CYCLE FACTORS AND RENEWAL THEORY

JEFF KAHN, EYAL LUBETZKY, AND NICHOLAS WORMALD

ABSTRACT. For which values of k does a uniformly chosen 3-regular graph G on n vertices typically contain n/k vertex-disjoint k-cycles (a k-cycle factor)? To date, this has been answered for $k = n$ and for $k \ll \log n$; the former, the Hamiltonicity problem, was finally answered in the affirmative by Robinson and Wormald in 1992, while the answer in the latter case is negative since with high probability (w.h.p.) most vertices do not lie on k-cycles.

A major role in our study of this problem is played by renewal processes without replacement, where one wishes to estimate the probability that in a uniform permutation of a given a set of positive integers, the partial sums hit a designated target integer. Using sharp tail estimates for these renewal processes, which may be of independent interest, we settle the cycle factor problem completely: the "threshold" for a k-cycle factor in G as above is $\kappa_0 \log_2 n$ with $\kappa_0 = [1 - \frac{1}{2}\log_2 3]^{-1} \approx 4.82$.

Precisely, G contains a k-cycle factor w.h.p. if $k \geq K_0(n) := \lceil \kappa_0 \log_2(2n/e) \rceil$ and w.h.p. does not contain one if $k < K_0(n) - \log^2 n/n$. Thus, for most values of n the threshold concentrates on the single integer $K_0(n)$. As a byproduct, we confirm the "Comb Conjecture," an old problem concerning the embedding of certain spanning trees in the random graph $\mathcal{G}(n,p)$.

1. INTRODUCTION

An H-factor of a graph G is a collection of vertex-disjoint copies of the graph H covering all vertices of G. Thresholds for the existence of H-factors in random graphs have been extensively studied — from classical works in the 1960's (for instance, perfect matchings [12]) to recent ones (such as triangle-factors and the related "Shamir's problem" of matchings in hypergraphs [19]). Here we consider the following question on k-cycle factors in random regular graphs.

> *(Cycle factors.)* For which values of $k = k(n)$ does a uniformly chosen 3-regular graph on n vertices contain n/k vertex-disjoint k-cycles with high probability[1]?

When $k = n$ this is the Hamiltonicity problem, which was finally answered in the affirmative in 1992 by Robinson and Wormald [28]. At the other extreme, for $k = O(1)$ it is known ([8, 32]) that the total number of k-cycles in G is asymptotically Poisson with bounded mean, and in particular there is no k-cycle factor w.h.p. (the total number of vertices on such cycles is uniformly bounded); moreover, the typical absence of a k-cycle factor extends to the range $k \ll \log n$, throughout which most vertices do not lie on k-cycles w.h.p. No results were known on intermediate values of k.

Somewhat surprisingly, a major role in our study of the above problem will be played by a question on tail probabilities of *renewal processes without replacement* — where the recurrence times from the classical setting of renewal processes, rather than being i.i.d. random variables, are drawn uniformly yet *without* replacement from a finite set. We now define this question formally.

J. Kahn is supported by NSF grant DMS0701175.

N. Wormald was supported in part by the Canada Research Chairs Program and NSERC and partly by an ARC Australian Laureate Fellowship.

[1] A sequence of events (A_n) is said to hold *with high probability* (w.h.p.) if $\mathbb{P}(A_n) \to 1$ as $n \to \infty$.

Let $X = \{x_1, \dots, x_m\}$ be a multiset of positive integers summing to n. Let (Y_i) be a sequence of i.i.d. uniform samples of the x_i's (recurrence times), and let $S_t = \sum_{i=1}^{t} Y_i$ be its partial sum sequence (the renewal process). By the classical Renewal Theorem (due to Erdős, Feller and Polard [11] in the discrete setting), the probability that the partial sums "hit" some integer k, denoted $R_k = \mathbb{P}(k \in \{S_1, S_2, \dots\})$, tends to m/n as $k \to \infty$ provided that $\gcd(x_1, \dots, x_m) = 1$ (see, e.g., [13], as well as the background in §2.1). We consider the following variant of R_k:

> *(Renewals without replacement.)* Let $X = \{x_1, \dots, x_m\}$ be a multiset of positive integers summing to n, take a uniform permutation $\sigma \in \mathcal{S}_m$ and let $S_t = \sum_{i=1}^{t} x_{\sigma(i)}$. What is the probability $P_k = \mathbb{P}(k \in \{S_1, S_2, \dots\})$ that the partial sums hit $k \in \mathbb{Z}$?

Our main result will hinge on sharp quantitative bounds for P_k (and a variant of it called Q_k), including asymptotic second order terms and correct exponential tails (see Theorem 3 below).

Going back to the main problem on cycle factors in random regular graphs, it is interesting to compare the situation for the Erdős-Rényi random graph $\mathcal{G}(n,p)$ (in which each edge appears independently with probability p), where for a given k one is interested in the threshold p_c at which the probability of a k-cycle factor is $\frac{1}{2}$ (say). Here it is often natural to expect that p_c coincides (up to a factor $(1+o(1))$) with the threshold for the property that every vertex lies on a k-cycle. For instance, for k fixed, the latter threshold has order $n^{-\frac{k-1}{k}}(\log n)^{\frac{1}{k}}$; indeed, it was shown in [19] that the threshold for k-cycle factors has the same order, though its asymptotics remain unknown.

For the property that every vertex lies on a k-cycle in the random 3-regular graph, the threshold ("threshold" now referring to k) is at $k = (1+o(1))\log_2 n$; this follows from the fact that a given vertex has at most (and typically also roughly) $3 \cdot 2^{k/2}$ vertices at distance $k/2$ from it, and edges between these vertices have probability of order $1/n$. (With slightly more care, the same argument shows that the threshold is $\log_2 n + \log_2 \log n + O(1)$.)

We now state our main result, which settles the problem completely and shows that here the preceding intuition is not quite correct: the phase transition from no k-cycle factor to the existence of one occurs around $[1 - \frac{1}{2}\log_2 3]^{-1} \log_2 n \approx 4.82 \log_2 n$. Furthermore, we establish a 2-point concentration result (a single point for most values of n).

Theorem 1. *Let G be a random 3-regular graph on n vertices and let*

$$K_0(n) = \frac{1}{1 - \frac{1}{2}\log_2 3} \log_2(2n/e)\,. \tag{1.1}$$

If $k \geq K_0(n)$ is a divisor of n then G contains a k-cycle factor w.h.p., and on the other hand if $k \leq K_0(n) - \frac{\log^2 n}{n}$ then w.h.p. there is no k-cycle factor in G.

Moreover, the number of k-cycle factors in G, denoted by CF_k, satisfies

$$\frac{\mathsf{CF}_k}{\mathbb{E}[\mathsf{CF}_k]} \xrightarrow{\mathrm{d}} W = \prod_{j=3}^{\infty} (1+\delta_j)^{Z_j} e^{-\delta_j \mathbb{E} Z_j} \qquad \text{as } n \to \infty \tag{1.2}$$

for any $k \geq K_0(n)$ that divides n, where $\delta_j = \frac{(-1)^j - 1}{2^j}$ and the Z_j's are i.i.d. Poisson$(\frac{2^j - 1}{j})$ *variables.*

Remark. The proof technique extends, with very few modifications, to yield the threshold (as well as a 2-point concentration) for k-cycles factors in a random d-regular graph for any fixed $d \geq 3$.

The proof of [28] that a random cubic graph is Hamiltonian introduced the *small subgraph conditioning method*, an interesting twist on the second moment method: upon calculating the second moment of H_n, the number of Hamilton cycles in that random graph, one finds that it is unfortunately (just barely) too large, namely $\mathbb{E}H_n^2/(\mathbb{E}H_n)^2 \to c$ for fixed $c > 0$. The culprit turns out to be the set of small cycles (those with bounded length) in the graph, which in a sense blow up the variance by allowing local detours along a Hamilton cycle. Luckily — and quite mysteriously (in various situations this fails, e.g., when half the degrees are 3 and half are 4) — the second moment drops to $\varepsilon(\mathbb{E}H_n)^2$ once we *condition on the joint cycle distribution* up to length $M(\varepsilon)$, implying Hamiltonicity with high probability (see [33] and [18, §9.3] for more information).

As our result gives Hamiltonicity for the special case $k = n$, naturally we follow the framework of small subgraph conditioning, which was highly nontrivial already for a single cycle. Though far more delicate to carry out in our setting (as explained below), this method enjoys two byproducts (implicit in [28] and formalized in [17] (and for (b) below also in [25]); cf. [18, §9.5],[33, §4]): (a) it gives the limiting law of the variable (as in Eq. (1.2) above), and (b) it further implies *contiguity*.

Definition (Contiguity of distributions)**.** Let $\mathbb{P}_n$ and $\mathbb{Q}_n$ be two sequences of probability measures defined on the same measurable spaces $(\Omega_n, \mathcal{F}_n)$. We say that $\mathbb{P}_n$ and $\mathbb{Q}_n$ are contiguous, denoted $\mathbb{P}_n \approx \mathbb{Q}_n$, if for any sequence of events (A_n) one has $\lim_{n\to\infty} \mathbb{P}_n(A_n) = 1 \;\Leftrightarrow\; \lim_{n\to\infty} \mathbb{Q}_n(A_n) = 1$.

In other words, events hold w.h.p. under $\mathbb{P}_n$ iff they hold w.h.p. under $\mathbb{Q}_n$. The first example of contiguity in our context ([17,25]) was that $\mathcal{G}(n,3)$, the uniform distribution on 3-regular graphs on n vertices, is contiguous to the union of a uniform Hamilton cycle and a uniform perfect matching (conditioned on no multiple edges in the union). Our work extends this result: roughly put, the next corollary says that one can distinguish with probability $1-o(1)$ between $\mathcal{G}(n,3)$ and the union of a k-cycle factor and a perfect matching if and only if $k \geq K_0(n)$ from Eq. (1.1).

Corollary 2. *Let $\mathcal{G}(n,3)$ be the uniform 3-regular graph on n vertices and, for $k \mid n$, let $\mathcal{G}(n,k,3)$ be the union of a uniform k-cycle factor and a uniform prefect matching, conditioned on no multiple edges. If $k \geq K_0(n)$ then $\mathcal{G}(n,3) \approx \mathcal{G}(n,k,3)$ whereas if $k \leq K_0(n) - \frac{\log^2 n}{n}$ then $\mathcal{G}(n,3) \not\approx \mathcal{G}(n,k,3)$.*

The most challenging hurdles in the proofs of these results arise in the second moment calculation, already before the small subgraph conditioning enters the picture. Indeed, calculating the second moment of the number of Hamilton cycles amounts to understanding the typical intersection of two cycles (a collection of paths): obtaining all Hamilton cycles that contain these paths amounts simply to orienting each path and ordering the sequence of paths, i.e., stitching them into an n-cycle (one can verify that these paths are a partition of all the vertices since the graph is 3-regular).

However, for k-cycle factors, the common intersection of two such factors must be stitched into n/k cycles, and now one seeks only those permutations of the m parts that form k cycles: letting $X = \{x_1, \ldots, x_m\}$ be the set of path lengths, we see the connection to the above question on renewals without replacement, as we wish to hit all multiples of k with the partial sums of the permutation. It is further seen that very sharp error estimates are needed, up to the correct exponential error-term; e.g., when $k \asymp \log n$, we must repeatedly hit k for about $\exp(ck)$ times (to build n/k cycles), along which these errors accumulate. Our next result establishes such estimates.

Recall that for a set $X = \{x_1, \dots, x_m\}$ of positive integers summing to n, we let P_k be the probability that k belongs to the set of partial sums $S_t = \sum_{i=1}^{t} x_{\sigma(i)}$, where $\sigma \in \mathcal{S}_m$ is uniform. For a reason to be later explained, knowing P_k would not suffice for deriving the asymptotic threshold $K_0(n)$, due to a second-order term of order $1/m$ in this probability which destroys our control over the second moment at some $k = O(\log n)$ still beyond above the desired threshold. Fortunately, our sampling procedure is a variant of the above, in which this second-order term vanishes:

- let $\sigma(1)$ be a *size-biased* sample of the elements, i.e., $\mathbb{P}(\sigma(1) = j) = x_j/n$;
- let $(\sigma(2), \dots, \sigma(m))$ be a uniform permutation in $\mathcal{S}_{m-1}$ over the remaining elements.

Let Q_n be the probability that $k \in \{S_1, S_2, \dots\}$ for this process, which we refer to as the size-biased renewal process (for brevity, while stressing that only the first step is size-biased).

Theorem 3. *Assume that $k \to \infty$ and $k = o(\sqrt{m})$ as $n \to \infty$. Let $R > 1$ and let $g(z)$ be a power series absolutely convergent for $|z| \le R$ with $g(z) \ne 1$ for $|z| \le R$ whenever $z \ne 1$, $z \in \mathbb{C}$. Also let $w(n) = o(1)$ as $n \to \infty$. Then for any $\varepsilon' > 0$ there exist functions*

$$q_1(n) = o(m^{-1}) + O(R^{-k} + k^4/m^2) \quad \text{and} \quad q_2(n) = o(m^{-1}) + O\big((R-\varepsilon')^{-k} + k^4/m^2\big)$$

such that the following holds. Let $x_1, \dots, x_m$, P_k, Q_k be as above, let $f(z) = \sum_\ell p_\ell z^\ell$ be the probability generating function of the (relative) frequencies $p_\ell = \frac{1}{m}\#\{j : x_j = \ell\}$, and assume that $|f(z) - g(z)| + |f'(z) - g'(z)| < w(n)$ for all $|z| \le R$.

(a) (Renewal without replacement.) Provided that $|f''(1) - g''(1)| < w(n)$,

$$\left| P_k - \frac{m}{n} + \frac{g'(1) - g'(1)^2 + g''(1)}{g'(1)^3 m} \right| \le q_1(n)\,;$$

(b) (Renewal without replacement, size-biased.)

$$\left| Q_k - \frac{m}{n} \right| \le q_2(n)\,.$$

Example (geometric distribution)**.** If $x_1 = 0$ and $x_\ell/m \sim 2^{1-\ell}$ for $\ell \ge 2$, then the function $f(z) = \sum p_\ell z^\ell$ is approximately $g(z) = z^2/(2-z)$, which satisfies $[g'(1) - g'(1)^2 + g''(1)]/g'(1)^3 = \frac{2}{27}$; our results then imply, for instance, that for any $1 \ll k \ll m^{1/4}$ and any fixed $\varepsilon > 0$ we have

$$\begin{aligned} R_k &= \tfrac{1}{3} + O\big((2-\varepsilon)^{-k}\big) && \textit{(with replacement)}\,, \\ P_k &= \tfrac{1}{3} - \tfrac{2/27 - o(1)}{m} + O\big((2-\varepsilon)^{-k}\big) && \textit{(without replacement)}\,, \\ Q_k &= \tfrac{1}{3} - o(1/m) + O\big((2-\varepsilon)^{-k}\big) && \textit{(size-biased without replacement)}\,. \end{aligned}$$

This example will be fundamental for the proof of Theorem 1. Recall that the x_i's correspond to the lengths (in vertices) of the paths that comprise the common intersection of two k-cycle factors. First, every such path has at least two vertices, whence $p_1 = 0$. Second, heuristically, suppose we are given a k-cycle factor F_1 and construct another, F_2, via a simple random walk (ignoring the many dependencies that exist in reality). While this random walk traverses on a common edge, there is a probability of $1/2$ that the next edge will extend the common path (i.e., follow the trace of F_1), and hence the geometric distribution with this parameter. Note that for $k \ge (2+\varepsilon')\log_2 n$ the error-terms are all $o(1/n)$, which will be crucial to our arguments.

1.1. **Applications for the Erdős-Rényi random graph.** An immediate corollary of Theorem 1 is that, for $k \geq K_0(n) \approx 4.82 \log_2 n$, the threshold for a k-cycle factor in the random graph $\mathcal{G}(n,p)$ has order $\frac{\log n}{n}$ (see Corollary 7.1), with a factor $2 - o(1)$ between the upper and lower bounds.

Another corollary is the "Comb Conjecture." A *comb of order* k, for some $k \mid n$, is a tree consisting of an (n/k)-vertex path P together with disjoint k-vertex paths beginning at the vertices of P. When $k = \sqrt{n}$ we will call this *the comb* and denote it Comb_n. The "Comb Conjecture" says that the threshold for the appearance of Comb_n in $\mathcal{G}(n,p)$ has order $\frac{\log n}{n}$ (the lower bound is obvious due to connectivity).

Consideration of the threshold for combs was suggested by the first author about 20 years ago as a test case for the more general conjecture (this was also proposed at that time by the first author, but, being a natural guess, is perhaps better regarded as folklore), that the same threshold statement holds for general (n-vertex) trees of bounded degree. The case $k = \sqrt{n}$ of this suggestion has come to be known as the "Comb Conjecture." The rationale for considering combs was the idea that they interpolated between instances for which the general conjecture was known to be true, namely Hamilton paths (for which the threshold $(1+o(1))\frac{\log n}{n}$ was established in [1] and [9]) and trees with order n leaves (which are easily handled via Hall's Theorem).

For related work on this topic, see, e.g., [3], which shows that, already at $p = O(\frac{1}{n})$, w.h.p. the random graph contains *every* bounded-degree tree on $(1-\varepsilon)n$ vertices, with an implicit constant depending on ε and Δ (see also the refined bounds in [4]); [7] which shows that $\mathcal{G}(n, \frac{c\log n}{n})$ w.h.p. contains almost every tree on n vertices; and [22], which shows that the threshold for any bounded-degree tree T is at most $n^{-1+o(1)}$. The latter was achieved by observing that such trees have either many leaves or long paths of degree-2 vertices, and then deploying a separate strategy in each case. The comb is indeed an extremal example as it precisely balances between these two elements.

We confirm the Comb Conjecture, and obtain the threshold of Comb_n up to a factor of $2+o(1)$.

Theorem 4. *For any $\varepsilon > 0$ the Erdős-Rényi random graph $\mathcal{G}(n,p)$ with $p = (2+\varepsilon)\frac{\log n}{n}$ contains a copy of Comb_n as a spanning subgraph w.h.p. In particular, the threshold for the appearance of Comb_n in $\mathcal{G}(n,p)$ is at $p \asymp \frac{\log n}{n}$.*

More generally, we get to within a factor of $2 + o(1)$ of the threshold for containing the comb of order k for every $k \geq K_0(n)$ (see Remark 7.2). In the companion paper [20] we treat the complementary range of k and conclude that for any $k = k(n)$ the threshold is $O(\frac{\log n}{n})$.

1.2. **Notation and organization.** On occasion we will write $f_n \lesssim g_n$ instead of $f_n = O(g_n)$ for brevity (similarly for $f_n \gtrsim g_n$); $f_n \sim g_n$ denotes $f_n = (1+o(1))g_n$, and $f \asymp g$ denotes $f_n \lesssim g_n \lesssim f_n$.

The rest of this paper is organized as follows. In §2 we establish Theorem 3, among other results on renewal processes with and without replacement. Sections 3–5 are devoted to the analysis of the second moment of the number of k-cycle factors. The application of the small subgraph conditioning method (and its consequences for contiguity) appears in §6, and concludes the proof of Theorem 1, as well as Corollary 2. Section 7 contains the proof of the Comb Conjecture (Theorem 4).

2. Renewal processes with and without replacement

In this section we obtain various results on the probability of the event $\mathscr{H}_k$ that a renewal process (S_t) hits a given value k (i.e., $S_t = k$ for some t), which in particular establish Theorem 3 that will be used in our later arguments; see §4. We first state the main results of this section. Note that we will require estimates of $\mathbb{P}(\mathscr{H}_k)$ in which the error terms are uniform over a range of sets X, which seems to require a new result even in the standard setting.

We approach the question by coupling the two models, with and without replacements, in Part (b) of the following theorem. Here, $[z^k]$ denotes extraction of the coefficient of z^k. Theorem 2.2 will provide estimates for the coefficients. We assume that all variables are functions of n.

Theorem 2.1. *Let $x_1, \dots, x_m$ be positive integers with $\sum_{j=1}^m x_j = n$. Write $p_\ell = \frac{1}{m}|\{j : x_j = \ell\}|$ and define $f(z) = \sum_{\ell \ge 1} p_\ell z^\ell$. For $\sigma : [m] \to [m]$ define*

$$Y_t = Y_t(\sigma) = \sum_{j \le t} x_{\sigma(j)} \quad \text{for } t = 1, \dots, m\,,$$

and let R_k equal $\mathbb{P}\big(k \in \{Y_1, \dots, Y_m\}\big)$ when σ is selected u.a.r., and let P_k equal $\mathbb{P}\big(k \in \{Y_1, \dots, Y_m\}\big)$ conditional upon σ being a permutation of $[m]$. Then

(a) $R_k = [z^k]\big(1 - f(z)\big)^{-1}$;
(b) as $n \to \infty$, provided that $k = o(\sqrt{m})$ we have

$$P_k = R_k - \big(m^{-1} + O(k^2/m^2)\big)[z^k]\frac{f(z^2) - f(z)^2}{\big(1 - f(z)\big)^3} + O(k^4/m^2)\,.$$

Here, the constants implicit in the $O(\cdot)$ are absolute, in particular not depending on the x_j's.

The list of constants and conditions in the following theorems is made longer than might be expected, because of our need to apply them uniformly to a class of sets of numbers $\{x_j\}$.

