(\widetilde{G})} \tag{3.1}$$

The coefficient n^2 is meant to ensure that ψ_n tends to a non-trivial limit as $n \to \infty$. To describe this limit, define a function $I : [0,1] \to \mathbb{R}$ as

$$I(u) := \frac{1}{2} u \log u + \frac{1}{2}(1-u)\log(1-u)$$

and extend I to $\widetilde{\mathcal{W}}$ in the usual manner:

$$I(\widetilde{h}) = \frac{1}{2} \iint_{[0,1]^2} I(h(x,y))\, dxdy \tag{3.2}$$

where h is a representative element of the equivalence class $\widetilde{h}$. As mentioned before, it follows from a result of [15] that I is well defined and lower semi-continuous on $\widetilde{\mathcal{W}}$. The following theorem is the first main result of this paper.

THEOREM 3.1. *If $T : \widetilde{\mathcal{W}} \to \mathbb{R}$ is a bounded continuous function and ψ_n and I are defined as above, then*

$$\psi := \lim_{n\to\infty} \psi_n = \sup_{\widetilde{h}\in\widetilde{\mathcal{W}}} (T(\widetilde{h}) - I(\widetilde{h})).$$

Proof. For each Borel set $\widetilde{A} \subseteq \widetilde{\mathcal{W}}$ and each n, define

$$\widetilde{A}_n := \{\widetilde{h} \in \widetilde{A} : \widetilde{h} = \widetilde{G} \text{ for some } G \in \mathcal{G}_n\}.$$

Let $\mathbb{P}_{n,p}$ be the Erdős–Rényi measure defined in Section 3. Note that $\widetilde{A}_n$ is a finite set and

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

Thus, if $\widetilde{F}$ is a closed subset of $\widetilde{\mathcal{W}}$ then by Theorem 2.4

$$\begin{aligned}\limsup_{n\to\infty} \frac{\log|\widetilde{F}_n|}{n^2} &\le \frac{\log 2}{2} - \inf_{\widetilde{h}\in\widetilde{F}} I_{1/2}(\widetilde{h}) \\ &= -\inf_{\widetilde{h}\in\widetilde{F}} I(\widetilde{h}). \end{aligned} \tag{3.3}$$

Similarly if $\widetilde{U}$ is an open subset of $\widetilde{\mathcal{W}}$,

$$\liminf_{n\to\infty} \frac{\log|\widetilde{U}_n|}{n^2} \ge -\inf_{\widetilde{h}\in\widetilde{U}} I(\widetilde{h}). \tag{3.4}$$

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

$$e^{n^2\psi_n} \le \sum_{a\in R} e^{n^2(a+\epsilon)}|\widetilde{F}^a_n| \le |R| \sup_{a\in R} e^{n^2(a+\epsilon)}|\widetilde{F}^a_n|.$$

By (3.3), this shows that

$$\limsup_{n\to\infty} \psi_n \le \sup_{a\in R}\bigl(a + \epsilon - \inf_{\widetilde{h}\in\widetilde{F}^a} I(\widetilde{h})\bigr).$$

Each $\widetilde{h} \in \widetilde{F}^a$ satisfies $T(\widetilde{h}) \ge a$. Consequently,

$$\sup_{\widetilde{h}\in\widetilde{F}^a} (T(\widetilde{h}) - I(\widetilde{h})) \ge \sup_{\widetilde{h}\in\widetilde{F}^a} (a - I(\widetilde{h})) = a - \inf_{\widetilde{h}\in\widetilde{F}^a} I(\widetilde{h}).$$

Substituting this in the earlier display gives

$$\begin{aligned}\limsup_{n\to\infty} \psi_n &\le \epsilon + \sup_{a\in R} \sup_{\widetilde{h}\in\widetilde{F}^a} (T(\widetilde{h}) - I(\widetilde{h})) \\ &= \epsilon + \sup_{\widetilde{h}\in\widetilde{\mathcal{W}}} (T(\widetilde{h}) - I(\widetilde{h})). \end{aligned} \tag{3.5}$$

For each $a \in R$, let $\widetilde{U}^a := T^{-1}((a, a+\epsilon))$. By the continuity of T, $\widetilde{U}^a$ is an open set. Note that

$$e^{n^2\psi_n} \geq \sup_{a \in R} e^{n^2 a} |\widetilde{U}^a_n|.$$

Therefore by (3.4), for each $a \in R$

$$\liminf_{n\to\infty} \psi_n \geq a - \inf_{\widetilde{h} \in \widetilde{U}^a} I(\widetilde{h}).$$

Each $\widetilde{h} \in \widetilde{U}^a$ satisfies $T(\widetilde{h}) < a + \epsilon$. Therefore,

$$\sup_{\widetilde{h} \in \widetilde{U}^a} (T(\widetilde{h}) - I(\widetilde{h})) \leq \sup_{\widetilde{h} \in \widetilde{U}^a} (a + \epsilon - I(\widetilde{h})) = a + \epsilon - \inf_{\widetilde{h} \in \widetilde{U}^a} I(\widetilde{h}).$$

Together with the previous display, this shows that

$$\begin{aligned} \liminf_{n\to\infty} \psi_n &\geq -\epsilon + \sup_{a \in R} \sup_{\widetilde{h} \in \widetilde{U}^a} (T(\widetilde{h}) - I(\widetilde{h})) \\ &= -\epsilon + \sup_{\widetilde{h} \in \widetilde{\mathcal{W}}} (T(\widetilde{h}) - I(\widetilde{h})). \end{aligned} \tag{3.6}$$

Since ϵ is arbitrary in (3.5) and (3.6), this completes the proof. □

Theorem 3.1 gives an asymptotic formula for ψ_n. However, it says nothing about the behavior of a random graph drawn from the exponential random graph model. Some aspects of this behavior can be described as follows. Let $\widetilde{F}^*$ be the subset of $\widetilde{\mathcal{W}}$ where $T(\widetilde{h}) - I(\widetilde{h})$ is maximized. By the compactness of $\widetilde{\mathcal{W}}$, the continuity of T and the lower semi-continuity of I, $\widetilde{F}^*$ is a non-empty compact set. Let G_n be a random graph on n vertices drawn from the exponential random graph model defined by T. The following theorem shows that for n large, $\widetilde{G}_n$ must lie close to $\widetilde{F}^*$ with high probability. In particular, if $\widetilde{F}^*$ is a singleton set, then the theorem gives a weak law of large numbers for G_n.

Theorem 3.2. *Let $\widetilde{F}^*$ and G_n be defined as the above paragraph. Then for any $\eta > 0$ there exist $C, \delta > 0$ such that for all n,*

$$\mathbb{P}(\delta_\square(\widetilde{G}_n, \widetilde{F}^*) > \eta) \leq C e^{-n^2\delta}.$$

Proof. Take any $\eta > 0$. Let

$$\widetilde{A} := \{\widetilde{h} : \delta_\square(\widetilde{h}, \widetilde{F}^*) \geq \eta\}.$$

It is easy to see that $\widetilde{A}$ is a closed set. By compactness of $\widetilde{\mathcal{W}}$ and $\widetilde{F}^*$, and upper semi-continuity of $T - I$, it follows that

$$2\delta := \sup_{\widetilde{h} \in \widetilde{\mathcal{W}}} (T(\widetilde{h}) - I(\widetilde{h})) - \sup_{\widetilde{h} \in \widetilde{A}} (T(\widetilde{h}) - I(\widetilde{h})) > 0.$$

Choose $\epsilon = \delta$ and define $\widetilde{F}^a$ and R as in the proof of Theorem 3.1. Let $\widetilde{A}^a := \widetilde{A} \cap \widetilde{F}^a$. Then

$$\mathbb{P}(G_n \in \widetilde{A}) \leq e^{-n^2\psi_n} \sum_{a \in R} e^{n^2(a+\epsilon)} |\widetilde{A}^a_n| \leq e^{-n^2\psi_n} |R| \sup_{a \in R} e^{n^2(a+\epsilon)} |\widetilde{A}^a_n|.$$

While bounding the last term above, it can be assumed without loss of generality that $\widetilde{A}^a$ is non-empty for each $a \in R$, for the other a's can be dropped without upsetting the bound. By (3.3) and Theorem 3.1 (noting that $\widetilde{A}^a$ is compact), the above display gives

$$\limsup_{n\to\infty} \frac{\log \mathbb{P}(G_n \in \widetilde{A})}{n^2} \le \sup_{a\in R}\bigl(a + \epsilon - \inf_{\widetilde{h}\in\widetilde{A}^a} I(\widetilde{h})\bigr) - \sup_{\widetilde{h}\in\widetilde{\mathcal{W}}} (T(\widetilde{h}) - I(\widetilde{h})).$$

Each $\widetilde{h} \in \widetilde{A}^a$ satisfies $T(\widetilde{h}) \ge a$. Consequently,

$$\sup_{\widetilde{h}\in\widetilde{A}^a} (T(\widetilde{h}) - I(\widetilde{h})) \ge \sup_{\widetilde{h}\in\widetilde{A}^a} (a - I(\widetilde{h})) = a - \inf_{\widetilde{h}\in\widetilde{A}^a} I(\widetilde{h}).$$

Substituting this in the earlier display gives

$$\begin{aligned}\limsup_{n\to\infty} \frac{\log \mathbb{P}(G_n \in \widetilde{A})}{n^2} &\le \epsilon + \sup_{a\in R}\sup_{\widetilde{h}\in\widetilde{A}^a} (T(\widetilde{h}) - I(\widetilde{h})) - \sup_{\widetilde{h}\in\widetilde{\mathcal{W}}} (T(\widetilde{h}) - I(\widetilde{h}))\\ &= \epsilon + \sup_{\widetilde{h}\in\widetilde{A}}(T(\widetilde{h}) - I(\widetilde{h})) - \sup_{\widetilde{h}\in\widetilde{\mathcal{W}}} (T(\widetilde{h}) - I(\widetilde{h})).\\ &= \epsilon - 2\delta = -\delta.\end{aligned}$$

This completes the proof. □

4. An Application

Let $H_1, \ldots, H_k$ be finite simple graphs, where H_1 is the complete graph on two vertices (i.e. just a single edge), and each H_i contains at least one edge. Let $\beta_1, \ldots, \beta_k$ be k real numbers. For any $h \in \mathcal{W}$, let

$$T(h) := \sum_{i=1}^{k} \beta_i t(H_i, h) \tag{4.1}$$

where $t(H_i, h)$ is the homomorphism density of H_i in h, defined in (2.10). Note that there is nothing special about taking H_1 to be a single edge; if we do not want H_1 in our sufficient statistic, we just take $\beta_1 = 0$; all theorems would remain valid.

