# ON ARRATIA'S COUPLING AND THE DIRICHLET LAW FOR THE FACTORS OF A RANDOM INTEGER

TONY HADDAD AND DIMITRIS KOUKOULOPOULOS

ABSTRACT. Let $x \geqslant 2$, let $N_x$ be an integer chosen uniformly at random from the set $\mathbb{Z} \cap [1, x]$, and let $(V_1, V_2, \dots)$ be a Poisson–Dirichlet process of parameter 1. We prove that there exists a coupling of these two random objects such that

$$\mathbb{E} \sum_{i \geqslant 1} \big| \log P_i - V_i \log x \big| \ll 1,$$

where the implied constant is absolute and $N_x = P_1 P_2 \cdots$ is the unique factorization of $N_x$ into primes or ones with the $P_i$'s being non-increasing. This establishes a conjecture of Arratia (2002), who proved that the right-hand side in the above estimate can be made $\ll \operatorname{loglog} x$. We also use this coupling to give a probabilistic proof of the Dirichlet law for the average distribution of the integer factorization into $k$ parts proved in 2023 by Leung and we improve on its error term.

## 1. INTRODUCTION

Let $N_x$ be an integer chosen uniformly at random from the set $\mathbb{Z} \cap [1, x]$. We may then factor it uniquely as $N_x = P_1 P_2 \cdots$ with the $P_i$'s forming a non-increasing sequence of primes or ones. In 1972, Billingsley [5] showed that, for any fixed positive integer $r$, the joint distribution of the random vector

$$\left( \frac{\log P_1}{\log x}, \dots, \frac{\log P_r}{\log x} \right)$$

converges in distribution as $x \to \infty$ to the first $r$ components of the Poisson–Dirichlet distribution (of parameter 1).

There are many ways to define the Poisson–Dirichlet distribution. One of the most intuitive ones involves a "stick-breaking" process that we will use throughout the paper. We start by sampling a sequence of i.i.d. random variables $(U_i)_{i \geqslant 1}$ that are all uniformly distributed in $[0, 1]$. We then define the sequence $(L_i)_{i \geqslant 1}$ in the following way:

$$L_1 := U_1 \quad \text{and} \quad L_j := U_j \prod_{i=1}^{j-1} (1 - U_i) \quad \text{for } j \geqslant 2.$$

The distribution of the process $\mathbf{L} = (L_1, L_2, \dots)$ is called the *GEM distribution* (of parameter 1). Lastly, we sort the components of $\mathbf{L}$ in non-increasing order to create $\mathbf{V} = (V_1, V_2, \dots)$. The distribution of this process is the *Poisson–Dirichlet distribution* (of parameter 1)[1]. We note that both $\sum_{i \geqslant 1} L_i$ and $\sum_{i \geqslant 1} V_i$ are equal to 1 almost surely.

In 2000, Tenenbaum [17] studied the rate of convergence in Billingsley's Theorem by providing an asymptotic series for the difference between the cumulative distribution functions of $\big(\frac{\log P_1}{\log x}, \dots, \frac{\log P_r}{\log x}\big)$ and of $(V_1, \dots, V_r)$.

*Date*: June 13, 2024.

[1]The GEM and Poisson–Dirichlet distributions have more general definitions involving typically a parameter $\theta$. In the rest of the paper, we will not be mentioning the parameter since we will always work with $\theta = 1$.

Another way to give a quantitative version of Billingsley's result is by constructing a *coupling* of $N_x$ and $\mathbf{V}$, i.e. a single probability space over which lives copies of $N_x$ and $\mathbf{V}$, such that the expectation

$$\mathbb{E}\sum_{i\geqslant 1}\big|\log P_i - V_i\log x\big| \tag{1.1}$$

is bounded by a positive monotone function that is $o(\log x)$ as $x\to\infty$. The random variables $N_x$ and $\mathbf{V}$ must be strongly correlated in this new probability space to achieve this. Indeed, if, for instance, $V_1$ and $P_1$ were independent, then we would have that $P_1\leqslant x^{1/3}$ and $V_1>2/3$ with positive probability, so that $\big|\log P_1 - V_1\log x\big|\geqslant \frac{\log x}{3}$ with positive probability. Hence, a coupling with $N_x$ and $\mathbf{V}$ being independent (also called a *trivial coupling*) makes (1.1) $\gg\log x$. In 2002, Richard Arratia [2] constructed a coupling satisfying

$$\mathbb{E}\sum_{i\geqslant 1}\big|\log P_i - V_i\log x\big| \ll \operatorname{loglog} x \tag{1.2}$$

for all $x\geqslant 3$. Moreover, he conjectured that there is a coupling for $N_x$ and $\mathbf{V}$ with the expectation above being $O(1)$. The main goal of this paper is to prove this conjecture:

**Theorem 1.** *There is a coupling of $N_x$ and $\mathbf{V}$ satisfying*

$$\mathbb{E}\sum_{i\geqslant 1}\big|\log P_i - V_i\log x\big| \ll 1$$

*for all $x\geqslant 1$.*

*Remarks.* (a) Theorem 1 is optimal. Indeed, since $\log P_i$ can never take any value in $(0,\log 2)$, we have

$$\mathbb{E}\sum_{i\geqslant 1}\big|\log P_i - V_i\log x\big| \geqslant \frac{\log 2}{3}\cdot\mathbb{E}\Big[\#\big\{i\geqslant 1 : V_i\in[a(x),2a(x)]\big\}\Big]$$

with $a(x):=\frac{\log 2}{3\log x}$ for any coupling of $N_x$ and $\mathbf{V}$. However, this last expectation is exactly equal to $\log 2$ no matter how we choose $x\geqslant 2$.