Theorem 2.2. *Define m, n and $f(z)$ as in Theorem 2.1. Suppose that for some real constants $r > 1$, $c > 0$ and $c_0 > 0$, we have*

(i) $\sum_{\ell \ge 1} p_\ell r^\ell \le c$,
(ii) $f(z) \ne 1$ for $|z| \le r$ whenever $z \ne 1$, $z \in \mathbb{C}$, and
(iii) $\min_{|z|=r} |f(z) - 1| \ge c_0^{-1}$ $(z \in \mathbb{C})$.

Then for all $k \ge 1$

$$|R_k - m/n| \le c_0 r^{-k}$$

and

$$\left|[z^k]\frac{f(z^2) - f(z)^2}{(1 - f(z))^3} - \frac{f'(1) - f'(1)^2 + f''(1)}{f'(1)^3}\right| \le c_1 r^{-k/2}\,,$$

where

$$c_1 = \max_{|z|=\sqrt{r}} \left|\frac{f(z^2) - f(z)^2}{(1 - f(z))^3}\right|\,.$$

Note that $f'(1) = n/m$.

After many twists and turns, it is an implication of the proof of our main result that the formulae as above cannot possibly be extended to the size-biased case without modification, because otherwise a certain random variable involving k-cycle factors would have a negative variance. The negative term of order m^{-1} in P_k is the problem. Fortunately, this term is cancelled out in the size-biased case, as follows.

Theorem 2.3. *Define $x_1, \dots, x_m$, p_ℓ, m, n and $f(z)$ as in Theorem 2.1. Let J be the random variable given by $\mathbb{P}(J = j) = x_j/n$ for $j = 1, \dots, m$, let σ be a random permutation of the indices $\{1, \dots, m\} \setminus \{J\}$, and define*

$$\hat{Y}_t = \hat{Y}_t(\sigma) = x_J + \sum_{j \le t-1} x_{\sigma(j)} \quad \text{for } t = 1, \dots, m\,.$$

Set $Q_k = \mathbb{P}\big(k \in \{\hat{Y}_1, \dots, \hat{Y}_m\}\big)$. Assume, for some positive constants r, c, c_0 and δ with $1+\delta < \sqrt{r}$,

(i) $\sum_{\ell \ge 1} p_\ell r^\ell \le c$,
(ii) $|f(z) - 1| \ge c_0^{-1}$ *if* $|z| \le r$ *and* $|z - 1| \ge \delta$, $z \in \mathbb{C}$,
(iii) $|f'(z) - f'(1)| < f'(1)/2$ *if* $|z - 1| < \delta$, $z \in \mathbb{C}$.

Then as $n \to \infty$, provided that $k \to \infty$ and $k = o(\sqrt{m})$, we have for any $\varepsilon > 0$

$$\left| Q_k - \frac{m}{n} \right| \le q(n)\,,$$

where $q(n)$ is a function of r, c, c_0, k and ε satisfying $q(n) \le 2c_0(r-\varepsilon)^{-k} + O\left(k^4/m^2\right) + o(1/m)$. In particular, $q(n)$ does not otherwise depend on the x_j's.

In the next subsection, we give some estimates for the generating function coefficients appearing in Theorem 2.1. In §2.2 we complete the proof of all three theorems and the corollary.

2.1. **Singularity analysis.** Assume that $p_0 = 0$ and $p_\ell \ge 0$ for $\ell \ge 1$, with $\sum_{\ell \ge 0} p_\ell = 1$. Let $f(z) = \sum_{\ell \ge 1} p_\ell z^\ell$. Note that $f(1) = 1$. We are interested in $u_n := [z^n](1-f)^{-1}$ for renewals without replacement, and some coefficients in related generating functions for renewals with replacement.

In the following, $z \in \mathbb{C}$. Suppose that $R > 1$, f is holomorphic in $|z| < R$ (equivalently, $p_\ell = O(R^{-\ell})$), and $\gcd\{\ell : p_\ell > 0\} = 1$. Kendall [21] showed that under these conditions, $u_\infty := \lim_{n\to\infty} u_n$ exists, and that $\sum_{n \ge 0}(u_n - u_\infty)z^n$ has radius of convergence strictly greater than 1. It follows that $|u_n - u_\infty| = O(r^{-n})$ for some $r > 1$. This can easily be proved using the method of 'subtracting the singularity.' (See Wilf [31, §5.2] for a description of this method.)

Baxendale [5, Theorem 3.2] examined the analyticity of the function $\sum_{n \ge 1}(u_n - u_{n-1})z^n$, and hence obtained explicit bounds on r and on $\sum_{n \ge 0}(u_n - u_\infty)z^n$ ($|z| = r$) under certain conditions. His conditions included $p_1 > 0$ so we cannot apply his general theorem to our case. The approach could however be adapted to our case; it is closely related to the method of subtracting the singularity. Such explicit results can be used to obtain bounds that hold uniformly for a family of functions, which is what we desire here. We will use the alternative approach of contour integration. Flajolet and Sedgewick [14, Lemma IX.2 (p. 668)] show how to obtain uniform estimates using this

approach, but their result is not quite suitable for our current purpose. Additionally, we have the opportunity to use a simpler contour in this case.

Lemma 2.4. *Let $r > 1$ and assume that*

(i) f is holomorphic in $|z| < r$ and continuous on $|z| \le r$;
(ii) $f(z) \neq 1$ for $|z| \le r$ when $z \neq 1$.

Then, with $u_\infty = \lim_{n\to\infty} u_n$,

(a) $u_\infty = f'(1)^{-1}$, and $\sum_{n\ge 0}(u_n - u_\infty)z^n$ has radius of convergence at least r.

If in addition

(iii) $\min_{|z|=r} |f(z) - 1| = c_0 > 0$,

then

(b) $|u_n - u_\infty| \le \dfrac{1}{c_0 r^n}$.

Proof. Define $g(z) = (1 - f(z))/(1 - z)$ as a formal power series, so $g(z) = \sum_{n\ge 0} g_n z^n$ with

$$g_n = 1 - \sum_{1\le \ell\le n} p_\ell = \sum_{\ell \ge n+1} p_\ell \qquad \text{for } n \ge 0\,.$$

By (i), this series for $g(z)$ has radius of convergence at least r, and $g(1) \ge g_0 = 1 \neq 0$.

Recall that $u_n = [z^n](1 - f)^{-1}$. For $|z| < r$, $z \neq 1$, we have

$$\frac{1}{1 - f(z)} = \frac{1}{g(z)(1 - z)}. \tag{2.1}$$

The method of subtracting the singularity now gives the conclusion on the radius of convergence in (a), but we omit details as we need the more precise bound in (b). For this, we use the Cauchy integral formula to extract the coefficient of z^n.

Let $\mathcal{C}$ be a contour that passes around the circle $|z| = r$ in a counterclockwise direction beginning at R, then along the real line from $z = r$ to $z = 1 + \varepsilon$ (for some $\varepsilon > 0$), then once around the circle $|z - 1| = \varepsilon$ clockwise, then back along the real line from $z = 1 + \varepsilon$ to $z = r$. By (i) and (ii), $(1 - f)^{-1}$ is holomorphic on the interior of $\mathcal{C}$ and continuous on its closure, so

$$[z^n](1 - f(z))^{-1} = \frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{z^{-n-1}}{1 - f(z)}\, dz.$$

First, note that $(1 - f(z))^{-1}$ is bounded above on the outer circle since $1 - f$ is non-zero there (and f is continuous). Hence, that part of the contour integral on the outer circle is $O(r^{-n})$. In fact, under assumption (iii), its absolute value is at most $(2\pi r)r^{-n}/c_0$. The part on the straight lines cancels since (2.1) gives a unique value for the function there. So

$$\frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{z^{-n-1}}{1 - f(z)}\, dz = O(r^{-n}) - \frac{1}{2\pi i} \oint_{|z-1|=\varepsilon} \frac{z^{-n-1}}{1 - f(z)}\, dz$$

where the direction of integration is counterclockwise. Since g is analytic near 1 and $g(1) \neq 0$, we may expand $1/g(z)$ as a power series about $z = 1$, and we see that the integrand has a simple pole

at $z = 1$ with principal part $(g(1)(1-z))^{-1}$. So

$$[z^n](1-f(z))^{-1} = O(r^{-n}) - \mathrm{Res}_1 \frac{1}{(1-z)g(1)} = O(r^{-n}) + g(1)^{-1}.$$

It follows that $u_\infty = g(1)^{-1}$. Noting that

$$g(1) = \sum_{n\geq 0} \sum_{\ell \geq n+1} p_\ell = \sum_{\ell \geq 1} \ell p_\ell = f'(1),$$

we deduce (a). Recalling the above given explicit bound on the error term $O(r^{-n})$ yields (b). ■

For f satisfying (i), Lemma 2.4 shows that the radius of convergence of $(1-f)^{-1}$ will be at least the smallest value of $|z| \leq r$ where $z \neq 1$ and $f(z) = 1$ (if any such z exist). It seems reasonable to suppose that this will be equality, though our argument does not prove this.

For the case of renewals with replacement, we will make use of the following, which again uses essentially standard singularity analysis.

Lemma 2.5. *Assume that f satisfies conditions (i) and (ii) of Lemma 2.4. Then*

$$[z^n]\frac{f(z^2)-f(z)^2}{(1-f(z))^3} = \frac{f'(1)-f'(1)^2+f''(1)}{f'(1)^3} + O(r^{-n})\,.$$

Moreover, the $O(r^{-n})$-term is in absolute value at most $c_1 r^{-n/2}$ where the constant c_1 is given by

$$c_1 = \max_{|z|=\sqrt{r}} \left| \frac{f(z^2)-f(z)^2}{(1-f(z))^3} \right|.$$

Proof. Note that c_1 exists by condition (ii) of Lemma 2.4 and the continuity of f. Since $f(z)$ is analytic near $z = 1$, and $f(1) = 1$, in any small neighborhood of 1 it can be expanded as

$$f(z) = 1 + f'(1)(z-1) + \frac{1}{2}f''(1)(z-1)^2 + O((z-1)^3)$$

and we have

$$\begin{aligned} f(z^2) &= 1 + 2f'(1)(z-1) + (f'(1)^2 + f''(1))(z-1)^2 + O((z-1)^3)\,, \\ f(z)^2 &= 1 + 2f'(1)(z-1) + (f'(1) + 2f''(1))(z-1)^2 + O((z-1)^3) \end{aligned}$$

in the same neighborhood. We evaluate the coefficient following the proof of Lemma 2.4, using almost the same contour: define $\mathcal{C}'$ the same as $\mathcal{C}$ but with $|z| = r$ replaced by $|z| = \sqrt{r}$. Then

$$[z^n]\frac{f(z^2)-f(z)^2}{(1-f(z))^3} = \frac{1}{2\pi i}\oint_{\mathcal{C}'} \frac{f(z^2)-f(z)^2}{(1-f(z))^3 z^{n+1}}\,dz\,.$$

The part of the integral on $|z| = \sqrt{r}$ is bounded above in absolute value by $c_1/r^{n/2}$. Near $z = 1$, after expanding $1/g(z)^3 = 1/g(1)^3 + O(z-1)$ and using the above expansions, we see the integrand has a simple pole, with residue $-\big(f'(1) - f'(1)^2 + f''(1)\big)/f'(1)^3$. The result follows. ■

2.2. **Proofs of renewal results.** Here we are interested in the hitting probability $\mathbb{P}(\mathscr{H}_k)$ (where $\mathscr{H}_k = \bigcup_{j\ge 0}\{S_j = k\}$) for the above defined renewal sequence $(S_i)_{i\ge 1}$ without replacement.

Proof of Theorem 2.1. Since the x_j are positive integers, to transfer results on $\mathbb{P}(\mathscr{H}_k)$ from 'with replacement' to 'without replacement', we can restrict ourselves to considering the first k holding times, $Y_1, \dots, Y_k$, regarded as a random sequence. We use ν to denote the probability measure in the case of renewals with replacement, where each $Y_j = x_{\sigma(j)}$ is independently chosen u.a.r. from $[m]$, and π in the case that σ is a random permutation of $[m]$. Define a *duplicate* in σ to be a pair (i,j), $i<j\le k$, for which $\sigma(i)=\sigma(j)$, and let D denote the number of duplicates. Then

$$\nu(\mathscr{H}_k) = \nu(\mathscr{H}_k \mid D=0)\nu(D=0) + \nu(\mathscr{H}_k \mid D\ge 1)\nu(D\ge 1)\,, \tag{2.2}$$

and clearly

$$\nu(\mathscr{H}_k \mid D=0) = \pi(\mathscr{H}_k)\,.$$

Note that $\mathbb{E}_\nu D = \binom{k}{2}/m$ and $\mathbb{E}_\nu\binom{D}{2} = O(k^4/m^2)$ which imply by inclusion-exclusion that (recalling $k=o(\sqrt{m})$ from the theorem's hypothesis)

$$\nu(D=1) = \binom{k}{2}m^{-1} + O(k^4/m^2)\,, \qquad \nu(D\ge 2) = O(k^4/m^2)\,. \tag{2.3}$$

To make use of (2.2), it only remains to estimate $\nu(\mathscr{H}_k \mid D\ge 1)$. For this, we define a modified probability space in which the probability of σ is weighted by $D(\sigma)$. Let $\Omega = \{(\sigma,i,j) : (i,j) \text{ is a duplicate in } \sigma\}$, and endow Ω with the uniform probability measure. By symmetry, to generate (σ,i,j) at random from Ω, we can first select (i,j) u.a.r. from $[k]$ and r u.a.r. from $[m]$, then set $\sigma(i)=\sigma(j)=r$, and finally generate the rest of σ by sampling independently from $[m]$ as with ν. Let μ denote this probability measure on Ω. Considering this generation procedure, and noting that it "plants" a duplicate, it is immediate that

$$\mu(D\ge 2) \le \mathbb{E}_\mu\binom{D}{2} = O(k^2/m) = o(1)$$

since the expected number of pairs of non-planted duplicates is $O(k^2/m)$ and $k=o(\sqrt{m})$. Noting the equivalence of μ and ν conditioned on the event $D=1$, this implies

$$\mu(\mathscr{H}_k) = \mu(\mathscr{H}_k \mid D=1) + O(k^2/m) = \nu(\mathscr{H}_k \mid D=1) + O(k^2/m) = \nu(\mathscr{H}_k \mid D\ge 1) + O(k^2/m)$$

in view of (2.3). Substituting this into (2.2), we find

$$\begin{aligned}\pi(\mathscr{H}_k) = \nu(\mathscr{H}_k \mid D=0) &= \frac{\nu(\mathscr{H}_k) - \nu(\mathscr{H}_k \mid D\ge 1)\nu(D\ge 1)}{\nu(D=0)} \\ &= \frac{\nu(\mathscr{H}_k) - \big(\mu(\mathscr{H}_k) + O(k^2/m)\big)\nu(D\ge 1)}{1-\nu(D\ge 1)}\,.\end{aligned} \tag{2.4}$$

So we may concentrate on estimating $\mu(\mathscr{H}_k)$. Define T on $\mathscr{H}_k$ such that $\sum_{s=1}^{T} x_{\sigma(s)} = k$, and partition the event $\mathscr{H}_k$ in $\Omega = \{(\sigma,i,j)\}$ into events A, B and C as follows.

$$A = \{i<j\le T\}\,, \quad B = \{i\le T<j\}\,, \quad C = \{T<i<j\}\,.$$

Letting $\mathscr{H}_k(t)$ denote the event that $\sum_{s=1}^{t} x_{\sigma(s)} = k$, we can sum over t and the value ℓ of $x_{\sigma(i)}$ to get

$$\binom{k}{2}\mu(A) = \sum_{t\geq 1}\sum_{l\geq 1}\binom{t}{2}\nu\big(\mathscr{H}_{k-2\ell}(t-2)\big)p_\ell\,.$$

Here the prefactor arises because $\binom{k}{2}^{-1}$ is the probability of picking the pair (i,j) in the above description of generating of a random element of Ω. It follows by elementary considerations that

$$\mu(A) = \binom{k}{2}^{-1}[z^k]\frac{f(z^2)}{\big(1-f(z)\big)^3}$$

with f as in Theorem 2.1. Similarly,

$$\binom{k}{2}\big(\mu(B)+\mu(C)\big) = \sum_{t\geq 1}\sum_{l\geq 1} t(k-t)\nu\big(\mathscr{H}_{k-\ell}(t-1)\big)p_\ell + \binom{k-t}{2}\nu\big(\mathscr{H}_k(t)\big)p_\ell\,.$$

Rewriting $\binom{k-t}{2}$ as $\binom{k}{2} - t(k-t) - \binom{t}{2}$ and using

$$\sum_{l\geq 1}\nu\big(\mathscr{H}_{k-\ell}(t-1)\big)p_\ell = \nu\big(\mathscr{H}_k(t)\big) = \sum_{l\geq 1}\nu\big(\mathscr{H}_k(t)\big)p_\ell\,,$$

we find that the terms containing $t(k-t)$ cancel. The remaining ones are

$$\sum_{t\geq 1}\sum_{l\geq 1}\binom{k}{2}\nu\big(\mathscr{H}_k(t)\big)p_\ell = \sum_{t\geq 1}\binom{k}{2}\nu\big(\mathscr{H}_k(t)\big) = \binom{k}{2}\nu(\mathscr{H}_k)$$

and

$$-\sum_{t\geq 1}\sum_{l\geq 1}\binom{t}{2}\nu\big(\mathscr{H}_k(t)\big)p_\ell = -\sum_{t\geq 0}\binom{t+2}{2}\nu\big(\mathscr{H}_k(t+2)\big) = -[z^k]\frac{f(z)^2}{\big(1-f(z)\big)^3}$$

with f as above. Assembling all this,

$$\mu(\mathscr{H}_k) = \mu(A)+\mu(B)+\mu(C) = \nu(\mathscr{H}_k) + \binom{k}{2}^{-1}[z^k]\frac{f(z^2)-f(z)^2}{\big(1-f(z)\big)^3}\,.$$

Substituting this into (2.4) we find

$$\pi(\mathscr{H}_k) = \nu(\mathscr{H}_k) - \frac{\nu(D\geq 1)}{1-\nu(D\geq 1)}\left(\binom{k}{2}^{-1}[z^k]\frac{f(z^2)-f(z)^2}{\big(1-f(z)\big)^3} + O(k^2/m)\right).$$

Recalling (2.3), we get

$$\frac{\nu(D\geq 1)}{1-\nu(D\geq 1)} = \binom{k}{2}m^{-1} + O(k^4/m^2)$$

and hence

$$\pi(\mathscr{H}_k) = \nu(\mathscr{H}_k) - \big(m^{-1}+O(k^2/m^2)\big)[z^k]\frac{f(z^2)-f(z)^2}{\big(1-f(z)\big)^3} + O(k^4/m^2)\,.$$

The constants implicit in the $O(\cdot)$ terms arise in (2.3) and consequently are functions of k and m alone. Since $p_0 = 0$, we have $P_k = \pi(\mathscr{H}_k)$ and $R_k = \nu(\mathscr{H}_k)$, and the theorem follows. ∎

Proof of Theorem 2.2. This follows from Theorem 2.1 combined with Lemmas 2.4 and 2.5, the only condition not explicitly assumed being condition (i) of Lemma 2.4. This follows from the fact that $\sum p_\ell z^\ell$ converges absolutely for $|z| \le r$ as we assumed in (i) that $p_\ell \le (1/c)r^{-\ell}$. ■

Proof of Theorem 2.3. Conditional upon $\hat{Y}_J = x_J = \ell$ say, the event that $k \in \{\hat{Y}_1, \dots, \hat{Y}_m\}$ has probability $\pi(\mathscr{H}_{k-\ell})$ where μ is defined using the random permutation σ of $[m] \setminus \{J\}$. Recall also that $\sum_\ell \ell p_\ell = n/m$. Hence

$$\mathbb{P}\left(k \in \{\hat{Y}_1, \dots, \hat{Y}_m\}\right) = \sum_{\ell \ge 1} \frac{\ell p_\ell}{n/m} \pi(\mathscr{H}_{k-\ell})\,. \tag{2.5}$$

We can ignore all terms in this summation with $\ell \ge \log_r(m \log m)$, as by condition (i) they are dominated by the $o(1/m)$ error term in the theorem's claim. For $\ell < \log_r(m \log m)$, we will apply Theorems 2.1 and 2.2 to the multiset of positive integers $\{x_1, \dots, x_m\} \setminus \{x_J\}$, noting that their sum is $n - \ell = n\big(1 + O(\log m)/m\big)$, f is replaced by $\hat{f} := f - z^\ell/m$, and the p_ℓ change accordingly, to values we call $\hat{p}_\ell$. Moreover, we use $\hat{r} = r - \varepsilon$ in place of r, where we choose $\varepsilon > 0$ such that $\sqrt{r - \varepsilon} > 1 + \delta$, i.e. $\varepsilon < r - (1+\delta)^2$. Truth of the theorem for such restricted ε implies that it holds for all larger ε.