As remarked in Section 2.3, T is continuous with respect to the cut distance on $\mathcal{W}$, and hence admits a natural definition on $\widetilde{\mathcal{W}}$. Note that for any finite simple graph G that has at least as many nodes as the largest of the H_i's,

$$T(\widetilde{G}) = \sum_{i=1}^{k} \beta_i t(H_i, G).$$

For example, if $k = 2$, and H_2 is a triangle, and G has at least 3 nodes, then

$$T(\widetilde{G}) = \frac{2\beta_1(\#\text{edges in } G)}{n^2} + \frac{6\beta_2(\#\text{triangles in } G)}{n^3}.$$

Let ψ_n be as in (3.1), and let G_n be the n-vertex exponential random graph with sufficient statistic T. Theorem 3.1 gives a formula for $\lim_{n\to\infty} \psi_n$ as the solution of a variational problem. Surprisingly the variational problem is explicitly solvable if $\beta_2, \ldots, \beta_k$ are non-negative.

THEOREM 4.1. *Let T, ψ_n and $H_1, \ldots, H_k$ be as above. Suppose $\beta_2, \ldots, \beta_k$ are non-negative. Then*

$$\lim_{n\to\infty} \psi_n = \sup_{0\le u\le 1} \left(\sum_{i=1}^k \beta_i u^{e(H_i)} - I(u)\right) \tag{4.2}$$

where $I(u) = \frac{1}{2}u\log u + \frac{1}{2}(1-u)\log(1-u)$ and $e(H_i)$ is the number of edges in H_i. Moreover, each solution of the variational problem of Theorem 3.1 *for this T is a constant function, where the constant solves the scalar maximization problem* (4.2).

Proof. By Theorem 3.1,

$$\lim_{n\to\infty} \psi_n = \sup_{h\in\mathcal{W}} (T(h) - I(h)). \tag{4.3}$$

By Hölder's inequality,

$$t(H_i, h) \le \iint_{[0,1]^2} h(x,y)^{e(H_i)}\, dxdy.$$

Thus, by the non-negativity of $\beta_2, \ldots, \beta_k$,

$$\begin{aligned} T(h) &\le \beta_1 t(H_1, h) + \sum_{i=2}^k \beta_i \iint_{[0,1]^2} h(x,y)^{e(H_i)}\, dxdy \\ &= \iint_{[0,1]^2} \sum_{i=1}^k \beta_i h(x,y)^{e(H_i)}\, dxdy. \end{aligned}$$

On the other hand, the inequality in the above display becomes an equality if h is a constant function. Therefore, if u is a point in $[0,1]$ that maximizes

$$\sum_{i=1}^k \beta_i u^{e(H_i)} - I(u),$$

then the constant function $h(x,y) \equiv u$ solves the variational problem (4.3). To see that constant functions are the only solutions, assume that there is at least one i such that the graph H_i has at least one vertex with two or more neighbors. The above steps show that if h is a maximizer, then for each i,

$$t(H_i, h) = \iint_{[0,1]^2} h(x,y)^{e(H_i)}\, dxdy. \tag{4.4}$$

In other words, equality holds in Hölder's inequality. By the assumed condition and the criterion for equality in Hölder's inequality, it follows that $h(x,y) = h(y,z)$ for almost every (x,y,z). From this one can easily conclude that h is almost everywhere a constant function.

If the condition does not hold, then each H_i is a union of vertex-disjoint edges. Assume that some H_i has more than one edge. Then again by (4.4) it follows that h must be a constant function.

Finally, if each H_i is just a single edge, then the maximization problem (4.3) can be explicitly solved and the solutions are all constant functions. $\square$

Theorem 4.1 gives the limiting value of ψ_n if $\beta_2, \ldots, \beta_k$ are non-negative. The next theorem describes the behavior of the exponential random graph G_n under this condition if n is large.

THEOREM 4.2. *For each n, let G_n be an n-vertex exponential random graph with sufficient statistic T defined in* (4.1). *Assume that $\beta_2, \ldots, \beta_k$ are non-negative. Then:*

(a) If the maximization problem in (4.2) *is solved at a unique value u^*, then G_n is indistinguishable from the Erdős–Rényi graph $G(n, u^*)$ in the large n limit, in the sense that $\widetilde{G}_n$ converges to the constant function u^* in probability as $n \to \infty$.*

(b) Even if the maximizer is not unique, the set U of maximizers is a finite subset of $[0, 1]$ *and*

$$\min_{u \in U} \delta_\square(\widetilde{G}_n, \widetilde{u}) \to 0 \quad \textit{in probability as } n \to \infty$$

where $\widetilde{u}$ denotes the image of the constant function u in $\widetilde{\mathcal{W}}$. In other words, G_n behaves like an Erdős–Rényi graph $G(n, u)$ where u is picked randomly from some probability distribution on U.

Proof. The assertions about graph limits in this theorem are direct consequences of Theorems 3.2 and 4.1. Since $\sum_{i=1}^k \beta_i u^{e(H_i)}$ is a polynomial function of u and $I(u)$ is sufficiently well-behaved, showing that U is a finite set is a simple analytical exercise. $\square$

It may be noted here that the conclusion of Theorem 4.2 was proved earlier by Bhamidi et al. [6] under certain restrictions on the parameters that they called a 'high temperature condition'. An important observation from [6] is that when $\beta_2, \ldots, \beta_k$ are non-negative, the model satisfies the so-called FKG property [23]. The FKG property has important consequences; for instance, it implies that the expected value of $t(H_i, G)$ is an increasing function of β_j for any i and j. We will see some further consequences of the FKG property in our proof of Theorem 5.1 in the next section.

5. PHASE TRANSITIONS AND NEAR-DEGENERACY

To illustrate the results of the previous section, recall the exponential random graph model (1.2) with edges and triangles as sufficient statistics:

$$T(\widetilde{G}) = 2\beta_1 \frac{\#\text{edges in } G}{n^2} + 6\beta_2 \frac{\#\text{triangles in } G}{n^3}. \tag{5.1}$$

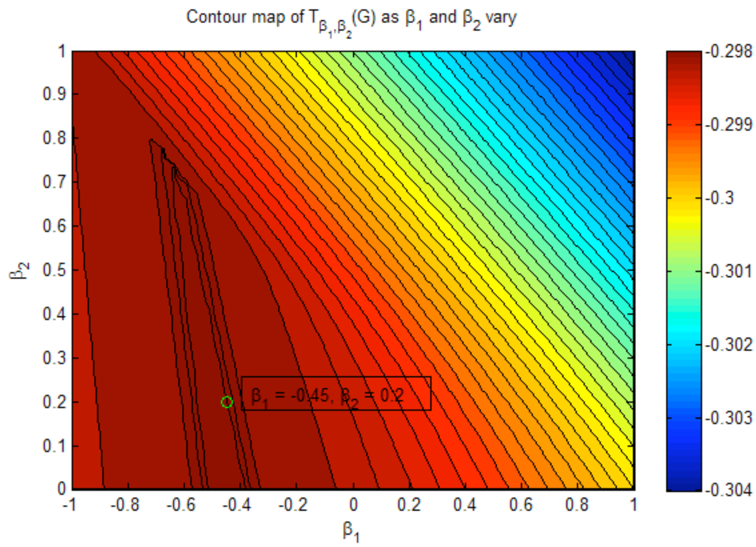


FIGURE 3. The contour plot of $\top_{\beta_1,\beta_2}(G)$. Here G is chosen from the distribution given by $\beta_1 = -0.45,\ \beta_2 = 0.2$. Given sample G it is most likely to be chosen from distributions given by parameters not too far from the original parameters β_1, β_2; this indicates that our approximation for p_{β_1,β_2} is good even when $n = 30$. (Picture by Sukhada Fadnavis.)

Let G_n be an n-vertex exponential random graph with sufficient statistic T. By Theorem 3.1, the probability mass function for this model can be approximated by $\hat{p}_{\beta_1,\beta_2}(G) = \exp(n^2 \top_{\beta_1,\beta_2}(G))$ with

$$\top_{\beta_1,\beta_2}(G) := \inf_{0\le u\le 1} \top_{\beta_1,\beta_2,G}(u),$$

where

$$\begin{aligned}\top_{\beta_1,\beta_2,G}(u) := 2\beta_1 \frac{\#\text{edges in } G}{n^2} + 6\beta_2 \frac{\#\text{triangles in } G}{n^3} \\ - \beta_1 u - \beta_2 u^3 + \frac{1}{2} u \log u + \frac{1}{2}(1-u)\log(1-u).\end{aligned}$$

The figures below have $n = 30$ and graphs are sampled from p_{β_1,β_2} using Glauber dynamics run for 10,000 steps. Figure 3 and Figure 4 show contour plots of $\top_{\beta_1,\beta_2}(G)$ as β_1 and β_2 vary, fixing a realization of G. Figure 5 and Figure 6 illustrate the behavior of $\top_{\beta_1,\beta_2,G}(u)$ as u varies. The captions explain the details.

Now fix β_1 and β_2 and let

$$\ell(u) := \beta_1 u + \beta_2 u^3 - I(u) \tag{5.2}$$

where $I(u) = \frac{1}{2} u \log u + \frac{1}{2}(1-u)\log(1-u)$, as usual. Let U be the set of maximizers of $\ell(u)$ in $[0,1]$. Theorem 4.2 describes the limiting behavior of G_n in terms of the set U. In particular, if U consists of a single point

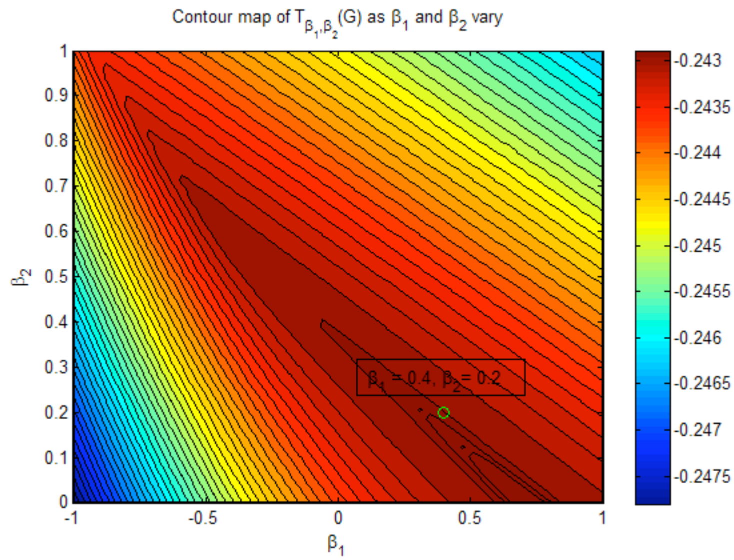


FIGURE 4. The contour plot of $\top_{\beta_1,\beta_2}(G)$. Here G is chosen from the distribution given by $\beta_1 = 0.4$, $\beta_2 = 0.2$. Again, given sample G it is most likely to be chosen from distributions given by parameters not too far from the original parameters β_1, β_2. (Picture by Sukhada Fadnavis.)

