THE DEGREE SEQUENCE OF A RANDOM GRAPH
I. THE MODELS

Brendan D. McKay+
Department of Computer Science
Australian National University
Canberra, ACT 0200, Australia
bdm@cs.anu.edu.au

Nicholas C. Wormald*
Department of Mathematics
University of Melbourne
Parkville, Vic 3052, Australia
nick@maths.mu.oz.au

Abstract.

We show that the joint distribution of the degrees of a random graph can be accurately approximated by several simpler models derived from a set of independent binomial distributions. On the one hand we consider the distribution of degree sequences of random graphs with n vertices and $\frac{1}{2}m$ edges. For a wide range of values of m, this distribution is almost everywhere in close correspondence with the conditional distribution $\{(X_1, \ldots, X_n) \mid \sum X_i = m\}$ where $X_1, \ldots, X_n$ are independent random variables, each having the same binomial distribution as the degree of one vertex. We also consider random graphs with n vertices and edge probability p. For a wide range of functions $p = p(n)$, the distribution of the degree sequence can be approximated by $\{(X_1, \ldots, X_n) \mid \sum X_i \text{ is even}\}$, where $X_1, \ldots, X_n$ are independent random variables each having the distribution $\mathrm{Binom}(n-1, p')$, where p' is itself a random variable with a particular truncated normal distribution. To facilitate computations, we demonstrate techniques by which statistics in this model can be inferred from those in a simple model of independent binomial random variables. Where they apply, the accuracy of our method is sufficient to determine asymptotically all probabilities greater than n^{-k} for any fixed k. In this first paper, we use the geometric mean of the degrees as a tutorial example. In the second paper, we will determine the asymptotic distribution of the t-th largest degree for all functions $t = t(n)$ as $n \to \infty$.

1. Introduction.

The distribution of the degrees of the vertices in a random graph on n vertices is one of the fundamental objects of study in random graph theory. In the commonly studied random graph model $\mathcal{G}(n,p)$, where edges occur independently and each with probability p, then the degree of a vertex is distributed binomially with mean $p(n-1)$. But the degrees of the vertices are not independent of each other, and it is this which makes the joint distribution of degrees interesting. Bollobás [2] and Palka [7] devote chapters to this topic.

Many results obtained so far on the degree sequence of a random graph concern the distribution of the t'th largest degree occurring in $G \in \mathcal{G}(n,p)$. In most results so far t is bounded;

+ Supported in part by Australian Research Council grant 90SARC1045

* Supported by the Australian Research Council

only quite weak bounds have been obtained for more general t. The main tool used for these results is the method of moments. There are some exceptions; for instance, Barbour et al. [1] use a more sophisticated approach to find the distribution of the number of vertices of degree exactly k or at least k. In this paper we give a model of the degree sequence of a random graph in $\mathcal{G}(n,p)$ which gives an entirely new access to this topic and permits computations for a vastly wider range of results, such as the distribution of the t-th largest degree for *all* t. As an example of such results, to appear in a subsequent paper, consider the distribution of the median of the degrees of the vertices in $\mathcal{G}(n,p)$ for constant p $(0 < p < 1)$. In the case of odd n, we will show that this is distributed asymptotically as a discretised normal with constant variance. That is, the probability that it equals j is

$$\sqrt{\frac{k}{\pi}}\int_j^{j+1} e^{-k(x-pn-c)^2}\,dx + o(1)$$

for functions $k = k(p)$ and $c = c(p)$ which we determine. Comparison with a sequence of independent binomial variables is interesting: if we assume each vertex degree has its true binomial distribution, but they are distributed independently of each other, then the median has a similar distribution but with different variance (i.e. different k).

Let the probability $p = p(n)$ satisfy $0 < p < 1$ for all n. Define $q = q(n)$ by $q = 1 - p$. The notation $\omega(n)$ refers to any function which tends to ∞ as $n \to \infty$, possibly a different such function at each appearance.

Our method rests on our earlier asymptotic enumeration results concerning graphs with a given degree sequence. These are incomplete in the sense that they do not translate to all values of p, but they do suffice for $p = o(1/\sqrt{n})$ and also for p approximately constant. For this reason we present our results for what we call *acceptable* values of p. These are values of p for which a certain asymptotic formula is valid. Whenever p is acceptable, our method will be valid. Our earlier results show that p is acceptable if either $\omega(n)\log n/n^2 \le pq \le o(n^{-1/2})$, or if $pq \ge c/\log n$ for some $c > \frac{2}{3}$. We conjecture that in fact only the condition $pq = \omega(n)\log n/n^2$ is needed.

In presenting our results, we define several probability spaces of integer vectors, to be used as stepping stones in calculating probabilities in the probability space of degree sequences. We call these spaces *models* since they are all approximations, of varying degrees of accuracy, to the space of degree sequences. Much of the present paper gives general relationships between these models. One of the models is a set of independent identically distributed binomial variables restricted to having even sum. Properties of such a distribution do not seem to have been studied before, and we do not have a very general result which gives a transparent relationship between this model and the others we consider. Thus, in order to demonstrate that our general theory is useful, a substantial part of the present paper is devoted to developing a technique for translating results concerning variables in a space of independent binomials to the corresponding variables in the restricted even-sum space.

In Section 2 we define the various models under consideration, and state the result (Theorem 2.5) which quantifies the relationship between our main model and the degree sequence of a random graph in $\mathcal{G}(n,p)$. The proof of this depends upon information about the relationship between our models, which we give in Section 3. Up to this point, all models are based on sequences with even sum. To complete any actual applications, we need to translate to a model without the even sum restriction, which we do in Section 4. Apart from the simple example presented in the final section, we defer detailed applications of our method to the second paper in this series [6].

2. Description of the models

In this section we will define our models and state their elementary properties. We will also give our main theorems.

For integer $n \geq 1$, define $N = \binom{n}{2}$ and $I_n = \{0, 1, \ldots, n-1\}^n$. A vector $\boldsymbol{d} \in I_n$ has components $d_1, \ldots, d_n$. Define $m = m(\boldsymbol{d}) = \sum_{i=1}^n d_i$, $\bar{d} = m/n$, $\lambda = \bar{d}/(n-1)$ and $\gamma_2 = (n-1)^{-2} \sum_{i=1}^n (d_i - \bar{d})^2$. Define $E_n = \{\boldsymbol{d} \in I_n \mid m(\boldsymbol{d}) \text{ is even}\}$ and, for $0 \leq m \leq 2N$, $I_{n,m} = \{\boldsymbol{d} \in I_n \mid m(\boldsymbol{d}) = m\}$.

If $\mathcal{M}$ is one of our models, $\mathrm{P}_{\mathcal{M}}(A)$ is the probability of event A in model $\mathcal{M}$, and $\mathrm{E}_{\mathcal{M}}(X)$ are $\mathrm{Var}_{\mathcal{M}}(X)$ are the expectation and variance of random variable X in model $\mathcal{M}$.

Binomial models $\mathcal{B}_p$ and $\mathcal{B}_m$.
The models $\mathcal{B}_p = \mathcal{B}_p(n)$ and $\mathcal{B}_m = \mathcal{B}_m(n)$ have domains I_n and $I_{n,m}$, respectively. For $\mathcal{B}_p$, the components d_i are independently distributed according to the binomial distribution $\mathrm{Binom}(n-1, p)$. *Model $\mathcal{B}_m$ is the restriction of $\mathcal{B}_p$ to $I_{n,m}$. Note that $\mathcal{B}_m$ is independent of p.*

Lemma 2.1.

$$\mathrm{P}_{\mathcal{B}_p}(\boldsymbol{d}) = p^m q^{2N-m} \prod_{i=1}^n \binom{n-1}{d_i}$$

and

$$\mathrm{P}_{\mathcal{B}_m}(\boldsymbol{d}) = \binom{2N}{m}^{-1} \prod_{i=1}^n \binom{n-1}{d_i}$$

for $\boldsymbol{d} \in I_n$ and $\boldsymbol{d} \in I_{n,m}$, respectively. ∎

Even-sum binomial models $\mathcal{E}_p$ and $\mathcal{E}'_p$.
Models $\mathcal{E}_p = \mathcal{E}_p(n)$ and $\mathcal{E}'_p = \mathcal{E}'_p(n)$ both have domain E_n. $\mathcal{E}_p$ is just the restriction of $\mathcal{B}_p$ to E_n. $\mathcal{E}'_p$ is constructed from $\mathcal{E}_p$ by weighting each $\boldsymbol{d} \in E_n$ with a weight depending only on m, such that the value $\frac{1}{2}m$ has the distribution $\mathrm{Binom}(N, p)$.