(b) The way we construct $N_x$ inside the coupling will be with a deterministic function of some random variables. These random variables stay unchanged as $x$ grows. The coupling actually generates a random process $(N_x)_{x\geqslant 1}$.

(c) Let $\sigma$ be a random permutation uniformly distributed in the permutation group $S_n$. It is well known that the factorization into primes of $N_x$ and the decomposition into disjoint cycles of $\sigma$ share similar statistics when $n\approx\log x$. In 2006, Arratia, Barbour and Tavaré [3] have proved that there exists a coupling between $\sigma$ and $\mathbf{V}$ such that

$$\mathbb{E}\sum_{i\geqslant 1}|C_i - nV_i| \sim \frac{\log n}{4}, \tag{1.3}$$

with $C_i$ being the number of cycles of length $i$ in $\sigma$. They showed that this (1.3) was optimal by using the inequality $|C_i - nV_i|\geqslant\|nV_i\|$ where $\|\cdot\|$ is the distance to the closest integer, and computing $\mathbb{E}\sum_{i\geqslant 1}\|nV_i\|$. This breaks the analogy between primes and permutations since Theorem 1 and (1.3) are not of the same order of magnitude when $n$ is replaced by $\log x$. The main reason why it is possible to get a better result in Theorem 1 is because the set $\{\log p : p \text{ primes}\}$ have much shorter gaps around $\log x$ than the gaps of $\mathbb{Z}$ around $n$.

1.1. **Application to distribution of factorizations of random integers.** We apply the coupling of Theorem 1 in the theory of divisors. Let $\Delta^{k-1}$ be the set of $k$-tuples $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k) \in \mathbb{R}^k_{\geqslant 0}$ satisfying $\alpha_1 + \cdots + \alpha_k = 1$. We also need to define a special class of functions:

**Definition 1.1** (The class of functions $\mathcal{F}_k(\boldsymbol{\alpha})$)**.** Given $k \in \mathbb{Z}_{\geqslant 2}$ and $\boldsymbol{\alpha} \in \Delta^{k-1}$, let $\mathcal{F}_k(\boldsymbol{\alpha})$ be the set of functions $f\colon \mathbb{N}^k \to \mathbb{R}_{\geqslant 0}$ satisfying the following three properties:

(a) For any fixed positive integer $n$, the function $f(\mathbf{d})$ is a probability mass function over all vectors $\mathbf{d} \in \mathbb{N}^k$ satisfying $d_1 \cdots d_k = n$, i.e. $\sum_{d_1 \cdots d_k = n} f(d_1, \dots, d_k) = 1$ for all $n \in \mathbb{N}$.

(b) Whenever $\mathbf{d}$ satisfies

$$d_i \coloneqq \begin{cases} p & \text{if } i = j \\ 1 & \text{if } i \neq j. \end{cases}$$

for some $1 \leqslant j \leqslant k$ and prime $p$, then $f(\mathbf{d}) = \alpha_j$.

(c) The function $f$ is *multiplicative*, i.e. for any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{N}^k$ such that $(a_1 \cdots a_k, b_1 \cdots b_k) = 1$, we have the property

$$f(a_1 b_1, \dots, a_k b_k) = f(a_1, \dots, a_k) \cdot f(b_1, \dots, b_k).$$

*Remark.* Let $\omega(n)$ denote the number of prime factors of $n$. If $n$ is square-free and $f \in \mathcal{F}_k$, then properties (b) and (c) of the definition above imply that

$$f(d_1, \dots, d_k) = \prod_{j=1}^{k} \alpha_j^{\omega(d_j)}. \tag{1.4}$$

We will use the class of functions $\mathcal{F}_k(\boldsymbol{\alpha})$ to define certain "random factorizations" of a random integer into $k$ parts. Specifically, let us fix $f \in \mathcal{F}(\boldsymbol{\alpha})$ and $x \geqslant 1$. We then define the *random $k$-factorization corresponding to $f$* to be a random variable $\mathbf{D}_x = (D_{x,1}, \dots, D_{x,k})$ to be a random variable taking values on $\mathbb{N}^k$ and satisfies the formula[2]

$$\mathbb{P}\Big[D_{x,i} = d_i \ \forall i \leqslant k \,\Big|\, N_x = n\Big] = f(d_1, \dots, d_k) \tag{1.5}$$

for all $k$-tuples $(d_1, \dots, d_k) \in \mathbb{N}^k$ with $d_1 \cdots d_k = n$.

Here are three examples of such random factorizations.