$u^* = u^*(\beta_1, \beta_2)$, then G_n behaves like the Erdős–Rényi graph $G(n, u^*)$ when n is large.

It is likely that $u^*(\beta_1, \beta_2)$ does not have a closed form expression, other than when $\beta_2 = 0$, in which case

$$u^*(\beta_1, 0) = \frac{e^{\beta_1}}{1 + e^{\beta_1}}.$$

It is, however, quite easy to numerically approximate $u^*(\beta_1, \beta_2)$. Figure 7 plots $u^*(\beta_1, \beta_2)$ versus β_2 for four different fixed values of β_1, namely, $\beta_1 = 0.2, -0.35, -0.45$, and -0.8. The figures show that u^* is a continuous function of β_2 as long as β_1 is not too far down the negative axis.

But for β_1 below a threshold (e.g., when $\beta_1 = -0.45$), u^* shows a single jump discontinuity in β_2, signifying a phase transition. In physical terms, this is a first order phase transition, by the following logic. By Theorem 4.2, our random graph behaves like $G(n, u^*)$ when n is large. On the other hand, by a standard computation the expect number of triangles is the first derivative of the free energy ψ_n with respect to β_2. Therefore in the large n limit, a discontinuity in u^* as a function of β_2 signifies a discontinuity in the derivative of the limiting free energy, which is the physical definition of a first order phase transition.

At the point of discontinuity, $\ell(u)$ is maximized at two values of u, i.e., the set U consists of two points. Lastly, as β_1 goes down the negative axis, the model starts to exhibit "near-degeneracy" in the sense of Handcock [31] (see also [45]) as seen in the last frame of Figure 7. This means that as β_2 varies, the model transitions from being a very sparse graph for low values of β_2, to a very dense (nearly complete) graph for large values of β_2, completely skipping all intermediate structures.

The following theorem gives a simple mathematical description of this phenomenon and hence the first rigorous proof of the degeneracy observed in exponential graph models. Related results are in Häggstrom and Jonasson [29].

THEOREM 5.1. *Let G_n be an exponential random graph with sufficient statistic T defined in* (5.1). *Fix any $\beta_1 < 0$. Let*

$$c_1 := \frac{e^{\beta_1}}{1+e^{\beta_1}}, \quad c_2 := 1 + \frac{1}{2\beta_1}.$$

Suppose $|\beta_1|$ is so large that $c_1 < c_2$. Let $e(G_n)$ be the number of edges in G_n and let $f(G_n) := e(G_n)/\binom{n}{2}$ be the edge density. Then there exists $q = q(\beta_1) \in [0,\infty)$ such that if $-\infty < \beta_2 < q$, then

$$\lim_{n\to\infty} \mathbb{P}(f(G_n) > c_1) = 0,$$

and if $\beta_2 > q$, then

$$\lim_{n\to\infty} \mathbb{P}(f(G_n) < c_2) = 0.$$

In other words, if β_1 is a large negative number, then G_n is either sparse (if $\beta_2 < q$) or nearly complete (if $\beta_2 > q$).

Remark. The difference in the values of c_1 and c_2 can be quite striking even for relatively small values of β_1. For example, $\beta_1 = -5$ gives $c_1 \simeq 0.007$ and $c_2 = 0.9$.

Proof. Fix $\beta_1 < 0$ such that $c_1 < c_2$. As a preliminary step, let us prove that for any $\beta_2 > 0$,

$$\lim_{n\to\infty} \mathbb{P}(f(G_n) \in (c_1, c_2)) = 0. \tag{5.3}$$

Fix $\beta_2 > 0$. Let u be any maximizer of ℓ. Then by Theorem 4.2, it suffices to prove that either $u \le e^{\beta_1}/(1+e^{\beta_1})$ or $u \ge 1 + 1/2\beta_1$. This is proved as follows. Define a function $g : [0,1] \to \mathbb{R}$ as

$$g(v) := \ell(v^{1/3}).$$

Then ℓ is maximized at u if and only if g is maximized at u^3. Since ℓ is a bounded continuous function and $\ell'(0) = \infty$, $\ell'(1) = -\infty$, ℓ cannot be maximized at 0 or 1. Therefore the same is true for g. Let v be a point in $(0,1)$ at which g is maximized. Then $g''(v) \le 0$. A simple computation shows that

$$g''(v) = \frac{1}{9v^{5/3}}\biggl(-2\beta_1 + \log\frac{v^{1/3}}{1-v^{1/3}} - \frac{1}{2(1-v^{1/3})}\biggr).$$

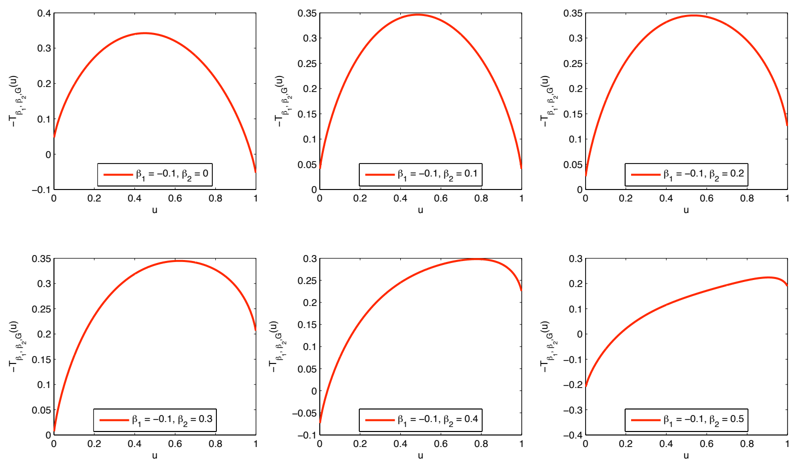


FIGURE 5. The plot of $-\top_{\beta_1,\beta_2,G}(u)$ versus u when β_1 is fixed at -0.1. For this choice of β_1 there is no phase transition and $\top$ has a unique maximum always. (Picture by Sukhada Fadnavis.)

Thus, $g''(v) \le 0$ only if

$$\log \frac{v^{1/3}}{1-v^{1/3}} \le \beta_1 \quad \text{or} \quad -\frac{1}{2(1-v^{1/3})} \le \beta_1.$$

This shows that $u \in (0,1)$ can be a maximizer of ℓ only if

$$u \le \frac{e^{\beta_1}}{1+e^{\beta_1}} \quad \text{or} \quad u \ge 1 + \frac{1}{2\beta_1}.$$

By Theorem 3.2, this completes the proof of (5.3) when $\beta_2 > 0$.

Now notice that as $\beta_2 \to \infty$, $\sup_{u \le a} \ell(u) \sim \beta_2 a^3$ for any fixed $a \le 1$. This shows that as $\beta_2 \to \infty$, any maximizer of ℓ must eventually be larger than $1 + 1/2\beta_1$. Therefore, for sufficiently large β_2,

$$\lim_{n\to\infty} \mathbb{P}(f(G_n) < c_2) = 0. \tag{5.4}$$

Next consider the case $\beta_2 \le 0$. Let $\widetilde{F}^*$ be the set of maximizers of $T(\widetilde{h}) - I(\widetilde{h})$. Take any $\widetilde{h} \in \widetilde{F}^*$ and let h be a representative element of $\widetilde{h}$. Let $p = c_1$. An easy verification shows that

$$T(h) - I(h) = \beta_2 t(H_2, h) - I_p(h),$$

where $I_p(h)$ is defined as in (2.15). Define a new function

$$h_1(x,y) := \min\{h(x,y),\ p\}.$$

Since the function I_p defined in (2.14) is minimized at p, it follows that for all $x, y \in [0,1]$, $I_p(h_1(x,y)) \le I_p(h(x,y))$. Consequently, $I_p(h_1) \le I_p(h)$.

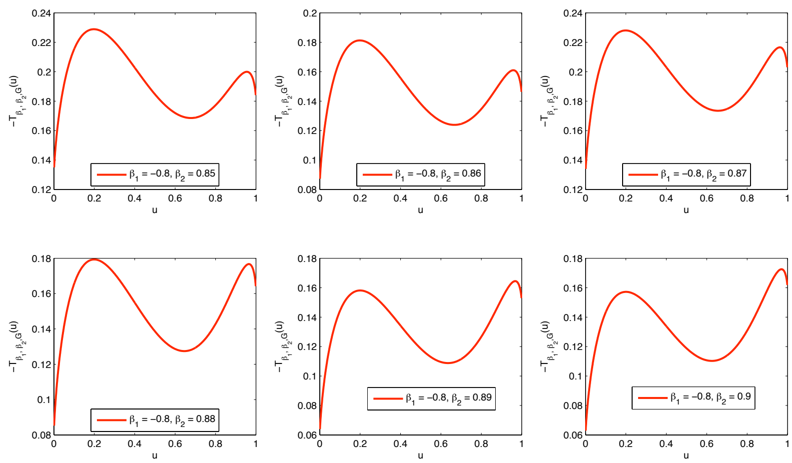


FIGURE 6. The plot of $-\top_{\beta_1,\beta_2,G}(u)$ versus u when β_1 is fixed at -0.8. For this choice of β_1 there is a phase transition and $-\top$ has two local maxima always. The left one starts as the global maxima; they become equal at phase transition, then the right maxima becomes the global maximum. This is the jump in the value of u^* observed in Figure 1. (Picture by Sukhada Fadnavis.)