Lemma 2.2. *For $\boldsymbol{d} \in E_n$,*

$$\mathrm{P}_{\mathcal{E}_p}(\boldsymbol{d}) = \left(\tfrac{1}{2} + \tfrac{1}{2}(q-p)^{2N}\right)^{-1} \mathrm{P}_{\mathcal{B}_p}(\boldsymbol{d})$$

and

$$\mathrm{P}_{\mathcal{E}'_p}(\boldsymbol{d}) = \binom{2N}{m}^{-1} \binom{N}{m/2} p^{m/2} q^{N-m/2} \prod_{i=1}^n \binom{n-1}{d_i}.$$

Proof. The first claim follows from the fact that $\mathrm{E}_{\mathcal{B}_p}(\xi^m) = (p\xi + q)^{2N}$, and so $\mathrm{P}_{\mathcal{B}_p}(E_n) = \frac{1}{2}(\mathrm{E}_{\mathcal{B}_p}(1^m) + \mathrm{E}_{\mathcal{B}_p}((-1)^m) = \frac{1}{2} + \frac{1}{2}(q-p)^{2N}$. The second claim follows from

$$\mathrm{P}_{\mathcal{E}_p}(m(\boldsymbol{d}) = m) = \left(\tfrac{1}{2} + \tfrac{1}{2}(q-p)^{2N}\right)^{-1} \binom{2N}{m} p^m q^{2N-m}. \quad ∎$$

Integrated model $\mathcal{I}_p$.
The model $\mathcal{I}_p = \mathcal{I}_p(n)$ has domain E_n. To generate a variate $\boldsymbol{d}$ in this model, first choose a value p' from the normal distribution with mean p and variance $pq/(2N)$, truncated to the unit interval $(0,1)$. Then generate a variate with the distribution of $\mathcal{E}_{p'}$.

Lemma 2.3. *Define*

$$K_p(p') = \sqrt{\frac{N}{\pi pq}} \exp\Bigl(-\frac{(p-p')^2 N}{pq}\Bigr)$$

and

$$V(p) = \int_0^1 K_p(p')\, dp'.$$

Then

$$\mathrm{P}_{\mathcal{I}_p}(\boldsymbol{d}) = \frac{2}{V(p)\bigl(1-(q-p)^{2N}\bigr)} \prod_{i=1}^{n} \binom{n-1}{d_i} \int_0^1 K_p(p')(p')^m(1-p')^{2N-m}\, dp'$$

for each $\boldsymbol{d} \in E_n$. ∎

We will show that $\mathcal{I}_p$ and $\mathcal{E}'_p$ are almost the same. The advantage of $\mathcal{I}_p$ is that, despite its apparent complexity, it allows easy transfer of computations from $\mathcal{E}_p$. In most cases we will have $V(p) \sim 1$, as will be shown later.

Lemma 2.4. *Let X be a random variable defined on E_n. Then*

$$\mathrm{E}_{\mathcal{I}_p}(X) = V(p)^{-1} \int_0^1 K_p(x) \mathrm{E}_{\mathcal{E}_x}(X)\, dx$$

and

$$\mathrm{Var}_{\mathcal{I}_p}(X) = V(p)^{-1} \int_0^1 K_p(x)\bigl(\mathrm{Var}_{\mathcal{E}_x}(X) + (\mathrm{E}_{\mathcal{E}_x}(X) - \mathrm{E}_{\mathcal{I}_p}(X))^2\bigr)\, dx. \quad ∎$$

Graph models $\mathcal{D}_p$ and $\mathcal{D}_m$.
The models $\mathcal{D}_p = \mathcal{D}_p(n)$ and $\mathcal{D}_m = \mathcal{D}_m(n)$ have domains E_n and $I_{n,m}$, respectively, where m is even. A variate in $\mathcal{D}_p$ is the sequence of degrees of a random graph with n vertices, generated by selecting each edge independently with probability p. $\mathcal{D}_m$ is the restriction of $\mathcal{D}_p$ to $I_{n,m}$ (but is independent of p); it corresponds to the degree sequences of random graphs with n vertices and $\frac{1}{2}m$ edges, with each such graph being equally likely.

We will call the probability $p = p(n)$ *acceptable* if there is a set-valued function $R_p(n) \subseteq E_n$ and a real function $\delta(n) \to 0$ such that the following conditions are satisfied. Recall that $\omega(n)$ can be any function with $\omega(n) \to \infty$.

A0. $pqN = \omega(n) \log n$.

A1. For each $\boldsymbol{d} \in R_p(n)$ there is a number $\delta_{\boldsymbol{d}}$ such that $|\delta_{\boldsymbol{d}}| \le \delta(n)$ and

$$\mathrm{P}_{\mathcal{D}_p}(\boldsymbol{d}) = \mathrm{P}_{\mathcal{E}'_p}(\boldsymbol{d}) \exp\Bigl(\frac{1}{4} - \frac{\gamma_2^2}{4\lambda^2(1-\lambda)^2} + \delta_{\boldsymbol{d}}\Bigr).$$

A2. In each of the models $\mathcal{E}_p$ and $\mathcal{D}_p$, we have $\mathrm{P}\bigl(R_p(n)\bigr) = 1 - n^{-\omega(n)}$.

Note that condition A0 is just that $\mathrm{Var}_{\mathcal{B}_p}(m) \to \infty$.

At the expense of a slight abuse of notation, we will also call a function $m = m(n)$ *acceptable* if $p = m/(2N)$ is acceptable and m is always an even integer.

Theorem 2.5. *The function $p(n)$ is acceptable if one of the following conditions holds.*
(i) $\omega(n)\log n/n^2 \le pq \le o(n^{-1/2})$;
(ii) $pq \ge c/\log n$ *for some* $c > \frac{2}{3}$.

Proof. For case (i) with $p \le q$, take $R_p(n)$ to consist of all $\boldsymbol{d} \in E_n$ such that $\frac{1}{2}pn \le \bar{d} \le \frac{3}{2}pn$ and $\max_i d_i - \min_i d_i \le p^{1/2}n^{1/2+\epsilon}$ (where $\epsilon > 0$ is sufficiently small). The validity of condition A1 follows from [5], while condition A2 follows by applying Lemma 3.3 (following) to $\bar{d}$ and to each d_i.

For case (i) with $p > q$, simply interchange p with q and d_i with $n-1-d_i$ for all i.

For case (ii), take $R_p(n)$ to consist of all $\boldsymbol{d} \in E_n$ such that $|d_i - np| \le n^{1/2+\epsilon}$, for sufficiently small $\epsilon > 0$ and all i. Condition A1 is proved in [4], and condition A2 follows by applying Lemma 3.3 to each d_i. ∎

We conjectured in [5] that Theorem 2.5 can be considerably strengthened. A corollary of our conjecture is the following.

Conjecture. *$p(n)$ is acceptable whenever condition A0 holds.* ∎

Our principal result is that $\mathcal{D}_m$ and $\mathcal{B}_m$ are closely related for acceptable m, as are $\mathcal{D}_p$, $\mathcal{E}'_p$ and $\mathcal{I}_p$, for acceptable p.

Theorem 2.6. *For $n \ge 1$, let $X_n : E_n \to \mathbb{S}$ be a random variable, where $\mathbb{S}$ is a linear space with a norm.*
(a) If $p = p(n)$ is acceptable, then

$$\|\mathrm{E}_{\mathcal{D}_p}(X_n) - \mathrm{E}_{\mathcal{E}'_p}(X_n)\| = o(1)\mathrm{E}_{\mathcal{E}'_p}(\|X_n\|) + n^{-\omega(n)} \max_{\boldsymbol{d}\in E_n} \|X_n(\boldsymbol{d})\|$$

and

$$\begin{aligned}\|\mathrm{E}_{\mathcal{D}_p}(X_n) - \mathrm{E}_{\mathcal{I}_p}(X_n)\| &= o(1)\mathrm{E}_{\mathcal{I}_p}(\|X_n\|)\\ &\quad + O\bigl(n^{-\omega(n)} + \exp(-\epsilon(n)(pqN)^{1/3})\bigr) \max_{\boldsymbol{d}\in E_n} \|X_n(\boldsymbol{d})\|,\end{aligned}$$

where $\epsilon(n)$ is any function with $\epsilon(n) \to 0$.
(b) If $m = m(n)$ is acceptable, then

$$\|\mathrm{E}_{\mathcal{D}_m}(X_n) - \mathrm{E}_{\mathcal{B}_m}(X_n)\| = o(1)\mathrm{E}_{\mathcal{B}_m}(\|X_n\|) + n^{-\omega(n)} \max_{\boldsymbol{d}\in I_{n,m}} \|X_n(\boldsymbol{d})\|.$$

Proof. The proof depends on results that we will not prove until the next section, but we give it here for convenience.