We first verify conditions (i–iii) of Theorem 2.2 for $\hat{f}$ etc. We have for all $j \ne \ell$ that

$$\hat{p}_j = p_j n/(n-\ell) = p_j\big(1 + O(\log m)/m\big) < p_j r/(r - \varepsilon)\,, \tag{2.6}$$

and $\hat{p}_\ell \le p_\ell$, which imply that $\sum_{\ell \ge 1} \hat{p}_\ell \hat{r}^\ell \le c$ (n sufficiently large), as required for (i). Note this implies that $\hat{f}(z) \le c$ for $|z| \le r - \varepsilon$, a fact that we will use before long. We turn next to (iii). For $|z| \le r - \varepsilon$, we have $z^\ell/m \le (r-\varepsilon)^{(\log m + \log\log m)/\log r}/m = o(1)$, and hence, using the second-last estimate in (2.6), $\hat{f}(z) = f(z) + o(1)$. So, by assumption (ii) of the present theorem and noting that $r - \varepsilon > 1 + \delta$, when $|z| = r - \varepsilon$, we have $|\hat{f}(z) - 1| = |f(z) - 1| + o(1) > \hat{c}_0^{-1}$ where $\hat{c}_0 = 2c_0$ say. This gives what is required for Theorem 2.2(iii) for $\hat{f}$ and $\hat{r}$. Regarding Theorem 2.2(ii), we first note that $\hat{f}'(1) = f'(1) + o(1)$ on $|z| \le r - \varepsilon$ for much the same reason as $\hat{f} = f + o(1)$. Hence (ii) implies for n sufficiently large that $\hat{f}(z) \ne 1$ when $|z| \le r - \varepsilon$ and $|z - 1| \ge \delta$. On the other hand, (iii) implies that the mean value of $f'(z)$ on any path in $|z - 1| < \delta$ has absolute value at least $f'(1)/2 \ge 1/2$, so the mean value of $\hat{f}'(z)$ on such a path cannot be zero. Thus, in this disc, the unique solution of $\hat{f}(z) = 1$ is at $z = 1$. We now have Theorem 2.2(ii) for $\hat{f}$ and $\hat{r}$.

Applying Theorems 2.1 and 2.2 in this context, we obtain

$$\pi(\mathscr{H}_{k-\ell}) = \frac{m-1}{n-\ell} - \big(m^{-1} + O(k^2/m^2)\big)\frac{\hat{f}'(1) - \hat{f}'(1)^2 + \hat{f}''(1)}{\hat{f}'(1)^3} + \hat{q}_k(n)$$

with

$$|\hat{q}_k(n)| \le \hat{c}_0(r-\varepsilon)^{-k} + O(m^{-1}c_1(r-\varepsilon)^{-k/2}) + O(k^4/m^2)\,.$$

Again by assumption (ii) of the present theorem, $c_1 \le \hat{c}_0^3 \max_{|z| = \sqrt{r-\varepsilon}} |\hat{f}(z^2) - \hat{f}(z)^2|$. Recalling that $|\hat{f}(z)| \le c$ and noting that $c \ge 1$ so $c^2 \ge c$, we get $c_1 \le c^2\hat{c}_0^3 \le 8c^2c_0^3$, a constant. Since $r - \varepsilon > 1$ and $k \to \infty$, and $\hat{c}_0 = 2c_0$, we obtain

$$|\hat{q}_k(n)| \le \hat{c}_0(r-\varepsilon)^{-k} + O\left(k^4/m^2\right) + o(1/m) \tag{2.7}$$

where $o(\cdot)$ depends on the rate that $k \to \infty$. Note that $k^2/m^2 = o(1/m)$, $\hat{f}'(1) = f'(1) + o(1)$, $\hat{f}''(1) = f''(1) + o(1)$, and $(m-1)/(n-\ell) = (m-1)/n + \ell m/n^2 + o(1/n)$. (Also, $n \geq m$ since each x_j is a positive integer.) So, recalling that $\sum_{\ell \geq 1} \ell p_\ell = n/m$, (2.5) gives

$$\mathbb{P}\left(k \in \{\hat{Y}_1, \dots, \hat{Y}_m\}\right) = \frac{m-1}{n} - \frac{f'(1) - f'(1)^2 + f''(1)}{m f'(1)^3} + q_k(n) + \sum_{\ell \geq 1} \frac{\ell^2 p_\ell m^2}{n^3},$$

where $|q_k(n)|$ has an upper bound of the same form as given in (2.7). Since $\sum_{\ell \geq 1} \ell^2 p_\ell = f'(1) + f''(1)$ and $f'(1) = n/m$, cancellation now gives the theorem. ∎

Proof of Theorem 3. Theorem 2.2(i) holds for n sufficiently large since $|f(r) - g(r)| < w(n) = o(1)$ and $g(r)$ is finite by the absolute convergence assumption on g. The convergence of f and g in $|z| \leq r$ also imply that they are analytic, and so all their derivatives exist, in $|z| < r$. We have $|f'(1) - g'(1)| < w(n) = o(1)$ and hence $g'(1) = n/m + o(1)$. Thus for some $\delta > 0$ with $1 + \delta + \varepsilon' < r$ (we can assume ε' is arbitrarily small, so this is always possible) we have $|g'(z) - g'(1)| < g'(1)/3$ if $|z - 1| < \delta$, $z \in \mathbb{C}$. Arguing as above, this implies Theorem 2.3(iii) for n sufficiently large. Since $g(z) \neq 1$ for $|z| \leq r$ whenever $z \neq 1$, and by the continuity of g, there exists $c_0 > 0$ such that $|g(z) - 1| \geq 2c_0^{-1}$ if $|z| \leq r$ and $|z - 1| \geq \delta$, $z \in \mathbb{C}$, similarly yielding both Theorem 2.2(iii) and Theorem 2.3(ii). Theorem 2.2(ii) follows from these, as was shown in the proof of Theorem 2.3. So the hypotheses of both of these theorems hold. In the conclusions, we note for (a) (the non-size-biased case) that the difference between $\frac{g'(1) - g'(1)^2 + g''(1)}{g'(1)^3}$ and the corresponding function of f is $o(1)$ by the hypotheses of the corollary and the fact that $g'(1) \neq 0$, which is absorbed in the error term $o(m^{-1})$ in q_1. So Theorem 2.1 completes the proof of (a). On the other hand, (b) (the size-biased case) follows immediately from Theorem 2.3. ∎

3. Second moment analysis of cycle factors: framework

In this section we set the framework for a delicate asymptotic analysis of the second moment of the number of k-cycle factors in a random 3-regular multigraph generated by the *configuration model* $\mathcal{P}_{n,3}$. In this model, introduced first by Bollobás [8] (a different flavor of it was implicitly used by Bender and Canfield [6]), one associates each of the n vertices with a triplet of distinct points (also referred to as "half-edges") and consider a uniform perfect matching (a pairing) on these points. The random 3-regular multigraph is obtained by contracting each of the triplets into a single vertex (possibly introducing multiple edges and self-loops). Clearly, on the event that this produces a simple graph, it is uniformly distributed among all cubic graphs, and furthermore, this event occurs with probability bounded away from 0. Hence, every event that occurs w.h.p. for this model, also occurs w.h.p. for a random 3-regular graph, and it will suffice to prove our results in this framework. See [10, 18, 33] for additional information.

Let CF_k denote the number of k-cycle factors in a configuration of $\mathcal{P}_{n,3}$. The bulk of the analysis of CF_k will be carried out in §4, where our goal is to establish that $\mathbb{E}[\mathsf{CF}_k^2] \leq (3 + o(1))\mathbb{E}[\mathsf{CF}_k]^2$ provided $k \geq K_0(n)$, with $K_0(n)$ as defined in Eq. (1.1). We first estimate the first moment, $\mathbb{E}[\mathsf{CF}_k]$, which is trivial by comparison.

Rather than working directly with CF_k, it will be more convenient to consider a re-scaled version of this variable. A k-cycle factor is *ordered* if its cycles are linearly ordered, *rooted* if each cycle is rooted at one of its vertices (i.e., a vertex of each cycle is distinguished as its *root*), and *directed* if each cycle is given one of the two possible orientations. Let

$$Y_k = \#\{\textit{rooted, ordered} \text{ and } \textit{directed } k\text{-cycle factors in } G \sim \mathcal{P}_{n,3}\}\,,$$

and call members of this set *ROD factors* for brevity. Observe that each k-cycle factor corresponds to precisely $(n/k)!(2k)^{n/k}$ distinct ROD factors, so that

$$\frac{Y_k}{\mathsf{CF}_k} = \left(\frac{n}{k}\right)!\,(2k)^{n/k} = (1+o(1))\sqrt{\frac{2\pi n}{k}}\left(\frac{2n}{e}\right)^{n/k}\,. \tag{3.1}$$

In the next subsections we establish the first and second moment of Y_k, followed by an analysis of Y_k given prescribed sequences of small cycle counts (to be used in the framework of the small subgraph conditioning method).

3.1. **Expected number of cycle factors.** It is straightforward to estimate the expectation of Y_k. A given rooted, ordered and directed cycle factor is in one-to-one correspondence with a permutation on the n vertices (we will often use this form), and in order to give rise to it in $\mathcal{P}_{n,3}$ we choose the incoming and outgoing half-edges of each vertex in its cycle (a total of 6^n options) and then add a perfect matching on the remaining half-edges, for which there are

$$\mathfrak{M}(n) := (n-1)!! = \frac{n!}{(\frac{n}{2})!\,2^{n/2}}$$

options. Recalling that $|\mathcal{P}_{n,3}| = \mathfrak{M}(3n)$ we can conclude that

$$\mathbb{E}Y_k = (1+o(1))\frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot 6^n \cdot \sqrt{2}\left(\frac{n}{e}\right)^{n/2}}{\sqrt{2}\left(\frac{3n}{e}\right)^{3n/2}} = (1+o(1))\sqrt{2\pi n}\left(\tfrac{4}{3}\right)^{n/2}\,. \tag{3.2}$$

It thus follows from (3.1) that

$$\mathbb{E}[\mathsf{CF}_k] = (1+o(1))\sqrt{k}\left(\frac{e(4/3)^{k/2}}{2n}\right)^{n/k}\,. \tag{3.3}$$

Observe that whenever $e(4/3)^{k/2}/(2n) \geq 1$, which occurs when k is at least

$$2\log_{4/3}(2n/e) = [1-\tfrac{1}{2}\log_2 3]^{-1}\log_2(2n/e)\,,$$

we have $\mathbb{E}[\mathsf{CF}_k] \geq (1+o(1))\sqrt{k} \to \infty$. On the other hand, if for instance $e(4/3)^{k/2}/(2n) < 1 - \frac{k\log k}{n}$ we get that $\mathbb{E}[\mathsf{CF}_k] \to 0$, which holds for any $k \leq K_0(n) - 2\log_{4/3}[1/(1-\frac{k\log k}{n})]$. As the last additive term is of order $\frac{k\log k}{n} = o(\frac{\log^2 n}{n})$, we conclude that $\mathbb{E}[\mathsf{CF}_k] = o(1)$ whenever $k \leq K_0(n) - \frac{\log^2 n}{n}$.

3.2. **Second moment via intersection patterns.** To estimate $\mathbb{E}Y_k^2$ we will count the triples in the set

$$\{(G, R_1, R_2) \ : \ G \in \mathcal{P}_{n,3} \text{ and } R_1, R_2 \text{ are ROD factors of } G\} \ , \tag{3.4}$$

where we say that R is a ROD factor of a pairing G if each of its underlying cycles belongs to G. Define the following quantity to capture the potential underlying intersection graph of two factors:

$$\mathcal{I}_{h,m} = \left\{ \begin{array}{c} \text{undirected graphs on } [n] \text{ with } h+m \text{ components:} \\ h \text{ cycles of length } k \text{ and } m \text{ nontrivial paths} \end{array} \right\} \tag{3.5}$$

(where here and in what follows, a path is nontrivial if it has at least 2 vertices). We refer to the graphs — usually denoted S — in any $\mathcal{I}_{h,m}$ as *intersection patterns*.

For two ROD factors R_1, R_2 and $S \in \mathcal{I}_{h,m}$ we say $R_1 \cap R_2$ *projects* to S, denoted by $R_1 \cap R_2 \hookrightarrow S$, if the underlying undirected graph of $R_1 \cap R_2$ is equal to S (that is, the undirected edges in common in R_1, R_2 give S).

Crucially, $R_1 \cap R_2$ necessarily projects to some $S \in \mathcal{I}_{h,m}$ for some h, m. Indeed, let R_1, R_2 be two ROD factors. Upon removing the ordering, rooting and directing of the cycles we are left with some number h of k-cycles that belong two both factors. On the remaining vertices we have two k-cycle factors H_1 and H_2 of a 3-regular multigraph on $n-kh$ vertices, whose intersection is a set of simple nontrivial paths with the following property: if we traverse a common path along a k-cycle of H_1, then as soon as this path ends we see an edge exclusive to H_1 followed immediately by a new nontrivial common path (and similarly for H_2). This is because H_2 forms a spanning 2-regular multigraph in our 3-regular multigraph, and therefore there cannot be two edges exclusive to H_1 incident to the same vertex.

Say that a ROD factor *contains* $S \in \mathcal{I}_{h,m}$ if its underlying undirected graph contains S, and set

$$\mathscr{N}(S) = \#\left\{\text{ROD factors containing } S\right\} \tag{3.6}$$

(for any h, m and $S \in \mathcal{I}_{h,m}$). Apart from the orienting of the paths, and the rooting and directing of the cycles, $\mathscr{N}(S)$ corresponds to the number of different ways one can weave the m paths in $S \in \mathcal{I}_{h,m}$ into a collection of $n/k - h$ disjoint k-cycles by repeatedly connecting the ends of two paths by an edge.

We now claim that the quantity defined in (3.4) satisfies

$$\#(G, R_1, R_2) = \sum_h \sum_m \sum_{S \in \mathcal{I}_{h,m}} \mathscr{N}(S)^2 6^n \mathfrak{M}(n-2m) \, . \tag{3.7}$$

Indeed, the factor $\mathscr{N}(S)^2$ accounts for choosing R_1, R_2 given the intersection pattern $S \in \mathcal{I}_{h,m}$. Once we compose the factors R_1 and R_2, the $2m$ endpoints of the m paths in S reach degree 3 (each having two exclusive edges in R_1, R_2 and one common edge in S) and all other vertices are of degree 2. Thus, it remains to complete the pairing by matching the latter $n-2m$ vertices, and finally to order the three half-edges of each vertex (the factor of 6^n).

Recalling that $|\mathcal{P}_{n,3}| = \mathfrak{M}(3n)$ it then follows from (3.7) that

$$\mathbb{E}Y_k^2 = \frac{6^n}{\mathfrak{M}(3n)} \sum_h \sum_m \mathfrak{M}(n-2m) \sum_{S \in \mathcal{I}_{h,m}} \mathscr{N}(S)^2 \, . \tag{3.8}$$

The lion's share of the work involved in estimating this quantity will be devoted to intersection patterns with $h = 0$ (i.e., cycle-free) and m close to $n/3$; these will turn out to account for most of the right-hand side of (3.8). The contribution of the remaining pairs h, m will then be shown, partly by reducing to the earlier analysis, to be negligible in comparison with these main terms.

The next theorem is the heart of the matter. Its proof, which relies, *inter alia*, on the renewal estimates of §2, is given in §4.

Theorem 3.1. *Let m and n be such that $|\frac{m}{n} - \frac{1}{3}| \le (\log n)^{-1/3}$, and define $\mathscr{N}(S)$ as in (3.6). If $k > (2+\varepsilon)\log_2 n$ for some fixed $\varepsilon > 0$ then*

$$\sum_{S \in \mathcal{I}_{0,m}} \mathscr{N}(S)^2 \le (9 + o(1))\,(m!\,2^m)^2\,|\mathcal{I}_{0,m}|\ .$$

What we will actually find is that most of this sum comes from those $S \in \mathcal{I}_{0,m}$ for which $\mathscr{N}(S) \le (3 + o(1))m!2^m$. We may contrast this with, for example, the crude bound

$$\mathscr{N}(S) \le k^{n/k} m!2^m\,. \tag{3.9}$$

Here we think of each member of $\mathscr{N}(S)$ as corresponding to a permutation of the m paths, gotten by listing the cycles in their ROD order and then, within each cycle, beginning with the path containing the root and listing the remaining paths in the order dictated by the orientation of the cycle. The factors 2^m and $k^{n/k}$ account for orienting the paths and rooting the cycles. Of course this is wasteful in two ways: first, we have overpaid for the roots (which are chosen from the first path in a cycle rather than from the entire cycle), and second, the only permutations arising from S are those with the property that all multiples of k are partial sums of their sequence of path sizes. In particular, consideration of the probability of the property just mentioned leads naturally to renewals and the material of §2.

3.3. **Properties of a typical intersection pattern.** To begin, we will argue that

$$|\mathcal{I}_{0,m}| = \binom{n-m-1}{m-1}\frac{n!}{m!2^m} \tag{3.10}$$

and that the multiset of path sizes of a uniform $S \in \mathcal{I}_{0,m}$ follows the same law as the multiset of part sizes of a uniformly chosen member of $\mathcal{C}_{0,m}$, the set of m-compositions of the integer n with parts of size at least 2. To see this, regard $T = (a_1, \dots, a_m) \in \mathcal{C}_{0,m}$ as a graph in $\mathcal{I}_{0,m}$ consisting of the paths $(1, \dots, b_1), (b_1 + 1, \dots, b_2), \dots, (b_{m-1} + 1, \dots, b_m)$, where $b_i = a_1 + \cdots + a_i$ (in particular $b_m = n$). Each relabeling of the vertices of T by $[n]$ gives an $S \in \mathcal{I}_{0,m}$ whose path sizes coincide with the part sizes of T, and, conversely, this procedure (choose T, relabel its vertices) gives rise to each $S \in \mathcal{I}_{0,m}$ precisely $m!2^m$ times. This gives the second assertion above, and also (3.10) once we recall that $|\mathcal{C}_{0,m}| = \binom{n-m-1}{m-1}$.

The next lemma, which will be used several times in §4, addresses large deviations for the path sizes of a typical intersection pattern S, as well as a random subset of S.

Lemma 3.2. *Let $|\frac{m}{n} - \frac{1}{3}| = o(1)$ and let S be chosen uniformly from $\mathcal{I}_{0,m}$. Then for any t and ℓ such that $t\ell = o(n)$, the probability that S contains at least t paths of length at least ℓ is at most*

$$\left(\frac{4em}{t(2-o(1))^\ell}\right)^t. \tag{3.11}$$

Moreover, if $|\frac{m}{n} - \frac{1}{3}| < \varepsilon$ for some $0 < \varepsilon = \varepsilon(n) = o(1)$ and p_ℓ is the fraction of ℓ-vertex paths among the m paths in S for some $\ell = \ell(n) \geq 2$, then

$$\mathbb{P}\left(\frac{2-\varepsilon}{(2+\varepsilon)^\ell} \leq p_\ell \leq \frac{2+\varepsilon}{(2-\varepsilon)^\ell}\right) \geq 1 - O\left(\sqrt{m}\ell e^{-\frac{1}{8}\varepsilon^2 m/(\ell 3^\ell)}\right). \tag{3.12}$$

Proof. According to the discussion preceding the lemma, it suffices to prove the statements in question for the distribution of part sizes of a uniformly chosen member of $\mathcal{C}_{0,m}$. Here it will be convenient to work with another (standard) reformulation, identifying each member of $\mathcal{C}_{0,m}$ with a binary string W of length $n-m-1$ and Hamming weight $m-1$; namely such a word with 1's in positions $a_1, \ldots, a_{m-1}$ corresponds to the composition $(a_1+1, a_2-a_1+1, \ldots, a_{m-1}-a_{m-2}+1, n-m-a_{m-1}+1)$. In particular, a maximal interval of zeros of length ℓ corresponds to a part of size $\ell+2$ in the composition.

We will first establish (3.11). Let Z^*_ℓ be the number of parts of size at least ℓ; equivalently, $Z^*_\ell = \sum_{i=0}^{n-m-\ell+1} I^*_i$ where

$$I^*_i = \mathbb{1}_{\{W_i=1 \text{ and } W_{i+1}=\ldots=W_{i+\ell-2}=0\}} \quad (i = 0, \ldots, n-m-\ell+1).$$

We can only have $I^*_i = I^*_j = 1$ if $|i-j| \geq \ell-1$ (i.e., if the sets $\{i, \ldots, i+\ell-2\}$ and $\{j, \ldots, j+\ell-2\}$ are disjoint), and it is easy to see that

(a) there are $\binom{n-m-1-t(\ell-2)}{t}$ possible t-subsets $T \subset [n-m-\ell+1]$ satisfying $|i-j| \geq \ell-1$ for all $i \neq j \in T$, and for each one there are $\binom{n-m-1-t(\ell-1)}{m-1-t}$ strings with Hamming weight $m-1$ such that $I^*_i = 1$ for all $i \in T$;

(b) there are $\binom{n-m-1-t(\ell-2)}{t-1}$ such t-subsets $T \subset \{0, \ldots, n-m-\ell+1\}$ that include 0, and for each one there are $\binom{n-m-t(\ell-1)}{m-t}$ strings as above.

Thus,

$$\begin{aligned}\mathbb{P}(Z^*_\ell \geq t) &\leq \frac{\binom{n-m-1-t(\ell-2)}{t}\binom{n-m-1-t(\ell-1)}{m-1-t} + \binom{n-m-1-t(\ell-2)}{t-1}\binom{n-m-t(\ell-1)}{m-t}}{\binom{n-m-1}{m-1}} \\ &= \left[1 + \frac{t}{m-t}\right]\frac{1}{t!}\frac{(m-1)_t\,(n-2m)_{t(\ell-2)}}{(n-m-1)_{t(\ell-2)}}.\end{aligned}$$

Since $t\ell = o(n)$ by hypothesis and $m \sim n/3$, we see that $\frac{t}{m-t} = o(1)$ and that each of the factors in the numerator of the last fraction is $(1+o(1))m$ and each one in its denominator is $(2-o(1))m$. Using $t! \geq (t/e)^t$, this gives

$$\mathbb{P}(Z^*_\ell \geq t) \leq \left(\frac{em}{t\,(2-o(1))^{\ell-2}}\right)^t,$$

which is (3.11).

