Bernoulli **26**(2), 2020, 1016–1043
https://doi.org/10.3150/19-BEJ1135

Degeneracy in sparse ERGMs with functions of degrees as sufficient statistics

SUMIT MUKHERJEE

Department of Statistics, Columbia University, New York, NY, USA. E-mail: sm3949@columbia.edu

A sufficient criterion for "non-degeneracy" is given for Exponential Random Graph Models on sparse graphs with sufficient statistics which are functions of the degree sequence. This criterion explains why statistics such as alternating k-star are non-degenerate, whereas subgraph counts are degenerate. It is further shown that this criterion is "almost" tight. Existence of consistent estimates is then proved for non-degenerate Exponential Random Graph Models.

Keywords: degeneracy; ERGM; normalizing constant; sparse graphs

1. Introduction

Exponential families are frequently used in social science literature to model social networks (see [11,12,14,15,18,22,24–27] and the references within). Such models are usually referred to as Exponential Random Graph Models, commonly abbreviated as ERGMs, in the social science community. Starting with [12] in 2003, it has been noted in the social science literature that ERGMs with subgraph counts do not behave in a nice manner in terms of sampling and estimation procedures. This phenomenon is typically referred to as degeneracy. Attempts have been made to characterize degeneracy (see, e.g., [21,23]) but there is no universally accepted definition for degeneracy. This paper will adopt a notion similar to [12,24], where the degeneracy of a model is attributed to the sufficient statistics of the model. That is, the model will be deemed non-degenerate if the model behaves "nicely" for all choices of the parameter values. Thus under this notion, a degenerate model is caused by one or more degenerate statistics, and so the term degenerate will be used for both the model as well as the statistic.

One of the features of degeneracy is that such models place most of their mass on a very small sub-collection of graphs. The intuitive idea behind their reasoning is that in such models the neighboring edges are highly correlated. This causes a cascading effect through the graph, and so the model ends up putting most of its mass on very sparse or very dense graphs. In a sense, such models capture "too much interaction". Thus, an MCMC sample from such a model almost invariably gives either a very sparse graph, or a very dense graph. Another feature of such models is that small changes in the parameter can cause a large change in the underlying model. As such parameter estimates obtained from such models are usually not stable.

It has been subsequently noted in [24] in 2006 that not all ERGMs exhibit degeneracy in empirical studies. In fact, in this paper the authors argue that using modified versions of subgraph counts can reduce this problem to a large extent. The modifications are specifically aimed at reducing correlations between edges, and simulations seem to confirm this intuition. This raises

1350-7265 © 2020 ISI/BS

the question of whether we can justify this empirical non degeneracy in a more rigorous setting, and whether we can develop an Inferential Framework for such models.

1.1. Outline of the paper

Section 1.2 describes some examples of concrete interest and introduces the theoretical set up, and Section 1.3 outlines the main results of this paper. Section 2 explains how one can use the results of this paper to compute normalizing constants using four examples.

The main tool for proving the results of this paper is a large deviation principle for the empirical degree distribution μ_n^G for a sparse Erdös–Renyi graph G with respect to weak topology, studied in [8], Corollary 2.2, and [2], Corollary 1.9. Their result is outlined in Section 3. Section 3.1 carries out the proofs of the main results of the paper (Theorem 1.4, Corollary 1.5, and Theorem 1.7), using auxiliary lemmas which are proved in Section 3.2. Finally, Section 3.3 proves existence of consistence estimates for the ERGMs proposed in this paper (cf. Theorem 1.9 and 1.11).

1.2. Examples

The following definition gives the necessary notations for introducing some of the examples from [24] which are non degenerate at an empirical level.

Definition 1.1. Let $\mathcal{G}_n$ denote the space of all simple labelled undirected graphs on n vertices. For any $G \in \mathcal{G}_n$ let $\mathbf{d}(G) = (d_1(G), \dots, d_n(G))$ denote the labeled degree sequence of G, i.e. $d_j(G)$ is the degree of vertex j. Also let $E(G) := \frac{1}{2}\sum_{j=1}^n d_j(G)$ denote the number of edges in G.

For $0 \le i \le n-1$, let $h_i(G) := \#\{1 \le j \le n : d_j(G) = i\}$ denote the number of vertices of degree i. Summing over i gives $\sum_{i=0}^{n-1} h_i(G) = n$, since the sum is over all the vertices of G. The quantity $\mathbf{h}(G) := \{h_i(G)\}_{i=0}^{n-1}$ will be referred to as the degree frequency vector.

Recall that a k-star has k edges and $k+1$ vertices. For any $k \ge 2$, let $T_k(G)$ denote the number of copies of k-stars in G. The counting scheme is such that all copies of the k-star are considered, and not just the induced ones. This counting scheme gives the following simple formula for $T_k(G)$ in terms of its degrees $\mathbf{d}(G)$, as well as the degree frequency vector $\mathbf{h}(G)$:

$$T_k(G) = \sum_{j=1}^{n} \binom{d_j(G)}{k} = \sum_{i=0}^{n-1} h_i(G) \binom{i}{k}.$$

This is because for any vertex j, there are $\binom{d_j(G)}{k}$ k-stars with j as the center vertex, and so adding over j gives the total number of k-stars. The second equality follows by rearranging the first sum.

We will now introduce some of the non-degenerate statistics defined in [24].

(a) *Geometrically weighted degree statistic*

The geometrically weighted degree statistic has the form

$$\mathrm{gwd}_\alpha(G) := \sum_{i=0}^{n-1} e^{-\alpha i} h_i(G),$$

where $\alpha > 0$ is known. The geometrically decaying weights ensure that the contribution of vertices with large degree is negligible. Thus as the degrees of the graph increase, the statistic does not grow too fast, and cascading effect of this statistic is reduced.

(b) *The alternating k-star*

For a fixed parameter $\lambda > 1$, the alternating k-star is defined as

$$\mathrm{aks}_\lambda(G) := \sum_{k=2}^{n-1} \frac{(-1)^k}{\lambda^{k-2}} T_k(G),$$

where T_k's are the k-star counts defined above. In this case again the geometrically decaying weights ensure that the cascading effects of higher star counts is reduced. Also because of the alternate signs the cascading effect of consecutive terms is cancelled to a large extent.

The authors in [24] note that using the formula for $T_k(G)$ in terms of the degree frequency vector $\mathbf{h}(G)$, the alternating k-star statistic can be written as

$$\mathrm{aks}_\lambda(G) = \lambda^2 \sum_{i=0}^{n-1} \left[\left(1 - \frac{1}{\lambda}\right)^i - 1 + \frac{i}{\lambda} \right] h_i(G) = \lambda^2 \mathrm{gwd}_\alpha(G) - n\lambda^2 + 2\lambda E(G)$$

with $e^{-\alpha} = 1 - 1/\lambda$. Thus the two statistics gwd_α and aks_λ are connected by a simple formula, and both these statistics are functions of the degree frequency vector $\mathbf{h}$.

(c) *The number of isolated nodes*

The statistic $h_0(G)$ which is the number of isolated vertices in the graph G. This statistic is obtained from the gwd_α statistic by letting $\alpha \to \infty$, or equivalently from the aks_λ by letting $\lambda \to 1$.

(d) *The Yule distribution statistic*

Another statistic which penalizes high degrees is the Yule distribution statistic, given by

$$yu(G) := \sum_{j=1}^{n} \frac{1}{(d_j + c)_r} = \sum_{i=0}^{n-1} \frac{1}{(i+c)_r} h_i(G), \qquad (d)_r := d(d+1)\cdots(d+r-1),$$

where r and c are both positive integers. In this case the penalty is polynomial as opposed to geometric as in gwd_α, but a similar non-degenerative effect is achieved.

In all the four examples above the statistic under consideration can be written as $\sum_{i=0}^{n-1} f(i) \times h_i(G)$ for some function $f : \mathbb{N}_0 \mapsto \mathbb{R}$, where $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. As an illustration, the gwd_α, aks_λ, the number of isolated vertices and the Yule distribution statistic fit this framework with

$$f(i) = e^{-\alpha i}, \qquad f(i) = \lambda^2\left[\left(1 - \frac{1}{\lambda}\right)^i - 1 + \frac{i}{\lambda}\right], \qquad f(i) = 1_{i=0}, \qquad f(i) = \frac{1}{(i+c)_r}$$

respectively. Restricting attention to statistics of this form, one can ask when is this statistic well behaved. The results of this paper gives a sufficient condition for this which is easy to check:

$$\lim_{i\to\infty} \frac{|f(i)|}{i \log i} = 0. \tag{1}$$

This will be made precise in Theorem 1.4 and Corollary 1.5.

In particular, (1) holds for gwd_α for $\alpha > 0$, and the number of isolated vertices, and the Yule distribution statistic, as in all these cases the function f is bounded. For aks_λ with $\lambda > 1$ the function

$$f(i) = \lambda^2\left[\left(1 - \frac{1}{\lambda}\right)^i - 1 + \frac{i}{\lambda}\right]$$

is unbounded, but dominated by the linear term. On the other hand, the number of k-stars also equals $\sum_{i=0}^{n-1} f(i)h_i(G)$ for the choice $f(i) = \binom{i}{k}$ which does not satisfy (1), as in this case $f(i)$ grows at a polynomial rate. Also in the alternating k-star statistic if the signs do not alternate, then one has

$$\sum_{k=2}^{n-1} \frac{1}{\lambda^{k-2}} T_k(G) = \sum_{i=0}^{n-1} f(i)h_i(G)$$

with $f(i) = \lambda^2[(1 + \frac{1}{\lambda})^i - 1 - \frac{i}{\lambda}]$ which does not satisfy (1), as in this case the exponential term dominates. Thus it is crucial that the signs in the alternating k-star statistic do alternate. It follows from Theorem 1.7 that both the number of k-stars and the non-alternating k-star statistics are degenerate, in a sense which is again made precise in Theorem 1.7.

The main tool for these results is the analysis of sparse graphs, as opposed to dense graphs as in [1,3,19]. Recall that in a dense graph on n vertices the number of edges is $O(n^2)$ and the degrees are $O(n)$. Here and henceforth in this paper, we use the notation $a_n = O(b_n)$ for two positive real sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, if there exists a constant C free of n such that $a_n \leq Cb_n$ for all $n \geq 1$. With this notation, the term sparse graphs will refer to graphs which have $O(n)$ edges and the degrees of the vertices are $O(1)$. One reason it is interesting to model sparse graphs is that most real life networks seem to be sparse. Another reason is that the dense graph theory does not provide a good explanation for why the modified versions of subgraph counts mentioned above (such as aks_λ) are non-degenerate, whereas the subgraph count statistics (such as star counts) are.