Again, since $\beta_2 \leq 0$ and $h_1 \leq h$ everywhere, $\beta_2 t(H_2, h_1) \geq \beta_2 t(H_2, h)$. Combining these observations, we see that $T(h_1) - I(h_1) \geq T(h) - I(h)$. Since h maximizes $T - I$ it follows that equality must hold at every step in the above deductions, from which it is easy to conclude that $h = h_1$ a.e. In other words, $h(x, y) \leq p$ a.e. This is true for every $\widetilde{h} \in \widetilde{F}^*$. Thus, Theorem 3.2 proves that when $\beta_2 \leq 0$,

$$\lim_{n\to\infty} \mathbb{P}(f(G_n) > c_1) = 0. \tag{5.5}$$

Recalling that β_1 is fixed, define

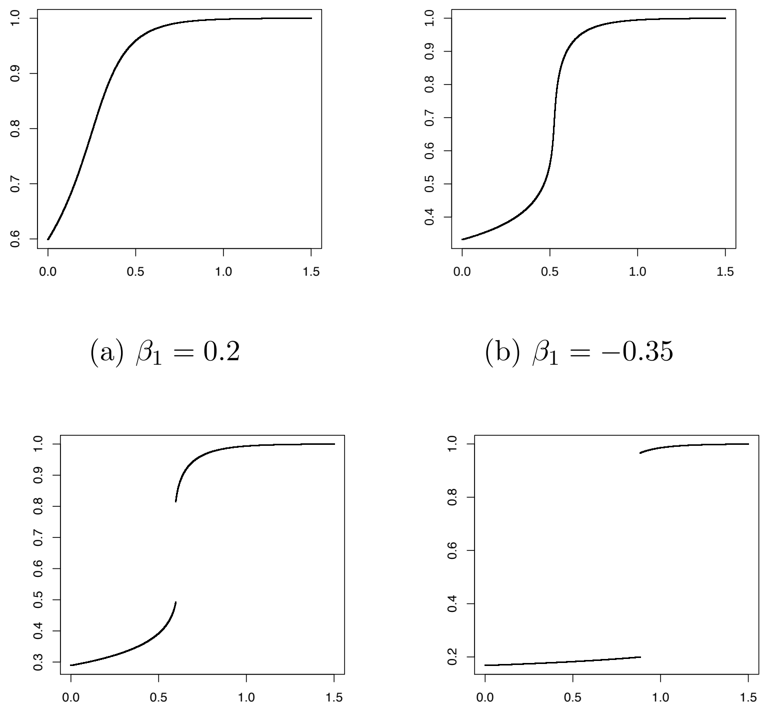
$$a_n(\beta_2) := \mathbb{P}(f(G_n) > c_1), \quad b_n(\beta_2) := \mathbb{P}(f(G_n) < c_2).$$

Let A_n and B_n denote the events in brackets in the above display. A simple computation shows that

$$a_n'(\beta_2) = \frac{6}{n}\mathrm{Cov}(1_{A_n}, \Delta(G_n)) \quad \text{and} \quad b_n'(\beta_2) = \frac{6}{n}\mathrm{Cov}(1_{B_n}, \Delta(G_n)),$$

where $\Delta(G_n)$ is the number of triangles in G_n. It is easy to see that the exponential random graph model with $\beta_2 \geq 0$ satisfies the FKG criterion [23]. Therefore the above identities show that on the non-negative axis, a_n is a non-decreasing function and b_n is a non-increasing function.

Let $q_1 := \sup\{x \in \mathbb{R} : \lim_{n\to\infty} a_n(x) = 0\}$. By equation (5.4), $q_1 < \infty$ and by equation (5.5) $q_1 \geq 0$. Similarly, if $q_2 := \inf\{x \in \mathbb{R} : \lim_{n\to\infty} b_n(x) = 0\}$,



(a) $\beta_1 = 0.2$ (b) $\beta_1 = -0.35$

(c) $\beta_1 = -0.45$ (d) $\beta_1 = -0.8$

FIGURE 7. Plot of $u^*(\beta_1, \beta_2)$ on y-axis vs β_2 on x-axis for different fixed values of β_1. Part (c) demonstrates a phase transition. Part (d) demonstrates near-degeneracy.

then $0 \le q_2 < \infty$. Also, clearly, $q_1 \le q_2$ since $a_n + b_n \ge 1$ everywhere. We claim that $q_1 = q_2$. This would complete the proof by the monotonicity of a_n and b_n.

To prove that $q_1 = q_2$, suppose not. Then $q_1 < q_2$. Then for any $\beta_2 \in (q_1, q_2)$, $\limsup a_n(\beta_2) > 0$ and $\limsup b_n(\beta_2) > 0$. A simple probability argument shows that

$$0 \le a_n(\beta_2) + b_n(\beta_2) - 1 \le \mathbb{P}(f(G_n) \in (c_1, c_2)).$$

Therefore by (5.3),

$$\lim_{n\to\infty} (a_n(\beta_2) + b_n(\beta_2) - 1) = 0.$$

Consequently,

$$\begin{aligned}
&\liminf_{n\to\infty} \int_{q_1}^{q_2} (1-a_n(\beta_2))(1-b_n(\beta_2))d\beta_2 \\
&\ge \int_{q_1}^{q_2} \liminf_{n\to\infty} (1-a_n(\beta_2))(1-b_n(\beta_2))d\beta_2 \\
&= \int_{q_1}^{q_2} \liminf_{n\to\infty} b_n(\beta_2)a_n(\beta_2)d\beta_2 > 0.
\end{aligned}$$

If G_n' is an independent copy of G_n, then a moment's thought shows that

$$\begin{aligned}
\mathrm{var}(f(G_n)) &= \frac{1}{2}\mathbb{E}(f(G_n) - f(G_n'))^2 \\
&\ge \frac{1}{2}(c_2-c_1)^2(1-a_n(\beta_2))(1-b_n(\beta_2)).
\end{aligned}$$

On the other hand, a simple computation gives

$$\frac{\partial^2 \psi_n}{\partial \beta_1^2} = \frac{4}{n^2}\,\mathrm{var}(e(G_n)) = (n-1)^2\,\mathrm{var}(f(G_n)).$$

Combining the last three displays gives

$$\begin{aligned}
&\liminf_{n\to\infty}\biggl(\frac{\partial \psi_n}{\partial \beta_1}(\beta_1, q_2) - \frac{\partial \psi_n}{\partial \beta_1}(\beta_1, q_1)\biggr) \\
&= \liminf_{n\to\infty} \int_{q_1}^{q_2} \frac{\partial^2 \psi_n}{\partial \beta_1^2}(\beta_1,\beta_2)d\beta_2 = \infty.
\end{aligned}$$

However, this is impossible, since for all (β_1, β_2),

$$\frac{\partial \psi_n}{\partial \beta_1} = \frac{2}{n^2}\mathbb{E}(e(G_n)) \le 1.$$

This completes the proof. □

6. The Symmetric Phase, Symmetry Breaking, and the Euler–Lagrange Equations

Borrowing terminology from spin glasses, we define the *replica symmetric phase* or simply the *symmetric phase* of a variational problem like maximizing $T(h) - I(h)$ as the set of parameter values for which all the maximizers are constant functions. When the parameters are such that all maximizers are non-constant functions we say that the parameter vector is in the region of broken replica symmetry, or simply broken symmetry. There may be another situation, where some optimizers are constant while others are non-constant, although we do not know of such examples. (This third region may be called a region of partial symmetry.)

Statistically, the exponential random graph behaves like an Erdős–Rényi graph in the symmetric region of the parameter space, while such behavior breaks down in the region of broken symmetry. This follows easily from Theorem 3.2.

Theorem 4.2 shows that for the sufficient statistic T defined in (4.1), every $(\beta_1, \beta_2, \ldots, \beta_k) \in \mathbb{R} \times \mathbb{R}_+^{k-1}$ falls in the replica symmetric region. Does symmetry hold only when $\beta_2, \ldots, \beta_k$ are non-negative? The following theorem (proven with the aid of the Euler–Lagrange equations of Theorem 6.3 below), shows that this is not the case; $(\beta_1, \ldots, \beta_k)$ is in the replica symmetric region whenever $|\beta_2|, \ldots, |\beta_k|$ are small enough. Of course, this does not supersede Theorem 4.2 since it does not cover large positive values of $\beta_2, \ldots, \beta_k$. However, it proves replica symmetry for small negative values of $\beta_2, \ldots, \beta_k$, which is not covered by Theorem 4.2.

THEOREM 6.1. *Consider the exponential random graph with sufficient statistic T defined in* (4.1). *Suppose $\beta_1, \ldots, \beta_k$ are such that*

$$\sum_{i=2}^{k} |\beta_i| e(H_i)(e(H_i) - 1) < 2$$

where $e(H_i)$ is the number of edges in H_i. Then the conclusions of Theorems 4.1 *and* 4.2 *hold true for this value of the parameter vector $(\beta_1, \ldots, \beta_k)$.*

Proof. It suffices to prove that the maximizer of $T(h) - I(h)$ as h varies over $\mathcal{W}$ is unique. This is because: if h is a maximizer, then so is $h_\sigma(x, y) := h(\sigma x, \sigma y)$ for any measure preserving bijection $\sigma : [0, 1] \to [0, 1]$. The only functions that are invariant under such transforms are constant functions.

Let $\|\cdot\|_\infty$ denote the L^∞ norm on $\mathcal{W}$ (that is, the essential supremum of the absolute value). Let h and g be two maximizers of $T - I$. For any finite simple graph H, a simple computation shows that

$$\begin{aligned}\|\Delta_H h - \Delta_H g\|_\infty &\le \sum_{(r,s)\in E(H)} \|\Delta_{H,r,s} h - \Delta_{H,r,s} g\|_\infty \\ &\le e(H)(e(H) - 1)\|h - g\|_\infty.\end{aligned}$$

Using the above inequality, Theorem 6.3 and the inequality

$$\left| \frac{e^x}{1 + e^x} - \frac{e^y}{1 + e^y} \right| \le \frac{|x - y|}{4}$$

(easily proved by the mean value theorem) it follows that for almost all x, y,

$$\begin{aligned}|h(x, y) - g(x, y)| &= \left| \frac{e^{2\sum_{i=1}^k \beta_i \Delta_{H_i} h(x,y)}}{1 + e^{2\sum_{i=1}^k \beta_i \Delta_{H_i} h(x,y)}} - \frac{e^{2\sum_{i=1}^k \beta_i \Delta_{H_i} g(x,y)}}{1 + e^{2\sum_{i=1}^k \beta_i \Delta_{H_i} g(x,y)}} \right| \\ &\le \frac{1}{2} \sum_{i=1}^{k} |\beta_i| \|\Delta_{H_i} h - \Delta_{H_i} g\|_\infty \\ &\le \frac{1}{2} \|h - g\|_\infty \sum_{i=1}^{k} |\beta_i| e(H_i)(e(H_i) - 1).\end{aligned}$$

If the coefficient of $\|h - g\|_\infty$ in the last expression is strictly less than 1, it follows that h must be equal to g a.e. $\square$

6.1. **Symmetry breaking.** Theorems 4.2 and 6.1 establish various regions of symmetry in the exponential random graph model with sufficient statistic T defined in (4.1). That leaves the question: is there a region where symmetry breaks? We specialize to the simple case where $k = 2$ and H_2 is a triangle, i.e., the example of Section 5. In this case, it turns out that replica symmetry breaks whenever β_2 is less than a sufficiently large negative number depending on β_1.