**Example 1** (Uniform sampling)**.** Let $f(d_1, \dots, d_k) = \tau_k(d_1 \cdots d_k)^{-1}$ with $\tau_k(n)$ being the number of $k$-factorizations of $n$. Then $f \in \mathcal{F}_k(\frac{1}{k}, \dots, \frac{1}{k})$. If $f$ is seen as a probability mass function as in (1.5), then we are sampling $D_{x,1} \cdots D_{x,k}$ uniformly among all $k$-factorizations of $N_x$.

**Example 2** (Recursive sampling)**.** Let $f(d_1, \dots, d_k) = \prod_{j=1}^{k-1} \tau(d_j \cdots d_k)^{-1}$ with $\tau(n)$ being the number of divisors of $n$. Then $f \in \mathcal{F}_k(\frac{1}{2}, \frac{1}{4}, \dots, \frac{1}{2^{k-1}}, \frac{1}{2^{k-1}})$. One way to realize this random $k$-factorization is by first sampling uniformly an divisor $D_{x,1}$ of $N_x$. Then, for all $j < k$, we recursively sample $D_{x,j}$ uniformly among the divisors of $\frac{N_x}{D_{x,1} \cdots D_{x,j-1}}$.

**Example 3** (Multinomial sampling)**.** For any fixed $\boldsymbol{\alpha} \in \Delta^{k-1}$, let $f(d_1, \dots, d_k) \coloneqq \prod_{i=1}^{k} \alpha_i^{\Omega(d_i)} \cdot \prod_{p \mid n} \binom{\nu_p(d_1 \cdots d_k)}{\nu_p(d_1), \dots, \nu_p(d_k)}$ with $\nu_p(d)$ being the $p$-valuation of $d$ and $\Omega(d)$ being the number of prime

[2] Strictly speaking, we only need property (a) of Definition 1.1 to define $\mathbf{D}_x$. But we will also need the other two properties when proving Theorem 2 below.

factors of $d$ counted with multiplicity. This function $f$ is in $\mathcal{F}_k(\boldsymbol{\alpha})$. This sampling can be understood as considering a sequence of i.i.d. random variables $(B_i)_{i\geqslant 1}$ satisfying $\mathbb{P}[B_i = j] = \alpha_j$ and constructing the $k$-factorization $D_{x,1}\cdots D_{x,k} = N_x$ as

$$D_{x,j} \coloneqq \prod_{i\geqslant 1:\ B_i=j} P_i,$$

where $P_1P_2\cdots$ is the prime factorization of $N_x$ as before.

When $k = 2$, we write for simplicity $D_x = D_{x,1}$. In Examples 1 and 2, the samplings are exactly the same when $k = 2$. In both cases, we are sampling uniformly a divisor $D_x$ of $N_x$. In 1979, Deshouillers, Dress and Tenenbaum [7] proved that

$$\mathbb{P}\big[D_x \leqslant N_x^u\big] = \frac{2}{\pi}\arcsin\sqrt{u} + O\Big((\log x)^{-\frac{1}{2}}\Big) \tag{1.6}$$

uniformly for $x \geqslant 2$ and $u \in [0,1]$. Their proof uses crucially the Landau–Selberg–Delange method to compute sums of multiplicative functions. Their result is optimal if we want an error term uniform in $u \in [0,1]$.

It turns out that for general values of $k$ and $\boldsymbol{\alpha}$, the distribution of $\mathbf{D}_x$ converges to the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$. Recall that if $\boldsymbol{\alpha} \in \Delta^{k-1}$ satisfies $\alpha_i > 0$ for all $i$, we say that a $\Delta^{k-1}$-valued random vector $\mathbf{Z}$ follows the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$ if

$$\mathbb{P}\big[Z_i \leqslant u_i\ \forall i < k\big] = F_{\boldsymbol{\alpha}}(\mathbf{u}) \coloneqq \prod_{i=1}^{k}\Gamma(\alpha_i)^{-1} \int\limits_{\substack{0\leqslant t_i\leqslant u_i\ \forall i<k\\ t_1+\cdots+t_{k-1}\leqslant 1}} \cdots \int \prod_{i=1}^{k} t_i^{\alpha_i-1}\,\mathrm{d}t_1\cdots\mathrm{d}t_{k-1}$$

with $t_k \coloneqq 1-(t_1+\cdots+t_{k-1})$ in the integrand, for any $\mathbf{u}\in[0,1]^{k-1}$.

In 2007, Bareikis and Manstavičius [4] proved that if $f \in \mathcal{F}_2(\alpha_1,\alpha_2)$ with $\alpha_1,\alpha_2 > 0$ and $\alpha_1+\alpha_2 = 1$, we have

$$\mathbb{P}[D_x \leqslant N_x^u] = F_{(\alpha_1,\alpha_2)}(u) + O_{\alpha_1,\alpha_2}\Big((1+u\log x)^{-\alpha_1}(1+(1-u)\log x)^{-\alpha_2}\Big).$$

As a matter of fact, their result covered a more general class of functions than $\mathcal{F}_2(\alpha_1,\alpha_2)$, with $f(p,1)$ being allowed to be *on average* $\alpha_1$, instead of being fixed.