For a unified treatment of these and other examples, consider an exponential family on $\mathcal{G}_n$ of the form

$$\mathbb{Q}_{n,\beta,f}(G) := \left(\frac{\beta}{n}\right)^{E(G)} \left(1 - \frac{\beta}{n}\right)^{\binom{n}{2} - E(G)} e^{\sum_{i=0}^{n-1} h_i(G) f(i) - Z_n(\beta, f)}, \tag{2}$$

where $f : \mathbb{N}_0 \mapsto \mathbb{R}$, β is a positive real valued parameter, and $Z_n(\beta, f)$ is the log normalizing constant. If f is either identically 0 or exactly linear, this model reduces to a sparse Erdös–Renyi model which puts most of its mass on sparse graphs. Thus the same should be true for functions $f(\cdot)$ which do not grow too fast. Since the model does not change if $f(\cdot)$ is replaced by $f(\cdot) + c$ for some constant c, without loss of generality we will assume $f(0) = 0$.

It should be noted at this point that $\mathbb{Q}_{n,\beta,f}$ is not the same as the $\boldsymbol{\beta}$ model studied in [4]. The $\boldsymbol{\beta}$ model is an exponential family on $\mathcal{G}_n$ whose probability mass function is proportional to $\exp\{\sum_{j=1}^n \beta_j d_j(G)\}$ where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)$ is an n dimensional parameter. In the β model, the labeled degree sequence $(d_1(G), \dots, d_n(G))$ is minimal sufficient. On the other hand in the model of (2) if the function f is assumed to be unknown the minimal sufficient statistics are the unlabeled degree sequence $(d_{(1)}(G) \ge d_{(2)}(G) \ge \cdots \ge d_{(n)}(G))$, or equivalently the degree frequency vector $(h_0(G), h_1(G), \dots, h_{n-1}(G))$. More importantly, model (2) introduces nontrivial dependence among the edges of the graph G, whereas under the $\boldsymbol{\beta}$ model the edges are mutually independent. In [4], the authors worked in the dense graph regime and showed that if the components of the parameter vector $\boldsymbol{\beta}$ stays uniformly bounded, then all entries of $\boldsymbol{\beta}$ can be simultaneously estimated consistently. In a similar manner, Theorem 1.11 shows that if the true function f is unknown and treated as a parameter, one can estimate the value of the function f consistently at every fixed i, under the assumption that f satisfies (1).

1.3. Statement of main results

For analyzing model (2) it suffices to study the degree sequence. The following definition encodes the entire degree sequence as one probability measure on non-negative integers.

Definition 1.2. Given the labelled degrees of a graph $(d_1(G), \dots, d_n(G))$, the empirical distribution of the degree sequence is defined by $\mu_n^G := \frac{1}{n}\sum_{j=1}^n \delta_{d_j(G)}$ i.e. μ_n^G is the measure which puts mass $1/n$ at each of the observed degree $d_j(G)$, and is a probability measure on $\mathbb{N}_0$. An equivalent definition of μ_n^G in terms of the degree frequency vector $\mathbf{h}(G)$ is the probability measure which puts mass $h_i(G)/n$ at i, for $0 \le i \le n-1$.

With this definition, any statistic of the form $\sum_{i=0}^{n-1} f(i)h_i(G)$ can be written as $n\mu_n^G[f]$, where $\mu[f]$ denotes the mean of f with respect to the measure μ (when it exists), i.e.

$$\mu[f] := \sum_{i=0}^{\infty} \mu(i) f(i).$$

In particular if $f(i) = i$ is the identity function, then define $\overline{\mu} := \sum_{i=1}^{\infty} i\mu(i)$ to be the mean of the measure μ.

The next definition gives all the necessary ingredients for expressing the asymptotic log normalizing constant as a one dimensional optimization problem.

Definition 1.3. Suppose the function $f : \mathbb{N}_0 \mapsto \mathbb{R}$ satisfies

$$\limsup_{i\to\infty} \frac{f(i)}{i \log i} = 0, \tag{3}$$

which is a slightly weaker condition than (1). For $u \geq 0$ define an exponential family on $\mathbb{N}_0$ with probability mass function

$$\sigma_{u,f}(i) = \frac{1}{i!} u^i e^{f(i) - Z(u,f)},$$

where $Z(u, f)$ is the log normalizing constant, i.e.

$$Z(u, f) := \log\left(\sum_{i=0}^{\infty} \frac{1}{i!} u^i e^{f(i)}\right).$$

Since f satisfies (3) we have that $Z(u, f) < \infty$. Let Ω_f denote the set of all probability measures of the form $\sigma_{u,f}$ for $u \geq 0$. Also let $m(u, f) := \overline{\sigma_{u,f}}$ denote the mean of $\sigma_{u,f}$. Finally, for $\beta > 0$ let $J(\beta, f)$ denote the solution to the following optimization problem

$$J(\beta, f) := \sup_{u \geq 0}\left\{Z(u, f) - m(u, f)\log u + \frac{m(u, f)}{2}\log\big(m(u, f)\beta\big) - \frac{m(u, f) + \beta}{2}\right\}. \quad (4)$$

The definition of $J(\beta, f)$ involves an optimization over the scalar non-negative variable u which can be computed numerically.

The first main result of this paper is the following theorem, which gives the asymptotics of the log normalizing constant for the model $\mathbb{Q}_{n,\beta,f}$ under assumptions on the growth rate of f. Existence of limiting log normalizing constant for a dependent system with growing number of variables governed by a Gibbs measure has attained considerable interest in Statistical Physics, where this is typically referred to as existence of the thermodynamic limit. Typically the limiting normalizing constant is expressed in terms of an optimization problem, and the optimizers represent the steady states of the distribution. See [20] for more on existence of thermodynamic limits and its properties in general.

Theorem 1.4. *Suppose either of these two conditions hold*:

(i) $f : \mathbb{N}_0 \mapsto \mathbb{R}$ *satisfies* (1),
or
(ii) $f : \mathbb{N}_0 \mapsto \mathbb{R}$ *is non increasing.*

Let G be a random graph from the exponential family $\mathbb{Q}_{n,\beta,f}$ as defined in (2).

(a) *Then as $n \to \infty$, the asymptotics of the log normalizing constant is given by*

$$\lim_{n\to\infty} \frac{1}{n} Z_n(\beta, f) = J(\beta, f),$$

with $J(\beta, f)$ as in definition 1.3.

(b) *The supremum in the definition of $J(\beta, f)$ is attained on a finite set of positive reals $\{u_1, u_2 \cdots, u_k\}$ satisfying the equation $u_l^2 = \beta\bar{\sigma}_{u_l,f}$ for $1 \leq l \leq k$, with $\sigma_{u,f}$ as in Defini-*

tion 1.3. Further, for any function ψ satisfying (1) one has

$$\min_{i=1}^{k}\left|\mu_n^G(\psi)-\sigma_{u_i,f}(\psi)\right| \xrightarrow{p} 0.$$

where μ_n^G is the empirical degree distribution of G.

Note that both the conditions (i) and (ii) considered in part (a) of Theorem 1.4 are sub-cases of the assumption (3), which is used to ensure that $J(\beta, f)$ introduced in (4) is well defined and finite. It is possible that the conclusion of Theorem 1.4 holds for all f satisfying (3).

An immediate application of the above theorem gives the following corollary.

Corollary 1.5. *Suppose $f : \mathbb{N}_0 \mapsto \mathbb{R}$ satisfy (1), and let G be a random graph from the exponential family $\mathbb{Q}_{n,\beta,\theta f}$, where $\mathbb{Q}_{n,\beta,f}$ is as defined in (2). Then the following conclusions hold:*

(a) *Both part (a) and part (b) of Theorem 1.4 hold with f replaced by θf, for all $\theta \in \mathbb{R}$. Also, the limiting log partition function*

$$J(\beta,\theta f)=\lim_{n\to\infty}\frac{1}{n}Z_n(\beta,\theta f)$$

is finite and continuous in θ.

(b) *There exists positive reals $m < M$ depending on (f, β, θ) such that*

$$\lim_{n\to\infty}\mathbb{Q}_{n,\beta,\theta f}\left(\frac{E(G)}{n}\in[m,M]\right)=1.$$

Remark 1.6. Part (a) of Corollary 1.5 says that if $|f|$ grows at a rate smaller than $i \log i$, then the corresponding model $\mathbb{Q}_{n,\beta,\theta f}$ is well behaved for both positive and negative θ, in the sense that the limiting log partition function is finite and continuous in θ. It also shows that the empirical degree distribution μ_n^G roughly behaves like a mixture of $\{\sigma_{u_i,\theta f}\}_{i=1}^k$ for large n. In particular if there is a unique optimizer u_0 to the optimization problem $J(\beta, \theta f)$, then the empirical degree distribution μ_n^G converges weakly to $\sigma_{u_0,\theta f}$, and $\bar{\mu}_n^G$ converges to $\bar{\sigma}_{u_0,\theta f}$.

Part (b) shows that irrespective of whether there is a “phase transition”, the number of edges is linear in the number of vertices for all parameter values θ (cf. [20] for details on phase transitions in models of Statistical Mechanics). Thus the level of sparsity of the graph does not change with the parameter.

Also, none of the limit points of the degree distribution is a Poisson, as $\sigma_{u,\theta f}$ is not a Poisson distribution unless f is identically 0 or linear, in which case the model $\mathbb{Q}_{n,\beta,\theta f}$ itself is a sparse Erdös–Renyi graph. On the other hand, the empirical degree distribution of a sparse Erdös–Renyi graph converges to Poisson. Thus unlike ERGMs on dense graphs as studied in [3], ERGMs on sparse graphs do not behave like mixture of Erdös–Renyi graphs. Also, in the case of sparse ERGMs, it is possible to estimate multiple parameters consistently from a large single graph. In particular, see Theorem 1.9 which constructs consistent estimates for (β, θ) when f is known, and Theorem 1.11 which constructs consistent estimates for the function f if f is unknown. It should be noted here that consistent estimation of parameters in ERGMs was achieved in [23], but under the assumption that the ERGM restricted to n vertices is a projection of the corresponding ERGM on $n+1$ vertices. Consistency results have also been obtained for sparse ERGMs in [16], but here the authors assume dyadic independence. In contrast, the models presented in this paper are neither projective nor have dyadic independence, and yet consistent estimation is possible in this case.

Since choosing a function f is equivalent in spirit to specifying the degree distribution of the graph, one can fit a wide class of degree distributions by choosing a corresponding function f. Of course restriction (1) ensures that the degree distribution will have a finite exponential moment, which rules out degree distribution with power law tails. Power law tails correspond to the case when $f(i)$ grows at the rate $i \log i$, which require a more delicate analysis and is not carried out in this paper.

The next theorem shows that some growth condition on f needs to satisfied for the model $\mathbb{Q}_{n,\beta,\theta f}$ to be well behaved for all values of θ.

Theorem 1.7. *Suppose $f : \mathbb{N}_0 \mapsto \mathbb{R}$ is a non-decreasing function, and G be a random graph from the exponential family $\mathbb{Q}_{n,\beta,\theta f}(\cdot)$, where $\mathbb{Q}_{n,\beta,f}$ is as defined in* (2).