It remains to establish (3.12). To this end, as before we can move to the setting where $\tilde{W}$ is a binary string of length $n-m-1$. In this case however we take each coordinate as an independent Bernoulli($\frac{m-1}{n-m-1}$) variable. We will account for $\mathbb{P}(\sum_i \tilde{W}_i = m-1) \gtrsim 1/\sqrt{m}$ later. For notational convenience, extend $\tilde{W}$ by defining $\tilde{W}_0 = \tilde{W}_{n-m} = 1$.

Let $\tilde{Z}_\ell$ ($\ell \geq 2$) count the number of parts of size ℓ in the composition corresponding to $\tilde{W}$. We can write $\tilde{Z}_\ell$ as a sum of indicators, $\tilde{Z}_\ell = \sum_{i=0}^{n-m-\ell} \tilde{I}_i$, where $\tilde{I}_i = \tilde{I}_i(\ell)$ is the event that coordinates $i, i+1, \ldots, i+\ell-1$ form a part of size exactly ℓ; that is,

$$\tilde{I}_i = \left\{ \begin{array}{c} \tilde{W}_i = \tilde{W}_{i+\ell-1} = 1\,, \\ \tilde{W}_{i+1} = \ldots = \tilde{W}_{i+\ell-2} = 0 \end{array} \right\} \quad (i = 0, \ldots, n-m-\ell+1)\,.$$

Then, for any $i \notin \{0, n-m-\ell+1\}$,

$$\mathbb{P}(\tilde{I}_i = 1) = \left(\frac{1}{2} + O\left(|\tfrac{m}{n} - \tfrac{1}{3}|\right) + O(1/n)\right)^\ell = \left(\tfrac{1}{2} + o(1)\right)^\ell\,. \tag{3.13}$$

Similarly the events $\tilde{I}_0, \tilde{I}_{n-m-\ell+1}$ have a probability of $(\frac{1}{2}+o(1))^{\ell-1}$ each. The events $\{\tilde{I}_i : i \in A\}$ are clearly mutually independent if A is a set of indices whose pairwise distances are all at least ℓ. We can thus partition $\tilde{Z}_\ell$ into $\tilde{Z}_\ell = \sum_{j=0}^{\ell-1} \tilde{Z}_\ell^{(j)}$ where

$$\tilde{Z}_\ell^{(j)} := \sum_{i \equiv j (\bmod \ell)} \tilde{I}_i$$

satisfies the following stochastic domination relations for large enough n:

$$\mathrm{Bin}\left(\lfloor \tfrac{n-m-\ell}{\ell} \rfloor, (2+\varepsilon)^{-\ell}\right) \preccurlyeq \tilde{Z}_\ell^{(j)} \preccurlyeq \mathrm{Bin}\left(\lceil \tfrac{n-m-\ell}{\ell} \rceil, (2-\varepsilon)^{-\ell}\right) + 2\,.$$

(The additive 2 on the right-hand side accounted for the events $\tilde{I}_0, \tilde{I}_{n-m-\ell}$.) The binomial variables on the left and right have means $(2-o(1))\frac{m}{\ell}(2+\varepsilon)^{-\ell}$ and $(2+o(1))\frac{m}{\ell}(2-\varepsilon)^{-\ell}$ respectively. Thus,

$$\mathbb{P}\left(\frac{m}{\ell}\frac{2-\varepsilon}{(2+\varepsilon)^\ell} < \tilde{Z}_\ell^{(j)} < \frac{m}{\ell}\frac{2+\varepsilon}{(2-\varepsilon)^\ell}\right) \geq 1 - O\left(e^{-(1-o(1))\frac{1}{6}\varepsilon^2 \frac{m}{\ell} 3^{-\ell}}\right) \geq 1 - O\left(e^{-\frac{1}{8}\varepsilon^2 m/(\ell 3^\ell)}\right)\,,$$

where the first inequality follows from standard large deviation estimates for the binomial (see, e.g., [18, Corollary 2.3]). A union bound over $j = 0, \ldots, \ell-1$ now completes the proof. ∎

As a corollary, we have the following rough bound, which for $m \sim n/3$ improves on the trivial estimate $\mathscr{N}(S) \leq k^{n/k} m! 2^m$ from (3.9) for all but an ultimately negligible proportion of the intersection patterns.

Lemma 3.3. *Let* $E = \{S \in \mathcal{I}_{0,m} \,:\, \mathscr{N}(S) > (\log k)^{3n/k} m! 2^m\}$ *for* m *satisfying* $m = (\frac{1}{3} + o(1))n$. *Then*

$$\sum_{S \in E} \mathscr{N}(S)^2 = o\left((m!2^m)^2 |\mathcal{I}_{0,m}|\right)\,.$$

Proof. Apply the first part of Lemma 3.2 with

$$t = \frac{n}{k}\frac{\log\log k}{\log k}\,, \quad \ell = \log^2 k$$

(noting that $t\ell = o(n)$ as needed in that lemma). as the base of the exponent in (3.11) for this choice of t and ℓ is $(2-o(1))^{-\log^2 k}$, the probability of at least t paths of length at least ℓ is at most

$$(2-o(1))^{-(n/k)\log k \log\log k} < k^{-5n/k}$$

for any sufficiently large k. Revisiting (3.9), such path partitions contribute $o((m!2^m)^2|\mathcal{I}_{0,m}|)$ to the summation of $\mathscr{N}(S)^2$. On the other hand, the partitions S with at most t paths of length at least ℓ have less than $k^t\ell^{n/k}$ ways to choose the root vertices from the initial paths of each cycle, and hence

$$\mathscr{N}(S) \le k^t\ell^{n/k}m!2^m = e^{(n/k)(\log\log k + 2\log\log k)}m!2^m\,,$$

and so no such partitions are in E. The result now follows. ∎

4. Cycle-free intersection patterns and renewals

4.1. **Normal and abnormal intersection patterns.** At its most naïve level, the proof of Theorem 3.1 would like to estimate (really meaning bound) $\mathscr{N}(S)$ for a given S by randomly ordering the paths and estimating — via the results of §2 — the probability that all multiples of k are partial sums of the resulting sequence of path sizes, as well as bounding the number of corresponding ROD factors. (Each ordering with this property — already mentioned following Theorem 3.1 — gives rise to a number, crudely between $2^{|S|}$ and $2^{|S|}k^{n/k}$, of ROD factors via directing paths and rooting cycles, while other orderings do not contribute to $\mathscr{N}(S)$. However, such bounds are too rough for our ultimate goal.)

The most obvious problem with this is that the statistics (p_ℓ) of S itself, or of some of the subsets of S remaining after some paths have been chosen ("suffixes"), might not support the use of Theorem 3. In extremely rough terms, one may say that much of the work below is aimed at reducing to situations where these results do apply (though even in the ideal situation where all such suffixes do behave nicely, treated in Theorem 4.4 below, the analysis is more delicate than the preceding description suggests).

We will use — with various values of δ — the following concrete conditions supporting use of our renewal estimates.

Definition 4.1 (δ-normal intersection patterns)**.** Let S be a set of m paths and let $x_1,\dots,x_m$ denote the lengths (numbers of vertices) of these paths. The *path-distribution* $(p_\ell)_{\ell\ge1}$ of S gives the relative frequencies of the x_i's, namely $p_\ell = \frac{1}{m}|\{j : x_j = \ell\}|$ for $\ell > 0$. Given $0 < \delta < 1$, we say that S is δ-*normal* if the following conditions hold:

(i) Short paths: $p_1 = 0$ and for all $\ell \le \mathscr{M} := \frac{1}{8}\log\log k$,

$$\left|p_\ell - 2^{1-\ell}\right| \le \varepsilon_\ell = \varepsilon_\ell(\delta) := \left(\ell^4(2-\delta)^\ell \log^{1/8} k\right)^{-1}\,. \tag{4.1}$$

(The lower bound on p_ℓ is meaningful since $\varepsilon_\ell \le e^{-\ell}$ for any $\ell \le \mathscr{M}$, recalling $k\to\infty$.)

(ii) Long paths: for all $\ell \ge \mathscr{M}$,

$$p_\ell \le \gamma_\ell = \gamma_\ell(\delta) := \left(\ell^4(2-\delta)^\ell\right)^{-1}\,. \tag{4.2}$$

Remark 4.2. For every δ-normal $S \in \mathcal{I}_{0,m}$, there are no paths of length $\ell \geq \log_{2-\delta} n$ (as $k \leq n$ so such paths are long, and then $\gamma_\ell \ll 1/n \leq 1/m$), and the following hold:

$$m = \left(\tfrac{1}{3} + o(1)\right) n \,, \tag{4.3}$$

$$\sum_{\ell=2}^{\mathscr{M}} \ell^2 (2-\delta)^\ell \varepsilon_\ell = o(1) \,, \tag{4.4}$$

$$\sum_{\ell \geq \mathscr{M}} \ell^2 (2-\delta)^\ell p_\ell = o(1) \,. \tag{4.5}$$

(Condition (4.3), in place of the explicit bound in Theorem 3.1, suffices for much of what we do.)

In line with our discussion above, we will be able to estimate the number of ROD factors arising from any δ-normal intersection pattern quite accurately, as the following theorem states.

Theorem 4.3. *For every $\varepsilon > 0$ there exists some $0 < \delta < 1$ such that the following holds: if $k \geq (2+\varepsilon)\log_2 n$ and $S \in \mathcal{I}_{0,m}$ is δ-normal then*

$$\mathscr{N}(S) \leq (3 + o(1)) m! 2^m \,.$$

Our plan for the remainder of §4 is as follows. We begin here with the aforementioned Theorem 4.4, which deals with situations in which we never encounter an abnormal path distribution; it is in the proof of this result, which will be central to our argument, that the results of §2 will come into play. We then, in §4.2, derive Theorem 3.1 from Theorem 4.3. The proof of Theorem 4.3 itself, the trickiest part of the whole business, is given in §4.3.

For $t \in [n/k]$, the *t-suffix* of a ROD factor is the sequence of paths in its last t cycles. (To define "sequence" we may regard the *first* path in a cycle as the one containing the root, but actually in what follows we will really only be interested in the *set* of paths in a suffix.)

Let $\mathscr{N}_\delta(S)$ be the number of ROD factors containing S in which each of the n/k suffixes has a δ-normal path distribution. (In particular, $\mathscr{N}_\delta(S) = 0$ unless S is δ-normal.) The following theorem, central to our proof, provides an estimate on $\mathscr{N}_\delta(S)$ using the estimates on renewal processes without replacement given in §2.

Theorem 4.4. *For every $\varepsilon > 0$ there exists some $0 < \delta < 1$ such that the following holds: if $k \geq (1+\varepsilon)\log_2 n$ then for any $S \in \mathcal{I}_{0,m}$,*

$$\mathscr{N}_\delta(S) \leq (3 + o(1)) m! 2^m \,.$$

The proof of this (somewhat paradoxically) requires bounding the probability of δ-abnormal suffixes. For this we use the next two assertions concerning δ'-normality of subsets of a δ-normal intersection pattern. The first of these addresses random (uniform) subsets.

Claim 4.5. *Fix $0 < \delta < \delta' < 1$. Let S' be a uniform m'-subset of a δ-normal set of paths $S \in \mathcal{I}_{0,m}$. The following hold.*
(1) If $m' \geq m/\log\log k$ then

$$\mathbb{P}(S' \text{ is } \delta'\text{-abnormal}) \leq e^{-mk^{-o(1)}} \,.$$

(2) If S has no path of length at least $\log_{2-\delta} k$ and we let Σ' denote the number of vertices in S' and $\theta_0 = 1 - \log_{2-\delta}(2-\delta')$ then for $n' \le n$,

$$\mathbb{P}(S' \text{ is } \delta'\text{-abnormal}, \Sigma' = n') \le \sqrt{n'} e^{-n' k^{-1+\theta_0 - o(1)}} .$$

Proof. Throughout this proof, write p_ℓ for the relative frequency of ℓ-vertex paths in S and let p'_ℓ be the analogous quantity for S'.

Consider some $\ell \le \mathscr{M}$, and note that for such ℓ we have $\left|p_\ell - 2^{1-\ell}\right| < \varepsilon_\ell(\delta)$ by the δ-normality of S. As the number of ℓ-paths in S' is hypergeometric with mean $m' p_\ell$, Hoeffding's inequality for hypergeometric variables [16, Theorem 2 and §6] implies that for any $\alpha > 0$,

$$\mathbb{P}\left(\left|p'_\ell - p_\ell\right| \ge \alpha\right) \le 2\exp(-2\alpha^2 m') .$$

Substituting $\alpha = \varepsilon_\ell(\delta')/\log k$, for instance, we get that with probability $1 - \exp(-m' k^{-o(1)})$ we have $|p'_\ell - p_\ell| = o(\varepsilon_\ell(\delta'))$. In this case, combining the facts $\left|p_\ell - 2^{1-\ell}\right| < \varepsilon_\ell(\delta)$ and $\varepsilon_\ell(\delta')/\varepsilon_\ell(\delta) = \left(1 + \frac{\delta'-\delta}{2-\delta'}\right)^\ell > 1$ implies (recalling δ and δ' are fixed) that $|p'_\ell - 2^{1-\ell}| < \varepsilon_\ell(\delta')$. A union bound now establishes the short-path condition for all $\ell \le \mathscr{M}$ with the same probability guarantee of $1 - \exp(-m' k^{-o(1)})$ (since we can assume $m' \ge 1$).

Part (1) of the claim now easily follows, since the long-path condition is fulfilled deterministically for S', as the following argument shows. Let $\ell \ge \mathscr{M}$. By hypothesis, there are at most $m\gamma_\ell(\delta)$ paths of length ℓ in S, and thus also in S', yielding that

$$p'_\ell \le (m/m') p_\ell \le \gamma_\ell(\delta) \log\log k$$

using our hypothesis on m' in this part. However, $\gamma_\ell(\delta')/\gamma_\ell(\delta) = \left(1 + \frac{\delta'-\delta}{2-\delta'}\right)^\ell$, which is at least $(\log k)^c$ for some $c = c(\delta, \delta') > 0$. For large enough n, this outweighs the $\log\log k$ factor from above and gives $p'_\ell \le \gamma_\ell(\delta')$ for all $\ell \ge \mathscr{M}$.

It remains to prove Part (2). First we will show that $\Sigma' \le 4m'$ with probability $1 - \exp(-n' k^{-o(1)})$. Recall that we can assume that S is δ-normal, and hence $m = m/3 + o(n)$ by (4.3). If m' differs from $n'm/n \sim n'/3$ by more than $n'/100$ (say) then, using the bound on the maximal length of a path, Hoeffding's inequality for the concentration of hypergeometric variables [16] again shows that for some absolute constant $c > 0$

$$\mathbb{P}(\Sigma' = n') \le \exp\left(- c\frac{(n')^2}{m'(\log_{2-\delta} k)^2}\right) \le \exp\left(-n' k^{-o(1)}\right) ,$$

using the fact that $m' \le n'/2$ (otherwise this probability is 0 since every path in S has length at least 2). Comparing this probability with the desired estimate in the statement of the claim, we may now assume $m' \ge n'/4$ (as $\theta_0 < 1$ is fixed).

It will be convenient to approximate S' via $\tilde{S}$ that contains each path from S independently with probability m'/m. As in the proof of Lemma 3.2, since $\mathbb{P}(|\tilde{S}| = m') \gtrsim \frac{1}{\sqrt{m'}} \gtrsim \frac{1}{\sqrt{n'}}$ and on this event $\tilde{S}$ has the same distribution as S', one can infer properties of S' from those of $\tilde{S}$ via a multiplicative cost of $O(\sqrt{n'})$ in the probability.

The number ℓ-vertex paths in $\tilde{S}$ is simply a $\mathrm{Bin}(mp_\ell, m'/m)$ variable, where p_ℓ is the relative frequency of ℓ-vertex paths in S. Consider $\mathscr{M} \le \ell \le \log_{2-\delta} k$ (with the upper bound justified by

our hypothesis in this part). The number of ℓ-vertex paths in S' would violate δ'-normality only if it should exceed

$$m'\gamma_\ell(\delta') = m'\gamma_\ell(\delta)\left(\frac{2-\delta}{2-\delta'}\right)^\ell .$$

As such, large deviation estimates for the binomial distribution, together with the fact that $p_\ell \le \gamma_\ell$ thanks to the δ-normality of S, show that the probability of this event is at most

$$\exp\left(-m'\gamma_\ell(\delta)\left(\frac{2-\delta}{2-\delta'}\right)^{\ell+o(\ell)}\right) = \exp\left(-m'(2-\delta')^{-(1+o(1))\ell}\right)$$

(the $o(\ell)$ term is an additive term of order $\log \ell$ working in our favor, but that will not be needed). Since $\ell \le \log_{2-\delta} k$ we have $(2-\delta')^\ell = (2-\delta)^{(1-\theta_0)\ell} \le k^{1-\theta_0}$, and now a union bound over $\mathscr{M} \le \ell \le \log_{2-\delta} k$, including the factor $\sqrt{n'}$ from above, completes the proof. ∎

The second of the above-mentioned supporting assertions for proving Theorem 4.4 says that for $\delta' > \delta$ and sufficiently large n, any large enough subset of a δ-normal path distribution is (deterministically) δ'-normal.

Claim 4.6. *Fix $0 < \delta < \delta' < 1$ and $\alpha > 0$. If $S \in \mathcal{I}_{0,m}$ is δ-normal and S' is an m'-subset of S for $m' \ge m - nk^{-\alpha}$ then S' is δ'-normal for large enough n.*

Proof. Let p_ℓ and p'_ℓ be the relative frequencies of ℓ-vertex paths in S and S', resp.

First consider the long-path condition. By the δ-normality of S, for any $\ell \ge \mathscr{M}$ we have at most $m\gamma_\ell(\delta)$ paths of length ℓ in S. Thus, using (4.3) to see that $m \sim m'$,

$$p'_\ell \le (m/m')\gamma_\ell(\delta) = (1+o(1))\gamma_\ell(\delta) < \gamma_\ell(\delta') ,$$

with the last inequality following from the fact, already used in the proof of Claim 4.5, that $\gamma_\ell(\delta')/\gamma_\ell(\delta) = \left(1+\frac{\delta'-\delta}{2-\delta'}\right)^\ell \to \infty$ with ℓ (as a poly-log).

Now take $\ell \le \mathscr{M}$. The δ-normality of S implies that $\left|p_\ell - 2^{1-\ell}\right| < \varepsilon_\ell(\delta)$. Observe that if $d_\ell \le nk^{-\alpha}$ is the number of ℓ-vertex paths in $S \setminus S'$ then

$$\left|p'_\ell - p_\ell\right| = \frac{1}{m'}\left|mp_\ell(1-m'/m) - d_\ell\right| \le \frac{m-m'}{m'} + \frac{d_\ell}{m'} = O(k^{-\alpha}) .$$

This term is therefore negligible compared to $\varepsilon_\ell(\delta) \ge k^{-o(1)}$, and so

$$|p'_\ell - 2^{1-\ell}| \le (1+o(1))\varepsilon_\ell(\delta) .$$

The proof is now concluded by the fact $\varepsilon_\ell(\delta')/\varepsilon_\ell(\delta) = \left(1+\frac{\delta'-\delta}{2-\delta'}\right)^\ell > 1$. ∎

With the above two claims at our disposal, we turn to Theorem 4.4.

Proof of Theorem 4.4. Fix $\varepsilon_0 > 0$ arbitrarily small. For sufficiently small δ, the upper bound on $\mathscr{N}_\delta(S)$ will be established by induction: we will show that if S' has m' paths and a total of $n' = rk$ vertices then deterministically

$$\mathscr{N}_\delta(S') \le (3+\varepsilon_0)e^{\sum_{t=2}^r \frac{\psi(t)}{kt}}(m')!2^{m'} ,$$

for some function

$$\psi(t) = \psi(k,t) = \begin{cases} O(k^{3/4}) & t < k^4 \\ o(1) & t \geq k^4 \,. \end{cases}$$

This will of course imply the desired statement on $\mathscr{N}_\delta(S)$ (recalling that ε_0 can be taken arbitrarily small) since then for $r = n/k$ the exponential pre-factor will be

$$1 + O\Big(\sum_{t<k^4} \frac{k^{3/4}}{kt}\Big) + o\Big(\sum_{t=k^4}^{n/k} \frac{1}{kt}\Big) = 1 + k^{-1/4+o(1)} + o\Big(\frac{\log n}{k}\Big) = 1 + o(1)\,,$$

where the last inequality used the fact that k has order at least $\log n$.

We may (and do) assume that S' is δ-normal, since otherwise $\mathscr{N}_\delta(S') = 0$ by definition. Hence, (4.3) implies that $n' \sim 3m'$, a fact we will use several times. For $r = 1$ clearly there are $(m'-1)!2^{m'}$ possibilities for ordering and directing the m' paths into an ordered and directed cycle, which then offers $n' = k$ possible roots to form a ROD factor. Since $k = n' \sim 3m'$, indeed $\mathscr{N}_\delta(S) \leq (3+\varepsilon_0)(m')!2^{m'}$ for n sufficiently large.