THEOREM 6.2. *Consider the exponential random graph with sufficient statistic T defined in* (5.1)*. Then for any given value of β_1, there is a positive constant $C(\beta_1)$ sufficiently large so that whenever $\beta_2 < -C(\beta_1)$, $T(h) - I(h)$ is not maximized at any constant function. Consequently, if G_n is an n-vertex exponential random graph with this sufficient statistic, then there exists $\epsilon > 0$ such that*

$$\lim_{n\to\infty} \mathbb{P}\left(\delta_\square(\widetilde{G}_n, \widetilde{C}) > \epsilon\right) = 1$$

where $\widetilde{C}$ is the set of constant functions. In other words, G_n does not look like an Erdős–Rényi graph in the large n limit.

Proof. Fix β_1. Let $p = e^{\beta_1}/(1 + e^{\beta_1})$ and $\gamma := -\beta_2$, so that for any $h \in \mathcal{W}$,

$$T(h) - I(h) = -\gamma t(H_2, h) - I_p(h).$$

Assume without loss of generality that $\beta_2 < 0$. Suppose u is a constant such that the function $h(x, y) \equiv u$ maximizes $T(h) - I(h)$, i.e., minimizes $\gamma t(H_2, h) + I_p(h)$. Note that

$$\gamma t(H_2, h) + I_p(h) = \gamma u^3 + I_p(u).$$

Clearly, the definition of u implies that $\gamma u^3 + I_p(u) \le \gamma x^3 + I_p(x)$ for all $x \in [0, 1]$. This implies that u must be in $(0, 1)$, because the derivative of $x \mapsto \gamma x^3 + I_p(x)$ is $-\infty$ at 0 and ∞ at 1. Thus,

$$0 = \frac{d}{dx}(\gamma x^3 + I_p(x))\bigg|_{x=u} = 3\gamma u^2 + \frac{1}{2}\log\frac{u}{1-u}$$

which shows that $u \le c(\gamma)$, where $c(\gamma)$ is a function of γ such that

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

This shows that

$$\lim_{\gamma\to\infty} \min_{0\le x\le 1} \left(\gamma x^3 + I_p(x)\right) = I_p(0) = \frac{1}{2}\log\frac{1}{1-p}. \tag{6.1}$$

Next let g be the function

$$g(x, y) := \begin{cases} 0 & \text{if } x, y \text{ on same side of } 1/2 \\ p & \text{if not.} \end{cases}$$

Clearly, for almost all (x, y, z), $g(x, y)g(y, z)g(z, x) = 0$. Thus, $t(H_2, g) = 0$. A simple computation shows that

$$I_p(g) = \frac{1}{4} \log \frac{1}{1-p}.$$

Thus, $\gamma t(H_2, g) + I_p(g) = \frac{1}{4} \log \frac{1}{1-p}$. This shows that if γ is large enough (depending on p and hence β_1), then $T - I$ cannot be maximized at a constant function. The rest of conclusion follows easily from Theorem 3.2 and the compactness of $\widetilde{\mathcal{W}}$. □

6.2. **Euler–Lagrange equations.** We return to the exponential random graph model with sufficient statistic T defined in (4.1) in terms of the densities of k fixed graphs $H_1, \ldots, H_k$, where H_1 is a single edge. Theorems 4.1 and 4.2 analyze this model when $\beta_2, \ldots, \beta_k$ are non-negative. What if they are not? One can still try to derive the Euler–Lagrange equations for the related variational problem of maximizing $T(\widetilde{h}) - I(\widetilde{h})$. The following theorem presents the outcome of this effort.

For a finite simple graph H, let $V(H)$ and $E(H)$ denote the sets of vertices and edges of H. Given a symmetric measurable function $h : [0, 1]^2 \to \mathbb{R}$, for each $(r, s) \in E(H)$ and each pair of points $x_r, x_s \in [0, 1]$, define

$$\Delta_{H,r,s} h(x_r, x_s) := \int_{[0,1]^{V(H)\setminus\{r,s\}}} \prod_{\substack{(r',s')\in E(H)\\(r',s')\neq(r,s)}} h(x_{r'}, x_{s'}) \prod_{\substack{v\in V(H)\\ v\neq r,s}} dx_v.$$

For $x, y \in [0, 1]$ define

$$\Delta_H h(x, y) := \sum_{(r,s)\in E(H)} \Delta_{H,r,s} h(x, y). \tag{6.2}$$

For example, when H is a triangle, then $V(H) = \{1, 2, 3\}$ and

$$\Delta_{H,1,2} h(x, y) = \Delta_{H,1,3} h(x, y) = \Delta_{H,2,3} h(x, y) = \int_0^1 h(x, z)h(y, z)\, dz$$

and therefore $\Delta_H h(x, y) = 3 \int_0^1 h(x, z)h(y, z)dz$. When H contains exactly one edge, define $\Delta_H h \equiv 1$ for any h, by the usual convention that the empty product is 1.

Theorem 6.3. *Let $T : \widetilde{\mathcal{W}} \to \mathbb{R}$ be defined as in* (4.1) *and the operator Δ_H be defined as in* (6.2)*. If $\widetilde{h} \in \widetilde{\mathcal{W}}$ maximizes $T(\widetilde{h}) - I(\widetilde{h})$, then any representative element $h \in \widetilde{h}$ must satisfy for almost all $(x, y) \in [0, 1]^2$,*

$$h(x, y) = \frac{e^{2\sum_{i=1}^k \beta_i \Delta_{H_i} h(x,y)}}{1 + e^{2\sum_{i=1}^k \beta_i \Delta_{H_i} h(x,y)}}.$$

Moreover, any maximizing function must be bounded away from 0 *and* 1.

Proof. Let g be a symmetric bounded measurable function from $[0,1]$ into $\mathbb{R}$. For each $u \in \mathbb{R}$, let

$$h_u(x,y) := h(x,y) + ug(x,y).$$

Then h_u is a symmetric bounded measure function from $[0,1]$ into $\mathbb{R}$. First suppose that h is bounded away from 0 and 1. Then $h_u \in \mathcal{W}$ for every u sufficiently small in magnitude. Since h maximizes $T(h) - I(h)$ among all elements of $\mathcal{W}$, therefore under the above assumption, for all u sufficiently close to zero,

$$T(h_u) - I(h_u) \le T(h) - I(h).$$

In particular,

$$\frac{d}{du}(T(h_u) - I(h_u))\bigg|_{u=0} = 0. \tag{6.3}$$

It is easy to check that $T(h_u) - I(h_u)$ is differentiable in u for any h and g. In particular, the derivative is given by

$$\frac{d}{du}(T(h_u) - I(h_u)) = \sum_{i=1}^{k} \beta_i \frac{d}{du} t(H_i, h_u) - \frac{d}{du} I(h_u).$$

Now,

$$\begin{aligned}\frac{d}{du} I(h_u) &= \iint \frac{d}{du} I(h(x,y) + ug(x,y))\, dydx \\ &= \frac{1}{2} \iint g(x,y) \log \frac{h_u(x,y)}{1 - h_u(x,y)}\, dydx.\end{aligned}$$

Consequently,

$$\frac{d}{du} I(h_u)\bigg|_{u=0} = \frac{1}{2} \iint g(x,y) \log \frac{h(x,y)}{1 - h(x,y)}\, dydx.$$

Next, note that

$$\begin{aligned}&\frac{d}{du} t(H_i, h_u) \\ &\quad= \int_{[0,1]^{V(H)}} \sum_{(r,s)\in E(H_i)} g(x_r, x_s) \prod_{\substack{(r',s')\in E(H_i) \\ (r',s')\neq(r,s)}} h_u(x_{r'}, x_{s'}) \prod_{v\in V(H)} dx_v \\ &\quad= \iint g(x,y) \Delta_{H_i} h_u(x,y)\, dydx.\end{aligned}$$

Combining the above computations and (6.3), we see that for any symmetric bounded measurable $g : [0,1] \to \mathbb{R}$,

$$\iint g(x,y) \left(\sum_{i=1}^{k} \beta_i \Delta_{H_i} h(x,y) - \frac{1}{2} \log \frac{h(x,y)}{1 - h(x,y)} \right) dydx = 0.$$

Taking $g(x,y)$ equal to the function within the brackets (which is bounded since h is assumed to be bounded away from 0 and 1), the conclusion of the theorem follows.

Now note that the theorem was proved under the assumption that h is bounded away from 0 and 1. We claim that this is true for any h that maximizes $T(h)-I(h)$. To prove this claim, take any such h. Fix $p\in(0,1)$. For each $u\in[0,1]$, let

$$h_{p,u}(x,y) := (1-u)h(x,y) + u\max\{h(x,y),p\}.$$

Then certainly, $h_{p,u}$ is a symmetric bounded measurable function from $[0,1]^2$ into $[0,1]$. Note that

$$\frac{d}{du}h_{p,u}(x,y) = \max\{h(x,y),p\} - h(x,y) = (p-h(x,y))_+.$$

Using this, an easy computation as above shows that

$$\begin{aligned}
&\frac{d}{du}\left(T(h_{p,u}) - I(h_{p,u})\right)\bigg|_{u=0}\\
&= \iint \left(\sum_{i=1}^k \beta_i \Delta_{H_i} h(x,y) - \frac{1}{2}\log\frac{h(x,y)}{1-h(x,y)}\right)(p-h(x,y))_+\,dydx\\
&\ge \iint \left(-C - \frac{1}{2}\log\frac{h(x,y)}{1-h(x,y)}\right)(p-h(x,y))_+\,dydx
\end{aligned}$$

where C is a positive constant depending only on $\beta_1,\ldots,\beta_k$ and $H_1,\ldots,H_k$ (and not on p or h). When $h(x,y)=0$, the integrand is interpreted as ∞, and when $h(x,y)=1$, the integrand is interpreted as 0.