(a) *If $\theta < 0$, then both parts* (*a*) *and part* (*b*) *of Theorem* 1.4 *hold with f replaced by θf. Also, the asymptotic log normalizing constant*

$$J(\beta, \theta f) = \lim_{n\to\infty} \frac{1}{n} Z_n(\beta, \theta f)$$

is finite and continuous in θ. Further, there exists positive constants $m < M$ depending on (θ, β, f) such that

$$\lim_{n\to\infty} \mathbb{Q}_{n,\beta,\theta f}\left(\frac{E(G)}{n} \in [m, M]\right) = 1.$$

(b) *If f further satisfies*

$$\liminf_{i\to\infty} \frac{f(i) - f(i-1)}{\log i} > 4, \tag{5}$$

then for $\theta > 0$ we have

$$\lim_{n\to\infty} \frac{1}{nf(n)} Z_n(\beta, \theta f) = \theta,$$

and $\lim_{n\to\infty} \mathbb{Q}_{n,\beta,\theta f}(G = K_n) = 1$.

Remark 1.8. Note that the assumption (5) automatically implies $f(i)$ is at least of order $i \log i$, that is, (1) does not hold. Under this assumption, Theorem 1.7 demonstrates degeneracy in the sense of [12,24] in two ways. First, in this case the behavior of the model $Q_{n,\beta,\theta f}$ changes drastically at the origin. For $\theta < 0$ the model puts all its mass on sparse graphs with $O(n)$ edges, whereas for $\theta > 0$ the model suddenly shifts all its mass to the complete graph where number of edges is $\binom{n}{2} \sim \frac{n^2}{2}$. Also, for $\theta > 0$ the model puts most of its mass on a very small subset of $\mathcal{G}_n$ (namely a subset of size 1). Thus model (2) can indeed be degenerate without any growth conditions on f.

In particular this happens for the choice $f(i) = \binom{i}{k}$ for any k fixed, for which the statistic $\sum_{i=0}^{n-1} f(i)h_i(G)$ becomes the number of k-stars, and for the choice $f(i) = \lambda^2[(1+\frac{1}{\lambda})^i - 1 - \frac{i}{\lambda}]$, for which the statistic $\sum_{i=0}^{n-1} f(i)h_i(G)$ is the non alternating k-star. Note that in both these cases, the function f is indeed non-decreasing, and satisfies (5).

1.4. Identifiability and estimating parameters

Since model $\mathbb{Q}_{n,\beta,\theta f}$ is well behaved for all θ when f satisfies (1), this subsection explores the estimation of parameters of the model, under the assumption that f satisfies (1). Assuming that f is known, one can focus on estimating the parameters (β, θ) in the model $\mathbb{Q}_{n,\beta,\theta f}$ from one sample G from this model. If f is exactly linear, that is, there exists a constant b such that $f(i) = bi$, then the model $\mathbb{Q}_{n,\beta,\theta f}$ is same as Erdös–Renyi with parameter

$$\frac{1}{1 + \frac{n-\beta}{\beta} e^{-\theta b}} \approx \frac{\beta e^{\theta b}}{n}.$$

This model is asymptotically not identifiable along the curve where $\beta e^{\theta b}$ is constant, and so joint estimation of both parameters (β, θ) is not possible. If f is not linear, consistent estimation of both the parameters is possible under this model.

In order to motivate our proposed estimates, recall the prediction of part (b) of Corollary 1.5, that for large n we have

$$\frac{h_i(G)}{n} \approx \frac{u^i}{i!} e^{\theta f(i) - Z(u,\theta f)},$$

if f satisfies (1). Taking this to be an exact equality, multiplying both sides by $i!$ and taking log gives

$$\log \frac{i! h_i(G)}{n} = -Z(u, \theta f) + \theta f(i) + i \log u.$$

Thus taking $x_1(i) = f(i)$, $x_2(i) = i$, $y(i) = \log \frac{i! h_i(G)}{n}$, we get a linear equation of the form

$$y(i) = -Z(u, \theta f) + \theta x_1(i) + (\log u) x_2(i),$$

and so by fitting a multiple linear regression model using least squares with y as response and $\{x_1, x_2\}$ as explanatory variables we can estimate θ and $\log u$. Finally, note that u, θ, β are connected by the equation $u^2 = \beta \bar{\sigma}_{u,\theta f}$, as shown in part (b) of Corollary 1.5. Since the empirical average of degrees $\bar{d}(G) = \frac{2E(G)}{n}$ converge to $\bar{\sigma}_{u,\theta f}$ in probability, one can use the approximate equation $nu^2 = 2\beta E(G)$ along with the least squares estimate of $\log u$ to get an estimate of β. We will now show that the estimates of (θ, β) outlined above are indeed consistent.

Theorem 1.9. *Let $f : \mathbb{N}_0 \mapsto \mathbb{R}$ be a known function which satisfies* (1), *and $\theta_0 \in \mathbb{R}$, $\beta_0 > 0$ be the true unknown parameters. Let L be a fixed positive integer free of n such that $f(i)/i$ is not constant for all $i \in [0, L]$. Let $(\hat{\theta}_n, \hat{u}_n)$ be the least square estimates of (θ, u) defined via the following optimization problem*:

$$(\hat{\theta}_n, \hat{u}_n) := \arg \inf_{c,\theta \in \mathbb{R}, u>0} \sum_{i=0}^{L} \left\{ \log \frac{i! h_i(G)}{n} - c - \theta f(i) - i \log u \right\}^2.$$

Then as $n \to \infty$, one has $\hat{\theta}_n \xrightarrow{p} \theta_0$. Further, the estimator $\hat{\beta}_n := \frac{n\hat{u}_n^2}{2E(G)} \xrightarrow{p} \beta_0$.

Remark 1.10. The estimates $(\hat{\theta}, \hat{\beta})$ of the previous theorem are motivated by the fact that

$$\frac{h_i(G)}{n} \approx \frac{1}{e^{Z(u,\theta f)}} e^{\theta f(i)} u^i,$$

as predicted in Corollary 1.5 under the assumption that f satisfies (1).

Even though one can use a larger value of L (for e.g. $L = n - 1$), estimates of $f(i)$ for large i are not as reliable. In particular, for large i it is possible to have $h_i(G) = 0$ which will give undefined values for $(\hat{\theta}_n, \hat{u}_n)$. This is the reason for choosing L fixed, free of n. Given a graph G, any valid choice of L must satisfy $L \le L_n(G) := \max_{1 \le j \le n} d_j(G)$. Indeed this is because $h_i(G) = 0$ for $i > L_n(G)$, and so the least square optimization problem in Theorem 1.9 is not defined. A natural choice of L is the maximum i such that $h_i(G)$ is non zero for all $i \in [0, L]$.

Frequently it is the case that an observed graph G has no isolated vertices. For example, any person in a social network has at least one friend. A natural model in this case is the same exponential family, but conditioned to have no isolated vertices. Since $h_0(G) = 0$, the estimator defined in Theorem 1.9 becomes undefined. In such cases, instead of starting at 0 one can consider the values $i \in [1, L]$ for the least squares procedure. The same proof shows that the resulting estimator is consistent, whenever f satisfies (1).

If there is no reasonable guess for the function f, then one can think of estimating the whole function f in the model $\mathbb{Q}_{n,\beta,f}$ using one large graph G. For any f the two models $\mathbb{Q}_{n,\beta,f}$ and $\mathbb{Q}_{n,1,\tilde{f}}$ are asymptotically unidentifiable, where

$$\tilde{f}(i) = f(i) + (i/2) \log \beta,$$

and so without loss of generality one may further assume $\beta = 1$. Under these assumptions, the next theorem reconstructs the whole function f.

Theorem 1.11. *Let $f : \mathbb{N}_0 \mapsto \mathbb{R}$ be such that f satisfy* (1), *and consider the model $\mathbb{Q}_{n,1,f}$ where $\mathbb{Q}_{n,\beta,f}$ is as defined in* (2). *Setting $\hat{u}_n := \sqrt{\frac{2E(G)}{n}}$ the function $\hat{f}_n : \mathbb{N}_0 \mapsto \mathbb{R}$ defined by*

$$\hat{f}_n(i) = \log \left[\frac{i! h_i(G)}{h_0(G)} \right] - i \log \hat{u}_n$$

satisfies

$$\hat{f}_n(i) \xrightarrow{p} f(i)$$

for $i \geq 1$, *as* $n \to \infty$.

1.5. Scope for future work

The statistics considered in this paper are functions of the degree sequence, or equivalently functions of 1 neighborhoods of the graph. The literature has also focused on statistics which cannot be expressed in terms of the degrees, for example the alternating k-triangle statistic (for more details on this statistic refer to [12,15,24]). The alternating k-triangle statistic depends on 2 neighborhoods of a vertex and not 1. The large deviation result of [2] applies for any finite neighborhood, and so it seems plausible that the two neighborhood can be dealt with a modified version of the strategy of this paper. Of course, for 2 (and general) neighborhoods the involved rate function will be more complicated.

2. Some examples

This section uses the results of this paper to analyze four ERGMs on sparse graphs from the probability mass function $\mathbb{Q}_{n,\beta,\theta f}$ of (2). To specify the model it suffices to choose the function f. Note that in none of these models a closed form expression for the normalizing constant $Z_n(\beta, \theta f)$ seems available. Using Corollary 1.5, one can get numerical approximations for the asymptotic normalizing constant.

2.1. Geometrically weighted degree

In this case, we have $\mathrm{gwd}_\alpha = \sum_{i=0}^{n-1} h_i(G) f(i)$ with $f(i) = e^{-\alpha i}$ for some $\alpha > 0$. An application of Corollary 1.5 gives the asymptotics of the log normalizing constant as

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} Z_n(\beta, \theta f)\\ = \sup_{u \geq 0} \left\{ Z(u, \theta f) - m(u, \theta f) \log u + \frac{m(u, \theta f)}{2} \log\bigl(m(u, \theta f)\beta\bigr) - \frac{m(u, \theta f) + \beta}{2} \right\},\end{aligned}$$

where $Z(u, \theta f)$ and $m(u, \theta f)$ are the log normalizing constant and mean respectively, of the probability mass function $\sigma_{u,\theta f}$ on non-negative integers given by

$$\sigma_{u,\theta f}(i) \propto \frac{1}{i!} u^i e^{\theta f(i)}.$$

Setting $\gamma := e^{-\alpha}$ one has

$$e^{Z(u,\theta f)} = \sum_{i=0}^{\infty} \frac{u^i}{i!} \exp\{\theta \gamma^i\} = \sum_{i=0}^{\infty} \frac{u^i}{i!} \sum_{j=0}^{\infty} \frac{\theta^j \gamma^{ij}}{j!} = e^{u+\theta} \sum_{i,j=0}^{\infty} \frac{e^{-u} u^i}{i!} \frac{e^{-\theta}\theta^j}{j!} \gamma^{ij} = e^{u+\theta} \mathbb{E}\gamma^{XY},$$

where X, Y are mutually independent and $X \sim \text{Pois}(u)$, $Y \sim \text{Pois}(\theta)$. By a similar calculation one has

$$m(u, \theta f) = \frac{\sum_{i=1}^{\infty} i \frac{u^i}{i!} \exp\{\theta \gamma^i\}}{\sum_{i=0}^{\infty} \frac{u^i}{i!} \exp\{\theta \gamma^i\}} = u \frac{\sum_{i=0}^{\infty} \frac{u^i}{i!} \exp\{\theta \gamma \gamma^i\}}{\sum_{i=0}^{\infty} \frac{u^i}{i!} \exp\{\theta \gamma^i\}} = u \frac{e^{u+\theta\gamma} E\gamma^{XZ}}{e^{u+\theta} E\gamma^{XY}} = u e^{\theta(\gamma-1)} \frac{\mathbb{E}\gamma^{XZ}}{\mathbb{E}\gamma^{XY}},$$

where $Z \sim \text{Poisson}(\theta\gamma)$ independent of X.