Suppose that $p = p(n)$ is acceptable. By conditions A1 and A2, and Corollary 3.5, $\mathcal{D}_p$ and $\mathcal{E}'_p$ agree within ratio $1 + o(1)$ except on an event of probability $n^{-\omega(n)}$ in both models. This gives the first equation of part (a); Theorem 3.6 then gives the second equation.

Part (b) follows in the same manner. ∎

Note that Theorem 2.6 can be applied immediately to probabilities of events, simply by considering the indicator functions of those events. In particular, applying the theorem to the indicator function the event $X_n \le x$ directly yields estimates for the distribution of X_n.

Theorem 2.6 can be applied to estimating variances, but a different approach gives good error terms more easily.

Theorem 2.7. *For $n \geq 1$, let $X_n : E_n \to \mathbb{R}$ be a random variable.*
(a) If $p = p(n)$ is acceptable, then

$$\mathrm{Var}_{\mathcal{D}_p}(X_n) = \mathrm{Var}_{\mathcal{E}'_p}(X_n)\big(1 + o(1)\big) + n^{-\omega(n)} \max_{\boldsymbol{d} \in E_n} X_n(\boldsymbol{d})^2$$

and

$$\mathrm{Var}_{\mathcal{D}_p}(X_n) = \mathrm{Var}_{\mathcal{I}_p}(X_n)\big(1 + o(1)\big) + O\big(n^{-\omega(n)} + \exp(-\epsilon(n)(pqN)^{1/3})\big) \max_{\boldsymbol{d} \in E_n} X_n(\boldsymbol{d})^2,$$

where $\epsilon(n)$ is any function with $\epsilon(n) \to 0$.
(b) If $m = m(n)$ is acceptable, then

$$\mathrm{Var}_{\mathcal{D}_m}(X_n) = \mathrm{Var}_{\mathcal{B}_m}(X_n)\big(1 + o(1)\big) + n^{-\omega(n)} \max_{\boldsymbol{d} \in I_{n,m}} X_n(\boldsymbol{d})^2.$$

Proof. Let Y_n be an independent copy of X_n. Then $\mathrm{Var}(X_n) = \frac{1}{2}\mathrm{E}\big((X_n - Y_n)^2\big)$. Now apply the same method as for Theorem 2.6. ∎

In the following sections, we show how, in many cases of practical interest, statistics in $\mathcal{E}'_p$ can be derived from those in $\mathcal{E}_p$, which in turn can be derived from those in $\mathcal{B}_p$. Meanwhile, the following weaker but more general theorem can be useful.

Theorem 2.8. *Let $p = p(n)$ be acceptable and, for each n, let $A_n \subseteq E_n$. If $\mathrm{P}_{\mathcal{B}_p}(A_n) \to 0$, then $\mathrm{P}_{\mathcal{D}_p}(A_n) \to 0$.*

Proof. In view of Theorem 2.6, it suffices to prove that $\mathrm{P}_{\mathcal{B}_p}(A_n) \to 0$ implies that $\mathrm{P}_{\mathcal{E}'_p}(A_n) \to 0$.

Define $y = y(n) = \min\{\sqrt{-\log \mathrm{P}_{\mathcal{B}_p}(A_n)}, (pqN)^{1/7}\}$ and $C_n = \big\{\boldsymbol{d} \in E_n \mid |m - 2pN| \leq y\sqrt{pqN}\big\}$. In $\mathcal{E}'_p$, $\frac{1}{2}m$ has the distribution $\mathrm{Binom}(N, p)$, so $\mathrm{P}_{\mathcal{E}'_p}(E_n - C_n) \to 0$. Also, by Lemmas 2.1 and 2.2, $\mathrm{P}_{\mathcal{E}'_p}(\boldsymbol{d})/\mathrm{P}_{\mathcal{B}_p}(\boldsymbol{d}) \leq \binom{N}{m/2}^{-1} p^{-m/2} q^{-N+m/2}$ since $\binom{N}{m/2}^2 \leq \binom{2N}{m}$, and so $\mathrm{P}_{\mathcal{E}'_p}(\boldsymbol{d}) = O(e^{y^2/4})\mathrm{P}_{\mathcal{B}_p}(\boldsymbol{d})$ uniformly for $\boldsymbol{d} \in C_n$, by the local normal approximation to the binomial distribution. Hence

$$\mathrm{P}_{\mathcal{E}'_p}(\boldsymbol{d}) \leq \mathrm{P}_{\mathcal{E}'_p}(E_n - C_n) + O(e^{y^2/4})\mathrm{P}_{\mathcal{B}_p}(A_n \cap C_n) \to 0. \quad ∎$$

If A_n is restricted to a limited number of $I_{n,m}$'s for each n, a more explicit form of Theorem 2.8 can be obtained using the inequality

$$\mathrm{P}_{\mathcal{E}'_p}(A_n \cap I_{n,m}) \leq \sqrt{\mathrm{P}_{\mathcal{B}_p}(A_n \cap I_{n,m})}\,.$$

3. Relationship between the models.

In the section we will analyse the relationships between most of our models. Corollary 3.5 relates $\mathcal{D}_p$ to $\mathcal{E}'_p$ and Theorem 3.6 relates $\mathcal{E}'_p$ to $\mathcal{I}_p$. Theorems 3.7 and 3.8 permit evaluation of the difference between $\mathcal{I}_p$ and $\mathcal{E}_p$. We leave the relationship between $\mathcal{E}_p$ and $\mathcal{B}_p$ to the next section. We begin with a few general bounds. The first has the same proof as Lemma 7 of [9].

Lemma 3.1. *Let $X_0, X_1, \ldots$ be a martingale such that, for all i, $|X_{i+1} - X_i| \le c$ with probability at least $1-r$, and $|X_{i+1} - X_i| \le K$ always. Then, for any $\lambda \ge 0$, $n \ge 0$ and $0 < p < 1$ we have*

$$\mathrm{P}\big(|X_n - X_0| > \lambda(c + Kp)n^{1/2} + nKp\big) < nr(1 + 1/p) + 2e^{-\lambda^2/2}. \quad \blacksquare$$

Corollary 3.2. *Under the conditions of Lemma 3.1, suppose $K \le cn^{O(1)}$ and $r = n^{-\omega(n)}$. Then*

$$\mathrm{P}\big(|X_n - X_0| > \omega(n)c\sqrt{n \log n}\,\big) = n^{-\omega(n)}.$$

Proof. Put $p = r^{1/2}$ and $\lambda = \omega(n)\sqrt{\log n}$. $\blacksquare$

Lemma 3.3. *Let B be a random variable with distribution* $\mathrm{Binom}(t, p)$*, where $p = p(t)$ is an arbitrary function such that $0 < p < 1$. Suppose*

$$\delta = \omega(t)\Big(\frac{\log t}{tpq} + \sqrt{\frac{\log t}{tpq}}\,\Big).$$

Then as $t \to \infty$,

$$\mathrm{P}\big(|B - tp| \ge \delta tpq\big) = t^{-\omega(t)},$$

where the two functions ω are different.

Proof. By Chernoff's bound,

$$\mathrm{P}(B - pt \ge \epsilon tpq) \le \exp\big(-tp(1 + q\epsilon)\log(1 + q\epsilon) - tq(1 - p\epsilon)\log(1 - p\epsilon)\big). \tag{3.1}$$

We now show that (3.1) is $t^{-\omega(t)}$ for some $\epsilon \le \delta$.

For this proof, we distinguish between various functions ω by using subscripts. Suppose firstly that $pq = \omega_1(t)\log t/t$. Then we can choose ϵ such that $\epsilon = o(1)$, $\epsilon \le \delta$ and $\epsilon = \omega_2(t)\sqrt{\log t/(tpq)}$. Now (3.1) becomes

$$\exp\big(-\tfrac{1}{2}tpq\epsilon^2 + O(tpq\epsilon^3)\big) = t^{-\omega(t)}.$$

Suppose on the other hand that $pq = O(\log t/t)$. Then $\delta \to \infty$, and so we are done if $p \ge \frac{1}{2}$. For $p < \frac{1}{2}$, we can choose ϵ so that $\epsilon \le \delta$, $p\epsilon = o(1)$, and $\epsilon = \omega_3(t)\log t/(pqt)$. Since $\epsilon \to \infty$, (3.1) is

$$\exp\big(-tp\epsilon(\log \epsilon + o(1))\big) = t^{-\omega(t)}.$$

By symmetry, $\mathrm{P}(B - pt \le pqt) = t^{-\omega(t)}$ as well. $\blacksquare$

Now we will show that the random variable γ_2 is sharply concentrated near just that value needed to make the argument of the exponential in condition A1 zero.