Now, if p is so small that

$$-C - \frac{1}{2}\log\frac{p}{1-p} > 0,$$

then the previous display proves that the derivative of $T(h_{p,u})-I(h_{p,u})$ with respect to u is strictly positive at $u=0$ if $h<p$ on a set of positive Lebesgue measure. Hence h cannot be a maximizer of $T-I$ unless $h\ge p$ almost everywhere. This proves that any maximizer of $T-I$ must be bounded away from zero. A similar argument shows that it must be bounded away from 1 and hence completes the proof of the theorem. □

6.3. **A solvable case with negative parameters.** A j-star is an undirected graph with one 'root' vertex and j other vertices connected to the root vertex, with no edges between any of these j vertices. Let H_j be a j-star for $j=1,\ldots,k$. Let T be the sufficient statistic

$$T(G) = \sum_{j=1}^k \beta_j t(H_j,G). \tag{6.4}$$

Theorems 4.1 and 4.2 describe the behavior of this model when $\beta_2,\ldots,\beta_k$ are all non-negative. The following theorem shows that the behavior is the same even if $\beta_2, \ldots, \beta_k$ are all non-positive. This phenomenon for j-star models was first observed in simulations by Sukhada Fadnavis.

THEOREM 6.4. *For the sufficient statistic T defined in* (6.4)*, the conclusions of Theorems* 4.1 *and* 4.2 *hold when $\beta_2, \ldots, \beta_k$ are all non-positive.*

Proof. Since $\beta_2, \ldots, \beta_k \le 0$ and I is a convex function, note that for any $h \in \mathcal{W}$

$$\begin{aligned} T(h) - I(h) &= \beta_1 \int h(x,y)dxdy + \sum_{j=2}^k \beta_j \int \biggl(\int h(x,y)dy\biggr)^j dx \\ &\quad - \int I(h(x,y))dxdy \\ &\le \beta_1 \int h(x,y)dxdy + \sum_{j=2}^k \beta_j \biggl(\iint h(x,y)dydx\biggr)^j \\ &\quad - I\biggl(\int h(x,y)dxdy\biggr) \\ &\le \sup_{0\le u\le 1} (\beta_1 u + \beta_2 u^2 + \cdots + \beta_k u^k - I(u)), \end{aligned}$$

with equality holding in all steps if and only if h is identically equal to a constant that solves the maximization problem in the last step. $\square$

Naturally, the question arises as to whether the conclusions of Theorems 4.1 and 4.2 continue to hold for all values of $\beta_1, \ldots, \beta_k$, even when some of them are positive and some negative. As of now, we do not know the answer.

7. Extremal behavior

In the sections above we have been assuming that $\beta_2, \ldots, \beta_k$ are positive or barely negative. In this section we investigate what happens when $k = 2$ and β_2 is large and negative. The limits are describable but far from Erdős-Rényi. Our work here is inspired by related results of Sukhada Fadvanis who has a different argument (using Turán's theorem) for the case of triangles.

Suppose H is any finite simple graph containing at least one edge. Let T be the sufficient statistic

$$T(\widetilde{G}) = 2\beta_1 \frac{\#\text{edges in } G}{n^2} + \beta_2 t(H, G).$$

Let G_n be the exponential random graph on n vertices with this sufficient statistic and let ψ_n be the associated normalizing constant as defined in (3.1). Then Theorem 3.1 gives

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

where I is defined in (3.2). We also know (by Theorem 3.2 that

$$\delta_\square(\widetilde{G}_n, \widetilde{F}^*) \to 0 \quad \text{in probability as } n \to \infty,$$

where $\widetilde{F}^*$ is the subset of $\widetilde{\mathcal{W}}$ where $T - I$ is maximized. (Note that $\widetilde{F}^*$ is a closed set since $T - I$ is an upper semicontinuos map.)

We can compute $\widetilde{F}^*$ and ψ when β_2 is positive, or negative with small magnitude. We are unable to carry out the explicit computation in the case of large negative β_2, unless H is a convenient object like a j-star. However, a qualitative description can still be given by analyzing the behavior of $\widetilde{F}^*$ and ψ as $\beta_2 \to -\infty$. Fixing β_1, we consider these objects as functions of β_2 and write $\widetilde{F}^*(\beta_2)$, $\psi(\beta_2)$ and T_{β_2} instead of $\widetilde{F}^*$, ψ and T.

Theorem 7.1. *Fixing H and β_1, let $\widetilde{F}^*(\beta_2)$ and $\psi(\beta_2)$ be as above. Let $\chi(H)$ be the chromatic number of H, and define*

$$g(x,y) := \begin{cases} 1 & \text{if } [(\chi(H)-1)x] \neq [(\chi(H)-1)y], \\ 0 & \text{otherwise,} \end{cases} \tag{7.1}$$

where $[x]$ denotes the integer part of a real number x. Let $p = e^{2\beta_1}/(1+e^{2\beta_1})$. Then

$$\lim_{\beta_2 \to -\infty} \sup_{\tilde{f} \in \widetilde{F}^*(\beta_2)} \delta_\square(\tilde{f}, p\tilde{g}) = 0$$

and

$$\lim_{\beta_2 \to -\infty} \psi(\beta_2) = \frac{(\chi(H)-2)}{2(\chi(H)-1)} \log \frac{1}{1-p}.$$

Intuitively, the above result means that if β_2 is a large negative number and n is large, then an exponential random graph G_n with sufficient statistic T looks roughly like a complete $(\chi(H)-1)$-equipartite graph with $1-p$ fraction of edges randomly deleted, where $p = e^{2\beta_1}/(1+e^{2\beta_1})$. In particular, if H is bipartite, then G_n must be very sparse, since a 1-equipartite graph has no edges. Figure 8 gives a simulation result for the triangle model with large negative β_2.

Theorem 7.1 is closely related to the Erdős-Stone theorem from extremal graph theory (or equivalently, Turán's theorem in the case of triangles as in the work of Fadnavis). Indeed, it may be possible to prove some parts of our theorem using the Erdős-Stone theorem, but we prefer the bare-hands argument given below. Due to this connection with extremal graph theory, we refer to behavior of the graph in the 'large negative β_2' domain as *extremal behavior*.

Lemma 7.2. *Let r be any integer $\geq \chi(H)$. Let K_r be the complete graph on r vertices. Then for any symmetric measurable $h : [0,1]^2 \to \{0,1\}$, if $t(K_r, h) > 0$ then $t(H, h) > 0$.*

Proof. Let $h_n(x,y)$ be the average value of h in the dyadic square of width 2^{-n} containing the point (x,y). A standard martingale argument implies

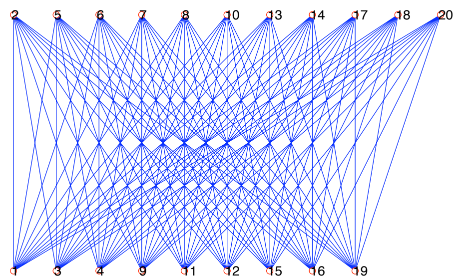


FIGURE 8. A simulated realization of the exponential random graph model on 20 nodes with edges and triangles as sufficient statistics, where $\beta_1 = 120$ and $\beta_2 = -400$. (Picture by Sukhada Fadnavis.)

that the sequence of functions $\{h_n\}_{n\geq 1}$ converges to h almost everywhere. For any positive integer u, let K_r^u denote the complete r-partite graph on ru vertices, where each partition consists of u vertices (so that $K_r^1 = K_r$). Since $r \geq \chi(H)$, it is easy to see that there exists u so large that H is a subgraph of K_r^u. Fix such a u.

By the almost everywhere convergence of h_n to h and the assumption that $t(K_r, h) > 0$, there is a set of r distinct points $x_1, \ldots, x_r \in [0,1]$ such that $h(x_i, x_j) > 0$ and $\lim_{n\to\infty} h_n(x_i, x_j) = h(x_i, x_j)$ for each $1 \leq i \neq j \leq r$. Since h is $\{0,1\}$-valued, $h(x_i, x_j) = 1$ for each $i \neq j$. Choose n so large that for each $i \neq j$,

$$h_n(x_i, x_j) \geq 1 - \epsilon,$$

where $\epsilon = 2/ru$. Let $(X_i^s)_{1\leq i\leq r,\, 1\leq s\leq u}$ be independent random variables, where X_i^s is uniformly distributed in the dyadic interval of width 2^{-n} containing x_i. Then for each $1 \leq i \neq j \leq r$, $1 \leq q, s \leq u$,

$$\mathbb{P}(h(X_i^q, X_j^s) = 1) = h_n(x_i, x_j) \geq 1 - \epsilon.$$

Therefore,

$$\mathbb{P}(h(X_i^q, X_j^s) = 1 \text{ for all } 1 \leq i \neq j \leq r,\ 1 \leq q, s \leq u) \geq 1 - ru\epsilon = 1/2.$$

Let $(Y_i^s)_{1\leq i\leq r,\, 1\leq s\leq u}$ be independent random variables uniformly distributed in $[0,1]$. Conditional on the event that Y_i^s belongs to the dyadic interval of width 2^{-n} containing x_i, Y_i^s has the same distribution as X_i^s. This shows that

$$t(K_r^u, h) \geq \mathbb{P}(h(Y_i^q, Y_j^s) = 1 \text{ for all } 1 \leq i \neq j \leq r,\ 1 \leq q, s \leq u) > 0.$$

Since H is a subgraph of K_r^u, therefore $t(H, h) > 0$. $\square$

THEOREM 7.3. *Let g be the function defined in (7.1). Take any $p \in (0,1)$. If f is any element of $\mathcal{W}$ that minimizes $I_p(f)$ among all f satisfying $t(H,f) = 0$, then $\widetilde{f} = p\widetilde{g}$.*

Proof. Take any minimizer f. (Minimizers exist due to the Lovász-Szegedy compactness theorem and the lower semicontinuity of I_p.) First, note that $f \le p$ almost everywhere: if not, then $I_p(f)$ can be decreased by replacing f with $\min\{f, p\}$, which retains the condition $t(H,f) = 0$.

Next, note that for almost all x, y, $f(x,y) = 0$ or p. If not, then redefine f to be equal to p wherever f was positive. This decreases the entropy while retaining the condition $t(H,f) = 0$.

Let $h = f/p$. Then h takes value 0 or 1 almost everywhere and h maximizes $\iint h(x,y)dxdy$ among all symmetric measurable $h : [0,1]^2 \to \{0,1\}$ satisfying $t(H,h) = 0$. Our goal is to show that $\widetilde{h} = \widetilde{g}$.