Since closed form expressions are not known for moment generating function of products of independent Poissons, further simplification is not possible in this case. Of course one can use numerical approximations by simulating an i.i.d. sample of products of Poissons, and then using strong law of large numbers to estimate the moment generating function.

2.2. Logarithmic model

For this model set

$$f(i) = -\log(i+1)_r = -\log(i+1)(i+2)\cdots(i+r),$$

where r is a positive integer. In this case $|f(i)|$ grows logarithmically, and so by Corollary 1.5 the asymptotics of the normalizing constant requires only the knowledge of $Z(u, \theta f)$ and $m(u, \theta f)$. For the special case $\theta = 1$, a direct computation shows that

$$e^{Z(u,f)} = \frac{1}{u^r} \left[e^u - \sum_{i=0}^{r-1} \frac{u^i}{i!} \right], \qquad m(u, f) = u - r + \frac{\frac{u^r}{(r-1)!}}{e^u - \sum_{i=0}^{r-1} \frac{u^i}{i!}}$$

Thus in this case both $Z(u, f)$ and $m(u, f)$ are explicit, and numerical optimization of $J(\beta, f)$ is easy to carry out. No such simple formula exists for $Z(u, \theta f)$ and $m(u, \theta f)$ for $\theta \neq 1$.

2.3. Sparse penalty model

For this model set $f(i) = 1_{i=0}$, for which the corresponding model $\mathbb{Q}_{n,\beta,\theta f}$ has sufficient statistic $h_0(G)$, the number of isolated vertices. This can be viewed as a penalty term which prefers or dislikes isolated vertices depending on whether $\theta > 0$ or $\theta < 0$. Since f is bounded, the asymptotics of the log normalizing constant follows from Corollary 1.5. For this particular choice of f, a direct calculation reveals that

$$e^{Z(u,\theta f)} = e^u + e^\theta - 1, \qquad m(u, \theta f) = \frac{u e^u}{e^u + e^\theta - 1}.$$

Computation of $J(\beta, \theta f)$ can then be carried out numerically in a straightforward manner.

2.4. Polynomial decay model

For this model set $f(i) = i^\alpha$ for some known $\alpha \in [0, 1]$. In this case the decay is at most linear by assumption, and so Corollary 1.5 applies. Proceeding to compute $Z(u, \theta f)$, we have

$$e^{Z(u,\theta f)} = \sum_{i=0}^{\infty} \frac{u^i}{i!} e^{\theta i^\alpha} = \sum_{i=0}^{\infty} \frac{u^i}{i!} \sum_{j=0}^{\infty} \frac{\theta^j i^{\alpha j}}{j!} = e^{u+\theta} \sum_{i,j=0}^{\infty} \frac{e^{-u} u^i}{i!} \frac{e^{-\theta} \theta^j}{j!} i^{\alpha j} = e^{u+\theta} \mathbb{E} X^{\alpha Y},$$

where $X \sim \text{Poisson}(u)$ and $Y \sim \text{Poisson}(\theta)$ are mutually independent. A similar computation gives

$$m(u, \theta f) = \frac{\sum_{i=1}^{\infty} \frac{u^i}{(i-1)!} e^{\theta i^\alpha}}{\sum_{i=0}^{\infty} \frac{u^i}{i!} e^{\theta i^\alpha}} = u \frac{\sum_{i=0}^{\infty} \frac{u^i}{i!} e^{\theta (i+1)^\alpha}}{\sum_{i=0}^{\infty} \frac{u^i}{i!} e^{\theta i^\alpha}} = u \frac{e^{u+\theta} \mathbb{E}(X+1)^{\alpha Y}}{e^{u+\theta} \mathbb{E} X^{\alpha Y}} = u \frac{\mathbb{E}(X+1)^{\alpha Y}}{\mathbb{E} X^{\alpha Y}}.$$

Further simplification is not possible in general, and one has to use numerical methods to compute both $Z(u, \theta f)$ and $m(u, \theta f)$.

3. Proofs of main results

The main tool for proving our results is a large deviation principle for the empirical degree distribution μ_n^G. To see how large deviation comes into the picture, note that the log normalizing constant of the model $\mathbb{Q}_{n,\beta,f}$ can be written as

$$Z_n(\beta, f) = \log \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{n \mu_n^G [f]},$$

where $\mathbb{P}_{n,\beta}$ is the Erdös–Renyi model with parameter (β/n). By Varadhan's lemma, this equates the problem to studying the large deviation of μ_n^G under the Erdös–Renyi (β/n) model. A large deviation for the whole graph G with respect to local weak convergence has recently been derived in [2], Theorem 1.8), which in particular gives a large deviation principle for μ_n^G with respect to the weak topology, as pointed out in [2], Corollary 1.9. The same large deviation was also obtained in [8], Corollary 2.2, while studying large deviation for colored random graphs.

The following definition introduces the rate function for this large deviation principle.

Definition 3.1. Let $\mathcal{S} \subset \mathbb{P}(\mathbb{N}_0)$ denote the set of all probability measures μ such that $\bar{\mu} = \sum_{i=1}^{\infty} i \mu(i) < \infty$. Set the function $I_\beta : \mathbb{P}(\mathbb{N}_0) \mapsto [0, \infty]$ to be $+\infty$ if $\mu \notin \mathcal{S}$, and for $\mu \in \mathcal{S}$ set

$$\begin{aligned} I_\beta(\mu) &:= \sum_{i=0}^{\infty} \mu(i) \log\bigl(i! \mu(i)\bigr) - \frac{\overline{\mu}}{2} \log(\overline{\mu} \beta) + \frac{\overline{\mu} + \beta}{2} \\ &= D(\mu \parallel p_\beta) + \frac{1}{2}(\overline{\mu} - \beta) + \frac{\overline{\mu}}{2} \log \beta - \frac{\overline{\mu}}{2} \log \overline{\mu} \end{aligned}$$

where $D(\cdot \parallel \cdot)$ is the Kullback Leibler divergence, and p_β is the Poisson distribution with parameter β.

The following large deviation follows from [2,8].

Theorem 3.2. *If G is an Erdös–Renyi random graph with parameter β/n, then μ_n^G satisfies a large deviation principle on $\mathbb{P}(\mathbb{N}_0)$ with respect to weak topology, with speed n and the good rate function $I_\beta(\cdot)$.*

A direct application of the above large deviations result can be used to prove that

$$\lim_{n\to\infty} \frac{1}{n} Z_n(\beta, f) = \sup_{\mu\in\mathcal{S}} \{\mu[f] - I_\beta(\mu)\}$$

when f is a bounded function. We now state three lemmas which will be used to extend this to all functions satisfying the conditions of Theorem 1.4.

Lemma 3.3. *For any function $f : \mathbb{N}_0 \mapsto \mathbb{R}$ satisfying (3) and any set $B \subset \mathbb{P}(\mathbb{N}_0)$ one has*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{\sum_{i=0}^{n-1} h_i(G) f(i)} 1\{\mu_n(G) \in B\} \le \sup_{\mu\in B\cap\mathcal{S}} \{\mu[f] - I_\beta(\mu)\}.$$

Lemma 3.4. *For finite positive real α and $f : \mathbb{N}_0 \mapsto \mathbb{R}$ satisfying (3) we have*

$$\sup_{\mu: I_\beta(\mu) - \mu[f] \le \alpha} \sum_{i=0}^{\infty} i \log i \mu(i) \le C,$$

where $C = C(\alpha, f, \beta)$ is a finite positive constant.

Lemma 3.5. *Let $f : \mathbb{N}_0 \mapsto \mathbb{R}$ satisfy (3).*

(a) *We have*

$$\sup_{\mu\in\mathcal{S}} \{\mu[f] - I_\beta(\mu)\} = J(\beta, f),$$

where $J(\beta, f)$ is as defined in (4). The supremum in this definition is finite, and is attained over a finite set of positive reals $\{u_1, \dots, u_k\}$. Further, any optimizing u satisfies the relation $u = \sqrt{\beta \overline{\sigma}_{u,f}}$.

(b) *For any $\varepsilon > 0$ and ψ satisfying (1) we have*

$$\sup_{\mu\in U^c} \{\mu[f] - I_\beta(\mu)\} < \sup_{\mu\in\mathcal{S}} \{\mu[f] - I_\beta(\mu)\},$$

where $U := \{\mu \in \mathbb{P}(\mathbb{N}_0) : \min_{1\le l\le k} |\mu(\psi) - \sigma_{u_l}(\psi)| < \varepsilon\}$.

3.1. Proofs of Theorem 1.4, Corollary 1.5, and Theorem 1.7

We now complete the proof of the main results of this paper, deferring the proof of the lemmas stated above to Section 3.2.

Proof of Theorem 1.4.

(a) To begin note that

$$e^{Z_n(\beta,f)} = \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{\sum_{i=0}^{n-1} h_i(G)f(i)}, \tag{6}$$

which on taking log, dividing by n and letting $n \to \infty$ along with Lemma 3.3 gives

$$\limsup_{n\to\infty} \frac{1}{n} Z_n(\beta, f) \le \sup_{\mu\in\mathcal{S}} \big\{\mu[f] - I_\beta(\mu)\big\},$$

and so we have verified the upper bound. The proof of the lower bound is split into two cases, depending on whether we are in case (i) or case (ii).