In 2013, Nyandwi and Smati [15] considered the 3-factorization $D_{x,1}D_{x,2}D_{x,3} = N_x$ with its distribution $f$ as in Example 1. They proved

$$\mathbb{P}\big[D_{x,1} \leqslant N_x^{u_1} \text{ and } D_{x,2} \leqslant N_x^{u_2}\big] = F_{(\frac{1}{3},\frac{1}{3},\frac{1}{3})}(u_1,u_2) + O\Big((\log x)^{-\frac{1}{3}}\Big).$$

In 2016, de la Bretèche and Tenenbaum [6] considered the 3-factorization $D_{x,1}D_{x,2}D_{x,3} = N_x$ with its distribution $f$ as in Example 2, and they found that

$$\mathbb{P}\big[D_{x,1} \leqslant N_x^{u_1} \text{ and } D_{x,2} \leqslant N_x^{u_2}\big] = F_{(\frac{1}{2},\frac{1}{4},\frac{1}{4})}(u_1,u_2) + O\Big((\log x)^{-\frac{1}{4}}\Big).$$

More generally, for each fixed $k \geqslant 2$ and each $\boldsymbol{\alpha} \in \Delta^{k-1}$, Leung [14, Theorem 7.1] proved in 2023 that for any $f \in \mathcal{F}_k(\boldsymbol{\alpha})$, we have

$$\mathbb{P}\big[D_{x,i} \leqslant N_x^{u_i}\ \forall i < k\big] = F_{\boldsymbol{\alpha}}(\mathbf{u}) + O_{\boldsymbol{\alpha}}\Big((\log x)^{-\min\{\alpha_1,\ldots,\alpha_k\}}\Big) \tag{1.7}$$

uniformly for $x \geqslant 2$ and $\mathbf{u} \in [0,1]^{k-1}$ satisfying $u_1+\cdots+u_{k-1} \leqslant 1$. Similarly to the result of Bareikis and Manstavičius in [4], Leung's theorem holds for a more general class of functions than

$\mathcal{F}_k(\boldsymbol{\alpha})$, where the quantities $f(1,\dots,1,p,1,\dots,1)$, with $p$ being at the $j^{\text{th}}$ coordinate, are allowed to equal $\alpha_j$ *on average* instead of pointwise.

All results mentioned above use Fourier-analytic techniques, such as the Landau–Selberg–Delange method, as their main ingredients to get to their results.

We have a new approach to this problem: Since the size of the divisors on a logarithmic scale of any integer is entirely determined by the size of its prime factors, we might expect Leung's theorem to be a consequence of some quantitative form of Billingsley's Theorem. Indeed, using the coupling from Theorem 1, we deduce an improved form Leung's theorem for the class of functions $\mathcal{F}_k(\boldsymbol{\alpha})$.

**Theorem 2** (Dirichlet law for the factorization into $k$ parts)**.** *Let $k \geqslant 2$, let $\boldsymbol{\alpha} \in \Delta^{k-1}$ be fixed with $\alpha_i > 0$ for all $i$, and let $f \in \mathcal{F}_k(\boldsymbol{\alpha})$. In addition, let $x > 1$ and let $\mathbf{D}_x$ be the random $k$-factorization corresponding to $f$.*

*For any $\mathbf{u} \in [0,1]^{k-1}$ with at least one $i \in \{1,\dots,k-1\}$ with $u_i \neq 1$, we have*

$$\mathbb{P}\big[D_{x,i} \leqslant N_x^{u_i} \ \forall i < k\big] = F_{\boldsymbol{\alpha}}(\mathbf{u}) + O\bigg(\sum_{\substack{1\leqslant i<k\\ u_i\neq 1}} \frac{1}{(1+u_i\log x)^{1-\alpha_i}(1+(1-u_i)\log x)^{\alpha_i}}\bigg);$$

*the implied constant in the big-Oh is completely uniform in all parameters.*

*Remark.* When $u \in [0,1/2]$, we have $(1+u_i\log x)^{1-\alpha_i}(1+(1-u_i)\log x)^{\alpha_i} \geqslant (0.5\log x)^{\alpha_i}$. Similarly, when $u \in [1/2,1]$, we have $(1+u_i\log x)^{1-\alpha_i}(1+(1-u_i)\log x)^{\alpha_i} \geqslant (0.5\log x)^{1-\alpha_i}$. We thus find that the expression in the big-Oh in Theorem 2 is $\leqslant (0.5\log x)^{-\min\{\alpha_1,\dots,\alpha_k,1-\alpha_1,\dots,1-\alpha_k\}}$ uniformly in $\mathbf{u} \in [0,1]^{k-1}$. Since $1-\alpha_j \geqslant \alpha_i$ for all $i \neq j$ by our assumption that $\alpha_1+\cdots+\alpha_k = 1$, we conclude that the error term in Theorem 2 is $\ll_{\boldsymbol{\alpha}} (\log x)^{-\min\{\alpha_1,\dots,\alpha_k\}}$, thus recovering Leung's estimate (1.7) when $f$ lies in the class $\mathcal{F}_k(\boldsymbol{\alpha})$.