Theorem 3.4. *In the model $\mathcal{B}_p$, $\gamma_2 = \lambda(1 - \lambda)(1 + o(1))$ with probability*

(i) $1 - n^{-\omega(n)}$*, for $pqN = \omega(n)\log n$;*

(ii) $1 - O(p^2q^2n^3)$*, for all p.*

Proof. Note that $\boldsymbol{d} = (d_1, \ldots, d_n) \in I_n$, λ and γ_2 are defined at the start of Section 2. By symmetry, we can assume that $p \le \frac{1}{2}$. Put $t = n-1$. We have

$$\frac{\lambda(1-\lambda)-\gamma_2}{\lambda(1-\lambda)} = \frac{Y(\boldsymbol{d})}{\bar{d}(t-\bar{d})}, \tag{3.2}$$

where $Y(\boldsymbol{d}) = \bar{d}^2 t - \bar{d} - \sum_i (d_i - d_i^2)$. Define the martingale $X_0, X_1, \ldots$ by $X_0 = E\big(Y(\boldsymbol{d})\big)$ and $X_i = E\big(Y(\boldsymbol{d}) \bigm| d_1, \ldots, d_i\big)$ for $i \ge 1$. It is clear that $|X_{i+1} - X_i| \le 2n^2$ always. Suppose that

$$\delta = \frac{(\log t)^{3/2}}{tpq} + \frac{\log t}{\sqrt{tpq}}.$$

Then by Lemma 3.3 we have that $|d_i - tp| \le \delta tpq$ for $1 \le i \le n$ with probability $1 - n^{-\omega(n)}$. So assume that these inequalities hold and let $\epsilon = (d_{i+1} - tp)/(tpq)$. Given $d_1, \ldots, d_i$, we have $X_{i+1} - X_i = F(D + tp + \epsilon tpq, n-i-1) - F(D, n-i) + \epsilon tpq(1-1/n) + tpq - 2\epsilon p^2 q t^2 - \epsilon^2 p^2 q^2 t^2$, where $D = \sum_{j=1}^i d_j$ and $F(x,k) = tn^{-2}\mathrm{E}\big((x + \sum_{j=1}^k B_j)^2\big)$ with $B_1, \ldots, B_k$ being independent binomials distributed $\mathrm{Binom}(t,p)$. Since $\sum_{j=1}^k B_j$ is distributed $\mathrm{Binom}(tk, p)$, we find that $F(x,k) = tn^{-2}(x^2 + 2xktp + ktpq + k^2t^2p^2)$. This gives

$$X_{i+1} - X_i = 2\epsilon pqt^2(D - ipt)/n + tpq(1 - t/n^2) + \epsilon t^2 pq(q-p)/n - \epsilon^2 p^2 q^2 t^2.$$

By assumption, $D = itp + O(\delta pqt^2)$. Furthermore, $\delta pqt \to \infty$ and $\delta^2 pqt \to \infty$ by the definition of δ, and $|\epsilon| \le \delta$. Thus, $X_{i+1} - X_i = O\big((\log t)^3 + (\log t)^2 tpq\big)$, and so, by Corollary 3.2,

$$\mathrm{P}\big(|X_{i+1} - X_i| > (\log n)^4 n^{1/2} + (\log n)^3 n^{3/2} pq\big) = n^{-\omega(n)}.$$

Note also that $X_0 = O(pq)$. Hence, for $pq > (\log n)^{9/2}/n^{3/2}$, $Y(\boldsymbol{d}) = o(pqn^2)$ with probability $1 - n^{-\omega(n)}$. For the same range of p, since $n\bar{d}$ is distributed $\mathrm{Binom}(nt, p)$, Lemma 3.3 shows that $\bar{d}(t - \bar{d}) > \frac{1}{3}pqn^2$ with probability $1 - n^{-\omega(n)}$. This gives part (i) of the theorem for such p.

Now suppose that $p \le n^{-3/2+\epsilon}$ for $0 < \epsilon < \frac{1}{3}$ and $pN \to \infty$. It is easy to see that $\bar{d} = o(t)$ with probability $1 - n^{-\omega(n)}$, in which case (3.2) is $o(1)$ provided $\sum_{d_i \ge 2} d_i = o(n\bar{d})$. Let S be the value of the sum, and let n_2 be its number of terms. Suppose that $K = K(n)$ is some positive integer function. Then

$$\mathrm{P}(d_i \ge K) = \sum_{j=K}^{t} \binom{t}{j} p^j q^{t-j} \le 2\Big(\frac{enp}{K}\Big)^K$$

for sufficiently large n. For $K = 2$ we have more precisely $\mathrm{P}(d_i \ge 2) \le n^2p^2/2$, so

$$\mathrm{P}(n_2 \ge K) \le \binom{n}{K} (n^2p^2/2)^K \le \Big(\frac{en^3p^2}{2K}\Big)^K.$$

Combining these two bounds, we have

$$\mathrm{P}(S \ge K^3) \le 2\Big(\frac{enp}{K}\Big)^K + \Big(\frac{en^3p^2}{2K}\Big)^K \tag{3.3}$$

for sufficiently large n.

If $\omega(n) n^{-2} \log n \le p \le n^{-3/2+\epsilon}$, then $\mathrm{P}(n\bar{d} \le \frac{1}{2}pn^2) = n^{-\omega(n)}$ by Lemma 3.3. Putting $K = (pn^2)^{1/4}$, we find from (3.3) that $\mathrm{P}\big(S = o(n\bar{d})\big) = 1 - n^{-\omega(n)}$ as required.

The general bound in part (ii) is obtained by noting that $\gamma_2 = \lambda(1-\lambda)\big(1 + o(1)\big)$ if $n_2 = 0$. This occurs with probability $1 - O(p^2n^3)$. ∎

Corollary 3.5. *If $m = m(n)$ is acceptable, then $\gamma_2 = \lambda(1-\lambda)(1+o(1))$ with probability $1-n^{-\omega(n)}$ in model $\mathcal{D}_m$. Similarly in model $\mathcal{D}_p$ if $p = p(n)$ is acceptable.*

Proof. Since the coefficient of γ_2^2 in condition A1 is negative, any event which has probability $n^{-\omega(n)}$ in model $\mathcal{E}'_p$ also has probability $n^{-\omega(n)}$ in model $\mathcal{D}_p$. This gives the claim about $\mathcal{D}_p$. For $\mathcal{D}_m$, note that $\mathrm{P}_{\mathcal{B}_p}(I_{n,m}) = n^{-O(1)}$ for $p = m/(2N)$. ∎

Next, we show the close relationship between models $\mathcal{I}_p$ and $\mathcal{E}'_p$.

Theorem 3.6. *Suppose $pqN \to \infty$ and $y = o\big((pqN)^{1/6}\big)$. Then*

$$\mathrm{P}_{\mathcal{I}_p}(\boldsymbol{d}) = \mathrm{P}_{\mathcal{E}'_p}(\boldsymbol{d})\Big(1 + O\Big(\frac{1+|y|^3}{\sqrt{pqN}}\Big)\Big)$$

uniformly over all $\boldsymbol{d} \in E_n$ such that $|\frac{1}{2}m - pN| \le y\sqrt{pqN}$.

Proof. Since $\mathrm{P}_{\mathcal{I}_p}$ and $\mathrm{P}_{\mathcal{E}'_p}$ are exactly proportional for each fixed m, it will suffice to find conditions on m under which $\mathrm{P}_{\mathcal{I}_p}(I_{n,m}) \approx \mathrm{P}_{\mathcal{E}'_p}(I_{n,m})$, of which the exact value is $\binom{N}{m/2}p^{m/2}q^{N-m/2}$.

Since $V(p) = 1 + o(e^{-2pqN})$ (from the standard bound $1 - \Phi(x) = o(e^{-x^2/2})$ as $x \to \infty$) and $1 + (q-p)^{2N} = 1 + O(e^{-2pqN})$, we have from Lemma 2.3 that

$$\mathrm{P}_{\mathcal{I}_p}(I_{n,m}) = 2\binom{2N}{m}\big(1 + O(e^{-2pqN})\big)T(p,m),$$

where

$$T(p,m) = \int_0^1 K_p(p')(p')^m(1-p')^{2N-m}\,dp'.$$

Suppose that $m = 2pN + 2y\sqrt{pqN}$, where $y = o\big((pqN)^{1/6}\big)$. Change variable in the integral from p' to x, where $p' = p + (\frac{1}{2}y + x)\sqrt{pq/N}$. Define $z = \frac{1}{2}y + x$, $x_0 = -\sqrt{Np/q} - \frac{1}{2}y$ and $x_1 = \sqrt{Nq/p} - \frac{1}{2}y$. Then

$$T(p,m) = \frac{1}{\sqrt{\pi}}\int_{x_0}^{x_1} t(p,y,x)\,dx$$

where

$$\begin{aligned} t(p,y,x) = \exp\big(&-z^2 + 2(pN + y\sqrt{pqN}\,)\log(p + z\sqrt{pq/N}\,) \\ &+ 2(qN - y\sqrt{pqN}\,)\log(q - z\sqrt{pq/N}\,)\big). \end{aligned}$$