(i) Define a function $T : \mathbb{P}(\mathbb{N}_0) \mapsto \mathbb{R}$ by

$$\begin{aligned} T(\mu) &= \mu[f] \quad \text{if } I_\beta(\mu) < \infty, \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

and use Lemma 3.4 to note that $I_\beta(\mu) < \infty$ implies $\sum_{i=0}^{\infty} i \log i \mu(i) < \infty$. Also, since f satisfies (1), there exists $C_0 < \infty$ such that $|f(i)| \le C_0 i \log i$ for all $i \ge 0$. This immediately gives

$$\big|T(\mu)\big| = \big|\mu(f)\big| \le C_0 \sum_{i=0}^{\infty} i \log i \mu(i) < \infty.$$

Also for every $m \ge 1$ define the function $T_m : \mathbb{P}(\mathbb{N}_0) \mapsto \mathbb{R}$ by setting $T_m(\mu) = \sum_{i=0}^{m} f(i)\mu(i)$, and note that T_m is continuous with respect to weak topology. We claim that for every positive real α and $\delta > 0$ we have

$$\lim_{m\to\infty} \sup_{\mu: I_\beta(\mu)\le\alpha} \big|T_m(\mu) - T(\mu)\big| = 0. \tag{7}$$

To see this, fixing $\delta > 0$ and invoking (1) we have $|f(i)| \le \delta i \log i$ for all $i > M(\delta)$. Thus for all $m \ge M(\delta)$ we have

$$\begin{aligned} \big|T_m(\mu) - T(\mu)\big| &= \left|\sum_{i=m+1}^{\infty} f(i)\mu(i)\right| \le \delta \sum_{i=M(\delta)+1}^{\infty} i \log i \mu(i) \\ &\le \delta \sum_{i=0}^{\infty} i \log i \mu(i) \le \delta C(\alpha, 0, \beta), \end{aligned}$$

where the existence of $C(\alpha, 0, \beta)$ follows from invoking Lemma 3.4 with $f \equiv 0$. Since $\delta > 0$ is arbitrary, this verifies (7).

We further claim that for every $\delta > 0$ we have

$$\lim_{m\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_{n,\beta}\big(\big|T_m\big(\mu_n^G\big) - T\big(\mu_n^G\big)\big| > \delta\big) = -\infty. \tag{8}$$

Indeed, with $\psi(i) := \sqrt{i \log i |f(i)|}$ we have

$$\big|f(i)\big| \ll \psi(i) \ll i \log i,$$

and so there exists $M(\delta)$ such that for all $i \geq M(\delta)$ we have $|f(i)| \leq \delta^2\psi(i)$. Thus for $m \geq M(\delta)$ we have

$$\big|T_m(\mu) - T(\mu)\big| \leq \delta^2 \sum_{i=M(\delta)+1} \psi(i)\mu(i) \leq \delta^2\mu[\psi],$$

and so Markov's inequality gives

$$\mathbb{P}_{n,\beta}\big(\big|T_m\big(\mu_n^G\big) - T\big(\mu_n^G\big)\big| > \delta\big) \leq \mathbb{P}\left(\mu_n^G[\psi] > \frac{1}{\delta}\right) \leq e^{-\frac{n}{\delta}} \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{n\mu_n^G[\psi]}.$$

On taking log, dividing by n and letting $n \to \infty$ along with Lemma 3.3 we get

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_{n,\beta}\big(\big|T_m\big(\mu_n^G\big) - T\big(\mu_n^G\big)\big| > \delta\big) \leq -\frac{1}{\delta} + J(\psi, \beta).$$

Since $J(\psi, \beta)$ is finite and δ is arbitrary, (8) follows.

Given (7) and (8), it follows by [7], Theorem 4.2.23, and Theorem 3.2 that $T(\mu_n^G) = \mu_n^G[f]$ satisfies a large deviation principle on $\mathbb{R}$ with the good rate function

$$\widetilde{I}(x) := \inf_{\mu\in\mathbb{P}(\mathbb{N}_0):\mu[f]=x} I_\beta(\mu).$$

Also Lemma 3.3 gives

$$\begin{aligned}\frac{1}{n} \log \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{2n\sum_{i=0}^{n-1} h_i(G) f(i)} &= \frac{1}{n} \log \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{2nT(\mu_n^G)} \\ &\leq \sup_{\mu\in\mathcal{S}} \big\{2\mu[f] - I_\beta(\mu)\big\} = J(\beta, 2f).\end{aligned}$$

The right hand side above is finite by part (a) of Lemma 3.5, as $2f(\cdot)$ satisfies (1). This verifies [7], (4.3.3.), with $\gamma = 2$, and so by [7], Theorem 4.3.1, with $\phi(x) = x$ we have

$$\begin{aligned}\frac{1}{n} \log \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{n\sum_{i=0}^{n-1} h_i(G) f(i)} &= \frac{1}{n} \log \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{nT(\mu_n^G)} = \sup_{x\in\mathbb{R}} \big\{\theta x - \widetilde{I}(x)\big\} \\ &= \sup_{\mu\in\mathcal{S}} \big\{\mu(f) - I_\beta(\mu)\big\} = J(\beta, f),\end{aligned}$$

where the last equality again uses part (a) of Lemma 3.5. This completes the proof of part (a).

(ii) Fixing $m \in \mathbb{N}$ one has

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_{n,\beta}} e^{n\mu_n^G[f]} &\geq \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{n\mu_n^G[f]} 1_{\max_{1\leq j\leq n} d_j(G)\leq m} \\ &= \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{\sum_{i=0}^m h_i(G) f(i)} 1_{\max_{1\leq j\leq n} d_j(G)\leq m} \\ &\geq \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{\sum_{i=0}^m h_i(G) f(i) + f(m)\sum_{i=m+1}^{n-1} h_i(G)} 1_{\max_{1\leq j\leq n} d_j(G)\leq m},\end{aligned}$$

where the last inequality uses the fact that $f(i) \leq f(0) = 0$ for all i, as f is non-increasing. This, along with an application of FKG inequality [10], Prop. 1, gives

$$\mathbb{E}_{\mathbb{P}_{n,\beta}} e^{n\mu_n^G[f]} \geq \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{\sum_{i=0}^m h_i(G) f(i) + f(m)\sum_{i=m+1}^{n-1} h_i(G)} \mathbb{P}_{n,\beta}(d_1 \leq m)^n,$$

where we use the fact that the function

$$G \mapsto \sum_{i=0}^{n-1} h_i(G) \tilde{f}_m(i), \qquad \tilde{f}_m(i) = \max\bigl(f(i), f(m)\bigr)$$

is non-increasing on the space of graphs $\mathcal{G}_n$, as $\tilde{f}_m$ is non-increasing. Since $f(0) \leq \tilde{f}_m(i) \leq f(m)$, it follows that $\tilde{f}_m$ is bounded, and so an application of part (i) gives

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{n\mu_n^G[\tilde{f}_m]} &\geq \sup_{\mu\in\mathcal{S}} \left\{ \sum_{i=0}^{\infty} \mu(i) \tilde{f}_m(i) - I_\beta(\mu) \right\} + \log p_\beta[0, m] \\ &\geq \sup_{\mu\in\mathcal{S}} \left\{ \sum_{i=0}^{\infty} \mu(i) f(i) - I_\beta(\mu) \right\} + \log p_\beta[0, m],\end{aligned}$$

where the last inequality uses the fact that $\tilde{f}_m \geq f$, and $p_\beta[0, m]$ is the probability that a Poisson random variable with parameter β is at most m. The lower bound follows on letting $m \to \infty$ and noting that $p_\beta[0, m] \to 1$. Combining the upper and lower bound gives

$$\lim_{n\to\infty} \frac{1}{n} Z_n(\beta, f) = \sup_{\mu\in\mathcal{S}} \bigl\{\mu[f] - I(\mu)\bigr\}.$$

(b) By part (a) of Lemma 3.5, the supremum in the right-hand side above is finite and equals $J(\beta, f)$ of (4), and the set of optimizing u in the definition of $J(\beta, f)$ has finite cardinality. Denoting this set by $\{u_1, u_2, \ldots, u_k\}$, let

$$U := \left\{ \mu \in \mathbb{P}(\mathbb{N}_0) : \min_{i=1}^{k} \bigl|\mu[\psi] - \sigma_{u_i, f}[\psi]\bigr| < \varepsilon \right\},$$

where $\varepsilon > 0$ is fixed. Thus we have

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{Q}_{n,\beta,f}\big(\mu_n^G \in U^c\big)\\
&\quad \le \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{n\mu_n^G[f]} 1_{\mu_n^G \in U^c} - \liminf_{n\to\infty} \frac{1}{n} \log Z_n(\beta, f)\\
&\quad \le \sup_{\mu \in U^c \cap \mathcal{S}} \big\{\mu[f] - I_\beta(\mu)\big\} - \sup_{\mu \in \mathcal{S}} \big\{\mu[f] - I_\beta(\mu)\big\},
\end{aligned}$$

where the last line uses Lemma 3.3 with $B = U^c$, and part (a). The last quantity above is negative by part (b) of Lemma 3.5, and so the conclusion follows. □

Proof of Corollary 1.5. Part(a) follows trivially from part (a) of Theorem 1.4 on noting that the function $\theta f(\cdot)$ satisfies (1) for all $\theta \in \mathbb{R}$. The continuity of the limiting log partition function follows from the fact that limit of convex functions is convex.

For part (b), setting $m := \frac{1}{4} \min_{i=1}^k \bar{\sigma}_{u_i,\theta f}$, $M := \max_{i=1}^k \bar{\sigma}_{u_i,\theta f}$, the desired conclusion follows from part (b) of Theorem 1.4. □

Proof of Theorem 1.7.

(a) Since f is non-decreasing and $\theta < 0$, the function θf is non-increasing and non-positive, and so an application of Theorem 1.4 proves part (a).

(b) It suffices to show that $\mathbb{Q}_{n,\beta,\theta f}(G = K_n)$ converges to 1, as the desired conclusion about the log normalizing constant immediately follows.

To this effect, using (5) there exists $M > 4$ such that $f(i) - f(i-1) \ge M \log i$ for all $i \ge k_n := \lfloor n/2 \rfloor$, for all n large enough. We now claim that for all $r \in [0, n-1]$ we have

$$f(n-1) - f(n-1-r) \ge \frac{1}{4} Mr \log n \tag{9}$$

Indeed, if $r \le k_n$, then we have

$$\begin{aligned}
f(n-1) - f(n-1-r) &= \sum_{i=1}^r \big(f(n-i) - f(n-i-1)\big)\\
&\ge M \sum_{i=1}^r \log(n-i) \ge Mr \log(n/2).
\end{aligned} \tag{10}$$

On the other hand if $r > k_n$, using the monotonicity of f along with (9) gives

$$f(n-1) - f(n-1-r) \ge f(n-1) - f(n-1-k_n) \ge Mk_n \log(n/2). \tag{11}$$

Combining (10) and (11), (9) follows.