The proof of Theorem 2 is based on a 1987 result of Donnelly and Tavaré [8], who proved the following probabilistic version of the Leung's theorem: If $\mathbf{V} = (V_1, V_2, \dots)$ is a Poisson–Dirichlet process and $(C_i)_{i\geqslant 1}$ is a sequence of i.i.d. random variables supported on $\{1,\dots,k\}$ satisfying $\mathbb{P}[C_i = j] = \alpha_j$, then

$$\bigg(\sum_{i:\ C_i=1} V_i, \dots, \sum_{i:\ C_i=k} V_i\bigg) \tag{1.8}$$

follows exactly the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$. In 1998, Arratia [1] used this result to show that the left-hand side of (1.6) is $\frac{2}{\pi}\arcsin\sqrt{u} + o(1)$ as $x \to \infty$ with probabilistic methods. We use the coupling to bridge between the distribution of (1.8) and Theorem 2 and get an explicit error term.

*Remarks.* Here is a brief heuristic about the shape of the error term we obtain in Theorem 2. Let $\boldsymbol{\delta}_x := (\frac{\log D_{x,1}}{\log N_x}, \dots, \frac{\log D_{x,k}}{\log N_x})$. There exists a coupling between $\boldsymbol{\delta}_x$ and the random vector (1.8) such that their distance is of typical size $\asymp \frac{1}{\log x}$. For each $j$, the marginal distribution of the $j^{\text{th}}$ component of $\mathrm{Dir}(\boldsymbol{\alpha})$ is $\mathrm{Beta}(\alpha_j, 1-\alpha_j)$, which is why we get the error term of Theorem 2 when none of the $u_i$'s are close to either 0 or 1.

When one of the $u_i$'s is close to 0 or 1, the behavior changes completely in the error term of Theorem 2, because the vector $\boldsymbol{\delta}_x$ has discrete distribution (compared to the continuous distribution $\mathrm{Dir}(\boldsymbol{\alpha})$). For instance, if $u_i < \log 2/\log x$ for some $i$, then the relation $D_{x,i} \leqslant N_x^{u_i}$ implies that $D_{x,i} = 1$.

1.2. **Structure of the paper.** We have organized the paper as follows in two main parts.

Part I contains the proof of Theorem 1. It is divided as follows:

- In Section 2, we present the coupling implicit in Theorem 1 and present a proof of the latter as a corollary of four key results (Lemmas 2.1-2.3 and Proposition 2.4). Lemma 2.2 is simple and proven right away.
- In Section 3, we prove Lemma 2.1.
- In Section 4, we explain another way to realize a GEM process that was presented by Arratia in [2], and we use it to prove Lemma 2.3. This alternative way of describing a GEM process will be also key in proving Proposition 2.4.
- Sections 5, 6 and 7 are reserved to prove Proposition 2.4.

Finally, Part II contains the proof of Theorem 2 and it is organized in the following way:

- In Section 8, we present an argument due to Donnelly–Tavaré showing that random $k$-partitions of the Poisson–Dirichlet distribution are distributed according to Dirichlet's law.
- In Section 9, we use the coupling of Section 2 to construct a coupling of the random $k$-factorization $\mathbf{D}_x$ and an analogous $k$-partition of the Poisson–Dirichlet distribution. We then use this coupling to reduce Theorem 2 to estimating certain boundary events, one involving number-theoretic objects and the other one pure probabilistic objects.
- In Section 10, we prove the necessary estimate for the number-theoretic boundary event, and in Section 11 we show the analogous probabilistic estimate.

*Remark.* The readers interested only in Theorem 2 need to only understand Section 2 from Part I.

1.3. **Notation.** We let $\log_j$ denote the $j$-iteration of the natural logarithm, meaning that $\log_1 = \log$ and $\log_j = \log \circ \log_{j-1}$ for $j \geqslant 2$.

Throughout the paper, the letter $p$ is reserved for prime numbers and the letter $n$ is reserved for natural numbers, unless stated otherwise. Given such $p$ and $n$, we write $\nu_p(n)$ for the $p$-adic valuation of $n$, that is to say the largest integer $v \geqslant 0$ such that $p^v | n$. In addition, we write $\omega(n)$ for the number of distinct prime factors of $n$.

Moreover, $n \in \mathbb{N}$, we let $s(n)$ denote its large square-full divisor. Also, we let $n^\flat = n/s(n)$ and note that $n^\flat$ is square-free and co-prime to $s(n)$.

We write $\pi(x)$ for the number of primes $\leqslant x$. We shall also use heavily Chebyshev's function $\theta(x) = \sum_{p \leqslant x} \log p$.

To describe various estimates, we use Vinogradov's notation $f(x) \ll g(x)$ or Landau's notation $f(x) = O(g(x))$ to mean that $|f(x)| \leqslant C \cdot g(x)$ for a positive constant $C$. If $C$ depends on a parameter $\alpha$, we write $f(x) \ll_\alpha g(x)$ or $f(x) = O_\alpha(g(x))$. If two positive functions $f, g$ have the same order of magnitude in the sense that $f(x) \ll g(x) \ll f(x)$, then we write $f(x) \asymp g(x)$.