Next, consider $n' = rk$ for some $r > 1$. A natural way to carry out the induction would be to count the ways to select some ordered subset $A \subset S'$ of paths which together sum up to k vertices, direct each of them, distinguish one of the k points as a root and proceed to select the remaining cycles recursively. We must then divide by j, being the number of ordered subsets that produce the same cyclic order prior to rooting. While the part involving the root is easy, it turns out that estimating the probability that some j random paths sum up to k, divided by j, would call for a delicate joint estimate of typical values of j and their corresponding probabilities of hitting k, and it is unclear how to estimate these probabilities to the desired accuracy.

Instead, here we will compose the first cycle out of some j paths differently. We first select its root (for which there are n' possible candidates) corresponding to some path in S', complement it with a selection of some $j-1$ paths in an ordered manner, multiply by the 2^j ways of directing the paths and finally weigh in the probability that the paths thus chosen contain k vertices. This corresponds to the renewal problem studied in §2: we choose an element of S' in a size-biased sample (the root path) and permute the remaining elements uniformly, then ask for the probability of hitting k via one of the partial sums. Let $\mathscr{H}_k$ denote this event, let $\mathscr{H}_k^{(j)}$ denote the event of having the first j elements sum to k and, for a given subset A, write $\mathscr{H}_k(A)$ for the event that its elements sum up to k. The astute reader will notice that we have just introduced notation for the very event whose probability we mentioned above is hard to estimate. Next we will proceed to wave our magic wand and make $\mathscr{H}_k^{(j)}$ disappear in a puff of summation signs.

By slight abuse of notation we write $v \in A$ for a vertex v to denote that A contains a path going through v. Immediately from the definition, any $(r-1)$-suffix of a ROD factor counted in $\mathscr{N}_\delta(S')$

is also counted in $\mathscr{N}_\delta(S' \setminus A)$. Hence, we can apply the induction hypothesis to get that

$$\begin{aligned}\mathscr{N}_\delta(S') &= \sum_j \sum_{v\in[n']} \sum_{\substack{A\subset S'\\ |A|=j,\, v\in A}} (j-1)!2^j \mathbb{1}_{\mathscr{H}_k(A)} \mathscr{N}_\delta(S'\setminus A)\\ &\le \sum_j (j-1)!2^j \sum_{v\in[n']} \sum_{\substack{A\subset S'\\ |A|=j,\, v\in A}} \mathbb{1}_{\mathscr{H}_k(A)} (3+\varepsilon) e^{\sum_{t=2}^{r-1}\frac{\psi(t)}{kt}} (m'-j)!2^{m'-j}\\ &\le (3+\varepsilon)e^{\sum_{t=2}^{r-1}\frac{\psi(t)}{kt}} n'(m'-1)!2^{m'}\mathbb{P}(\mathscr{H}_k)\,,\end{aligned}$$

where the last inequality is by the fact that the number of choices for a distinguished vertex v, followed by $j-1$ *ordered* elements of S' together summing to k, is precisely $n'(m'-1)\cdots(m'-j+1)\mathbb{P}\big(\mathscr{H}_k^{(j)}\big)$, and so we can collect the factorial terms independently of j and remain with $\sum_j \mathbb{P}\big(\mathscr{H}_k^{(j)}\big) = \mathbb{P}(\mathscr{H}_k)$.

We now wish to show that $\mathbb{P}(\mathscr{H}_k) \le \left(1+\frac{\psi(r)}{n'}\right)\frac{m'}{n'}$. If $r \ge k^4$ we appeal to the estimate in Theorem 3(b) for the size-biased renewal process without replacement. Let $g(z) = \sum_{\ell\ge 2} 2^{1-\ell} z^\ell$ and let $R = 2-\delta$. Then $g(z)$ is absolutely convergent for $|z| \le R$, where its value is $z^2/(2-z)$, and hence the unique solution of $g(z) = 1$ in $|z| \le R$ is at $z = 1$. Also let $f(z) = \sum_{\ell\ge 2} p_{-\ell} z^\ell$. Since S' is δ-normal we have $|f(z)-g(z)| + |f'(z)-g'(z)| + |f''(z)-g''(z)| < w(n)$ for $|z| \le R$, where $w(n) = o(1)$ by (4.4) and (4.5) and w depends only on δ and k. Thus, f and g satisfy the requirements of Theorem 3, and so by (b) we have for any $\varepsilon' > 0$

$$|Q_k - m'/n'| \le o(1/m') + O\big((R-\varepsilon')^{-k} + k^4/(m')^2\big)\,.$$

Recalling (4.3), $m' \asymp n'$, and since $k^4 \le r = n/k$ we have $k^4/(m')^2 = o(1/n')$. To make the other error term $o((1/n')$ as well, since $k > (1+\varepsilon)\log_2 n$ we can simply choose $\delta = \delta(\varepsilon)$ sufficiently small such that $(1+\varepsilon)\log_2(2-\delta) > 1$, and then choose ε' to be sufficiently small enough such that $(R-\varepsilon')^{-k} < (2-\delta-\varepsilon')^{(1+\varepsilon)\log_2 n} = o(1/n')$. That is, we have $\psi(r) = o(1)$ in this case.

When $2 \le r \le k^4$ (equivalently, $n' \le k^5$) we modify this strategy to bypass the error term $O(k^4(n')^{-2})$ (which is large enough to foil the entire framework already when $n' \asymp k^2$). Again we choose the first path in a size-biased way, so as to determine the root of the next cycle, but now we condition on this path. Denoting its length by ℓ_0, observe that $\ell_0 \le \log_{2-\delta} n' \le 5\log_{2-\delta} k$ by the assumption that $n' = rk \le k^5$.

Note that the remaining paths have a deterministically δ_0-normal path distribution for any $\delta_0 > \delta$ and large enough n thanks to Claim 4.6 (as we move from a δ-normal set of m' paths to an $(m'-1)$-subset of it).

Subsequently, we choose paths uniformly (without replacement), but this time we do so until reaching (or exceeding) a total of $y = \lfloor k - 10\log_{2-\delta} k\rfloor$ vertices in these paths (including the first one). Let m_0 and n_0 denote the numbers of remaining paths and vertices at that point, and write

$$\Delta_m = m' - 1 - m_0\,, \quad \Delta_n = n' - \ell_0 - n_0$$

for the numbers of paths and vertices sampled uniformly without replacement en route. Again, the maximal path length assumption implies that

$$n' - y - 5\log_{2-\delta} k < n_0 \leq n' - y\,, \tag{4.6}$$

whence $\Delta_n = y - \ell_0 + O(\log k)$.

We now claim that $|\Delta_m - \frac{m'-1}{n'-\ell_0}(y-\ell_0)| \leq k^{3/4}$ except with probability, say, $O(k^{-100})$. Indeed, the sum Σ of the lengths of the paths $2,\ldots,w$ sampled as above is hypergeometric with mean $(w-1)(n'-\ell_0)/(m'-1)$. Thus for a specific value of w where $2 \leq w \leq k$ with $|w - \frac{m'-1}{n'-\ell_0}(y-\ell_0)| > k^{3/4}$, and a specific v with $0 \leq v \leq 5\log_{2-\delta} k$, we have that

$$\mathbb{P}(\Sigma = n' - y - v) \leq \exp\left(-(1-o(1))\frac{k^{3/2}}{2w}\right) \leq \exp\left(-(1/2-o(1))\sqrt{k}\right)$$

by Hoeffding's inequality (using our bound on the maximal path length). Summing this probability over w and v is easily $O(k^{-100})$, and we may assume henceforth that this event does not occur.

Finally, we wish to infer from Claim 4.5 that the path distribution on the remaining m_0 paths (uniformly chosen out of the $m'-1$ paths that are left after positioning the leading size-biased one) is δ'-normal for any $\delta' > \delta_0$ except with probability $O(k^{-100})$. This is achieved as follows:

- If $r \geq \log k$ then $m_0 \sim m'$ and so Part (1) of that claim implies that the probability of a δ'-abnormal path distribution is at most $\exp(-k^{1-o(1)})$ (even after enumerating over the possible values of m_0).
- If $r \leq \log k$ then $\log_{2-\delta} k \sim \log_{2-\delta} n'$, and since the maximal length of a path in S is $o(\log_{2-\delta} n')$, we may appeal to Part (2) of that claim. This results in a probability of $\exp(-k^{\theta_0+o(1)})$ for a δ'-abnormal path distribution, even after enumerating over the (polynomial in k) number of possibilities for m_0 and the total number of paths in our m_0-subset.

We now proceed to choose paths uniformly from the remaining m_0 paths without replacement. Our target to hit is $k - \Delta_n$. Defining g, R and f as for our application of Theorem 3(b) above, we may appeal to the renewal without replacement estimate from Theorem 3(a), and deduce that the probability of hitting $k - \Delta_n$ is

$$\frac{m_0}{n_0} + O\left((2-\delta')^{-(k-(n'-n_0))}\right) + O(1/n')\,.$$

It is easy to see (using (4.6)) that

$$\frac{m_0}{n_0} = \frac{m'-1-\Delta_m}{n'-\ell_0-\Delta_n} = \frac{m'-1}{n'-\ell_0} + O\left(\frac{k^{3/4}}{n'}\right) + O\left(\frac{\log k}{n'}\right)\,.$$

Additionally,

$$\frac{m'-1}{n'-\ell_0} = \frac{m'}{n'} + O\left(\frac{\ell_0}{n'}\right) = \frac{m'}{n'} + O\left(\frac{\log k}{n'}\right)\,,$$

and combining these gives that the hitting probability is $\frac{m'}{n'} + O(k^{3/4}/n')$. That is, $\psi(r) = O(k^{3/4})$ in this case, as required. ∎

4.2. **Proof of Theorem 3.1 modulo Theorem 4.3.** Fix $0 < \delta' < \delta < 1$. For the reduction to δ-normality (and applicability of Theorem 4.3), it will turn out that long paths, and in particular violations of (4.2), are our main concern, since the short-path condition (4.1), even in its more restrictive version with δ', is satisfied with high enough probability that we may more or less ignore path distributions that violate it. We will deal with the long paths by disposing of them (that is, choosing the cycles containing them) first. Of course, this entails making sure that this "preprocessing" (i) is affordable and (ii) doesn't (usually) too badly distort the path distribution of what remains (a point that will exploit the slack between δ' and δ).

Recall that the short-path condition (4.1) for δ' says that

$$\left|p_\ell - 2^{1-\ell}\right| < \varepsilon_\ell = \varepsilon_\ell(\delta') = \frac{1}{\ell^4(2-\delta')^\ell \log^{1/8} k}$$

for $\ell \le \mathscr{M} = \frac{1}{8}\log\log k$. Notice that if $\left|x - 2^{1-\ell}\right| > \varepsilon_\ell$ for some $\ell \le \mathscr{M}$, then we violate the estimate $\frac{2-\varepsilon}{(2+\varepsilon)^\ell} < x < \frac{2+\varepsilon}{(2-\varepsilon)^\ell}$ (from (3.12)) for $\varepsilon = (\log n)^{-1/3}$, since the terms sandwiching x also sandwich $2^{1-\ell}$ and differ by a factor $1 + O(\varepsilon\ell) = 1 + (\log n)^{-1/3+o(1)}$, whereas $\varepsilon_\ell > (\log k)^{-1/4} > (\log n)^{-1/4}$ for large enough n.

Plugging the hypothesis $|\frac{m}{n} - \frac{1}{3}| < (\log n)^{-1/3}$ of Theorem 3.1 into (3.12) establishes that the probability of violating the short-path condition w.r.t. δ' is at most $\exp(-nk^{-o(1)}\log^{-2/3} n) \le \exp(-nk^{-2/3-o(1)})$. Comparing this to the factor $k^{n/k}$ in (3.9), we see that the total contribution to $\sum \mathscr{N}(S)^2$ from intersection patterns violating the short-path condition is $o\left((m!2^m)^2|\mathcal{I}_{0,m}|\right)$. (We could alternatively appeal to Lemma 3.3 for a factor of $(\log k)^{3n/k}$ but that would make little difference here.) In what follows, we may thus confine our attention to intersection patterns satisfying the short-path condition w.r.t. δ'.

Next, consider an intersection pattern S violating the long-path condition w.r.t. δ'. (Recall this says that $p_\ell \le \gamma_\ell(\delta') = 1/\left(\ell^4(2-\delta')^\ell\right)$ for every $\ell \ge \mathscr{M}$.)

We first claim that, following the same line of argument used above, we can reduce to the situation where no p_ℓ violates this condition for any $\mathscr{M} \le \ell \le \mathscr{M}^*_{\delta'}$, with

$$\mathscr{M}^*_{\delta'} := \log_{2-\delta'}\left(k/\log^5 k\right) .$$

Indeed, for such ℓ we have $\gamma_\ell(\delta') \ge c\frac{\log k}{k}$ for $c = c(\delta') > 0$, thus Eq. (3.11) with $t = m\gamma_\ell$ and (noting that $\ell\gamma_\ell \to 0$ as $\ell \to \infty$ and in particular $t\ell = o(m)$) yields

$$\mathbb{P}\left(p_\ell \ge \gamma_\ell(\delta')\right) \le \left(c'\ell^4(1-\delta'/2+o(1))^\ell\right)^{cm\frac{\log k}{k}} \le e^{-c''(n/k)\log k\log\log k} .$$

As before, the probability of encountering such an S nullifies its contribution to $\mathscr{N}(S)^2$, even via the rough overestimate of $k^{n/k}$ from (3.9).

At this point, only ℓ-vertex paths with $\ell \ge \mathscr{M}^*_{\delta'}$ can potentially violate the conditions required for δ-normality. For such ℓ with $p_\ell \ge \gamma_\ell(\delta')$, label an arbitrary set of $\lceil (p_\ell - \gamma_\ell(\delta'))m \rceil$ paths of length ℓ as *excess*.

Roughly speaking, our goal at this point is to somehow eliminate the excess paths and reduce to a δ-normal (as opposed to a δ'-normal) path distribution, to which we can, finally, apply Theorem 4.3. A key point is that we can achieve this by revealing the cycles (of the ROD factor) containing the

excess paths *first*. The number of ways to choose these early cycles will be somewhat larger than we would like, but this is more than compensated for by the fact that excesses are rare.

Let $R = R(S, \delta')$ denote the number of excess paths in S.

Claim 4.7. *Let $S \in \mathcal{I}_{0,m}$ be a uniformly chosen intersection pattern and let $\delta' > 0$. Then*

$$\mathbb{P}(R \geq r) \leq k^{-\frac{\delta' - o(1)}{2\log(2-\delta')} r} \quad \text{for any } r \geq 1 \,.$$

Proof. If there are r excess paths then for some $\ell \geq \mathscr{M}^*_{\delta'}$ there are at least $t = \lfloor m\gamma_\ell \rfloor + r$ paths of length at least ℓ in S; namely, for the least ℓ for which there is an excess, we have $\lfloor m\gamma_\ell \rfloor + s$ paths of length ℓ for some $s \geq 1$, and, if $s < r$, at least $r - s$ excess paths of length greater than ℓ. By (3.11), the probability of this event is at most

$$\left(\frac{4em}{m\gamma_\ell \left(2 - o(1)\right)^\ell} \right)^r = \left(1 - \delta'/2 + o(1)\right)^{\ell r} \leq e^{(-\delta'/2 + o(1))\ell r} \,,$$

where the first equality absorbed the factor $c\ell^4$ into the $o(1)$-term in the base of the exponent, justified by the fact that $\ell \to \infty$. The fact $\ell \geq \mathscr{M}^*_{\delta'} = (1 + o(1)) \log_{2-\delta'} k$ completes the proof. ∎

Claim 4.7 allows us to reduce to the case

$$R \leq \frac{n}{k \log\log k} \,, \tag{4.7}$$

since applying it with $r = n/(k \log\log k)$ gives $\mathbb{P}(R \geq r) \leq e^{-c\frac{n}{k}\frac{\log k}{\log\log k}}$, which outweighs the factor $(\log k)^{3n/k}$ from the bound on $\mathscr{N}(S)$ given by Lemma 3.3.

Claim 4.8. *Let $\delta > \delta' > 0$, and suppose $S \in \mathcal{I}_{0,m}$ satisfies the δ'-normality conditions for path lengths $\ell \leq \mathscr{M}^*_{\delta'}$ and has $r \leq n/k$ excess paths. For some $\bar{m} \sim m$ with $\bar{m} \leq m - r$, let $\bar{S}$ be a uniform $\bar{m}$-subset of the $m - r$ non-excess paths in S. Then*

$$\mathbb{P}\left(\bar{S} \text{ is } \delta\text{-normal}\right) \geq 1 - \exp\left(-\bar{m} k^{-o(1)}\right) \,.$$

Proof. Let U be the set of non-excess paths, let $u = m - r$ denote its size and let μ_ℓ denote the relative frequency of ℓ-vertex paths in U. We will show that, deterministically, U is δ'-normal for any $\delta' > \delta$, at which point the statement of the claim will follow immediately from Part (1) Claim 4.5 since $\bar{m} \sim u$ (both are asymptotically m).

First consider the long-path condition. By our definition of excess paths, for any $\ell \geq \mathscr{M}$ we have at most $m\gamma_\ell(\delta) \sim u\gamma_\ell(\delta)$ of length ℓ. As in the proof of Claim 4.5, we have $\gamma_\ell(\delta)/\gamma_\ell(\delta') = \left(1 + \frac{\delta - \delta'}{2 - \delta}\right)^\ell$, which diverges with ℓ (as a poly-log even), easily implying that $\mu_\ell < \gamma_\ell(\delta')$ for large enough n. To consider short paths, take $\ell \leq \mathscr{M}$. By hypothesis, $\left|p_\ell - 2^{1-\ell}\right| < \varepsilon_\ell(\delta')$ in S. Since by definition all excess paths have length greater than $\mathscr{M}^*_{\delta'}$, the set U contains all the ℓ-vertex paths in S and exactly $m(1 - p_\ell) - r$ others. It follows that μ_ℓ, the relative frequency of ℓ-vertex paths in U, satisfies

$$|\mu_\ell - p_\ell| \leq r/m = O(1/k) = o\left(\varepsilon_\ell(\delta')\right) \,,$$

where the last inequality holds for $\ell \leq \mathscr{M}$ since $\varepsilon_\ell(\delta') \geq \varepsilon_{\mathscr{M}}(\delta') = k^{-o(1)}$. The fact that $\varepsilon_\ell(\delta)/\varepsilon_\ell(\delta') = \left(1 + \frac{\delta - \delta'}{2 - \delta}\right)^\ell > 1$ now implies the short-path condition and completes the proof. ∎

Claim 4.9. *Let $\delta' > 0$, and suppose $S \in \mathcal{I}_{0,m}$ satisfies the δ'-normality conditions for path lengths $\ell \le \mathscr{M}^*_{\delta'}$ and has $r \le n/(k \log\log k)$ excess paths. Then*

$$\mathscr{N}(S) \le (3+o(1))^{r+1} 2^m m!\,.$$

Proof. To bound the number of ROD factors that S gives rise to, we first account for the cycles containing the r excess paths, and then bound the number of ways to fill in the remaining cycles and produce a ROD factor.

We will use q to denote the number of cycles that contain excess paths in the ROD factor, and let x denote the total number of paths in these q cycles. Note that trivially

$$\lceil r/k \rceil \le q \le r \qquad \text{and} \qquad r \le x \le qk = o(m)\,.$$

We will begin by counting all the ROD factors in which, once we filter out the above mentioned q cycles containing all excess paths, the remaining $m-x$ paths form a δ-abnormal distribution. For given q and x, one first chooses these $m-x$ paths out of the $m-r$ possible non-excess ones. As $x = o(m)$, Claim 4.8 shows that an $e^{-mk^{-o(1)}}$ fraction of these choices results in a δ-abnormal path distribution. For each such choice, there are at most $2^{m-x}(m-x)!k^{n/k-q}$ ways to order and orient the paths, and then root the cycles. As for the remaining q cycles, we select an excess path for each of these, order and orient the remaining $x-q$ paths, then root the cycles and insert them into the list of n/k cycles with a final factor of $(n/k)_q$. The number of such ROD factors (i.e., with a δ-abnormal suffix) is thus at most

$$\begin{aligned} &e^{-mk^{-o(1)}} \binom{m-r}{x-r}\binom{r}{q}(x-q)!\binom{n}{k}_q (m-x)!k^{n/k}2^m \\ =&e^{-mk^{-o(1)}} \binom{r}{q}\frac{(m-r)_{x-r}}{(m)_x}\frac{(x-q)!}{(x-r)!}\binom{n}{k}_q k^{n/k} m!2^m\,. \end{aligned} \tag{4.8}$$

As $(m)_x/(m-r)_{x-r} = \big((\frac{1}{3}-o(1))n\big)^r$, this expression (slightly rearranged) is at most

$$e^{-mk^{-o(1)}}2^r\left(\frac{3+o(1)}{k}\right)^q\left(\frac{(3+o(1))x}{n}\right)^{r-q} k^{n/k}m!2^m\,.$$

As $2^r k^{n/k} = e^{O((n/k)\log k)}$, even after summing over q and x this is $o(m!2^m)$. We may thus restrict our attention to ROD factors for which the cycles that do not contain excess paths induce a δ-normal distribution.