Thus if $G \in \mathcal{G}_n$ is a graph with degree sequence $(d_1(G), \dots, d_n(G))$, then setting $r_j(G) := n-1-d_j(G)$ for $G \in \mathcal{G}_n$ we have

$$\sum_{j=1}^n f(n-1) - \sum_{j=1}^n f\big(d_j(G)\big) \geq \frac{M \log n}{4} \sum_{j=1}^n r_j(G) = \frac{M \log n}{2}\bigg(\frac{n(n-1)}{2} - E(G)\bigg),$$

which immediately gives

$$\frac{\mathbb{Q}_{n,\beta,\theta f}(G)}{\mathbb{Q}_{n,\beta,\theta f}(K_n)} \leq \bigg(\frac{n}{\beta}\bigg)^{R(G)} e^{-\frac{MR\log n}{2}R(G)},$$

where $R(G) := \binom{n}{2} - E(G)$ for $G \in \mathcal{G}_n$. This on summing gives

$$\mathbb{Q}_{n,\beta,\theta f}\big(R(G) \geq 1\big) \leq \sum_{r=1}^{\binom{n}{2}} \left(\binom{\binom{n}{2}}{r}\right) e^{-\frac{Mr\log n}{2}} \leq \sum_{r=1}^{\binom{n}{2}} n^{2r} e^{-\frac{Mr\log n}{2}} \leq \sum_{r=1}^{\infty} \big(n^2 e^{-\frac{M\log n}{2}}\big)^r,$$

which converges to 0 as $n \to \infty$, as $M > 4$. □

3.2. Proofs of Lemmas 3.3, 3.4 and 3.5

Proof of Lemma 3.3. Let $\mathcal{H}_n$ to be the set of all degree frequency vectors $h(G) = (h_0(G), \dots, h_{n-1}(G))$ on n vertices as the graph G varies in $\mathcal{G}_n$. Fixing $\delta > 0$ arbitrary we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_{n,\beta}} e^{\sum_{i=0}^{n-1} h_i(G) f(i)} 1\big\{\mu_n(G) \in B\big\} \\
&\quad = \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{\sum_{i=0}^{n-1} h_i(G) f(i)} 1\big\{\mu_n(G) \in B, E(G) \leq \delta a_n\big\} \\
&\qquad + \mathbb{E}_{\mathbb{P}_{n,\beta}} e^{\sum_{i=0}^{n-1} h_i(G) f(i)} 1\big\{\mu_n^G \in B, E(G) > \delta a_n\big\},
\end{aligned} \tag{12}$$

By (1) there exists $N = N(\delta)$ such that $f(i) \leq \frac{\delta}{4} i \log i$ for all $i > N(\delta)$, and so with $M := M(\delta) = \max_{0 \leq i \leq N} f(i)$ and $a_n := \frac{n(n-1)}{2}$ the second term in the right-hand side of (12) can be bounded by

$$\begin{aligned}
e^{nM + \frac{\delta}{4} n^2 \log n} \mathbb{P}_{n,\beta}\big(E(G) > \delta a_n\big) &\leq e^{nM + \frac{\delta}{4} n^2 \log n} \sum_{r=\delta a_n}^{a_n} \binom{a_n}{r} \bigg(\frac{\beta}{n}\bigg)^r \\
&\leq e^{nM + \frac{\delta}{4} n^2 \log n} \times a_n 2^{a_n} \bigg(\frac{\beta}{n}\bigg)^{\delta a_n},
\end{aligned}$$

which on taking log, dividing by n, and letting $n \to \infty$ gives $-\infty$, and so we can ignore this term. The first term on the right-hand side of (12) can be written as

$$= \sum_{\mathbf{h}\in\mathcal{H}_n} N_n(\mathbf{h}) e^{\sum_{i=0}^{n-1} h_i(G) f(i)} \left(\frac{\beta}{n}\right)^{E(G)} \left(1-\frac{\beta}{n}\right)^{\binom{n}{2}-E(G)} 1\{\mu_n^G \in B, E(G) \le \delta a_n\},$$

where $N_n(\mathbf{h})$ is the number of labeled graphs in $\mathcal{G}_n$ whose degree frequency vector is $\mathbf{h}$. It follows from [17] that

$$N_n(\mathbf{h}) \le \frac{(2r)!}{r! 2^r \prod_{i=0}^{n-1} i!^{h_i}} \times \frac{n!}{\prod_{i=0}^{n-1} h_i!},$$

where the extra factor $\frac{n!}{\prod_{i=0}^{n-1} h_i!}$ accounts for the fact that for labeled graphs any relabeling between vertices with the same degree needs to be taken into account. Thus one has the following bound on the first term of the right-hand side of (12):

$$\begin{aligned} &\mathbb{E}_{\mathbb{P}_{n,\beta}} e^{\sum_{i=0}^{n-1} h_i(G) f(i)} 1\{\mu_n(G) \in B, E(G) \le \delta a_n\} \\ &\quad \le \sum_{\mathbf{h}\in\mathcal{H}_n} \overline{N}_n(\mathbf{h}) 1\{\mu_n(G) \in B, E(G) \le \delta a_n\}, \end{aligned} \tag{13}$$

where

$$\overline{N}_n(\mathbf{h}) := e^{\sum_{i=0}^{n-1} h_i f(i)} \left(\frac{\beta}{n}\right)^{r} \left(1-\frac{\beta}{n}\right)^{\binom{n}{2}-r} \frac{(2r)!}{r! 2^r \prod_{i=0}^{n-1} i!^{h_i}} \times \frac{n!}{\prod_{i=0}^{n-1} h_i!}$$

with $r = \sum_{i=0}^{n-1} i h_i$. Using Stirling's approximation one has

$$C_2 e^{-n} n^{n+1/2} \le n! \le C_1 e^{-n} n^{n+1/2}$$

for all $n \ge 1$, for some positive constants C_1, C_2 free of n. Using this, a direct computation gives

$$\overline{N}_n(\mathbf{h}) \le e^{n(1+o_n(1))\{\mu_n^G[f] - I_\beta(\mu_n^G)\}},$$

which along with (13) gives

$$\begin{aligned} &\sum_{\mathbf{h}\in\mathcal{H}_n} \overline{N}_n(\mathbf{h}) 1\{\mu_n(G) \in B, E(G) \le \delta a_n\} \\ &\quad \le e^{n(1+o_n(1)) \sup_{\mu\in B\cap\mathcal{S}}\{\mu[f] - I_\beta(\mu)\}} \sum_{r=0}^{\delta a_n} \left| \mathbf{h} \in \mathcal{H}_n : \sum_{i=0}^{n-1} i h_i = r \right|. \end{aligned} \tag{14}$$

Letting $p(r)$ denote the number of un-ordered partitions of r, we have the upper bound

$$\left| \mathbf{h} \in \mathcal{H}_n : \sum_{i=0}^{n-1} i h_i = r \right| \le p(2r).$$

This is because given any such degree frequency vector $\mathbf{h}$, the corresponding unordered degree sequence sums up to $2r$, and so one can get a partition of $2r$ by dropping the vertices with degree 0. Since

$$\lim_{r\to\infty}\frac{1}{\sqrt{r}}\log p(r)=\pi\sqrt{\frac{2}{3}},$$

(for a proof of this classical result see [9] or [13]), taking logs, dividing by n and taking $n\to\infty$ gives

$$\limsup_{n\to\infty}\frac{1}{n}\log\left(\sum_{r=0}^{\delta a_n}\left|\mathbf{h}\in\mathcal{H}_n:\sum_{i=0}^{n-1}ih_i=r\right|\right)\le\pi\sqrt{\frac{2\delta}{3}}.$$

This, along with (12) and (14) gives

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{E}_{\mathbb{P}_{n,\beta}}e^{\sum_{i=0}^{n-1}h_i(G)f(i)}1\{\mu_n^G\in B\}\le\sup_{\mu\in B\cap\mathcal{S}}\{\mu[f]-I(\mu)\}+\pi\sqrt{\frac{2\delta}{3}},$$

from which the desired conclusion follows since $\delta>0$ is arbitrary. □

Proof of Lemma 3.4. Since $\log(i!)=\sum_{k=1}^{i}\log k\ge\int_{x=0}^{i}\log x\,dx=i\log i-i$, we have

$$\sum_{i=0}^{\infty}\log(i!)\mu(i)\ge\sum_{i=0}^{\infty}i\log i\mu(i)-\overline{\mu}\tag{15}$$

Also define $\sigma\in\mathcal{S}$ by $\sigma(i):=2^{-(i+1)}$ for $i\in\mathbb{N}_0$ and note that

$$\sum_{i=0}^{\infty}\mu(i)\log\mu(i)=D(\mu\parallel\sigma)+\sum_{i=0}^{\infty}\mu(i)\log\sigma(i)\ge-(\overline{\mu}+1)\log 2.\tag{16}$$

Finally by (1) there exists $M<\infty$ such that $f(i)\le M+\frac{1}{4}i\log i$ for all $i\ge 0$. This gives

$$\sum_{i=0}^{\infty}i\log i\mu(i)-\mu[f]\ge-M+\frac{3}{4}\sum_{i=0}^{\infty}i\log i\mu(i)\ge-M+\frac{3}{4}\bar{\mu}\log\bar{\mu},\tag{17}$$

where the last step uses Jensen's inequality. Combining (15), (16) and (17) gives

$$\begin{aligned}I_\beta(\mu)-\mu[f]&=\sum_{i=0}^{\infty}\log(i!)\mu(i)-\mu[f]+\sum_{i=0}^{\infty}\mu(i)\log\mu(i)-\frac{\bar{\mu}}{2}\log(\bar{\mu}\beta)+\frac{\bar{\mu}+\beta}{2}\\&\ge\sum_{i=0}^{\infty}i\log i\mu(i)-\bar{\mu}-\mu[f]-(\bar{\mu}+1)\log 2-\frac{\bar{\mu}}{2}\log(\bar{\mu}\beta)+\frac{\bar{\mu}+\beta}{2}\\&\ge-M+\frac{3}{4}\bar{\mu}\log\bar{\mu}-\bar{\mu}-(\bar{\mu}+1)\log 2-\frac{\bar{\mu}}{2}\log(\bar{\mu}\beta)+\frac{\bar{\mu}+\beta}{2}\end{aligned}$$

$$= -M + \frac{1}{4}\bar{\mu}\log\bar{\mu} - \overline{\mu}\left(\log 2 + \frac{3+\log\beta}{2}\right) + \frac{\beta}{2} - \log 2 = \phi_1(\overline{\mu}), \qquad (18)$$

where

$$\phi_1(x) := -M + \frac{1}{4}x\log x - x\left(\log 2 + \frac{3+\log\beta}{2}\right) + \frac{\beta}{2} - \log 2.$$

Since $\phi_1(x)$ is continuous and diverges to ∞ as $x \to \infty$, it follows that $\phi_1(\overline{\mu}) \le \alpha$ implies $\overline{\mu} \le K(\alpha)$ for some $K(\alpha) < \infty$. Thus, we have

$$\begin{aligned}
\alpha \ge I_\beta(\mu) - \mu[f] &= \sum_{i=0}^{\infty}\log(i!)\mu(i) - \mu[f] + \sum_{i=0}^{\infty}\mu(i)\log\mu(i) - \frac{\bar{\mu}}{2}\log(\bar{\mu}\beta) + \frac{\bar{\mu}+\beta}{2} \\
&\ge \sum_{i=0}^{\infty} i\log i\,\mu(i) - \bar{\mu} - \mu[f] - (\bar{\mu}+1)\log 2 - \frac{\bar{\mu}}{2}\log(\bar{\mu}\beta) + \frac{\bar{\mu}+\beta}{2} \\
&\ge -M + \frac{3}{4}\sum_{i=0}^{\infty} i\log i\,\mu(i) - \bar{\mu} - (\bar{\mu}+1)\log 2 - \frac{\bar{\mu}}{2}\log(\bar{\mu}\beta) + \frac{\bar{\mu}+\beta}{2} \\
&= \frac{3}{4}\sum_{i=0}^{\infty} i\log i\,\mu(i) - \phi_2(x),
\end{aligned}$$

where $\phi_2(x) := M + x + (x+1)\log 2 + \frac{x}{2}\log(x\beta) - \frac{x+\beta}{2}$. Thus, we have

$$\frac{3}{4}\sup_{\mu: I_\beta(\mu)-\mu[f]\le\alpha}\sum_{i=0}^{\infty} i\log i\,\mu(i) \le \alpha + \sup_{0\le x\le K(\alpha)}\phi_2(x),$$

from which the conclusion of the lemma follows. □

Proof of Lemma 3.5.