Suppose further that $x = o\big((pqN)^{1/2}\big)$. Then, by Taylor expansion,

$$t(p,y,x) = p^m q^{2N-m}\exp\Big(\tfrac{1}{2}y^2 - 2x^2 + O\Big(\frac{|y|^3 + |x|^3}{\sqrt{pqN}}\Big)\Big). \tag{3.4}$$

To bound $t(p,y,x)$ for larger x, note that the function $(pN + y\sqrt{pqN})\log(p + z\sqrt{pq/N}) + (qN - y\sqrt{pqN})\log(q - z\sqrt{pq/N})$ has its maximum with respect to z when $z = y$. Thus, using the calculation above,

$$t(p,y,x) \le p^m q^{2N-m}\exp\big(y^2 - (\tfrac{1}{2}y + x)^2 + o(1)\big) \tag{3.5}$$

for $x_0 \le x \le x_1$. Using (3.4) for $|x| \le (pqN)^{1/3}$ and (3.5) for $|x| > (pqN)^{1/3}$, we find

$$T(p,m) = e^{y^2/2} p^m q^{2N-m} \Big(\frac{1}{\sqrt{2}} + O\Big(\frac{1+|y|^3}{\sqrt{pqN}} \Big) \Big).$$

By Stirling's formula,

$$\binom{2N}{m} = \frac{1}{\sqrt{2}} \binom{N}{m/2} \Big(\frac{m}{2N} \Big)^{-m/2} \Big(\frac{2N-m}{N} \Big)^{-(N-m/2)} \Big(1 + O\Big(\frac{N}{m(2N-m)} \Big) \Big).$$

Application of the same argument that gave (3.4) now gives the theorem. ∎

As illustrated by Lemma 2.4, calculations in $\mathcal{E}_p$ can be carried into $\mathcal{I}_p$ by means of an integration. Usually the integrals can be approximated easily, because $K_p(x)$ is rather sharply concentrated near $x = p$. In the case where the integrands vary smoothly with x, the following may be useful. Recall from the proof of Theorem 3.6 that $V(p) \sim 1$ if $pqN \to \infty$.

Theorem 3.7. *Suppose $pqN \to \infty$. Let $y = y(n)$ be such that $y(n) \to \infty$ and $0 < y < \min\big(\sqrt{pN/q}, \sqrt{qN/p}\big)$. Let $t \ge 1$ be a fixed integer. Define the interval $I = [p - y\sqrt{pq/N}, p + y\sqrt{pq/N}]$, and the function $K_p(x)$ as in Lemma 2.3. Let $f(x) = f_n(x)$ be an integrable function from $(0,1)$ to $\mathbb{R}$ satisfying*

(i) $|f(x)| \le B_1(n)$ for $x \in (0,1)$,

(ii) $f(x), f'(x), \ldots, f^{(2t)}(x)$ exist and are continuous for $x \in I$,

(iii) $|f^{(2t)}(x)| \le B_2(n)$ for $x \in I$.

Define

$$\bar{f}(p) = \int_0^1 K_p(x) f(x)\, dx.$$

Then

$$\bar{f}(p) = \sum_{k=0}^{t-1} \frac{1}{k!} f^{(2k)}(p) \Big(\frac{pq}{4N} \Big)^k \big(1 + O(y^{2k-1} e^{-y^2/2})\big) + O\Big(B_2(n) \Big(\frac{pq}{4N} \Big)^t + B_1(n) y^{-1} e^{-y^2/2} \Big),$$

where the error terms are uniform over p and y for fixed t.

Proof. By Taylor's theorem we have, for $x \in I$,

$$f(x) = \sum_{k=0}^{2t-1} \frac{1}{k!} (x-p)^k f^{(k)}(p) + O\big((x-p)^{2t} B_2(n)\big),$$

and so

$$\begin{aligned} \bar{f}(p) &= \sum_{k=0}^{2t-1} \frac{1}{k!} f^{(k)}(p) \int_I K_p(x)(x-p)^k\, dx \\ &\quad + O\big(B_2(n)\big) \int_I K_p(x)(x-p)^{2t}\, dx + O\big(B_1(n)\big) \int_{(0,1)-I} K_p(x)\, dx. \end{aligned}$$

By symmetry, the odd terms in the summation are zero. To complete the proof, note that

$$\int_{-\infty}^{\infty} K_p(x)(x-p)^{2k} = \frac{(2k)!}{k!} \Big(\frac{pq}{4N} \Big)^k$$

and

$$\int_y^\infty x^{2k} e^{-x^2/2}\,dx < \frac{y^{2k} e^{-y^2/2}}{y - 2k/y},$$

provided $2k < y^2$. To prove the latter, note that $(y+u)^{2k} \le y^{2k} \exp(2ku/y)$ for $u, y \ge 0$ and apply the standard inequality $\int_y^\infty e^{-x^2/2}\,dx < e^{-y^2/2}/y$ for $y > 0$. ∎

A common occurence is that variables have asymptotic normality in $\mathcal{E}_p$. If the mean and variance vary smoothly with p, we are likely have normality in $\mathcal{I}_p$ as well, though perhaps with a different variance. Several of our intended applications will fit the following theorem.

Theorem 3.8. *Let X be a random variable defined on E_n. Suppose $pqN \to \infty$ and let $a = a(n)$, $y = y(n)$ and $\epsilon = \epsilon(n)$ be functions such that $y \to \infty$, $y < \min\bigl(\sqrt{pN/q}, \sqrt{qN/p}\bigr)$, and $\epsilon \to 0$. Suppose that for $|x - p| \le y\sqrt{pq/N}$ we have that X has an asymptotically normal distribution in $\mathcal{E}_x$ with mean $\mu(x)$ and variance $\sigma(x)^2$ in the sense that*

$$\mathrm{P}_{\mathcal{E}_x}(X \le t) = O(\epsilon) + \frac{1}{\sigma(x)\sqrt{2\pi}} \int_{-\infty}^t \exp\Bigl(-\frac{(\mu(x) - z)^2}{2\sigma(x)^2}\Bigr)\,dz \tag{3.6}$$

uniformly over x and t. Suppose further that if in addition $|t - \mu(x)| \le y\sigma(x)$ we have

$$\frac{t - \mu(x)}{\sigma(x)} = \frac{t - \mu(p)}{\sigma(p)} - \frac{a}{\sigma(p)}(x - p) + O(\epsilon) \tag{3.7}$$

uniformly over x and t. Then

$$\mathrm{P}_{\mathcal{I}_p}(X \le t) = O(\epsilon + e^{-y^2/2}) + \frac{1}{s\sqrt{2\pi}} \int_{-\infty}^t \exp\Bigl(-\frac{(\mu(p) - z)^2}{2s^2}\Bigr)\,dz,$$

uniformly over t, where

$$s^2 = \sigma^2(p) + \frac{a^2 pq}{2N}.$$

Proof. Write $\mu = \mu(p)$ and $\sigma = \sigma(p)$. For $|x - p| \le y\sqrt{pq/N}$, we have

$$\mathrm{P}_{\mathcal{E}_x}(X \le t) = O(\epsilon + e^{-y^2/2}) + \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^t \exp\Bigl(-\frac{(z - \mu - (x - p)a)^2}{2\sigma^2}\Bigr)\,dz \tag{3.8}$$

uniformly for all t. For $|t - \mu(x)| \le y\sigma(x)$, this is immediate from (3.6) and (3.7) using the simple fact that $\Phi(D + \delta) = \Phi(D) + O(\delta)$ uniformly for all D, δ. For $t < \mu(x) - y\sigma(x)$ and $t > \mu(x) - y\sigma(x)$, we have $\mathrm{P}_{\mathcal{E}_x}(X \le t) = O(\epsilon + e^{-y^2/2})$ and $\mathrm{P}_{\mathcal{E}_x}(X \le t) = 1 - O(\epsilon + e^{-y^2/2})$, respectively. These are easily seen to hold also for the right side of (3.8).