Let $r := \chi(H)$. Let $X_0, X_1, X_2, \ldots$ be a sequence of i.i.d. random variables uniformly distributed in $[0,1]$. Let

$$\mathcal{R} := \{i : h(X_i, X_j) = 1 \text{ for all } 1 \le j < i\},$$

and let $R := |\mathcal{R}|$. Let $\lambda(x) := \int h(x,y)dy$, so that for any given i,

$$\mathbb{P}(h(X_i, X_j) = 1 \text{ for all } 1 \le j < i) = \mathbb{E}(\lambda(X_i)^{i-1}) = \mathbb{E}(\lambda(X_0)^{i-1}).$$

Thus,

$$\begin{aligned}
\mathbb{E}(R) &= \sum_{i=1}^\infty \mathbb{P}(h(X_i, X_j) = 1 \text{ for all } 1 \le j < i) \\
&= \sum_{i=1}^\infty \mathbb{E}(\lambda(X_0)^{i-1}) \\
&\ge \sum_{i=1}^\infty (\mathbb{E}\lambda(X_0))^{i-1} = \frac{1}{1 - \mathbb{E}\lambda(X_0)} = \frac{1}{1 - \iint h(x,y)dxdy}. \qquad (7.2)
\end{aligned}$$

Let g be the function defined in (7.1). Suppose the vertex set of H is $\{1, \ldots, k\}$ for some integer k. If $t(H,g) > 0$, then there exist $x_1, \ldots, x_k$ such that $g(x_i, x_j) = 1$ whenever (i,j) is an edge in H. By the nature of g, this implies that H can be colored by $r - 1$ colors; since this is false, therefore $t(H,g)$ must be zero. By the optimality property of h, this gives

$$\iint h(x,y)dxdy \ge \iint g(x,y)dxdy = 1 - \frac{1}{r-1}.$$

Therefore by (7.2),

$$\mathbb{E}(R) \ge r - 1.$$

Again by Lemma 7.2, $t(K_r, h) = 0$. Therefore, $R \le r - 1$ almost surely. Combined with the above display, this shows that equality must hold in

(7.2) and $R = r-1$ almost surely. In particular, $\mathbb{E}(\lambda(X_0)^2) = (\mathbb{E}\lambda(X_0))^2$ and $\mathbb{E}\lambda(X_0) = 1 - 1/(r-1)$, which shows that

$$\lambda(x) = 1 - \frac{1}{r-1} \quad \text{a.e.}$$

For each x, let $A(x) := \{y : h(x,y) = 0\}$. Then $|A(x)| = 1/(r-1)$ a.e., where $|A(x)|$ denotes the Lebesgue measure of $A(x)$.

Define a random graph G on $\{0,1,2,\ldots\}$ by including the edge (i,j) if and only if $h(X_i, X_j) = 1$. Since $t(K_r, h) = 0$, G cannot contain any copy of K_r. Thus, with probability 1, $h(X_0, X_i) = 0$ for some $i \in \mathcal{R}$. In other words, $\bigcup_{i\in\mathcal{R}} A(X_i)$ cover almost all of $[0,1]$. Again, $|A(X_i)| = 1/(r-1)$ for all $i \in \mathcal{R}$ and $|\mathcal{R}| = r-1$ almost surely. All this together imply that with probability 1, $A(X_i) \cap A(X_j)$ has Lebesgue measure zero for all $i \neq j \in \mathcal{R}$, since

$$\sum_{i,j\in\mathcal{R},\ i<j} |A(X_i) \cap A(X_j)| \le \sum_{i\in\mathcal{R}} |A(X_i)| - \biggl|\bigcup_{i\in\mathcal{R}} A(X_i)\biggr| = 0.$$

Let $Y_1, Y_2, \ldots$ and $Z_1, Z_2, \ldots$ be i.i.d. random variables uniformly distributed in $[0,1]$, that are independent of the sequence $X_1, X_2, \ldots$. Since $t(K_r, h) = 0$, with probability 1 there cannot exist l and a set B of integers of size $r-2$ such that $h(Y_l, X_i) = h(Z_l, X_i) = 1$ for all $i \in B$, $h(X_i, X_j) = 1$ for all $i \neq j \in B$, and $h(Y_l, Z_l) = 1$.

Now fix a realization of $X_1, X_2, \ldots$. This fixes the set $\mathcal{R}$. Take any $i \in \mathcal{R}$. Let I be the smallest integer such that both Y_I and Z_I are in $A(X_i)$. Clearly Y_I and Z_I are independent and uniformly distributed in $A(X_i)$, conditional on the sequence $X_1, X_2, \ldots$ and our choice of $i \in \mathcal{R}$. By the observation from the preceding paragraph, $h(Y_I, Z_I) = 0$ with probability 1, since the set $\mathcal{R}\backslash\{i\}$ serves the role of B.

This shows that given $X_1, X_2, \ldots$, the sets $A(X_i)$ have the property that for almost all $y, z \in A(X_i)$, $h(y,z) = 0$. Since $\lambda(x) = 1 - 1/(r-1)$ a.e. and $|A(X_i)| = 1/(r-1)$, this shows that for almost all $y \in A(X_i)$ and almost all $z \notin A(X_i)$, $h(y,z) = 1$.

The properties of $(A(X_i))_{i\in\mathcal{R}}$ that we established can be summarized as follows: the sets $A(X_i)$ are disjoint up to errors of measure zero; each $A(X_i)$ has Lebesgue measure $1/(r-1)$ and together they cover the whole of $[0,1]$; for almost all $y, z \in [0,1]$, $h(y,z) = 0$ if they belong to the same $A(X_i)$, and $h(y,z) = 1$ if $y \in A(X_i)$ and $z \in A(X_j)$ for some $i \neq j$. These properties immediately show that h is the same as the function g up to a rearrangement; the formal argument can be completed as follows.

Given $X_1, X_2, \ldots$, let $u : [0,1] \to [0,1]$ be the map defined as

$$u(x) := \text{minimum } i \in \mathcal{R} \text{ such that } x \in A(X_i).$$

Note that with probability 1, for almost all x there is a unique $i \in \mathcal{R}$ such that $x \in A(X_i)$. Let $\sigma : [0,1] \to [0,1]$ be a measure-preserving bijection such that $x \mapsto u(\sigma x)$ is a non-increasing (we omit the construction). Then σ maps the intervals $[0, 1/(r-1)]$, $[1/(r-1), 2/(r-1)]$, $\ldots$, $[(r-2)/(r-1), 1]$

onto the sets $(A(X_i))_{i\in\mathcal{R}}$ up to errors of measure zero. By the properties of $A(X_i)$ established above, this shows that $h(\sigma x, \sigma y)$ is the same as $g(x,y)$ up to an error of measure zero. $\square$

Proof of Theorem 7.1. First, note that

$$T_{\beta_2}(h) - I(h) = \beta_2 t(H,h) - I_p(h) - \frac{1}{2}\log(1-p),$$

where $p = e^{2\beta_1}/(1+e^{2\beta_1})$. Take a sequence $\beta_2^{(n)} \to -\infty$, and for each n, let $\widetilde{h}_n$ be an element of $\widetilde{F}^*(\beta_2^{(n)})$. Let $\widetilde{h}$ be a limit point of $\widetilde{h}_n$ in $\widetilde{\mathcal{W}}$. If $t(H,h) > 0$, then by the continuity of the map $t(H,\cdot)$ and the boundedness of I_p,

$$\lim_{n\to\infty} \psi(\beta_2^{(n)}) = -\infty.$$

But this is impossible since $\psi(\beta_2^{(n)})$ is uniformly bounded below, as can be easily seen by considering the function g defined in (7.1) as a test function in the variational problem. Thus, $t(H,h) = 0$. If f is a function such that $t(H,f) = 0$ and $I_p(f) < I_p(h)$, then for all sufficiently large n,

$$T_{\beta_2^{(n)}}(h_n) - I(h_n) < T_{\beta_2^{(n)}}(f) - I(f)$$

contradicting the definition of $\widetilde{F}^*(\beta_2)$. Thus, if f is a function such that $t(H,f) = 0$, then $I_p(f) \ge I_p(h)$. By Theorem 7.3, this shows that $\widetilde{h} = p\widetilde{g}$. The compactness of $\widetilde{\mathcal{W}}$ now proves the first part of the theorem.

For the second part, first note that

$$\begin{aligned}\liminf_{n\to\infty} \psi(\beta_2^{(n)}) &\ge \lim_{n\to\infty} (T_{\beta_2^{(n)}}(g) - I(g)) \\ &= -I_p(g) - \frac{1}{2}\log(1-p) \\ &= \frac{(\chi(H)-2)}{2(\chi(H)-1)}\log\frac{1}{1-p}.\end{aligned}$$

Next, note that by the lower-semicontinuity of I_p and the fact that $\beta_2^{(n)}$ is eventually negative,

$$\begin{aligned}\limsup_{n\to\infty} \psi(\beta_2^{(n)}) &= \limsup_{n\to\infty} (\beta_2^{(n)} t(H,h_n) - I_p(h_n)) - \frac{1}{2}\log(1-p) \\ &\le \limsup_{n\to\infty}(-I_p(h_n)) - \frac{1}{2}\log(1-p) \\ &\le -I_p(g) - \frac{1}{2}\log(1-p).\end{aligned}$$

The proof is complete. $\square$

8. Transitivity and clumping

In the social networks literature, one of the key motivations for considering exponential random graphs is to develop models of random graphs that exhibit 'transitivity'. In simple terms, this means that a friend of a friend is more likely to be a friend than a random person. Presence of transitivity gives rise to 'clumps' of nodes that have higher connectivity between themselves. Since transitivity is closely related to the presence of 'triads' (i.e. triangles) researchers initially tried to model transitivity by the exponential random graph with edges and triangles as sufficient statistics. Sometimes, j-stars were thrown in for additional effect. For a history of such attempts and their experimental outcomes, see the discussion in Snijders et. al. [51].

However, as seen in experiments and through heuristics [46] and proved in Theorems 4.1 and 7.1, it is futile to model transitivity with only edges and triangles as sufficient statistics. If β_2 is positive, the graph is essentially behaving like an Erdős-Rényi graph, while if β_2 is negative, it becomes roughly bipartite. The degeneracy observed in experiments and proved in Theorem 5.1 also renders this model quite useless.