(a) It suffices to consider the minimization of $\mu \mapsto \{I_\beta(\mu) - \mu[f]\}$ over $\mathcal{S}$. To this effect, first note that

$$\alpha := \inf_{\mu\in\mathcal{S}}\left\{I_\beta(\delta_0) - \delta_0[f]\right\} + 1 < \infty.$$

Indeed, taking $\mu = \delta_0$ gives

$$I_\beta(\delta_0) - \delta_0[f] = \beta/2 - f(0) < \infty.$$

Thus it suffices to minimize $\mu \mapsto \{I_\beta(\mu) - \mu[f]\}$ over the set $B_\alpha := \{\mu : I_\beta(\mu) - \mu[f] \le \alpha\}$. By Lemma 3.4 we have

$$\sup_{\mu\in B_\alpha}\sum_{i=0}^{\infty} i\log i\,\mu(i) \le C(\alpha) < \infty, \qquad (19)$$

and so by Markov's inequality the set B_α is tight with respect to weak topology. Let $\{\nu_k\}_{k\ge 1}$ be a sequence of measures in B_α such that

$$\lim_{k\to\infty}\big\{I_\beta(\nu_k)-\nu_k[f]\big\}=\inf_{\mu\in B_\alpha\cap U^c}\big\{I_\beta(\mu)-\mu[f]\big\}.$$

Then by tightness of B_α, there exists a subsequence which converges weakly to ν, say. Without loss of generality, assume the original sequence $\{\nu_k\}_{k\ge 1}$ converges weakly to ν. Since f satisfies (1), invoking uniform integrability implied by (19) it follows that $\nu_k(f)$ converges to $\nu(f)$. This, along with the observation that $I_\beta(\cdot)$ is lower semi continuous gives

$$\inf_{\mu\in\mathcal{S}}\big\{I_\beta(\mu)-\mu[f]\big\}=\lim_{k\to\infty}\big\{I_\beta(\nu_k)-\nu_k[f]\big\}\ge\big\{I_\beta(\nu)-\nu[f]\big\},$$

and so ν attains the infimum. Let $A\subset\mathcal{S}$ be the set of all probability measures where the infimum is attained. Then for any $\mu\in\mathcal{S}$ and $\nu\in A$, by convexity of $\mathcal{S}$ we have $(1-t)\nu+t\mu\in\mathcal{S}$ for any $t\in[0,1]$. Thus with $u:=\sqrt{\overline{\mu}\beta}$ we have

$$\begin{aligned}
&\frac{\partial}{\partial t}\big[I_\beta\big((1-t)\nu+t\mu\big)-(1-t)\nu[f]-t\mu[f]\big]_{t=0}\ge 0\\
\Leftrightarrow\quad&\sum_{i=0}^{\infty}\bigg(1+\log\nu(i)+\log i!-\frac{i}{2}(1+\log\overline{\nu})-\frac{i}{2}\log\beta+\frac{i}{2}-f(i)\bigg)\\
&\quad\times\big(\mu(i)-\nu(i)\big)\ge 0\\
\Leftrightarrow\quad&\sum_{i=0}^{\infty}\big(\log\nu(i)+\log i!-i\log u-f(i)\big)\big(\mu(i)-\nu(i)\big)\ge 0\\
\Leftrightarrow\quad&D(\nu\parallel\sigma_{u,f})+D(\mu\parallel\nu)\le D(\mu\parallel\sigma_{u,f}).
\end{aligned}$$

where $\sigma_{u,f}$ is as defined in definition (1.3). Since this holds for all $\mu\in\mathcal{S}$, setting $\mu=\sigma_{u,f}$ gives $D(\sigma_{u,f}\parallel\nu)=0$, and so $\nu=\sigma_{u,f}$. Thus $A\subset\Omega_f$, and consequently

$$\sup_{\mu\in\mathcal{S}}\big\{\mu[f]-I_\beta(\mu)\big\}=\sup_{u\ge 0}\big\{\sigma_{u,f}[f]-I_\beta(\sigma_{u,f})\big\}=J(\beta,f),$$

where the last equality follows by a simple algebra. It also follows from the proof that any $\sigma_{u,f}\in A$ must satisfy $u=\sqrt{\beta\overline{\sigma_{u,f}}}$.

Finally, to solve the optimization $u\mapsto\phi_1(u):=\sigma_{u,f}[f]-I_\beta(\sigma_{u,f})$ over $u\ge 0$, differentiating with respect to u gives

$$\phi_1'(u)=-m'(u,f)\log\frac{u}{\sqrt{m(u,f)\beta}}.$$

Also setting $\phi_2(u) := \sum_{i=0}^{\infty} \frac{e^{f(i)}}{i!} u^i = e^{Z(u,f)}$ we have $m(u, f) = u\frac{\phi_2'(u)}{\phi_2(u)}$, which on differentiating with respect to u gives

$$m'(u, f) = \frac{\phi_2(u)\phi_2'(u) + u\phi_2(u)\phi_2''(u) - u\phi_2'(u)^2}{\phi_2(u)^2},$$

and so

$$\lim_{u\to 0} m'(u, f) = \lim_{u\to 0} \frac{m(u, f)}{u} = \frac{\phi_2'(0)}{\phi_2(0)} = e^{f(1)-f(0)} > 0.$$

This gives $\lim_{u\to 0} \phi_1'(u) = +\infty$, and so $u = 0$ is not a local maxima of $\phi_1(\cdot)$. Also it follows from (19) that optimizing measure μ satisfies

$$\bar{\mu} \log \bar{\mu} \leq \sum_{i=0}^{\infty} i \log i \mu(i) \leq C(\alpha),$$

and so $m(u, f) \leq C'$ for some finite constant C'. This along with the relation $u^2 = m(u, f)\beta$ implies any optimizing u is at most $\sqrt{C'\beta}$. Thus, denoting $\widetilde{A}$ denote the subset of all positive reals u which are global maximizers of the function $u \mapsto \phi_1(u)$, it follows that the set $\widetilde{A}$ is compact. Since an analytic non constant function on a bounded domain cannot have infinitely many minimizers, the set $\widetilde{A}$ must have finite cardinality. This completes the proof of part (a).

(b) If $\inf_{\mu\in U^c}\{I_\beta(\mu) - \mu[f]\} = \infty$ then there is nothing to show. Assuming that

$$\alpha' := \inf_{\mu\in U^c}\bigl\{I_\beta(\mu) - \mu[f]\bigr\} + 1 < \infty,$$

it suffices to minimize $\mu \mapsto \{I_\beta(\mu) - \mu[f]\}$ over $B_{\alpha'} \cap U^c$. Letting $\{\nu_k\}_{k\geq 1}$ be a sequence of measures in $B_{\alpha'} \cap U^c$ such that

$$\lim_{k\to\infty}\bigl\{I_\beta(\nu_k) - \nu_k[f]\bigr\} = \inf_{\mu\in B_{\alpha'}\cap U^c}\bigl\{I_\beta(\mu) - \mu[f]\bigr\},$$

by a similar tightness and uniform integrability argument as in part (a) it follows that there exists a measure $\nu \in \mathcal{S}$ such that $\{\nu_k\}_{k\geq 1}$ converges to ν weakly, and

$$\lim_{k\to\infty} \nu_k(f) = \nu(f), \qquad \lim_{k\to\infty} \nu_k[\psi] = \nu[\psi].$$

Since $\nu_k \in U^c$ and $\nu_k(\psi)$ converges to $\nu(\psi)$, we have $\nu \in U^c$. Since U contains all the global minimizers of $\mu \mapsto \{I_\beta(\mu) - \mu[f]\}$, we have

$$\begin{aligned}\inf_{\mu\in U^c}\bigl\{I_\beta(\mu) - \mu[f]\bigr\} &= \lim_{k\to\infty}\bigl\{I_\beta(\nu_k) - \nu_k(f)\bigr\} \quad \bigl[\text{By choice of } \{\nu_k\}_{k\geq 1}\bigr]\\ &\geq I_\beta(\nu) - \nu[f] \quad \bigl[\text{By lower semi continuity of } I_\beta(\cdot)\bigr]\\ &> \inf_{\mu\in\mathcal{S}}\bigl\{I_\beta(\mu) - \mu[f]\bigr\},\end{aligned}$$

where the last step uses the fact that $\nu \in U^c$ is not in a global minimizer of $\mu \mapsto \{I_\beta(\mu) - \mu[f]\}$. This completes the proof of part (b). □