If $P$ is some proposition, then the indicator function $\mathbb{1}_P$ will be equal to 1 if $P$ is true and 0 if $P$ is false. For a set or an event $A$, we will sometimes write $\mathbb{1}_A(\omega)$ to mean $\mathbb{1}_{\omega \in A}$.

## ACKNOWLEDGEMENTS

The authors would like to thank Matilde Lalín and Gérald Tenenbaum for their helpful comments on the paper.

TH is supported by the Courtois Chair II in fundamental research.

DK is supported by the Courtois Chair II in fundamental research, by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-05699 and RGPIN-2024-05850) and by the Fonds de recherche du Québec - Nature et technologies (2022-PR-300951 and 2025-PR-345672).

# PART I. SHARPENING ARRATIA'S COUPLING

## 2. THE COUPLING

In this section, we describe the coupling behind Theorem 1. To construct it, we begin with an ambient probability space $\Omega$ containing the following objects:

- a GEM process $\mathbf{L} = (L_1, L_2, \ldots)$;
- three mutually independent random variables $U_1'$, $U_2'$ and $U_3'$ that are also independent from $\mathbf{L}$, and which are uniformly distributed in the open interval $(0, 1)$.

We shall extract an integer $N_x$ and a Poisson–Dirichlet process $\mathbf{V}$ as deterministic functions of these random objects. Extracting $\mathbf{V}$ is done by sorting the components of $\mathbf{L}$ in non-increasing order. The extraction of $N_x$ is more complicated, and we need to introduce additional notation to describe it.

Let $(\lambda_j)_{j\geqslant 0}$ be the increasing sequence of positive real numbers defined by $\lambda_0 := e^{-\gamma}$ and

$$\lambda_j := \exp\biggl(-\gamma + \sum_{i\leqslant j} \frac{1}{v_i q_i}\biggr) \quad \text{for } j \geqslant 1$$

with $\gamma$ being the Euler-Mascheroni constant and $q_j = p_j^{v_j}$ being the $j^{\text{th}}$ smallest prime power, i.e., $(q_j)_j$ is the sequence $2, 3, 2^2, 5, 7, 2^3, 3^2, \ldots$ Note that

$$\lambda_j = \log q_j + O(1/(\log q_j)^2). \tag{2.1}$$

Indeed, Mertens's theorems and the Prime Number Theorem [12, Theorems 3.4 and 8.1] yield

$$\sum_{p^k\leqslant y} \frac{1}{kp^k} = \log\log y + \gamma + O(1/(\log y)^3) \qquad (y \geqslant 2).$$

A detailed proof of this estimate is given in Proposition A.2. Taking $y = q_j$ in it proves (2.1).

Moreover, we have that

$$q_{j+1} = q_j + O(q_j/(\log q_j)^3). \tag{2.2}$$

Indeed, the Prime Number Theorem (Proposition A.1) implies that $\psi(q_j + Cq_j/(\log q_j)^3) - \psi(q_j) > 0$ if $C$ is large enough, whence $q_{j+1} \leqslant q_j + Cq_j/(\log q_i)^3$, as needed.

Next, we define the step-function $h : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ by

$$h(t) := \sum_{j\geqslant 1} (\log q_j) \cdot \mathbb{1}_{\lambda_{j-1} < t \leqslant \lambda_j} \tag{2.3}$$

In particular, (2.1) and (2.2) imply that, if $r(t) := |h(t) - t|$, then

$$r(t) \ll \min\{t, t^{-2}\} \quad \text{for all } t > 0. \tag{2.4}$$

Using the above notation, here is how to extract $N_x$ from $\mathbf{L}$, $U_1'$, $U_2'$ and $U_3'$:

(1) Construct the sequence of random prime powers or ones $(Q_i)_{i\geqslant 1}$ by letting $Q_i := e^{h(L_i \log x)}$. Note that only finitely many $Q_i$ will be prime powers.

(2) Define the random integer $J_x := \prod_{j\geqslant 2} Q_j$.

(3) Define the *extra prime* $P_{\texttt{extra}}$ as the smallest element of the set $\{1\} \cup \{\text{primes}\}$ such that $\theta(P_{\texttt{extra}}) \geqslant U_1'\theta(x/J_x)$, where $\theta(y) = \sum_{p \leqslant y} \log p$ is Chebyshev's function. (In particular, we have $P_{\texttt{extra}} = 1$ when $J_x > x/2$; otherwise, $P_{\texttt{extra}}$ is a prime $\leqslant x/J_x$.)

(4) Let $\mu_x$ be the probability measure induced by the random variable $M_x \coloneqq J_x P_{\texttt{extra}}$, and let $\nu_x$ be the uniform counting measure on $\mathbb{Z} \cap [1, x]$. Then, by Lemma B.2 and our assumption that $U_2'$ and $U_3'$ exist on $\Omega$, there exists a random variable $N_x$ on $\Omega$ such that:
- $N_x$ is uniformly distributed on $\mathbb{Z} \cap [1, x]$,
- $\mathbb{P}[M_x \neq N_x] = d_{\mathrm{TV}}(\mu_x, \nu_x)$

with $d_{\mathrm{TV}}$ being the total variation distance defined in (B.1).