To count these, we number our excess paths from 1 to r (in an arbitrary way) and proceed as follows:

(i) Repeat the following steps until all excess paths are exhausted:
- Select a location for a new cycle (amongst the n/k slots), the lowest numbered remaining excess path to be a part of it (which we direct, as usual), and its root (given by its offset $i \in [k]$ from the start of the chosen path).
- Complete the cycle that contains this path (including directions) using any choice of the remaining paths, including excess ones.

(ii) Order (and direct) the remaining paths.

This gives an arrangement consisting of (i) a set of rooted, directed cycles containing all excess paths (where each cycle contains at least one excess path), each with its position in one of n/k cycle slots specified, and (ii) an ordering of the remaining (non-excess) paths. We claim the total number of arrangements is at most

$$(3+o(1))^r m!2^m\,. \tag{4.9}$$

Indeed, there are $m-i+1$ choices for the i-th path unless we are at the beginning of a new cycle, say the j-th cycle. In the latter case, the next path is dictated by the ordering of the excess paths, while we have $(n/k-j)k$ possibilities for positioning and rooting the new cycle. Since the excess paths are exhausted after some $q \le r$ cycles, at which point the number of remaining paths is at least $m-qk \ge m-rk = (1-o(1))m$, we can replace each such term $(n/k-j)k$ by the "default" term $m-i+1$ at a cost of $(1+o(1))m/n = 3+o(1)$. This establishes (4.9).

To complete the proof, partition the arrangements into classes according to the output of Step (i). For any given class, if x is the total number of paths in its set of excess cycles, by definition there are $(m-x)!2^{m-x}$ arrangements in the class, corresponding to Step (ii). On the other hand, we may assume the remaining $m-x$ paths have a δ-normal distribution by the previous discussion. Thus, we know by Theorem 4.3 that the number of ROD factors whose cycles with excess paths agree with this class, is at most $(3+o(1))(m-x)!2^{m-x}$, so at most $3+o(1)$ times the number of arrangements in the class. Altogether, $\mathscr{N}(S)$ is at most $3+o(1)$ times the total number of arrangements, as required. ∎

We are now in a position to complete the proof of Theorem 3.1. Let $\mathcal{E} = \mathcal{E}(S)$ be the event that $S \in \mathcal{I}_{0,m}$ satisfies the δ'-normality conditions for path lengths $\ell \le \mathscr{M}^*_{\delta'}$ and has at most $n/(k\log\log k)$ excess paths. Thus far, we have reduced the proof of the theorem to showing that

$$\mathbb{E}\left[\mathscr{N}(S)^2\mathbb{1}_{\mathcal{E}}(S)\right] \le (3+o(1))m!2^m\,,$$

where S is a uniformly chosen intersection pattern. Combining Claim 4.9 with the estimate in Claim 4.7 on the number of excess paths R, we get

$$\begin{aligned}\mathbb{E}\left[\mathscr{N}(S)^2\mathbb{1}_{\mathcal{E}}(S)\right] &= \sum_r \mathbb{E}\left[\mathscr{N}(S)^2\mathbb{1}_{\mathcal{E}}(S)\mid R=r\right]\mathbb{P}(R=r)\\ &\le (3+o(1))m!2^m\left(1+\sum_{r\ge 1}\left[(3+o(1))k^{-c(\delta')}\right]^r\right) = (3+o(1))m!2^m\end{aligned}$$

(where $c(\delta')$ is the constant from Claim 4.7 and the summand for $r=0$ is given by Theorem 4.3 as S is then δ'-normal), as required. ∎

4.3. **Abnormal suffix: proof of Theorem 4.3.** For an inductive proof based on placing cycles, the main problem is to bound number of ways a ROD factor can be formed such that the t-suffix is δ-normal for each $t > i$ and yet δ-abnormal for $t = i$. These two events (normal when more than i cycles remain, and abnormal when i remain) each have very small probability and we cannot assume independence. Moreover, given an abnormal i-suffix, a complementary prefix (of total length $n-ik$) might not be δ'-normal as a stand-alone path distribution for any useful δ'. These features make this argument rather twisted.

By the δ-normality of S, the number of paths that are of length at least $\log_{2-\delta} k$ is at most $m\sum_{\ell\geq\log_{2-\delta}k}\gamma_\ell(\delta)$, which is $O(m/(k\log^4 k)) < n/(k\log k)$ for large enough n. The first step in the proof is to reduce to the case where there are no such paths whatsoever, after which the following lemma would complete the proof.

Lemma 4.10. *For any fixed $\delta > 0$, if $S \in \mathcal{I}_{0,m}$ is δ-normal and contains no path of length at least $\log_{2-\delta} k$ then*

$$\mathscr{N}(S) \leq (3+o(1))m!2^m\,.$$

We postpone the proof of the above lemma in favor of first showing how to reduce to its setting, in a way similar to our treatment of excess paths in the proof of Claim 4.9.

Fix $\delta' > \delta$. Let T be the subset of all paths of length at least $\log_{2-\delta} k$ in S, and let $t = |T| \leq n/(k\log k)$. We first consider all the ROD factors containing S for which there is a subset $C_1,\ldots,C_{q-1}$ of the cycles and C_q, which is part (possibly all) of another cycle, such that every C_i contains some path in T and the distribution of $S\setminus\cup_{i\leq q}C_i$ is δ'-abnormal. To estimate the number of such factors, we sum over $q \leq t = o(n/k)$ and select q such paths (each for a separate cycle). Next, we sum over $x \leq kq = o(m)$, the total number of paths in these cycles, and run over all $\binom{m-q}{m-x}$ subsets for the remaining cycles. The combination of Claim 4.6 and Part (1) of Claim 4.5 implies that for any $\delta' > \delta$ an $O(\exp(-mk^{-o(1)}))$-fraction of these will result in a δ'-abnormal path distribution (the former addresses the normality of the $(m-q)$-subset and the latter treats its $(m-x)$-subsets). Thus, similarly to Eq. (4.8), the total number of such ROD factors is at most

$$\begin{aligned}&e^{-mk^{-o(1)}}\binom{t}{q}(m-q)_{x-q}\left(\frac{n}{k}\right)_q(m-x)!k^{n/k}2^m\\ \leq&e^{-mk^{-o(1)}}2^t\left(\frac{3+o(1)}{k}\right)^q k^{n/k}m!2^m\,,\end{aligned}$$

where we used the fact that $m-i=(1-o(1))m$ for $1\leq i\leq x$ to replace the term n^q by $(3+o(1))^q$. Since $q\leq t$ and $(6+o(1))^t k^{n/k} = \exp(mk^{-1+o(1)})$, the entire expression is $o(m!2^m)$, as required.

The other ROD factors (in which no set $C_1,\ldots,C_q$ as above leaves behind a δ'-abnormal path distribution) are handled by prioritizing the paths of length at least $\log_{2-\delta} k$, as done before with the excess paths for the proof of Claim 4.9. This time, however, it is crucial to estimate the probability that these align to k-cycles, because we cannot afford to give away any constant factor (let alone a larger term such as the $(3+o(1))^q$ in (4.9)). Using the same procedure as in that claim (order the paths in T arbitrarily, repeatedly select the lowest numbered such path and complete it into a cycle, and finally order the remaining paths), we now argue that in lieu of the estimate (4.9), the total number of arrangements is at most

$$(1+o(1))m!2^m\,.$$

To see this, as in the proof of Theorem 4.4, we appeal to one of two strategies depending on the relation between k and n:

- If $k = o(\sqrt{n})$, we will appeal to our renewal estimate after placing each of the q leading paths in $C_1,\ldots,C_q$. Namely, recall that upon forming the $(j+1)$-st cycle with a leading path from

T — letting ℓ_0 denote the length of this path — the set of remaining paths (excluding the leading path) is δ'-normal by our current assumption. Implementing Theorem 3(a) as before, we see that the probability of hitting the partial sum $k-\ell_0$ in a random permutation over the remaining elements is $\frac{m'-1}{n'-\ell_0}+O(1/n')+O((2-\delta')^{-(k-\ell_0)})$, where m' and $n'=n-jk$ are the numbers of paths and vertices left after the first j cycles, respectively. Thus, the $n'=(n/k-j)k$ choices for positioning and rooting the cycle can be replaced by the "default" term m' at a multiplicative cost of $1+O(\ell_0/n')=1+O(k/n)$ (as the total number of vertices in these q cycles is at most $qk\le tk=o(n)$). Repeating this procedure for all $q\le t$ cycles, then finally ordering and directing all remaining paths, bounds the number of arrangements by $\exp(O(tk/n))m!2^m=(1+o(1))m!2^m$, as claimed.

- If $k\gtrsim\sqrt{n}$, we tweak the above approach by revealing the paths that follow the leading path in C_{j+1} until reaching at least $y=\lfloor k-10\log_2 k\rfloor$ vertices in that cycle. Let m_0 and n_0 be the numbers of paths and vertices remaining at that point. By our assumption, these remaining paths form a δ'-normal distribution; thus, the renewal estimate from Theorem 3(a) is $m_0/n_0+O(1/n')$. As $n_0=n'-y+O(\log k)$, this is $m'/n'+O(k^{3/4}/n')+O(k^{-100})$ due to the variability in m_0. Hence, we are again entitled to replace the term $n'=(n/k-j)k$ by m', this time at a cost of $1+O(\ell_0/n')+O(k^{3/4}/n')$. Accumulating these errors over the $q\le t$ cycles gives a factor of $\exp(O(tk/n)+k^{-1/4+o(1)})=1+o(1)$, and therefore a total of at most $(1+o(1))m!2^m$ arrangements.

The final step is to partition the arrangements into classes according to the cycles involving T; if these contain a total of x paths in some given class then this class contains $(m-x)!2^{m-x}$ arrangements. The path distribution on the remaining $m-x$ paths is δ'-normal and contains no path of length at least $\log_{2-\delta}k$, and so Lemma 4.10 guarantees that the number of ROD factors agreeing with this class is at most $(3+o(1))(m-x)!2^{m-x}$. Altogether, $\mathscr{N}(S)\le(3+o(1))m!2^m$ modulo Lemma 4.10.

With the values of n and k understood, we say that a path distribution P on $n-kr$ points is *r-close to δ-normal* if there is some set of paths on a total of rk points which, when added to P, gives a δ-normal path distribution with maximum length at most $\log_{2-\delta}k$ (in particular, no paths of greater length exist in P itself). Such path distributions are the ones that can conceivably result when deleting (the last) r cycles from the path distributions under consideration.

To complete the proof of Lemma 4.10 we will use the following.

Lemma 4.11. *Fix $\delta'>\delta>0$ and let $0<\theta<\theta_0$ for $\theta_0=\theta_0(\delta,\delta')$ as was given in Claim 4.5. For all $r=r(n)\ge1$,*

(i) If S' is a set of m' paths on $n-rk$ vertices that is r-close to δ-normal, then for any sufficiently large n, the number of ROD *factors (consisting of $n/k-r$ cycles) that contain S' is at most $(m')!2^{m'}\exp\left(rk^\theta\right)$.*

(ii) For any δ-normal $S\in\mathcal{I}_{0,m}$ with no paths of length at least $\log_{2-\delta}k$, if n is large enough then the number of ROD *factors arising from S which have a δ'-abnormal r-suffix is at most $m!2^m\exp\left(-rk^\theta\right)$.*

(For our purposes, the bound in Part (ii) could be replaced by any other that would give $o(2^m m!)$ when summed over r.)

Proof. We first prove Part (i) by downward induction on r. Let

$$r_0 = \frac{n}{k^{1+\theta}} \log k \,.$$

First consider $r \geq r_0$. There are $(m')!2^{m'}$ possibilities for ordering and directing the paths (ignoring the need to hit multiples of k). The choices of roots in the lead paths give an extra factor at most $(\log_{2-\delta} k)^{n/k}$ by the upper bound on path length. Since $r \geq r_0$, the claimed upper bound is at least $(m')!2^{m'} \exp[(n/k)\log k]$ and so Part (i) holds in this case.

Now take $1 \leq r < r_0$ while assuming that Part (i) holds for all $r < r' \leq r_0$. Fix $\tilde{\delta} > \delta'$ and, for $\tilde{r} \geq 1$, let $\mathcal{R}_{\tilde{r}} = \mathcal{R}_{\tilde{r}}(S')$ be the set of ROD factors arising from S' where the $\tilde{r}$-suffix is $\tilde{\delta}$-abnormal but all shorter suffixes are $\tilde{\delta}$-normal. (It suffices to treat this case since the number of ROD factors in which all suffixes are $\tilde{\delta}$-normal is $\mathscr{N}_{\tilde{\delta}}(S) \leq (3+o(1))(m')!2^{m'}$ by Theorem 4.4, a fraction of $O(\exp(-k^\theta)) = o(1)$ out of the desired upper bound.)

As S' is r-close to δ-normal, it is deterministically δ'-normal by Claim 4.6 using $rk \leq r_0 k = nk^{-\theta+o(1)}$. Claim 4.5 therefore implies (via Part (2) of that claim, noting that S has no paths of length at least $\log_{2-\delta'} k \geq \log_{2-\delta} k$) that if $\tilde{S}$ is a uniform $\tilde{m}$-subset of S' and $\tilde{\Sigma}$ is its number of vertices then

$$\mathbb{P}\left(\tilde{S} \text{ is } \tilde{\delta}\text{-abnormal}, \, \tilde{\Sigma} = \tilde{r}k\right) \leq \exp\left(-\tilde{r}k^{\theta_0 - o(1)}\right) \tag{4.10}$$

(where we absorbed the prefactor $\sqrt{\tilde{r}k}$ from that bound into the $o(1)$-term). We will now argue that the number of ROD factors in $\mathcal{R}_{\tilde{r}}$ that have $\tilde{m}$ paths in the $\tilde{r}$-suffix and x paths in the $(n/k - \tilde{r}+1)$-st cycle is at most

$$\begin{aligned} &\left[\binom{m'}{\tilde{m}} e^{-\tilde{r}k^{\theta_0-o(1)}}\right]\left[(m'-\tilde{m})!2^{m'-\tilde{m}} e^{(r+\tilde{r})k^\theta}\right] \\ &\cdot \left[(\tilde{m})_x 2^x \log_{2-\delta} k\right]\left[(3+o(1))(\tilde{m}-x)!2^{\tilde{m}-x}\right]. \end{aligned} \tag{4.11}$$

The first expression in brackets corresponds to choosing $\tilde{m}$ paths for the $\tilde{\delta}$-abnormal $\tilde{r}$-suffix, as estimated above. The second expression bounds the number of ways to form a ROD factor out of the remaining $m' - \tilde{m}$ paths (the first $n/k - \tilde{r}$ cycles) via the induction hypothesis (since the set of $m' - \tilde{m}$ paths in the the $n/k - \tilde{r}$ prefix are $(r+\tilde{r})$-close to δ-normal and $\tilde{r} \geq 1$). The third expression treats the $(n/k - \tilde{r} + 1)$-st cycle, namely, ordering and directing its paths and selecting a root out of the first path (whose length is at most $\log_{2-\delta} k$). Finally, the last expression treats the $(\tilde{r}-1)$-suffix, in which all suffixes are $\tilde{\delta}$-normal by assumption, via an application of Theorem 4.4.

Since $\tilde{m} \leq \tilde{r}k$ and $x \leq k$, rearranging (4.11) gives, for each $\tilde{r} \geq 1$,

$$\begin{aligned} |\mathcal{R}_{\tilde{r}}| &\leq \tilde{r}k^2 (m')!2^{m'} e^{rk^\theta - \tilde{r}(k^{\theta_0-o(1)} - k^\theta)}(3+o(1))\log_{2-\delta} k \\ &= e^{-\tilde{r}k^{\theta_0-o(1)}}(m')!2^{m'}e^{rk^\theta} \end{aligned}$$

(using $\theta_0 > \theta$), and summing this over $\tilde{r} \geq 1$ now gives $o\big((m')!2^{m'}e^{rk^\theta}\big)$. This establishes Part (i).

It remains to prove Part (ii). Since S is δ-normal, similar to (4.10), the number of m'-subsets of S which are δ'-abnormal and contain exactly rk vertices is at most

$$\binom{m}{m'} \exp\left(-rk^{\theta_0-o(1)}\right) . \tag{4.12}$$

First, take $r \geq r_0$. By ordering the m' paths of the suffix as well as the remaining $m-m'$ paths, directing all m paths and choosing a root for each cycle at a multiplicative cost of $(\log_{2-\delta} k)^{n/k}$ we find that the total number ROD factors arising from S and having a δ'-abnormal r-suffix is at most

$$rk \cdot 2^m m! (\log_{2-\delta} k)^{n/k} \exp\left(-rk^{\theta_0-o(1)}\right) ,$$

where the prefactor rk bounds the number of choices for m'. Since $r \geq r_0$, we have $rk^{\theta_0-o(1)} \geq nk^{-1+\theta_0-o(1)}$, which outweighs the $\exp(nk^{-1+o(1)})$ factor from rooting the cycles (as well as the factor rk), so the above upper bound is at most

$$2^m m! \exp\left(-rk^{\theta_0-o(1)}\right) < 2^m m! \exp\left(-rk^{\theta}\right)$$

for large enough n, as required.

When $r < r_0$, for each of the choices for a δ'-abnormal r-suffix with m' paths (as estimated in (4.12)) we order and direct the paths of the suffix, then root its cycles at a multiplicative cost of $(m')!2^{m'}(\log_{2-\delta} k)^r$. As for the first $n/k-r$ cycles, the $m-m'$ paths used for these induce a path distribution with is r-close to δ-normal; thus, Part (i) bounds the number of ROD factors arising from these by $(m-m')!2^{m-m'}\exp(rk^{\theta})$. Overall we get the upper bound

$$rk \cdot 2^m m! \exp\left(r(k^{\theta} - k^{\theta_0-o(1)})\right) (\log_{2-\delta} k)^r \leq 2^m m! e^{-rk^{\theta}}$$

for large n (where the prefactor rk again accounts for the choice of m'). This establishes Part (ii) and completes the proof of the lemma. ∎

Finally, returning to the proof of Theorem 4.3, we sum the expression from Lemma 4.11(ii) over $r \geq 1$ and find that the contribution to $\mathscr{N}(S)$ of ROD factors with δ'-abnormal suffixes is $o(m!2^m)$. Consequently, Theorem 4.4 implies the statement of Lemma 4.10, which, as already noted, completes the proof of Theorem 4.3. ∎

5. Second moment of cycles factors via Theorem 3.1

Our main result in this section is the promised upper bound on $\mathbb{E}[\mathsf{CF}_k^2]$ (see §3), which is based on our estimate for the number of ROD factors arising from cycle-free intersection patterns.

Theorem 5.1. *If $k \geq K_0(n)$ with $K_0(n)$ as in (1.1), then the number of k-cycle factors in $G \sim \mathcal{P}(n,3)$ satisfies $\mathbb{E}[\mathsf{CF}_k^2] \leq (3+o(1))\mathbb{E}[\mathsf{CF}_k]^2$.*

Proof. Since $\mathsf{CF}_k/Y_k = (n/k)!(2k)^{n/k}$ (see (3.1)), we need to show that

$$\mathbb{E}[Y_k^2] \leq (3+o(1))\mathbb{E}[Y_k]^2 .$$

Recall also (see (3.8)) that estimating $\mathbb{E}[Y_k^2]$ amounts to estimating

$$\sum_h \sum_m \mathfrak{M}(n-2m) \sum_{S \in \mathcal{I}_{h,m}} \mathscr{N}(S)^2 \,. \tag{5.1}$$

We begin with what will turn out to be the dominant contributions to this sum, those with $h = 0$ and m close to $n/3$. (The analysis of this simpler case is similar to the corresponding analysis of Hamilton cycles in cubic graphs [27, Theorem 2.4] and in r-regular graphs [15], where $h = 0$ by definition.) Set

$$\mathcal{J} := \left\{ \lceil (\tfrac{1}{3} - \delta)n \rceil, \ldots, \lfloor (\tfrac{1}{3} + \delta)n \rfloor \right\} \qquad \text{where } \delta = \tfrac{1}{2}(\log n)^{-1/3} \,,$$

and recall that Theorem 3.1 says that if $|\frac{m}{n} - \frac{1}{3}| < (\log n)^{-1/3}$, then

$$\sum_{S \in \mathcal{I}_{0,m}} \mathscr{N}(S)^2 \le (9 + o(1)) \, (m!\, 2^m)^2 \, |\mathcal{I}_{0,m}| \tag{5.2}$$

(the slight difference between the bounds on $|\frac{m}{n} - \frac{1}{3}|$, here and in $\mathcal{J}$, will be helpful later).