From Lemma 2.4,

$$\begin{aligned}
\mathrm{P}_{\mathcal{I}_p}(X \le t) &= V(p)^{-1} \sqrt{\frac{N}{\pi pq}} \int_0^1 \exp\Bigl(-\frac{(p - x)^2 N}{pq}\Bigr) \mathrm{P}_{\mathcal{E}_x}(X \le t)\,dx \\
&= O(\epsilon + e^{-y^2/2}) + \sqrt{\frac{N}{\pi pq}} \int_{p - y\sqrt{pq/N}}^{p + y\sqrt{pq/N}} \exp\Bigl(-\frac{(p - x)^2 N}{pq}\Bigr) \\
&\qquad \times \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^t \exp\Bigl(-\frac{(z - \mu - (x - p)a)^2}{2\sigma^2}\Bigr)\,dz\,dx.
\end{aligned}$$

The contraction of the limits on x is justified since $|\mathrm{P}_{\mathcal{E}_x}(X \le t)| \le 1$. After replacing $\mathrm{P}_{\mathcal{E}_x}(X \le t)$ by its approximation from (3.8), we note that the replacement is bounded by 1 for all x. This enables us to expand the limits of the outside integral to $(-\infty, \infty)$ while absorbing the difference in value into the error term. Interchanging the order of integration now allows the integration over x to be carried out, with the desired result. ∎

4. Relationship between $\mathcal{B}_p$ and $\mathcal{E}_p$; a new model $\mathcal{Q}_p$.

In this section we demonstrate a strong correspondence between the models $\mathcal{B}_p$ and $\mathcal{E}_p$ for functions which do not change too rapidly. Roughly speaking, the technique used is to cover the space $\mathcal{B}_p$ with positively weighted hypercubes, such that sum of the weights of the hypercubes containing a point is equal to the probability of that point. This permits us to take advantage of any equality or near equality of the values of a function on adjacent even and odd points.

We need a preliminary result before defining the weights. For $k \le n+1$, define

$$w(k,n,p) = (-1)^k(q-p)^n + \sum_{i=0}^{k-1} (-1)^{k-1-i} \binom{n}{i} p^i q^{n-i}.$$

Lemma 4.1.

(i) For $0 \le k \le n+1$,

$$w(k,n,p) = \sum_{i=k}^{n} (-1)^{k-i} \binom{n}{i} p^i q^{n-i}.$$

(ii) For $1 \le k \le n+1$, $w(k-1,n,p) + w(k,n,p) = \binom{n}{k-1} p^{k-1} q^{n-k+1}$.

(iii) For $p \le \frac{1}{2}$ *and* $1 \le k \le n+1$,

$$0 \le w(k,n,p) \le \min\left\{ \binom{n}{k-1} p^{k-1} q^{n-k+1}, \binom{n}{k} p^k q^{n-k} \right\}.$$

(iv)

$$\sum_{k=0}^{n} w(k,n,p) = \tfrac{1}{2} + \tfrac{1}{2}(q-p)^n.$$

Proof. We note from the binomial theorem that $\sum_{i=0}^{n} (-1)^i \binom{n}{i} p^i q^{n-i} = (q-p)^n$ and (i) follows. Part (ii) is immediate. It is possible to see that (iii) follows from Bonferroni's inequalities, but it is simpler give a direct verification. Let j denote the greatest i for which $\binom{n}{i} p^i q^{n-i} \ge \binom{n}{i-1} p^{i-1} q^{n-i+1}$. First take $k \le j$. Then since $0 \le (q-p)^n \le q^n$, the definition of $w(k,n,p)$ is an alternating sum of nonincreasing nonnegative terms, beginning with $\binom{n}{k-1} p^{k-1} q^{n-k-1} \le \binom{n}{k} p^k q^{n-k}$ (if $j \ge 1$), or is just $(q-p)^n < (1-p)^n$ (if $j = 0$). This gives (iii) in this case. For $k > j$, (iii) follows from (i) since then $\binom{n}{k} p^k q^{n-k} \le \binom{n}{k-1} p^{k-1} q^{n-k+1}$. Finally, (iv) follows from (i):

$$\sum_{k=0}^{n} w(k,n,p) = \sum_{k=0}^{n+1} w(k,n,p) = \sum_{i=0}^{n} \tfrac{1}{2}(1+(-1)^i) \binom{n}{i} p^i q^{n-i} = \tfrac{1}{2} + \tfrac{1}{2}(q-p)^n. \quad ∎$$

We require further notation to proceed. Define $w(k) = w(k, n-1, p)$ for short. For $\boldsymbol{x} \in I_n$ and $S \subseteq \{1, \ldots, n\}$, define

$$w_{\boldsymbol{x},S} = \prod_{i \in S} w(x_i) \prod_{i \notin S} \binom{n-1}{x_i} p^{x_i} q^{n-1-x_i},$$

define $w_{\boldsymbol{x},S} = 0$ for $\boldsymbol{x} \notin I_n$ and all S, and put

$$w = \tfrac{1}{2} + \tfrac{1}{2}(q-p)^{n-1}.$$

Noting from Lemma 4.1(iv) that $\sum_{\boldsymbol{x} \in I_n} w_{\boldsymbol{x},S} = w^{|S|}$, and also Lemma 4.1(iii), provided $p \le \frac{1}{2}$ we can turn I_n into a probability space $\mathcal{Q}_{S,p}$ for which

$$\mathrm{P}_{\mathcal{Q}_{S,p}}(\boldsymbol{x}) = \frac{w_{\boldsymbol{x},S}}{w^{|S|}}.$$

For $p > \frac{1}{2}$ define $\mathcal{Q}_{S,p} = \mathcal{Q}_{S,q}$. For any function $f : I_n \to \mathbb{R}$, define

$$\tilde{f}(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}), & \boldsymbol{x} \in E_n; \\ -f(\boldsymbol{x}), & \text{otherwise}, \end{cases}$$

and for $\boldsymbol{x} \in I_n$ and $S \subseteq \{1, \ldots, n\}$ define

$$Q_{\boldsymbol{x},S} = \{\boldsymbol{y} \in I_n \mid y_i = x_i - 1 \text{ or } x_i \text{ for } i \in S, \text{ and } y_i = x_i \text{ otherwise}\},$$

and

$$f_{S,p}(\boldsymbol{x}) = w^{|S|} \sum_{\boldsymbol{y} \in Q_{\boldsymbol{x},S}} f(\boldsymbol{y}).$$

Finally, $\boldsymbol{e}_i$ denotes the elementary vector with 1 in the i-th component and 0 elsewhere.

Theorem 4.2. *For any function $f : I_n \to \mathbb{R}$*

$$\mathrm{E}_{\mathcal{B}_p} f = \sum_{\boldsymbol{x} \in I_n,\, \boldsymbol{y} \in Q_{\boldsymbol{x},S}} w_{\boldsymbol{x},S} f(\boldsymbol{y}) = \mathrm{E}_{\mathcal{Q}_{S,p}} f_{S,p}.$$

Proof. We assume without loss of generality that $p \le \frac{1}{2}$.

The right-hand equality is from the definitions, so we prove the one on the left by induction on the cardinality of S. It is immediate for $S = \emptyset$. Pick $i \in S$ and write

$$I_n(i,k) = \{\boldsymbol{x} \in I_n \mid x_i = k\}.$$

Then noting that

$$\frac{w_{\boldsymbol{x},S}}{w_{\boldsymbol{x},S-\{i\}}} = \frac{w_{x_i}}{\binom{n-1}{k} p^k q^{n-k-1}},$$

we have

$$\begin{aligned}
\sum_{\substack{\boldsymbol{x}\in I_n\\ \boldsymbol{y}\in Q_{\boldsymbol{x},S}}} w_{\boldsymbol{x},S} f(\boldsymbol{y}) &= \sum_{\substack{\boldsymbol{x}\in I_n\\ \boldsymbol{y}\in Q_{\boldsymbol{x},S-\{i\}}}} w_{\boldsymbol{x},S}(f(\boldsymbol{y})+f(\boldsymbol{y}-\boldsymbol{e}_i))\\
&= \sum_{\substack{\boldsymbol{x}\in I_n\\ \boldsymbol{y}\in Q_{\boldsymbol{x},S-\{i\}}}} (w_{\boldsymbol{x},S}+w_{\boldsymbol{x}+\boldsymbol{e}_i,S}) f(\boldsymbol{y})\\
&= \sum_{k=0}^{n-1} \frac{w(k)+w(k+1)}{\binom{n-1}{k} p^k q^{n-k-1}} \sum_{\substack{\boldsymbol{x}\in I_n(i,k)\\ \boldsymbol{y}\in Q_{\boldsymbol{x},S-\{i\}}}} w_{\boldsymbol{x},S-\{i\}} f(\boldsymbol{y})\\
&= \sum_{k=0}^{n-1} \sum_{\substack{\boldsymbol{x}\in I_n(i,k)\\ \boldsymbol{y}\in Q_{\boldsymbol{x},S-\{i\}}}} w_{\boldsymbol{x},S-\{i\}} f(\boldsymbol{y})\\
&= \sum_{\substack{\boldsymbol{x}\in I_n\\ \boldsymbol{y}\in Q_{\boldsymbol{x},S-\{i\}}}} w_{\boldsymbol{x},S-\{i\}} f(\boldsymbol{y})\\
&= \mathrm{E}_{\mathcal{B}_p} f,
\end{aligned}$$

where the third-last step is by Lemma 4.1(ii), and the last is by induction on $|S|$. ∎

Theorem 4.2, together with Lemma 2.2, gives the following immediately from the observation that

$$\mathrm{E}_{\mathcal{B}_p} f + \mathrm{E}_{\mathcal{B}_p} \tilde{f} = (2\mathrm{E}_{\mathcal{E}_p} f)\mathrm{P}_{\mathcal{B}_p}(E_n).$$

Corollary 4.3. *For any function* $f : I_n \to \mathbb{R}$,

$$\mathrm{E}_{\mathcal{E}_p} f = \frac{\mathrm{E}_{\mathcal{B}_p} f + \mathrm{E}_{\mathcal{Q}_{S,p}} \tilde{f}_{S,p}}{1+(q-p)^{2N}}.$$

Corollary 4.3 will be used to measure the difference between $\mathrm{E}_{\mathcal{E}_p} f$ and $\mathrm{E}_{\mathcal{B}_p} f$. By letting f be the indicator function of any subset A of E_n we obtain:

Corollary 4.4. *If* $pqN \to \infty$ *then* $\mathrm{P}_{\mathcal{E}_p}(A) \sim 2\mathrm{P}_{\mathcal{B}_p}(A)$ *for all* $A \subseteq E_n$.