This completes the definition of our coupling, since the space $\Omega$ contains a Poisson–Dirichlet process $\mathbf{V}$ and also a random variable $N_x$ with distribution $\nu_x$.

In Section 2.1, we show how to use the coupling to prove Theorem 1. Lastly, in Section 2.2, we make some technical remarks on the coupling.

2.1. **Reducing Theorem 1 to three lemmas and a proposition.** With the following four key results, we directly get Theorem 1. Recall that $s(n)$ denotes the largest square-full divisor of the integer $n$. In addition, let

$$\Theta_x \coloneqq \sum_{i \geqslant 1} r(V_i \log x).$$

**Lemma 2.1** (The $\ell^1$ distance within the coupling)**.** *When $M_x = N_x$, we have the inequality*

$$\sum_{i \geqslant 1} |\log P_i - V_i \log x| \leqslant \log(x/N_x) + 2 \cdot \log s(N_x) + 2 \cdot \Theta_x.$$

**Lemma 2.2** (Properties of $N_x$)**.** *Fix $\alpha \in [0, 1)$ and $\beta \in [0, 1/2)$. Uniformly over $x \geqslant 1$, we have*

$$\mathbb{E}\big[(x/N_x)^\alpha s(N_x)^\beta\big] \ll_{\alpha,\beta} 1.$$

*Proof.* We must show that

$$S \coloneqq \sum_{n \leqslant x} (x/n)^\alpha s(n)^\beta \ll_{\alpha,\beta} x.$$

Indeed, if we let $a = n^\flat$ and $b = s(n)$, then

$$S \leqslant \sum_{\substack{b \leqslant x \\ b \text{ square-full}}} x^\alpha b^{\beta-\alpha} \sum_{a \leqslant x/b} a^{-\alpha} \ll_\alpha \sum_{\substack{b \leqslant x \\ b \text{ square-full}}} x^\alpha b^{\beta-\alpha} (x/b)^{1-\alpha} \leqslant x \sum_{b \text{ square-full}} b^{\beta-1},$$

where we used our assumption that $\alpha < 1$. Since we have assumed that $\beta < 1/2$, the sum over $b$ converges, thus completing the proof. □

**Lemma 2.3** (Properties of $\Theta_x$)**.** *Fix $\alpha \geqslant 0$. Uniformly over $x \geqslant 1$, we have*

$$\mathbb{E}\big[e^{\alpha\Theta_x}\big] \ll_\alpha 1.$$

**Proposition 2.4** (Total variation distance between $M_x$ and $N_x$)**.** *For $x \geqslant 2$, we have*

$$\mathbb{P}\big[M_x \neq N_x\big] \ll \frac{1}{\log x}.$$

The proof of Lemma 2.1 is given in Section 3, and the proof of Lemma 2.3 is given in Section 4. The proof of Proposition 2.4 is the longest part. We set it up in Sections 5 and 6 to eventually give it in Section 7. Here is how we get Theorem 1 with these results:

*Proof of Theorem 1.* Let $S := \sum_{i\geqslant 1} |\log P_i - V_i \log x|$. We always have the trivial bound $S \leqslant 2\log x$. This bound and Lemma 2.1 gives us

$$S \leqslant \mathbb{1}_{M_x\neq N_x}\cdot(2\log x) + \mathbb{1}_{M_x=N_x}\cdot\big(\log(x/N_x) + 2\cdot\log s(N_x) + 2\cdot\Theta_x\big).$$

Taking expectations on both sides, we get $\mathbb{E}[S] \ll 1$ with Lemmas 2.2-2.3 and with Proposition 2.4. □

In fact, if we condition on the event $M_x = N_x$, we can obtain a much stronger bound:

**Proposition 2.5.** *Fix $\alpha \in [0, 1/4)$. For $x \geqslant 2$, we have*

$$\mathbb{E}\left[\exp\Big(\alpha\sum_{i\geqslant 1}|\log P_i - V_i\log x|\Big)\,\Big|\, M_x = N_x\right] \ll_\alpha 1.$$

*Proof.* This follows readily by Hölder's inequality and by Lemmas 2.1-2.3. □

2.2. **Remarks on the coupling.** (a) As discussed previously, we have $\lambda_{j-1} \approx \lambda_j \approx \log q_j$, and thus $(\log x)L_i \approx \log Q_i$ as long as $L_i$ is not too small. In particular, we expect that $\sum_{i\geqslant 1}\log Q_i$ would be too close to $\log x$, and thus $\prod_{i\geqslant 1} Q_i$ cannot serve as a proxy of $N_x$. This is the reason we have to delete $Q_1$ from the factors of $J_x$, and we insert instead an extra random prime $P_{\texttt{extra}}$ conveniently chosen so that $J_x P_{\texttt{extra}}$ has a distribution close to $\nu_x$.

As we already remarked, we have $P_{\texttt{extra}} = 1$ if, and only if, $J_x > x/2$ (which happens rarely); otherwise, $P_{\texttt{extra}}$ is a prime $\leqslant x/J_x$. As a matter of fact, for all $j \in \mathbb{Z}\cap[1, x/2]$, we have

$$\mathbb{P}\big[P_{\texttt{extra}} = p \,|\, J_x = j\big] = \frac{\mathbb{1}_{p\leqslant x/j}\cdot\log p}{\theta(x/j)}. \tag{2.5}$$

This is the crucial property that will allow us to show that $M_x = J_x P_{\texttt{extra}}$ is close to being uniformly distributed.