3.3. Proof of Theorems 1.9 and 1.11

Proof of Theorem 1.9. Differentiating with respect to θ, $\log u$, c and eliminating c gives the least square equations

$$\theta \sum_{i=0}^{L} \big(f(i) - \bar{f}\big)^2 + \log u \sum_{i=0}^{L} \Big(i - \frac{L}{2}\Big)\big(f(i) - \bar{f}\big) = \sum_{i=0}^{L} \big(f(i) - \bar{f}\big) \log \frac{i!h_i(G)}{n}$$

$$\theta \sum_{i=0}^{L} \Big(i - \frac{L}{2}\Big)\big(f(i) - \bar{f}\big) + \log u \sum_{i=0}^{L} \Big(i - \frac{L}{2}\Big)^2 = \sum_{i=0}^{L} \Big(i - \frac{L}{2}\Big) \log \frac{i!h_i(G)}{n},$$

where $\bar{f} := \frac{1}{L+1} \sum_{i=0}^{L} f(i)$. Thus, we have the following matrix equation for the least square estimates:

$$(\hat{\theta}_n, \log \hat{u}_n) A = \left[\sum_{i=0}^{L} \big(f(i) - \bar{f}\big) \log \frac{i!h_i(G)}{n}, \sum_{i=0}^{L} \Big(i - \frac{L}{2}\Big) \log \frac{i!h_i(G)}{n} \right], \tag{20}$$

where A is a 2×2 matrix defined by

$$A =: \begin{bmatrix} \sum_{i=0}^{L} \big(f(i) - \bar{f}\big)^2 & \sum_{i=0}^{L} \Big(i - \frac{L}{2}\Big)\big(f(i) - \bar{f}\big) \\ \sum_{i=0}^{L} \Big(i - \frac{L}{2}\Big)\big(f(i) - \bar{f}\big) & \sum_{i=0}^{L} \Big(i - \frac{L}{2}\Big)^2 \end{bmatrix}.$$

Now, by part (b) of Theorem 1.4 it follows that there exists a finite set $\{u_1, u_2, \dots u_k\}$ with $u_l > 0$ such that any limit point of the measure μ_n^G is of the form $\sigma_{u_l, \theta f}$ for some l, $1 \le l \le k$. This implies that there exists a random variable U_n taking values in $\{u_1, u_2, \dots, u_k\}$ such that for all $i \in [0, L]$ we have

$$\frac{h_i(G)}{n} - \frac{U_n^i}{i!} e^{\theta f(i) - Z(U_n, \theta f)} = o_P(1). \tag{21}$$

Plugging this estimate, Slutsky's theorem implies

$$\sum_{i=0}^{L} \big(f(i) - \bar{f}\big) \log \frac{i!h_i(G)}{n} = \theta \sum_{i=0}^{L} \big(f(i) - \bar{f}\big)^2 + \log U_n \sum_{i=0}^{L} i\big(f(i) - \bar{f}\big) + o_P(1),$$

$$\sum_{i=0}^{L} \Big(i - \frac{L}{2}\Big) \log \frac{i!h_i(G)}{n} = \theta \sum_{i=0}^{L} \Big(i - \frac{L}{2}\Big) f(i) + \log U_n \sum_{i=0}^{L} \Big(i - \frac{L}{2}\Big) i + o_P(1),$$

which along with (20) gives

$$(\hat{\theta}_n - \theta, \log \hat{u}_n - \log U_n)A = o_P(1). \tag{22}$$

We now claim that the minimum eigenvalue $\lambda_{\min}(A)$ is not 0. Deferring the proof of the claim, let us first complete the proof of the theorem. Given this claim, (22) implies

$$\|\hat{\theta}_n - \theta, \log \hat{u}_n - \log U_n\|_2 \le \frac{1}{\lambda_{\min}(A)} \left\|(\hat{\theta}_n - \theta, \log \hat{u}_n - \log U_n)A\right\|_2 = o_P(1),$$

thus proving that $\hat{\theta}_n$ is consistent for θ, and $\hat{u}_n = U_n + o_P(1)$.

Since part (b) of Theorem 1.4 with $\psi(i) = i$ implies

$$m(U_n, \theta f) - \frac{2E(G)}{n} = o_P(1),$$

we have

$$\frac{n\hat{u_n^2}}{2E(G)} = \frac{U_n^2}{m(U_n, \theta f)} + o_P(1) = \beta + o_P(1).$$

where the last equality invokes the relation $U_n^2 = \beta m(U_n, \theta f)$. This shows consistency of $\hat{\beta}_n$ for β as well.

It thus remains to verify the claim that $\lambda_{\min}(A)$ is not 0. To see this, note that if $\lambda_{\min}(A) = 0$, then $|A| = 0$, which gives

$$\sum_{i=0}^{L}(f(i) - \bar{f})^2 \sum_{i=0}^{L}\left(i - \frac{L}{2}\right)^2 = \left[\sum_{i=0}^{L}\left(i - \frac{L}{2}\right)(f(i) - \bar{f})\right]^2.$$

Thus equality holds in the Cauchy–Schwarz inequality, which implies $f(i) - \bar{f} = b(i - \frac{L}{2})$ for some $b \in \mathbb{R}$. But then $f(0) = 0$ forces $f(i) = bi$ for all $i \ge 0$, a contradiction. This completes the proof of the theorem. □

Proof of Theorem 1.11. As before there exists a random variable U_n taking values in a finite set $\{u_1, \dots, u_k\}$ such that (21) holds, which gives

$$\frac{i!h_i(G)}{h_0(G)} - e^{f(i)} U_n^i = o_P(1).$$

On taking log and using the definition of $\hat{f}_n(i)$ as in the theorem, this gives

$$\hat{f}_n(i) + i \log \hat{u}_n = \log\left[\frac{i!h_i(G)}{h_0(G)}\right] = f(i) + i \log U_n + o_P(1).$$

Thus to complete the proof it suffices to show that $\hat{u}_n - U_n = o_P(1)$. To prove this, first note that part (b) of Theorem 1.7 gives

$$\frac{2E(G)}{n} - m(U_n, f) = o_P(1).$$

Since one has $U_n^2 = m(U_n, f)$ as well, it readily follows that

$$\hat{u}_n = \sqrt{\frac{2E(G)}{n}} = \sqrt{m(U_n, f)} + o_P(1) = U_n + o_P(1),$$

thus completing the proof of the theorem. □

Acknowledgements

The content of this paper also appears in my Ph.D. thesis under the guidance of Prof. Persi Diaconis. I would also like to thank Prof. Sourav Chatterjee and Prof. Amir Dembo for helpful discussions on this paper. This paper also benefitted from some helpful discussions during an AIM conference on ERGMs held in June 2013, for which I would like to thank all participants of the conference. The presentation of this paper has also benefitted from the recommendations of an anonymous reviewer, and the Associate Editor. Research Partially Supported by NSF Grant DMS-1712037.

References

[1] Bhamidi, S., Bresler, G. and Sly, A. (2011). Mixing time of exponential random graphs. *Ann. Appl. Probab.* **21** 2146–2170. MR2895412 https://doi.org/10.1214/10-AAP740

[2] Bordenave, C. and Caputo, P. (2015). Large deviations of empirical neighborhood distribution in sparse random graphs. *Probab. Theory Related Fields* **163** 149–222. MR3405616 https://doi.org/10.1007/s00440-014-0590-8

[3] Chatterjee, S. and Diaconis, P. (2013). Estimating and understanding exponential random graph models. *Ann. Statist.* **41** 2428–2461. MR3127871 https://doi.org/10.1214/13-AOS1155

[4] Chatterjee, S., Diaconis, P. and Sly, A. (2011). Random graphs with a given degree sequence. *Ann. Appl. Probab.* **21** 1400–1435. MR2857452 https://doi.org/10.1214/10-AAP728

[5] Chatterjee, S. and Varadhan, S.R.S. (2011). The large deviation principle for the Erdős–Rényi random graph. *European J. Combin.* **32** 1000–1017. MR2825532 https://doi.org/10.1016/j.ejc.2011.03.014

[6] Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158. MR0365798 https://doi.org/10.1214/aop/1176996454

[7] Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*, 2nd ed. *Applications of Mathematics* (*New York*) **38**. New York: Springer. MR1619036 https://doi.org/10.1007/978-1-4612-5320-4

[8] Doku-Amponsah, K. and Mörters, P. (2010). Large deviation principles for empirical measures of colored random graphs. *Ann. Appl. Probab.* **20** 1989–2021. MR2759726 https://doi.org/10.1214/09-AAP647

[9] Erdös, P. (1942). On an elementary proof of some asymptotic formulas in the theory of partitions. *Ann. of Math.* (2) **43** 437–450. MR0006749 https://doi.org/10.2307/1968802

[10] Fortuin, C.M., Kasteleyn, P.W. and Ginibre, J. (1971). Correlation inequalities on some partially ordered sets. *Comm. Math. Phys.* **22** 89–103. MR0309498

[11] Frank, O. and Strauss, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.* **81** 832–842. MR0860518

[12] Handcock, M. Assessing degeneracy in statistical models of social networks. Technical Report, Center for Statistics and Social Sciences, University of Washington, Seattle, WA.

[13] Hardy, G.H. and Ramanujan, S. (1918). Asymptotic formulaae in combinatory analysis. *Proc. London Math. Soc.* (2) **17** 75–115. MR1575586 https://doi.org/10.1112/plms/s2-17.1.75

[14] 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. MR0608176

[15] Hunter, D.R. and Handcock, M.S. (2006). Inference in curved exponential family models for networks. *J. Comput. Graph. Statist.* **15** 565–583. MR2291264 https://doi.org/10.1198/106186006X133069

[16] Krivitsky, P.N. and Kolaczyk, E.D. (2015). On the question of effective sample size in network modeling: An asymptotic inquiry. *Statist. Sci.* **30** 184–198. MR3353102 https://doi.org/10.1214/14-STS502

[17] McKay, B.D. (1985). Asymptotics for symmetric 0–1 matrices with prescribed row sums. *Ars Combin.* **19** 15–25. MR0790916

[18] Morris, M., Handcock, M.S. and Hunter, D.R. (2008). Specification of exponential-family random graph models: Terms and computational aspects. *J. Stat. Softw.* **24** 1548–7660.

[19] Radin, C. and Yin, M. (2013). Phase transitions in exponential random graphs. *Ann. Appl. Probab.* **23** 2458–2471. MR3127941 https://doi.org/10.1214/12-AAP907

[20] Ruelle, D. (1999). *Statistical Mechanics: Rigorous Results*. River Edge, NJ: World Scientific Co., Inc.; Imperial College Press, London. Reprint of the 1989 edition. MR1747792 https://doi.org/10.1142/4090

[21] Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *J. Amer. Statist. Assoc.* **106** 1361–1370. MR2896841 https://doi.org/10.1198/jasa.2011.tm10747

[22] Schweinberger, M. and Handcock, M.S. (2015). Local dependence in random graph models: Characterization, properties and statistical inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 647–676. MR3351449 https://doi.org/10.1111/rssb.12081

[23] Shalizi, C.R. and Rinaldo, A. (2013). Consistency under sampling of exponential random graph models. *Ann. Statist.* **41** 508–535. MR3099112 https://doi.org/10.1214/12-AOS1044

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

[25] Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge Univ. Press.

[26] Watts, D.J. and Strogatz, S. (1998). Collective dynamics of "small-world" networks. *Nature* **393** 440–442.

[27] Yin, M., Rinaldo, A. and Fadnavis, S. (2016). Asymptotic quantization of exponential random graphs. *Ann. Appl. Probab.* **26** 3251–3285. MR3582803 https://doi.org/10.1214/16-AAP1175

Received July 2018 and revised April 2019