Setting

$$\Psi_0(m) = \mathfrak{M}(n-2m) n!\, m!\, 2^m \binom{n-m-1}{m-1} \tag{5.3}$$

and recalling from (3.10) that $|\mathcal{I}_{0,m}| = \frac{n!}{m!\, 2^m} \binom{n-m-1}{m-1}$, we rewrite (5.2) as

$$\mathfrak{M}(n-2m) \sum_{S \in \mathcal{I}_{0,m}} \mathscr{N}(S)^2 \le (9 + o(1)) \Psi_0(m) \,.$$

Notice that for any m,

$$\frac{\Psi_0(m)}{\Psi_0(m+1)} = \frac{m(n-m-1)}{2(m+1)(n-2m)} \,.$$

If $x = 1/3 - m/n > 0$ then this is at most $1 - \frac{9x}{2(1+6x)}$, and so in this case for any $m = n/3 - \gamma n$ with $0 < \gamma < 1/3$,

$$\frac{\Psi_0(m)}{\Psi_0(\lfloor \frac{n}{3} \rfloor)} \le e^{-\frac{9}{2}\left(\int_0^\gamma \frac{x\,dx}{1+6x}\right)n} = e^{-\left[\frac{3}{4}\gamma - \frac{1}{8}\log(1+6\gamma)\right]n} \le e^{-\left(\frac{9}{4}\gamma^2 - 9\gamma^3\right)n} \,. \tag{5.4}$$

Similarly, if $x = 1/3 - m/n < 0$ then

$$\frac{\Psi_0(m+1)}{\Psi_0(m)} \le (2 + O(1/n)) \frac{n-2m}{n-m} = (1 + O(1/n)) \left(1 - \frac{9|x|}{2 - 3|x|}\right) \,,$$

whence, for $m = n/3 + \gamma n$ with $0 < \gamma < 1/6$ (recall $m \le n/2$),

$$\frac{\Psi_0(m)}{\Psi_0(\lfloor \frac{n}{3} \rfloor)} \lesssim e^{-9\left(\int_0^\gamma \frac{x\,dx}{2-3x}\right)n} = e^{\left[3\gamma + 2\log(1-\frac{3}{2}\gamma)\right]n} \le e^{-\frac{9}{4}\gamma^2 n} \tag{5.5}$$

(where the multiplicative constant implicit in the first inequality accounts for accumulating the $1 + O(1/n)$ error factor over $O(n)$ values of m). Finally, in both cases, when $x = o(1)$ the upper bounds are essentially tight (using $1 - x \ge \exp[-x/(1-x)]$, for instance); namely,

$$\frac{\Psi_0(m)}{\Psi_0(\lfloor \frac{n}{3} \rfloor)} = e^{-(\frac{9}{4} + o(1))\gamma^2 n} \qquad \text{for } \gamma = o(1) \,. \tag{5.6}$$

From this last estimate we have

$$\sum_{m\in\mathcal{J}} \Psi_0(m) \sim \left(\int_{-\infty}^{\infty} e^{-(\frac{9}{4}-o(1))x^2/n}dx\right)\Psi_0(\lfloor \tfrac{n}{3}\rfloor) \sim \frac{2}{3}\sqrt{\pi n}\Psi_0(\lfloor \tfrac{n}{3}\rfloor)\,. \tag{5.7}$$

We next want to show that the remaining terms in (3.8) (those with $m \notin \mathcal{J}$ or $h \neq 0$) are negligible in comparison with those in (5.7).

For $h \neq 0$, we may choose $S \in \mathcal{I}_{h,m}$ by first choosing the h cycles — this can be done in $\frac{(n)_{kh}}{h!(2k)^h}$ ways — and then choosing the set, say S', of paths of S from the remaining $n' = n - kh$ vertices. We then have

$$\mathscr{N}(S) = (n/k)_h(2k)^h\mathscr{N}(S')\,,$$

where the first two factors count the number of ways to order (in the ROD factor), root and direct the cycles. Therefore, with $\mathcal{I}^h_{0,m}$ the number of cycle-free, m-path intersection patterns on $n - kh$ vertices, we have

$$\sum_{S\in\mathcal{I}_{h,m}} \mathscr{N}(S)^2 = \frac{(n)_{kh}}{h!(2k)^h}\left((n/k)_h(2k)^h\right)^2 \sum_{S'\in\mathcal{I}^h_{0,m}} \mathscr{N}(S')^2\,. \tag{5.8}$$

The natural benchmark for $\mathfrak{M}(n-2m)\sum_{S'\in\mathcal{I}^h_{0,m}}\mathscr{N}(S')^2$, analogous to $\Psi_0(m)$ (see (5.3)), is

$$\mathfrak{M}(n-2m)(n-kh)!m!2^m\binom{n-kh-m-1}{m-1}\,.$$

Combining the last two displays, we find that the counterpart of $\Psi_0(m)$ in consideration of the contribution of $\mathcal{I}_{h,m}$ is

$$\Psi_h(m) := \mathfrak{M}(n-2m)n!\,m!\,2^m\binom{n-kh-m-1}{m-1}\binom{n/k}{h}^2(2k)^h h!\,, \tag{5.9}$$

though here we will sometimes need to retreat to

$$\hat{\Psi}_h(m) := \Psi_h(m)k^{2(n/k-h)}\,. \tag{5.10}$$

(It may be worth observing, though we do not make explicit use of this, that the preceding description of the number of choices for the cycles of $S \in \mathcal{I}_{h,m}$, combined with (3.10), gives $|\mathcal{I}_{h,m}| = \frac{1}{h!(2k)^h}\frac{n!}{m!\,2^m}\binom{n-kh-m-1}{m-1}$.)

As we will soon see, the main task remaining is to control $\sum_{h>0}\sum_m \Psi_h(m)$. This is achieved by the following lemma.

Lemma 5.2. *For $\Psi_h(m), \hat{\Psi}_h(m)$ as in (5.9) and $h^* = \lceil n/(k\sqrt{\log k})\rceil$,*

$$\sum_{h=1}^{h^*}\sum_{m\in\mathcal{J}} \Psi_h(m) = o\Big(\sum_{m\in\mathcal{J}} \Psi_0(m)\Big)\,, \tag{5.11}$$

$$\sum_{h=h^*}^{n/k}\sum_{m\in\mathcal{J}} \hat{\Psi}_h(m) = o\Big(\sum_{m\in\mathcal{J}} \Psi_0(m)\Big)\,, \tag{5.12}$$

$$\sum_{h=0}^{n/k}\sum_{m\notin\mathcal{J}} \hat{\Psi}_h(m) = o\Big(\sum_{m\in\mathcal{J}} \Psi_0(m)\Big)\,. \tag{5.13}$$

The reason for the division at h^* is that for $h < h^*$ we will be able to apply Theorem 3.1 (with $n - kh$ vertices) to say that the inner sum on the left of (5.11) bounds, within a constant factor, the corresponding inner sum over $m \in \mathcal{J}$ in (5.1). This is in contrast to the remaining terms (those in (5.12) and (5.13)), for which we must pay the extra factor $k^{2(n/k-h)}$ appearing in (5.10).

Before proving Lemma 5.2, we show that it gives our earlier assertion that the terms with $h = 0$ and $m \in \mathcal{J}$ dominate the sum in (5.1); that is,

$$\sum_h \sum_m \mathfrak{M}(n-2m) \sum_{S\in\mathcal{I}_{h,m}} \mathscr{N}(S)^2 \le (1+o(1)) \sum_{m\in\mathcal{J}} \Psi_0(m)\,. \tag{5.14}$$

For $m \in \mathcal{J}$ and $1 \le h \le h^*$, setting $n' = n - kh$, we have

$$\left|\frac{m}{n'} - \frac{1}{3}\right| \le \delta + O(kh/n) = \delta + O\left(1/\sqrt{\log n}\right) < (\log n')^{-1/3}$$

for large enough n. Theorem 3.1 thus says that $\sum_{S'\in\mathcal{I}^h_{0,m}} \mathscr{N}(S')^2$ is at most $(9+o(1))\,(m!2^m)^2$. Hence, using (5.8) (and slightly simplifying) we have

$$\sum_{h=1}^{h^*}\sum_{m\in\mathcal{J}} \mathfrak{M}(n-2m) \sum_{S\in\mathcal{I}_{h,m}} \mathscr{N}(S)^2 \lesssim \sum_{h=1}^{h^*}\sum_{m\in\mathcal{J}} \Psi_h(m)\,,$$

which is $o\left(\sum_{m\in\mathcal{J}} \Psi_0(m)\right)$ by (5.11). For the terms in (5.14) corresponding to $h > h^*$ or $m \notin \mathcal{J}$, we can use the trivial bound $\mathscr{N}(S') \le k^{n'/k} m! 2^m$, whence (5.12) and (5.13) imply a total contribution of $o\left(\sum_{m\in\mathcal{J}} \Psi_0(m)\right)$. So, Lemma 5.2 indeed allows us to estimate (5.1).

Proof of Lemma 5.2. For any m and h, we have

$$\begin{aligned}\frac{\Psi_h(m)}{\Psi_0(m)} &= \binom{n/k}{h}^2 (2k)^h h! \prod_{i=1}^{m-1}\left(1 - \frac{kh}{n-m-i}\right)\\ &\le \frac{(2n^2/k)^h}{h!} e^{-kh\log\frac{n-m-1}{n-2m}} \le \left(2e\frac{n^2}{kh} e^{-k\log\frac{n-m-1}{n-2m}}\right)^h .\end{aligned} \tag{5.15}$$

The two cases with $m \in \mathcal{J}$ ((5.11) and (5.12)) are easy. Here we observe that, even with our lower bound on k relaxed to $k > (2+\varepsilon)\log_2 n$ for some $\varepsilon > 0$, the right-hand side of (5.15) is at most

$$\left[2e\frac{n^2}{kh}(2-o(1))^{-k}\right]^h < n^{-(\varepsilon-o(1))h}\,.$$

This gives (5.11) (actually the finer $\sum_{h=1}^{h^*}\Psi_h(m) = o(\Psi_0(m))$ for $m \in \mathcal{J}$). Similarly, (5.12) follows once we observe that when $h > h^*$ and $m \in \mathcal{J}$ one has $kh\log\frac{n-m-1}{n-2m} \gtrsim n/\sqrt{\log n}$. Thus, the term $\exp(-kh\log\frac{n-m-1}{n-2m})$ in (5.15) easily overpowers the extra $k^{2(n/k-h)}$ in $\hat\Psi_h(m)$.

For the more interesting (5.13), we write $m = n/3 + \gamma n$ and proceed as follows. First suppose $m \ge \frac{2}{9}n$. When estimating $\Psi_h(m)/\Psi_0(m)$ via (5.15), the two opposing factors in that bound are $\exp\big[-kh\log\frac{2-3\gamma-O(1/n)}{1-6\gamma}\big]$ vs. $\exp\big[h(2\log n - \log k + O(1))\big]$; thus, if

$$k\log\frac{2-3\gamma-O(1/n)}{1-6\gamma} \ge (2+\varepsilon)\log n \tag{5.16}$$

for some fixed $\varepsilon > 0$ then $\Psi_h(m)/\Psi_0(m) \le \exp[-O(n^{-\varepsilon})]$. Since $k \ge K_0(n) = 2\log_{4/3}(2n/e)$, it follows that

$$\exp\left(2\frac{\log n}{k}\right) \le \exp\left(2\frac{\log n}{K_0(n)}\right) = \frac{4}{3} - o(1)\,,$$

and so (5.16) easily holds as long as $\frac{2-\gamma}{1-6\gamma} \ge \frac{4}{3} + \varepsilon'$ for some $\varepsilon' > 0$, and in particular it holds for any $\gamma \ge -\frac{1}{9}$ (with room to spare). Overall we have

$$\sum_{h=1}^{n/k}\sum_{\substack{m\ge\frac{2}{9}n\\ m\notin\mathcal{J}}}\hat\Psi_h(m) = o\Big(\sum_{\substack{m\ge\frac{2}{9}n\\ m\notin\mathcal{J}}}\hat\Psi_0(m)\Big) = o\Big(\sum_{m\in\mathcal{J}}\Psi_0(m)\Big)\,,$$

with the last equality using (5.4)–(5.5), in which the deviation of δn in m, compared to the interval $\mathcal{J}$, translates into a bound of $\exp(-cn\log^{-2/3} n)$ for an absolute $c > 0$ and overtakes the factor $k^{2n/k} = \exp[O(n\frac{\log k}{k})]$.

It remains to establish (5.13) when $m < \frac{2}{9}n$. For such values of m we get from (5.4) that, for some absolute constant $c_0 > 0$,

$$\Psi_0(m)/\Psi_0(\lfloor\tfrac{n}{3}\rfloor) \le e^{-c_0 n}\,. \tag{5.17}$$

Observe now that if $h < n/\log^2 n$ then $n^{2h} = \exp(O(n/\log n)) = \exp(o(n))$, whereas if $h \ge n/\log^2 n$,

$$\frac{(2n^2/k)^h}{h!} = \left[(2e+o(1))\frac{n^2}{hk}\right]^h \le e^{(1+o(1))h\log n}\,.$$

Immediately we see that if $k > (2/c_0)\log n$ (say) then the last expression is at most $\exp[(c_0/2 + o(1))n]$, outweighed by the factor $\exp(-c_0 n)$ from (5.17); Similarly, if $hk/n \le 5c_0$ then $\frac{h\log n}{n} \le \frac{hk\log n}{nK_0(n)} < \frac{5}{6}c_0$ (using $K_0(n) \ge 6\log n$), again outweighed by the aforementioned factor $\exp(-c_0 n)$. Setting

$$\alpha := m/n\,, \quad \theta = hk/n\,,$$

it therefore remains to show (5.13) when

$$\theta > c > 0\,, \quad \frac{2(1-\theta)}{k} \le \alpha \le \frac{1-\theta}{2} \wedge \frac{2}{9}\,, \quad k \le c' \log n \tag{5.18}$$

for some absolute constants $c, c' > 0$ (the upper bound on α used that every path has length at least 2, while the lower bound on α used that each of the $(1-\theta)n/k$ cycles unaccounted for by h must contain at least 2 paths).

We next treat the range $c < \theta \le 3/4$. With the above notation for m,

$$\begin{aligned}
\Psi_0(m) &= \mathfrak{M}(n-2m)m!2^m \binom{n-m-1}{m-1} \\
&\asymp \left(\frac{(1-2\alpha)n}{e}\right)^{(1-2\alpha)n/2} \sqrt{\alpha n}\left(\frac{\alpha n}{e}\right)^{\alpha n} 2^{\alpha n} O\Big(\frac{1}{\sqrt{\alpha n}}\Big) e^{(1-\alpha)H_e\left(\frac{\alpha}{1-\alpha}\right)n} \\
&\asymp (n/e)^{n/2}\, e^{\left[(1-\alpha)H_e\left(\frac{\alpha}{1-\alpha}\right)+\frac{1-2\alpha}{2}\log(1-2\alpha)+\alpha\log(2\alpha)\right]n}\,,
\end{aligned}$$

where $H_e(x) = -x\log x - (1-x)\log(1-x)$ is the natural entropy function, and using the fact $\binom{y}{\alpha y} \asymp (\alpha y)^{-1/2}\exp(H_e(\alpha)y)$, valid for any $\alpha \in (0,1)$ and y. Letting

$$g(\alpha) := (1-\alpha)H_e\left(\frac{\alpha}{1-\alpha}\right) - \frac{1}{2}H_e(2\alpha)\,, \tag{5.19}$$

we see that

$$\Psi_0(m) \asymp (n/e)^{n/2}\exp\left[g(\alpha)n\right]\,. \tag{5.20}$$

Combining this with (5.15) (using $hk\log(\frac{n-m-1}{n-2m}) = hk\log(\frac{n-m}{n-2m}) + O(1)$) gives

$$\begin{aligned}
\frac{\hat\Psi_h(m)}{\Psi_0(\lfloor\frac{n}{3}\rfloor)} &= \frac{\hat\Psi_h(m)}{\Psi_0(m)}\frac{\Psi_0(m)}{\Psi_0(\lfloor\frac{n}{3}\rfloor)} \\
&\lesssim k^{2n/k}\exp\left[(\theta n/k)(\log n + O(1)) - \theta n\log\frac{1-\alpha}{1-2\alpha}\right] \\
&\qquad \cdot \exp\Big[\left(g(\alpha) - g(1/3)\right)n\Big]\,.
\end{aligned} \tag{5.21}$$

If $\log\frac{1-\alpha}{1-2\alpha} \ge \frac{\log n}{k}$ then the first exponent in the final expression is at most $\exp(o(n))$ and the entire expression is less than $\exp(-cn)$ for some fixed $c > 0$ thanks to the term $\exp[(g(\alpha)-g(1/3))n]$ (as $\alpha \le \frac{2}{9}$ and $g'(\alpha) = \log\frac{2-4\alpha}{1-\alpha}$, so g is increasing for $\alpha \le \frac{1}{3}$). Assume therefore that

$$\log\frac{1-\alpha}{1-2\alpha} < \frac{\log n}{k}\,.$$

Then the right-hand side of (5.21) is increasing in θ and decreasing in k, so we can take $\theta = \frac{3}{4}$ (its maximum in our present regime) and $k = K_0(n)$ to get

$$\frac{\hat\Psi_h(m)}{\Psi_0(\lfloor\frac{n}{3}\rfloor)} \lesssim k^{2n/k} e^{\left[\frac{3}{4}\frac{\log n + O(1)}{K_0(n)} - \frac{3}{4}\log\frac{1-\alpha}{1-2\alpha} + g(\alpha) - g(\frac{1}{3})\right]n}\,. \tag{5.22}$$

Since $g(1/3) = \frac{1}{2}\log(4/3)$ and $K_0(n) = 2\log_{4/3}(2n/e)$, clearly

$$\frac{\log n}{K_0(n)} = g(1/3) + O(1/\log n)\,, \tag{5.23}$$

and we further claim that

$$\beta \log \frac{1-\alpha}{1-2\alpha} > g(\alpha) \qquad \text{for any } \alpha \in (0, \tfrac{1}{2}) \text{ and } \beta > \log 2\,. \tag{5.24}$$

Indeed, letting $f(\alpha)$ denote the left-hand side, we have $f(0) = g(0) = 0$, so (5.24) will follow from showing that $f'(\alpha) > g'(\alpha)$ for $\alpha \in (0, 1/2)$. Along this interval $f'(\alpha) = \beta/(1-3a+2a^2)$ is increasing, while $g'(\alpha) = \log(\frac{2-4\alpha}{1-\alpha})$ is decreasing, and using $f'(0) = \beta > \log 2 = g'(0)$ this implies (5.24).

Plugging (5.23) and (5.24) (with $\beta = 3/4$) in (5.22) yields

$$\frac{\hat\Psi_h(m)}{\Psi_0(\lfloor \frac{n}{3} \rfloor)} \lesssim k^{2n/k} e^{-(\frac{1}{4} - o(1))g(\frac{1}{3})n}\,,$$

which clearly suffices to show that

$$\sum_{h \le \frac{3}{4}n/k} \sum_{m \notin \mathcal{J}} \hat\Psi_h(m) = o\Big(\sum_{m \in \mathcal{J}} \Psi_0(m) \Big)\,.$$

We proceed to the case $\frac{3}{4} \le \theta \le 1 - k/n$. Here we will compare $\Psi_h(m)$ to $\Psi_{n/k}(0)$ (rather than $\Psi_0(\lfloor \frac{n}{3} \rfloor)$), and later show that $\Psi_{n/k}(0)$ is small. Observe that

$$\begin{aligned}
\frac{\hat\Psi_h(m)}{\hat\Psi_{n/k}(0)} &= \frac{\mathfrak{M}(n-2m)}{\mathfrak{M}(n)} m! 2^m \binom{n-kh-m-1}{m-1} \frac{\binom{n/k}{h}^2 h! (2k)^{2n/k-h}}{(n/k)!(2k)^{n/k}} \\
&\lesssim \left(\frac{n-2m}{e}\right)^{\frac{n-2m}{2}} \left(\frac{n}{e}\right)^{-\frac{n}{2}} 2^m (n-kh-m)^m \frac{\binom{n/k}{h}}{(n/k-h)!} (2k)^{n/k-h}\,,
\end{aligned}$$

using $m \le n - kh - m$. Writing $\binom{n/k}{h} \le (\frac{en}{k(n/k-h)})^{n/k-h}$, we find the last expression to be at most

$$\begin{aligned}
&O(1)\left(1 - \frac{2m}{n}\right)^{n/2} \left(\frac{2e(n-kh-m)}{n-2m}\right)^m \frac{\left[\frac{e}{1-\theta}\right]^{(1-\theta)\frac{n}{k}} \left[\frac{2ek^2}{(1-\theta)n}\right]^{(1-\theta)\frac{n}{k}}}{\sqrt{(1-\theta)\frac{n}{k}}} \\
&\le \left(\frac{2(1-\theta-\alpha)}{1-2\alpha}\right)^{\alpha n} \left[\frac{2e^2k^2}{(1-\theta)^2 n}\right]^{(1-\theta)\frac{n}{k}}\,,
\end{aligned}$$

using $\sqrt{(1-\theta)n/k} \ge 1$ since $\theta \le 1 - k/n$. For any $\theta > 1/2$ we know that $\frac{2(1-\theta-\alpha)}{1-2\alpha}$ is at most $2(1-\theta)$. Recalling from (5.18) that $\alpha n \ge 2(1-\theta)n/k$, we can infer that the right-hand side of the last inequality is at most

$$\left(\frac{8e^2k^2}{n}\right)^{(1-\theta)\frac{n}{k}}\,.$$

Since $\theta = ik/n$ for some integer $1 \le i \le \frac{1}{4}n/k$, in which case $m \le ik/2$, summing over the last expression over m and θ amounts to

$$\frac{1}{2}\sum_{i=1}^{\frac{1}{4}n/k} ik\left(\frac{8e^2k^2}{n}\right)^i = O(k^3/n)$$

(here we used the fact $k = O(\log n)$ from (5.18)).