Two major special cases of Corollary 4.3 will be useful: one with $|S|$ fairly small and the other with $S = \{1, \dots, n\}$. In this paper we develop a result more useful for the former. The latter is deferred to the next paper.

Define an operator Δ_S recursively by

$$\Delta_S f(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}), & x_i = 0 \text{ for all } i \in S;\\ \Delta_{S-\{i\}} f(\boldsymbol{x}) - \Delta_{S-\{i\}} f(\boldsymbol{x}-\boldsymbol{e}_i), & \text{otherwise;}\end{cases}$$

where in the second case i is chosen from S with $x_i > 0$. Note that for $\boldsymbol{x} \in I_n$,

$$\Delta_S f(\boldsymbol{x}) = w^{-|S|}(-1)^{x_1+\cdots+x_n} \tilde{f}_{S,p}(\boldsymbol{x}). \tag{4.1}$$

In the following, $\bar{A}$ denotes the complement of A in I_n.

Corollary 4.5. *Suppose $pqN \to \infty$, $S = S(n) \subseteq \{1, \dots, n\}$, and $A = A(n) \subseteq I_n$. Let f be any real-valued function defined on I_n. Then*

$$\mathrm{E}_{\mathcal{E}_p} f = \mathrm{E}_{\mathcal{B}_p} f + O\big((q-p)^{2N} + 2^{|S|}\mathrm{P}_{\mathcal{B}_p}(\bar{A})\big) \sup_{I_n} |f| + O(\sup_{\boldsymbol{x}\in A} |\Delta_S f(\boldsymbol{x})|)$$

uniformly over S.

Proof. Without loss of generality take $p \le \frac{1}{2}$. Since $w \ge \frac{1}{2}$, we have $\mathrm{P}_{\mathcal{Q}_{S,p}}(\boldsymbol{x}) = O(\mathrm{P}_{\mathcal{B}_p}(\boldsymbol{x}))$, and so the corollary follows from Corollary 4.3 and (4.1). ∎

If f has mixed partial derivatives which are reasonably well behaved, then the following result can be used in conjunction with Corollary 4.5. For simplicity of notation we now let $S = \{1, \dots, k\}$, although this is no significant restriction. Let $f_{12\cdots k}$ denote the mixed partial derivative of f with respect to the variables $x_1, \dots, x_k$. Also, let $c(Q_{\boldsymbol{x},S})$ denote the convex hull of $Q_{\boldsymbol{x},S}$. For the following lemma we need to restrict A to the subset of I_n in which the first k coordinates are strictly positive, which we denote by I_n^{+k}.

Lemma 4.6. *Let $A \subseteq I_n^{+k}$. If $f_{12\cdots k}$ is defined on $\hat{A} = \bigcup_{\boldsymbol{x}\in A} c(Q_{\boldsymbol{x},\{1,2,\dots,k\}})$, then*

$$\sup_{\boldsymbol{x}\in A} |\Delta_S f(\boldsymbol{x})| \le \sup_{\boldsymbol{x}\in \hat{A}} |f_{12\cdots k}(\boldsymbol{x})|$$

provided the supremum on the right exists.

Proof. Recall that a function $g(x)$ is *absolutely continuous* if for all $\epsilon > 0$ there exists $\delta > 0$ such that $\sum_{i=1}^{j} |g(\beta_i) - g(\alpha_i)| < \epsilon$ for any j and any disjoint collection $\{\alpha_i, \beta_i\}$ of segments with total length less than δ. By the Mean Value Theorem, if g' bounded on a closed interval $[a, b]$ then g is absolutely continuous on $[a, b]$. Then by Theorem 7.20 of Rudin [8], $g(b) - g(a) = \int_a^b g'(x)\, dx$.

In the context of this lemma, we can take n to be fixed. Let C denote the supremum of $|f_{12\cdots k}|$ on $\hat{A}$. From the above remark, for $\boldsymbol{y} \in \hat{A}$ with y_k an integer at least 1,

$$\Delta_{\{k\}} f_{12\cdots k-1}(\boldsymbol{y}) = \int_{y_k-1}^{y_k} f_{12\cdots k}(y_1, \dots, y_{k-1}, x_k, y_{k+1}, \dots, y_n)\, dx_k. \tag{4.2}$$

Here we use the assumption that $A \subseteq I_n^{+k}$.

Since $|f_{12\cdots k}| \le C$ on $\hat{A}$, we now observe by the Mean Value Theorem that the absolute value of difference function on the left side of (4.2) is bounded above by C on that subset of $\hat{A}$ in which $y_k \ge 1$. Moreover this function is the partial derivative of $\Delta_{\{k\}} f_{12\cdots k-1}(\boldsymbol{y})$ with respect to y_{k-1}. Hence we can iterate this integration process to obtain, for $\boldsymbol{y} \in A$,

$$\Delta_{\{1\}}\Delta_{\{2\}} \cdots \Delta_{\{k\}} f(\boldsymbol{y}) \le C,$$

from which the lemma follows. ∎

As a simple example of the use of Corollary 4.5, we can we put $S = \{1\}$ and let A denote the set of $\boldsymbol{x} \in I_n$ for which $|f(\boldsymbol{x}) - f(\boldsymbol{x} - \boldsymbol{e}_1)| \le C|f(\boldsymbol{x})|$, to obtain the following.

Corollary 4.7. *If $pqN \to \infty$ and $\mathrm{P}_{\mathcal{B}_p}(|f(\boldsymbol{x}) - f(\boldsymbol{x} - \boldsymbol{e}_1)| > C \sup_{I_n} |f|) = o(1)$ for all $C > 0$ then $\mathrm{E}_{\mathcal{E}_p} f = \mathrm{E}_{\mathcal{B}_p} f + o(\sup_{I_n} |f|)$.*

In the second paper of this series, we will illustrate the use of Corollary 4.3 with $|S|$ large. For this paper, however, we will be content with demonstrating Corollary 4.7.

5. A simple example: geometric mean degree

In the section we will illustrate the use of our techniques by investigating the geometric mean of the degrees in model $\mathcal{D}_p$. Let $\bar{d} = (d_1 + \cdots + d_n)/n$ and $\hat{d} = d_1^{1/n} \cdots d_n^{1/n}$. For suitable p, our methods could be used to estimate the distribution of $\hat{d}$ to high accuracy. Specifically, we could asymptotically determine the probability in $\mathcal{D}_p$ of any event of the form $A \le \hat{d} \le B$ so long as that probability is greater than n^{-k} for some fixed k. However, in this paper we will be content with some simpler computations.

We know that $\hat{d} \le \bar{d}$ and can reasonably expect both to be close to the mean of $\bar{d}$, namely $p(n-1)$, so we will estimate the mean and variance of the two random variables

$$G_1 = \bar{d} - \hat{d},$$

and

$$G_2 = p(n-1) - \hat{d}.$$

We will assume that $pn = \omega(n)\log n$. This is a natural boundary for the problem because $\hat{d} = 0$ almost surely if $pn = o(\log n)$, for this is where the random graph almost surely contains an isolated vertex (see Bollobás [2]).