Recently, Snijders et. al. [51] have suggested a certain class of models that exhibit the desired transitivity and clumping properties in simulations. These models are of the type (4.1), where H_j is a j-star (or 'j-triangle', as defined in [51]) for $j = 1, \ldots, k-1$ and H_k is a triangle. The crucial assumption is that the parameters $\beta_1, \beta_2, \ldots, \beta_k$ have *alternating signs*. Usually, there is a single unknown parameter λ and β_j is taken to be $(-1)^{j-1}\lambda^{-j}$ for $j = 1, \ldots, k-1$. Based on simulations and heuristics, the authors of [51] claim that this class of models should demonstrate transitivity and clumping properties.

Although we do not yet have a general understanding as to why alternating sign models should give rise to transitivity, we can prove it in a certain special case. In this model, $k = 3$ and $H_1 =$ a single edge, $H_2 =$ a 2-star and $H_3 =$ a triangle. There is a single unknown (positive) parameter β, and the sufficient statistic is defined as

$$T(\widetilde{f}) := 3\beta\, t(H_1, \widetilde{f}) - 3\beta\, t(H_2, \widetilde{f}) + \beta\, t(H_3, \widetilde{f}).$$

Let $\widetilde{F}^* = \widetilde{F}^*(\beta)$ be as in Theorem 3.2. Of course, if β is sufficiently small, $\widetilde{F}^*(\beta)$ consists of a single constant function (and hence the model is effectively Erdős-Rényi) by Theorem 6.1. However, as following theorem shows, all elements of $\widetilde{F}^*(\beta)$ exhibit two clumps of roughly equal size when β is large.

Theorem 8.1. *In the setting described above,*

$$\lim_{\beta\to\infty} \sup_{\widetilde{f}\in\widetilde{F}^*(\beta)} \delta_{\square}(\widetilde{f}, \widetilde{h}) = 0,$$

where

$$h(x,y) = \begin{cases} 1 & \text{if } x, y \text{ on same side of } 1/2 \\ 1/2 & \text{if not.} \end{cases}$$

Proof. Note that $T(f)$ can be alternately written as a constant plus $S(1-f)$, where

$$S(g) := -\beta\, t(H_3, g).$$

The proof is now complete by Theorem 7.1 applied to the model with sufficient statistic S. $\square$

Intuitively, the function h in the above theorem represents connectivities between people in a population divided into two equal parts, say democrats and republicans, where all democrats are friends with each other, as are republicans; and there is a probability $1/2$ of friendship between a democrat and republican. It is clear that this arrangement automatically gives rise to transitivity.

Acknowledgements

We thank Hans Anderson, Charles Radin, Susan Holmes, Austin Head, Sumit Mukherjee and especially Sukhada Fadnavis for their substantial help with this paper.

References

[1] Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.*, **11** 581–598.
[2] Austin, T. (2008). On exchangeable random variables and the statistics of large graphs and hypergraphs. *Probab. Surv.*, **5** 80–145.
[3] Austin, T. and Tao, T. (2010). Testability and repair of hereditary hypergraph properties. *Random Structures Algorithms*, **36** 373–463.
[4] Bassetti, F. and Diaconis, P. (2006). Examples comparing importance sampling and the Metropolis algorithm. *Illinois J. Math.*, **50** 67–91 (electronic).
[5] Besag, J. (1975). Statistical analysis of non-lattice data. *Statistician*, **24** 179–195.
[6] Bhamidi, S., Bresler, G. and Sly, A. (2008). Mixing time of exponential random graphs. In *2008 IEEE 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 803–12.
[7] Bollobás, B. (2001). *Random graphs*, vol. 73 of *Cambridge Studies in Advanced Mathematics*. 2nd ed. Cambridge University Press, Cambridge.
[8] Bollobás, B. and Riordan, O. (2009). Metrics for sparse graphs. In *Surveys in combinatorics 2009*, vol. 365 of *London Math. Soc. Lecture Note Ser.* Cambridge Univ. Press, Cambridge, 211–287.

[9] Borgs, C., Chayes, J., Lovász, L., Sós, V. T. and Vesztergombi, K. (2006). Counting graph homomorphisms. In *Topics in discrete mathematics*, vol. 26 of *Algorithms Combin.* Springer, Berlin, 315–371.

[10] Borgs, C., Chayes, J. T., Lovász, L., Sós, V. T. and Vesztergombi, K. (2007). Convergent sequences of dense graphs. II. Multiway cuts and statistical physics. Preprint.

[11] Borgs, C., Chayes, J. T., Lovász, L., Sós, V. T. and Vesztergombi, K. (2008). Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing. *Adv. Math.*, **219** 1801–1851.

[12] Chatterjee, S. (2007). Estimation in spin glasses: A first step. *Ann. Statist.*, **35** no. 5, 1931–1946.

[13] Chatterjee, S. and Dey, P. S. (2010). Applications of Stein's method for concentration inequalities. *Ann. Probab.*, **38** 2443–2485.

[14] Chatterjee, S., Diaconis, P. and Sly, A. (2010). Random graphs with a given degree sequence. To appear.

[15] Chatterjee, S. and Varadhan, S. R. S. (2010). The large deviation principle for the Erdős-Rényi random graph. To appear in *European J. Combin.*

[16] Comets, F. and Janžura, M. (1998). A central limit theorem for conditionally centred random fields with an application to Markov fields. *J. Appl. Probab.*, **35** no. 3, 608–621.

[17] Corander, J., Dahmström, K. and Dahmström, P. (2002). Maximum likelihood estimation for exponential random graph models. In *Contributions to social network analysis, information theory and other topics in statistics: A festschrift in honour of Ove Frank* (J. Hagberg, ed.). Department of Statistics, University of Stockholm, Stockholm.

[18] Diaconis, P. and Janson, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl. (7)*, **28** 33–61.

[19] Diaconis, P. and Saloff-Coste, L. (1993). Comparison techniques for random walk on finite groups. *Ann. Probab.*, **21** 2131–2156.

[20] Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **5** 17–61.

[21] Fienberg, S. E. (2010). Introduction to papers on the modeling and analysis of network data. *Ann. Appl. Statist.*, **4** 1–4.

[22] Fienberg, S. E. (2010). Introduction to papers on the modeling and analysis of network data II. *Ann. Appl. Statist.*, **4** 533–534.

[23] Fortuin, C. M., Kasteleyn, P. W. and Ginibre, J. (1971). Correlation inequalities on some partially ordered sets. *Comm. Math. Phys.* **22** 89–103.

[24] Frank, O. and Strauss, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.*, **81** 832–842.

[25] Freedman, M., Lovász, L. and Schrijver, A. (2007). Reflection positivity, rank connectivity, and homomorphism of graphs. *J. Amer. Math. Soc.*, **20** 37–51 (electronic).

[26] Frieze, A. and Kannan, R. (1999). Quick approximation to matrices and applications. *Combinatorica*, **19** 175–220.
[27] Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, **13** 163–185.
[28] Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. Ser. B*, **54** no. 3, 657–699.
[29] Häggström, O. and Jonasson, J. (1999). Phase transition in the random triangle model. *J. Appl. Probab.*, **36** no. 4, 1101–1115.
[30] Hammersley, J. M. and Handscomb, D. C. (1965). *Monte Carlo Methods*. Methuen & Co. Ltd., London.
[31] Handcock, M. S. (2003). Assessing degeneracy in statistical models of social networks, Working Paper 39. Tech. rep., Center for Statistics and the Social Sciences, University of Washington.
[32] Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.*, **76** 33–65. With comments by Ronald L. Breiger, Stephen E. Fienberg, Stanley Wasserman, Ove Frank and Shelby J. Haberman and a reply by the authors.
[33] Hoover, D. N. (1982). Row-column exchangeability and a generalized model for probability. In *Exchangeability in probability and statistics (Rome, 1981)*. North-Holland, Amsterdam, 281–291.
[34] Janson, S., Łuczak, T. and Rucinski, A. (2000). *Random graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley-Interscience, New York.
[35] Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles*. Springer, New York.
[36] Kou, S. C., Zhou, Q. and Wong, W. H. (2006). Equi-energy sampler with applications in statistical inference and statistical mechanics. *Ann. Statist.*, **34** 1581–1652. With discussions and a rejoinder by the authors.
[37] Lovász, L. (2006). The rank of connection matrices and the dimension of graph algebras. *European J. Combin.*, **27** 962–970.
[38] Lovász, L. (2007). Connection matrices. In *Combinatorics, complexity, and chance*, vol. 34 of *Oxford Lecture Ser. Math. Appl.* Oxford Univ. Press, Oxford, 179–190.
[39] Lovász, L. and Sós, V. T. (2008). Generalized quasirandom graphs. *J. Combin. Theory Ser. B*, **98** 146–163.
[40] Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *J. Combin. Theory Ser. B*, **96** 933–957.
[41] Lovász, L. and Szegedy, B. (2007). Szemerédi's lemma for the analyst. *Geom. Funct. Anal.*, **17** 252–270.
[42] Lovász, L. and Szegedy, B. (2007). Testing properties of graphs and functions. To appear.

[43] Lovász, L. and Szegedy, B. (2009). Contractors and connectors of graph algebras. *J. Graph Theory*, **60** 11–30.

[44] Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica*, **6** 831–860.

[45] Park, J. and Newman, M. E. J. (2004). Solution of the two-star model of a network. *Phys. Rev. E (3)*, **70** 066146, 5.

[46] Park, J. and Newman, M. E. J. (2005). Solution for the properties of a clustered network. *Phys. Rev. E (3)*, **72** 026136, 5.

[47] Rinaldo, A., Fienberg, S. E. and Zhou, Y. (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Stat.*, **3** 446–484.

[48] Saloff-Coste, L. (1997). Lectures on finite Markov chains. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1996)*, vol. 1665 of *Lecture Notes in Math.* Springer, Berlin, 301–413.

[49] Sanov, I. N. (1961). On the probability of large deviations of random variables. In *Select. Transl. Math. Statist. and Probability, Vol. 1.* Inst. Math. Statist. and Amer. Math. Soc., Providence, R.I., 213–244.

[50] Snijders, T. A. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *J. Soc. Structure*, **3**.

[51] Snijders, T. A. B., Pattison, P. E., Robins, G. L. and Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociol. Method.*, **36** 99–153.

[52] Strauss, D. (1986). On a general class of models for interaction. *SIAM Rev.*, **28** 513–527.

[53] Talagrand, M. (2003). *Spin glasses: a challenge for mathematicians. Cavity and mean field models.* Springer-Verlag, Berlin.

[54] Wasserman, S. and Faust, K. (2010). *Social Network Analysis: Methods and Applications.* 2nd ed. Structural Analysis in the Social Sciences, Cambridge University Press, Cambridge.

Courant Institute of Mathematical Sciences, New York University

Departments of Mathematics and Statistics, Stanford University