To complete the proof, we analyze $h = n/k$. By (5.19)–(5.20), together with the fact that $g(1/3) = \frac{1}{2}\log(4/3)$,

$$\frac{\hat\Psi_{n/k}(0)}{\Psi_0(\lfloor\frac{n}{3}\rfloor)} \asymp \frac{(n/e)^{n/2}(n/k)!(2k)^{n/k}}{(n/e)^{n/2}\exp[g(\frac{1}{3})n]} \asymp \sqrt{n/k}\left(\frac{2n}{e}\right)^{\frac{n}{k}}\left(\frac{3}{4}\right)^{\frac{n}{2}} \le \sqrt{n/k}\,,$$

using the fact $(2n/e)^{n/k} \le (4/3)^{n/2}$ since $k \ge K_0(n) = 2\log_{4/3}(2n/e)$. Recalling from (5.7) that $\sum_{m\in\mathcal{J}}\Psi_0(m) \asymp \sqrt{n}\Psi_0(\lfloor\frac{n}{3}\rfloor)$ now gives

$$\sum_{h\ge\frac{3}{4}n/k}\sum_{m\notin\mathcal{J}}\hat\Psi_h(m) = O(1/\sqrt{k})\sum_{m\in\mathcal{J}}\Psi_0(m)\,.$$

This establishes (5.13) and completes the proof of the lemma. ∎

To complete the proof of the theorem, we revisit the definition of Ψ_0 in (5.9), and see that $\Psi_0(\lfloor\frac{n}{3}\rfloor)$ can readily be estimated as

$$\begin{aligned}\Psi_0(\tfrac{n}{3}) &\sim \sqrt{2\pi n}(n/e)^n(n/3)2^{n/6}\frac{\sqrt{\frac{4}{3}\pi n}\,(\frac{2}{3}n-1)^{\frac{2}{3}n-1}}{\sqrt{\frac{1}{3}\pi n}\,(n/6)^{n/6}}e^{1-n/2}\\ &\sim \sqrt{2\pi n}(n/e)^{3n/2}\left[(12)^{\frac{1}{6}}(\tfrac{2}{3})^{\frac{2}{3}}\right]^n = \sqrt{2\pi n}(n/e)^{3n/2}(4/3)^{n/2}\,,\end{aligned}$$

where we used the fact that $(\frac{2}{3}n-1)^{\frac{2}{3}n-1} \sim (\frac{2}{3}en)^{-1}(\frac{2}{3}n)^{\frac{2}{3}n}$. Combining Lemma 5.2 with (5.7) yields

$$\sum_h\sum_m \mathfrak{M}(n-2m)\sum_{S\in\mathcal{I}_{h,m}}\mathscr{N}(S)^2 \le \Big(6\pi\sqrt{2}+o(1)\Big)\,n\Big(\frac{n}{e}\Big)^{3n/2}\Big(\frac{4}{3}\Big)^{n/2}\,,$$

and altogether we have

$$\begin{aligned}\mathbb{E}[Y_k^2] &= \frac{6^n}{\mathfrak{M}(3n)}\sum_m \mathfrak{M}(n-2m)\sum_{S\in\mathcal{I}_{h,m}}\mathscr{N}(S)^2\\ &\le 6\pi\sqrt{2}\frac{6^n}{\sqrt{2}(\frac{3n}{e})^{\frac{3n}{2}}}n\Big(\frac{n}{e}\Big)^{3n/2}\Big(\frac{4}{3}\Big)^{n/2} = 6\pi n\frac{6^n}{3^{3n/2}}\Big(\frac{4}{3}\Big)^{n/2} = 6\pi n\Big(\frac{4}{3}\Big)^n.\end{aligned}$$

By (3.2) we have thus arrived at the promised

$$\mathbb{E}[Y_k^2] = (3+o(1))(\mathbb{E}Y_k)^2\,,$$

as required. ∎

6. Small subgraph conditioning and contiguity

In this section we derive Theorem 1 and Corollary 2 from Theorem 5.1, which was the upshot of the second moment analysis of the previous sections. For this we use what is called the small subgraph conditioning method in [33]. This calls for estimating the joint moments $\mathbb{E}\left[Y_k \prod_{i=1}^{j} [X_i]_{r_i}\right]$ where X_i is the number of i-cycles in a pairing in $\mathcal{P}_{n,3}$ (and $[X]_r = X!/(X-r)!$) for each fixed vector $(r_1, \dots, r_j)$. The computation here is almost the same as the corresponding original computation, the first time small subgraph conditioning was used, when the random variable of concern was the number of Hamilton cycles, and this is presented for the configuration model in [33, Proof of Theorem 4.5]. The argument is short so we include a complete version here.

We first show that for any fixed $i \geq 1$

$$\frac{\mathbb{E}[Y_k X_i]}{\mathbb{E}[Y_k]} \to \lambda_i(1+\delta_i) \tag{6.1}$$

where

$$\lambda_i = \frac{2^i}{2i} \quad \delta_i = \frac{(-1)^i - 1}{2^i}.$$

The fact that $\lambda_i = \mathbb{E}[X_i]$ was one of the first results Let D be some fixed set of pairs forming a ROD cycle factor in pairings in $\mathcal{P}_{n,3}$. By symmetry all copies of D are equivalent and so

$$\mathbb{E}[Y_k X_i]/\mathbb{E}[Y_k] = \mathbb{E}\left[X_i \mid D \subseteq \mathcal{P}_{n,3}\right].$$

If C is the set of pairs corresponding to an i-cycle (in which case we also call C itself an i-cycle), since $k \to \infty$ and i is fixed, we can assume $D \cap C$ forms a configuration of paths. We will classify C according to this configuration. Give these paths a consistent orientation along C and distinguish one path as first. This induces a linear ordering of paths around C, and considerting the overcounting we just introduced, it is now clear that

$$\frac{\mathbb{E}[Y_k X_i]}{\mathbb{E}[Y_k]} = \sum_{Q} \frac{1}{2|Q|} \mathbb{E}[X_i(Q) \mid D \subseteq \mathcal{P}_{n,3}] \tag{6.2}$$

where Q denotes the sequence of lengths of paths, $|Q|$ is the number of paths in Q and $X_i(Q)$ is the number of i-cycles in $\mathcal{P}_{n,3}$ consistent with such a configuration Q. Fix on such a Q with $|Q| = j$. There are asymptotically n^j ways to choose the starting points of the paths on D together with their directions along D. Almost all choices of such starting points are such that each two starting points are at distance greater than i along the cycles in D. Furthermore, once they are chosen, the pairs in C are determined if it creates an i-cycle yielding Q. The probability that these pairs all occur in $\mathcal{P}_{n,3}$ conditional upon $D \subseteq P$ is asymptotically n^{-j}. Hence $\mathbb{E}[X_i(Q) \mid D \subseteq P] \to 2^{|Q|}$. Now (6.2) becomes

$$\frac{\mathbb{E}[Y_k X_i]}{\mathbb{E}[Y_k]} \to \sum_{j \geq 1} \frac{2^j}{2j} \mid \{Q : \ |Q| = j\} \mid. \tag{6.3}$$

Note that every Q must have i vertices in total.

The ordinary generating function for the number of configurations Q with x marking the total number of vertices involved and y marking the number of paths is $\frac{g(x,y)}{1-g(x,y)}$ where $g(x,y)$ is the generating function for one path; that is, $yx^2/(1-x)$. Thus, with square brackets denoting extraction of coefficients,

$$\frac{\mathbb{E}[Y_k X_i]}{\mathbb{E}[Y_k]} \to \sum_{j\ge 1} \frac{2^j}{2j}[x^i y^j]\frac{yx^2}{1-x-yx^2}$$

and standard generating function manipulations (shown in detail in [33]) now give (6.1). This argument is easily generalised to give

$$\mathbb{E}\big[Y_k \prod_{i=1}^{j}[X_i]_{r_i}\big] \to \prod_{i=1}^{j} \lambda_i^{r_i}(1+\delta_i)^{r_i}\,.$$

By definition of λ_i and δ_i we have $\sum_{i\ge 1}\lambda_i\delta_i^2 = \log 3$. Since we showed in Theorem 5.1 that $\mathbb{E}Y_k^2 \le (3+o(1))(\mathbb{E}Y_k)^2$, all the requirements of small subgraph conditioning are satisfied (see [33, Theorems 4.1 and 4.3] or [18, Theorem 9.12 and Remark 9.16]) provided that $\mathbb{E}[Y_k]\to\infty$. To state the conclusion of this, we first note that $\delta_i = -1$ iff $i = 1$. Several things may now be concluded when $\mathbb{E}Y_k \to \infty$. One is that $\mathcal{P}_{n,3}$, conditioned on no loops, is contiguous to the superposition of a random ROD factor and a random perfect matching. Another is that in the same conditional space, the variable CF_k converges to a distribution of the type stated for W given in Theorem 1, but with the product starting at $j = 2$ rather than $j = 3$. The convergence in that statement and of the X_i to Poisson with means λ_i hold jointly. Hence, conditioning on $X_2 = 0$, i.e., the multigraph has no multiple edges either, we obtain the statement in the theorem (as $\mathbb{P}(X_2 = 0)$ is bounded away from 0).

7. The Comb Conjecture via cycle factors in regular graphs

In this section we derive the proof of Theorem 4 using our main result on cycle factors in random cubic graphs, Theorem 1. As a preliminary step we establish the next corollary on the threshold for a k-cycle factor in $\mathcal{G}(n,p)$.

Corollary 7.1. *Fix $\varepsilon > 0$. For any k, n such that $k \mid n$ and $k \ge K_0(n)$ as given in (1.1), the random graph $\mathcal{G}(n,p)$ with $p = (2+\varepsilon)\frac{\log n}{n}$ has a k-cycle factor w.h.p. In particular, for any $k \ge K_0(n)$ dividing n, the threshold for the existence of a k-cycle factor in $\mathcal{G}(n,p)$ is at $p \asymp \frac{\log n}{n}$.*

Proof. To simplify the exposition, we first give a proof establishing this fact for $p = (3+\varepsilon)\frac{\log n}{n}$. Further note that since the threshold for connectivity in $\mathcal{G}(n,p)$ is at the edge-probability $p = (1+o(1))\frac{\log n}{n}$, the above fact is already sufficient for establishing $p \asymp \frac{\log n}{n}$ as the threshold for a k-cycle factor in $\mathcal{G}(n,p)$.

Fix $\varepsilon > 0$ and consider the random graphs $G_i \sim \mathcal{G}(n,p')$ for $i = 1,2,3$, where $p' = (1+\varepsilon/3)\frac{\log n}{n}$. Clearly, by stochastic domination, for any given *simple* graph F the probability that $G' \sim \mathcal{G}(n,3p')$ contains F as a subgraph is at least the probability that F appears as a subgraph of the multigraph H comprised of the union of G_1, G_2, G_3.

Since the threshold for the appearance of a perfect matching in $\mathcal{G}(n,p)$ is at $p=(1+o(1))\frac{\log n}{n}$, w.h.p. we can extract an independent uniform perfect matching on M_i from each of the G_i's and denote the multigraph formed by the union of M_1,M_2,M_3 by $H_0\subset H$.

By the contiguity results of [17, 25], it is known that the multigraph H_0 is contiguous to a uniformly chosen 3-regular multigraph $\mathcal{P}_{n,3}$ conditioned to have no loops (see [18, Theorem 9.40], as well as [33] for further information). We can therefore invoke Theorem 1 (and the remark following that theorem concerning multigraphs) and gather that H_0 contains a k-cycle factor with high probability. Carrying this to $G'\sim\mathcal{G}(n,3p')$ concludes the proof.

To obtain the same result for $p=(2+\varepsilon)\frac{\log n}{n}$, rather than using three uniform independent perfect matchings we instead take a uniform perfect matching $\mathcal{M}$ and an independent uniform Hamilton cycle $\mathcal{H}$, each of which has a threshold of $(1+o(1))\frac{\log n}{n}$ in $\mathcal{G}(n,p)$. A delicate point one should note is the following: it is known that a random 3-regular graph is contiguous to $\mathcal{H}\oplus\mathcal{M}$, the union of $\mathcal{H}$ and $\mathcal{M}$ conditioned on having no self-loops or multiple edges. However, in our case we wish to address the multigraph formed by the union of $\mathcal{H}$ and $\mathcal{M}$ conditioned only to have no self-loops. The fact that this multigraph is contiguous to a random 3-regular multigraph $G\sim\mathcal{P}_{n,3}$ conditioned on having no self-loops (addressed by our cycle factor results) similarly holds (for example, see the second-to-last conclusion in the proof [33, Theorem 4.5]) and follows from the exact same arguments in the framework of [15], as well as from [17, Theorem 3]. ■

Proof of Theorem 4. As already mentioned, a prerequisite for containing the comb as a spanning subgraph is connectivity, whose threshold in $\mathcal{G}(n,p)$ is at $p=(1+o(1))\frac{\log n}{n}$; hence, establishing the threshold for the appearance of the comb will readily follow from showing that w.h.p. $\mathcal{G}(n,p)$ contains the comb at $p=(2+\varepsilon)\frac{\log n}{n}$.

First consider $G'\sim\mathcal{G}(n,p')$ for $p'=c/n$ with some large $c>0$. By the results of Ajtai, Komlós and Szemerédi [2], as well as de la Vega [30], in this regime w.h.p. the random graph contains a path of length $c'n$ for some absolute $c'>0$. In particular, there exists some path P of length $\sqrt{n}$ w.h.p.; let M denote the remaining $m=n-\sqrt{n}$ vertices. The path P will serve as the spine of the comb whereas the vertices of M will produce its $\sqrt{n}$ teeth.

Fix $\varepsilon>0$ and consider the random graph $G''\sim\mathcal{G}(n,(2+\varepsilon/2)\frac{\log n}{n})$. We first examine the induced subgraph on the vertices of M, which is a random graph $\mathcal{G}(m,p'')$ with $p''=(2+\varepsilon/2-o(1))\frac{\log m}{m}$. Applying Corollary 7.1 for $k=\sqrt{n}-1=(1-o(1))\sqrt{m}$ we deduce that w.h.p. this random graph contains a k-cycle factor. In other words, w.h.p. we can partition the vertices of M into $m/k=\sqrt{n}$ disjoint k-cycles, which we denote by $C_1,\dots,C_{\sqrt{n}}$.

Observe that in our analysis of G'' we have thus far only addressed the edges within the induced subgraph on M. Of the remaining edges, every edge between P and M appears in G'' independently with probability p''. Therefore, the random bipartite graph whose sides are the vertices of P vs. the cycles $C_1,\dots,C_{\sqrt{n}}$, with an edge connecting a vertex $u\in P$ and the cycle C_i iff they are connected in G'' as above, has edge probability $1-(1-p'')^k=(2-o(1))\frac{\log\sqrt{n}}{\sqrt{n}}$. By results of Erdős and Rényi [12] this edge probability is asymptotically twice than the threshold for a perfect bipartite matching in this bipartite random graph; thus, w.h.p. we can match each vertex of P to an exclusive $(\sqrt{n}-1)$-cycle.

Unraveling the cycles in the obvious manner now produces the comb (with P as its backbone). Altogether, the comb appears as a spanning subgraph of $\mathcal{G}(n,p)$ with $p = p' + p'' = (2 + \varepsilon/2 + o(1))\frac{\log n}{n}$, as required. ∎

Remark 7.2. The analogue of Theorem 4 holds for generalized combs $\mathsf{Comb}_{n,k}$ (in which the spine has n/k vertices, and the teeth are of length k) as long as $k \geq K_0(n)$. Indeed, via the exact same proof, we ultimately seek a perfect matching between n/k vertices (the spine) and k-cycles, which exists w.h.p. since the edge-probability in this bipartite graph is $(1 + o(1))pk \geq (2 + o(1))\frac{\log(n/k)}{n/k}$.

Acknowledgment. A major part of this work was carried out while the first and third authors were visiting the Theory Group of Microsoft Research, Redmond. They would like to thank the Theory Group for its hospitality and for creating a stimulating research environment.

References

[1] J. Komlós and E. Szemerédi, *Limit distribution for the existence of Hamiltonian cycles in a random graph*, Discrete Math. **43** (1983), no. 1, 55–63.
[2] M. Ajtai, J. Komlós, and E. Szemerédi, *The longest path in a random graph*, Combinatorica **1** (1981), no. 1, 1–12.
[3] N. Alon, M. Krivelevich, and B. Sudakov, *Embedding nearly-spanning bounded degree trees*, Combinatorica **27** (2007), no. 6, 629–644.
[4] J. Balogh, B. Csaba, M. Pei, and W. Samotij, *Large bounded degree trees in expanding graphs*, Electron. J. Combin. **17** (2010), no. 1, Research Paper 6, 9.
[5] P. H. Baxendale, *Renewal theory and computable convergence rates for geometrically ergodic Markov chains*, Ann. Appl. Probab. **15** (2005), no. 1B, 700–738.
[6] E. A. Bender and E. R. Canfield, *The asymptotic number of labeled graphs with given degree sequences*, J. Combinatorial Theory Ser. A **24** (1978), no. 3, 296–307.
[7] E. A. Bender and N. C. Wormald, *Random trees in random graphs*, Proc. Amer. Math. Soc. **103** (1988), no. 1, 314–320.
[8] B. Bollobás, *A probabilistic proof of an asymptotic formula for the number of labelled regular graphs*, European J. Combin. **1** (1980), no. 4, 311–316.
[9] B. Bollobás, *The evolution of sparse graphs*, Graph theory and combinatorics (Cambridge, 1983), Academic Press, London, 1984, pp. 35–57.
[10] B. Bollobás, *Random graphs*, 2nd ed., Cambridge Studies in Advanced Mathematics, vol. 73, Cambridge University Press, Cambridge, 2001.
[11] P. Erdös, W. Feller, and H. Pollard, *A property of power series with positive coefficients*, Bull. Amer. Math. Soc. **55** (1949), 201–204.
[12] P. Erdős and A. Rényi, *On the existence of a factor of degree one of a connected random graph*, Acta Math. Acad. Sci. Hungar. **17** (1966), 359–368.
[13] W. Feller, *An introduction to probability theory and its applications. Vol. II.*, Second edition, John Wiley & Sons Inc., New York, 1971. MR0270403 (42 #5292)
[14] P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge University Press, Cambridge, 2009.
[15] A. Frieze, M. Jerrum, M. Molloy, R. Robinson, and N. Wormald, *Generating and counting Hamilton cycles in random regular graphs*, J. Algorithms **21** (1996), no. 1, 176–198.
[16] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc. **58** (1963), 13–30.
[17] S. Janson, *Random regular graphs: asymptotic distributions and contiguity*, Combin. Probab. Comput. **4** (1995), no. 4, 369–405.

[18] S. Janson, T. Łuczak, and A. Rucinski, *Random graphs*, Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley-Interscience, New York, 2000.
[19] A. Johansson, J. Kahn, and V. Vu, *Factors in random graphs*, Random Structures Algorithms **33** (2008), no. 1, 1–28.
[20] J. Kahn, E. Lubetzky, and N. Wormald, *The threshold for combs in random graphs*, Preprint, available at arXiv:1401.2710.
[21] D. G. Kendall, *Unitary dilations of Markov transition operators, and the corresponding integral representations for transition-probability matrices*, Probability and statistics: The Harald Cramér volume (edited by Ulf Grenander), Almqvist & Wiksell, Stockholm, 1959, pp. 139–161.
[22] M. Krivelevich, *Embedding spanning trees in random graphs*, SIAM J. Discrete Math. **24** (2010), 1495–1500.
[23] T. Lindvall, *Lectures on the coupling method*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1992. A Wiley-Interscience Publication.
[24] T. Łuczak, *Cycles in a random graph near the critical point*, Random Structures Algorithms **2** (1991), no. 4, 421–439.
[25] M. S. O. Molloy, H. Robalewska, R. W. Robinson, and N. C. Wormald, 1-*factorizations of random regular graphs*, Random Structures Algorithms **10** (1997), no. 3, 305–321.
[26] L. Pósa, *Hamiltonian circuits in random graphs*, Discrete Math. **14** (1976), no. 4, 359–364.
[27] R. W. Robinson and N. C. Wormald, *Existence of long cycles in random cubic graphs*, Enumeration and Design (D. M. Jackson and S. A. Vanstone, eds.), Academic Press, Toronto, 1984, pp. 251–270.
[28] R. W. Robinson and N. C. Wormald, *Almost all cubic graphs are Hamiltonian*, Random Structures Algorithms **3** (1992), no. 2, 117–125.
[29] R. W. Robinson and N. C. Wormald, *Almost all regular graphs are Hamiltonian*, Random Structures Algorithms **5** (1994), no. 2, 363–374.
[30] W. Fernandez de la Vega, *Long paths in random graphs*, Studia Sci. Math. Hungar. **14** (1979), 335–340.
[31] H. S. Wilf, *generatingfunctionology*, 3rd ed., A K Peters Ltd., Wellesley, MA, 2006.
[32] N. C. Wormald, *The asymptotic distribution of short cycles in random regular graphs*, J. Combin. Theory Ser. B **31** (1981), no. 2, 168–182.
[33] N. C. Wormald, *Models of random regular graphs*, Surveys in combinatorics, 1999 (Canterbury), London Math. Soc. Lecture Note Ser., vol. 267, Cambridge Univ. Press, Cambridge, 1999, pp. 239–298.

Jeff Kahn
Department of Mathematics, Rutgers, Piscataway, NJ 08854, USA.
E-mail address: jkahn@math.rutgers.edu

Eyal Lubetzky
Courant Institute, New York University, 251 Mercer Street, New York, NY 10012, USA.
E-mail address: eyal@courant.nyu.edu

Nicholas Wormald
School of Mathematical Sciences, Monash University, Clayton, Victoria 3800, Australia.
E-mail address: nick.wormald@monash.edu