Theorem 5.1. *Let $p = p(n)$ be acceptable and $pn = \omega(n)\log n$. Then*

$$\begin{aligned} \mathrm{E}_{\mathcal{D}_p}(G_1) = \mathrm{E}_{\mathcal{D}_p}(G_2) &\sim \tfrac{1}{2}q, \\ \mathrm{Var}_{\mathcal{D}_p}(G_1) &\sim \tfrac{1}{2}q^2/n + \tfrac{1}{4}q/n^2, \end{aligned}$$

and

$$\mathrm{Var}_{\mathcal{D}_p}(G_2) \sim 2pq.$$

Proof. We begin by applying the model $\mathcal{B}_p$. For $r \ge 0$, define

$$f_1(r) = \sum_{k=0}^{n-1} \binom{n-1}{k} p^k q^{n-k-1} k^{r/n}$$

and

$$f_2(r) = \sum_{k=0}^{n-1} \binom{n-1}{k} p^k q^{n-k-1} k^{1+r/n}.$$

The functions $f_1(r)$ and $f_2(r)$ can be estimated by expanding $k^{r/n}$ in a Taylor series about $k = pn$ then performing the sum. For example, we have

$$\begin{aligned} \mathrm{E}_{\mathcal{B}_p}(G_1) = \mathrm{E}_{\mathcal{B}_p}(G_2) &= p(n-1) - f_1(1)^n \\ &= \tfrac{1}{2}q + O\Big(\frac{q}{pn}\Big). \end{aligned}$$

Similarly, $\mathrm{E}_{\mathcal{B}_p}(\hat{d}^2) = f_1(2)^n$, $\mathrm{E}_{\mathcal{B}_p}(\hat{d}\bar{d}) = f_1(1)^{n-1} f_2(1)$ and $\mathrm{E}_{\mathcal{B}_p}(\bar{d}^2) = p(n-1)(pn - p + q/n)$, so we obtain

$$\mathrm{Var}_{\mathcal{B}_p}(G_1) \sim \tfrac{1}{2}q^2/n + \tfrac{1}{4}q/n^2$$

and

$$\mathrm{Var}_{\mathcal{B}_p}(G_2) \sim pq.$$

Transferral of these results to $\mathcal{E}_p$ is easy using Corollary 4.5. In the notation used there, define $A = A(n) = \{\boldsymbol{d} \mid d_i \ge \frac{1}{2}np \text{ for all } i\}$. Then $\mathrm{P}_{\mathcal{B}_p}(\bar{A}) = n^{-\omega(n)}$ by Lemma 3.3. Using Lemma 4.6, we find for $k = |S| \ge 3$ that $\Delta_S \hat{d} = o(n^{-k+1})$, $\Delta_S \bar{d} = 0$, $\Delta_S \hat{d}^2 = o(n^{-k+2})$, $\Delta_S \bar{d}\hat{d} = o(n^{-k+2})$, and $\Delta_S \bar{d}^2 = 0$. Corollary 4.5 immediately implies that the expectation and variance of G_1 and G_2 agree in $\mathcal{B}_p$ and $\mathcal{E}_p$ to within a factor of $1 + O(n^{-t})$ for any t.

From $\mathcal{E}_p$, we can use Theorem 3.7 to take the results to the model $\mathcal{I}_p$ then apply Theorem 2.6 to obtain the claimed result. The only significant difference between $\mathcal{E}_p$ and $\mathcal{I}_p$ is for $\mathrm{Var}_{\mathcal{D}_p}(G_2)$. In the notation of Lemma 2.4, we have $\mathrm{Var}_{\mathcal{E}_x}(G_2) \sim x(1-x)$ and $\mathrm{E}_{\mathcal{E}_x}(G_2) - \mathrm{E}_{\mathcal{I}_p}(G_2) \sim (x-p)n$. Now Theorem 3.7 gives the required answer. ∎

Theorem 5.2. *Let $p = p(n)$ be acceptable, with $pn = \omega(n)\log n$ and $qn \to \infty$. Then $\hat{d}$ is asymptotically normal in $\mathcal{D}_p$ with mean $p(n - \frac{1}{2}) - \frac{1}{2} + O(q/pn)$ and variance $2pq$.*

Proof. Take $t > 0$, write f for $\hat{d}$, and let $\chi^{(t)}$ denote the indicator variable of the event $\{f \ge t\}$.

Take $S = \{1\}$. For integers $z_2, \ldots, z_n$, let

$$T = T_{z_2,\ldots,z_n} = \{\boldsymbol{x} \mid x_i = z_i, 2 \le i \le n\}.$$

There is at most one $\boldsymbol{x} \in T$ for which

$$\chi^{(t)}(\boldsymbol{x}) \ne \chi^{(t)}(\boldsymbol{x} - \boldsymbol{e}_1).$$

Thus, for $\boldsymbol{x} \in T$, $\tilde{f}_{S,p}(\boldsymbol{x})$ is either 0 or w, the latter only for at most one value of x_1. Hence

$$\begin{aligned}
\mathrm{E}_{\mathcal{Q}_{S,p}}(\tilde{f}_{S,p}(\boldsymbol{x}) \mid \boldsymbol{x} \in T) &\le \frac{w \max_{\boldsymbol{x}\in T} \mathrm{P}_{\mathcal{Q}_{S,p}}(\boldsymbol{x})}{\mathrm{P}_{\mathcal{Q}_{S,p}}(T)} \\
&= \frac{w \max_{\boldsymbol{x}\in T} w_{\boldsymbol{x},S}}{\sum_{\boldsymbol{x}\in T} w_{\boldsymbol{x},S}} \\
&= \frac{w \max_{x_i=0}^{n-1} w(x_i)}{\sum_{x_i=0}^{n-1} w(x_i)} \\
&= \max_{k=0,\ldots,n-1} w(k)
\end{aligned}$$

by Lemma 4.1 (iv)

$$\le \max_{k=0,\ldots,n-1} \binom{n-1}{k} p^k q^{n-1-k}$$

by Lemma 4.1 (ii)

$$= o(1)$$

since $pqn \to \infty$. Thus $\mathrm{E}_{\mathcal{Q}_{S,p}} \tilde{f}_{S,p} = o(1)$, and so by Corollary 4.3, $\mathrm{E}_{\mathcal{E}_p}\chi(t) = \mathrm{E}_{\mathcal{B}_p}\chi(t) + o(1)$; that is $\mathrm{P}_{\mathcal{E}_p}(\hat{d} \le t) = \mathrm{P}_{\mathcal{B}_p}(\hat{d} \le t) + o(1)$.

In $\mathcal{B}_p$, $\log \hat{d}$ is the sum of n iid random variables. Application of the Berry-Essèen inequality [3; page 542] shows that $\log \hat{d}$ is asymptotically normal and gives a uniform error estimate. This implies that $\hat{d}$ is asymptotically log-normal, but since the variance of $\log \hat{d}$ tends to 0 (in fact it is $O\big(1/(pn^2)\big)$) this is the same as being asymptotically normal. From the calculations above, we have that the mean of $\hat{d}$ in $\mathcal{B}_x$ is $f_1(1)^n$ and the variance is $x(1-x)(1+o(1))$ for x close to p, where f_1 is as defined above but for x instead of p. Now it is routine to check that (3.7) holds for $a = n$ and any $y \to \infty$ sufficiently slowly, with the uniformity of the approximation following from the continuity of the error term. Consequently, Theorem 3.8 gives that

$$\mathrm{P}_{\mathcal{I}_p}(\hat{d} \le t) = o(1) + \frac{1}{2\sqrt{\pi pq}} \int_{-\infty}^{t} \exp\Big(-\frac{(z-\mu(\hat{d}))^2}{4pq}\Big)\, dz$$

for all t. The mean $\mu(\hat{d})$ in any of $\mathcal{B}_p$, $\mathcal{E}_p$ or $\mathcal{D}_p$ can be used in this formula, as they are all the same to this accuracy. ∎

References.

[1] A.D. Barbour, L. Holst and S. Janson, Poisson Approximation, Clarendon, Oxford, 1992.

[2] B. Bollobás, *Random graphs*, Academic Press, London, 1985.

[3] W. Feller, An Introduction to Probability Theory and its Applications, Volume II, Wiley, New York, 1966.

[4] B. D. McKay and N. C. Wormald, Asymptotic enumeration by degree sequence of graphs of high degree, *European J. Combinatorics*, **11** (1990) 565–580.

[5] B. D. McKay and N. C. Wormald, Asymptotic enumeration by degree sequence of graphs with degrees $o(n^{1/2})$, *Combinatorica* **11** (1991) 369–382.

[6] B. D. McKay and N. C. Wormald, The degree sequence of a random graph, II. Applications. In preparation.

[7] Z. Palka, Asymptotic properties of random graphs, Dissertationes Mathematicae CCLXXV, Polska Akademia Nauk: Instytut Matematyczny, Warsaw, 1988.

[8] W. Rudin, *Real and complex analysis*, McGraw-Hill, New York, 1986.

[9] E. Shamir and J. Spencer, Sharp concentration of the chromatic number on random graphs $G_{n,p}$, *Combinatorica* **7** (1987) 121–129.