(b) The coupling we defined above is a modification of Arratia's coupling in [2]. Some of the differences in our definition are purely aesthetic. The one major difference is within step (3), which is the whole reason why we obtain a stronger bound than (1.2). The construction of Arratia's extra prime $P_{\texttt{Arratia}}$ had a different distribution which satisfied

$$\mathbb{P}\big[P_{\texttt{Arratia}} = p \,|\, J_x = j\big] = \frac{1}{1+\pi(x/j)} \tag{2.6}$$

for all $j \leqslant x$ and all $p \in \{1\}\cup\{\text{primes } \leqslant x/j\}$. It is possible to get the inequality in Lemma 2.1 with Arratia's original coupling. However, it would be impossible to get a version of Proposition 2.4 with a bound better than $\frac{\log_2 x}{\log x}$.

## 3. The $\ell^1$ distance within the coupling

In this section we establish Lemma 2.1. We need the following rearrangement inequality.

**Lemma 3.1** (Rearrangement inequality)**.** *For two non-increasing sequences $(x_i)_{i\geqslant 1}$ and $(y_i)_{i\geqslant 1}$ of real numbers, and for any two permutations $\sigma, \rho\colon \mathbb{N}\to\mathbb{N}$, we have*

$$\sum_{i\geqslant 1}|x_i - y_i| \leqslant \sum_{i\geqslant 1}|x_{\sigma(i)} - y_{\rho(i)}|.$$

*Proof.* See [3, Lemma 3.2]. □

*Proof of Lemma 2.1.* Recall that the definitions of the sequence $(Q_i)_{i\geqslant 1}$, the extra prime $P_{\mathtt{extra}}$ and $M_x$. We create another sequence of primes or ones $(\widetilde{P}_i)_{i\geqslant 1}$ in the following way:

- We set $\widetilde{P}_1 := P_{\mathtt{extra}}$.
- If $i \geqslant 2$ with $Q_i = 1$, then we set $\widetilde{P}_i := 1$.
- If $i \geqslant 2$ with $Q_i > 1$, we set $\widetilde{P}_i$ to be the only prime dividing $Q_i$.

We let $(\widehat{P}_i)_{i\geqslant 1}$ be the sequence $(\widetilde{P}_i)_{i\geqslant 1}$ in non-increasing order, and we set

$$\widehat{M}_x := \prod_{i\geqslant 1} \widetilde{P}_i = \prod_{i\geqslant 1} \widehat{P}_i.$$

Since $M_x = N_x$, we have $P_i \geqslant \widehat{P}_i$ for all $i$, because $(\widehat{P}_i)_{i\geqslant 1}$ is a subsequence of the non-increasing sequence $(P_i)_{i\geqslant 1}$. Therefore,

$$\sum_{i\geqslant 1} |\log P_i - V_i \log x| \leqslant \sum_{i\geqslant 1} |\log \widehat{P}_i - V_i \log x| + \sum_{i\geqslant 1} \log(P_i/\widehat{P}_i)$$

We have that $\prod_{i\geqslant 1} P_i/\widehat{P}_i = M_x/\widehat{M}_x$. Furthermore, the integer $M_x/\widehat{M}_x$ only contains prime factors whose square divides $N_x$. Thus, $M_x/\widehat{M}_x$ divides $s(N_x)$, and

$$\sum_{i\geqslant 1} \log(P_i/\widehat{P}_i) = \log(M_x/\widehat{M}_x) \leqslant \log s(N_x), \tag{3.1}$$

and thus

$$\sum_{i\geqslant 1} |\log P_i - V_i \log x| \leqslant \sum_{i\geqslant 1} |\log \widehat{P}_i - V_i \log x| + \log s(N_x). \tag{3.2}$$

Next, we use the rearrangement inequality (Lemma 3.1) to find that

$$\begin{aligned}\sum_{i\geqslant 1} |\log \widehat{P}_i - V_i \log x| &\leqslant |\log P_{\mathtt{extra}} - L_1 \log x| + \sum_{i\geqslant 2} |\log \widetilde{P}_i - L_i \log x| \\ &\leqslant |\log P_{\mathtt{extra}} - L_1 \log x| + \sum_{i\geqslant 2} |\log Q_i - L_i \log x| + \sum_{i\geqslant 2} \log(Q_i/\widetilde{P}_i),\end{aligned}$$

where we used that $\widetilde{P}_i \leqslant Q_i$ for each $i$. Moreover, we have $\prod_{i\geqslant 2} Q_i/\widetilde{P}_i = M_x/\widehat{M}_x$, so (3.2) implies that

$$\sum_{i\geqslant 1} |\log \widehat{P}_i - V_i \log x| \leqslant |\log P_{\mathtt{extra}} - L_1 \log x| + \sum_{i\geqslant 2} |\log Q_i - L_i \log x| + \log s(N_x).$$

Finally, we note as well as that $\log P_{\mathtt{extra}} = \log M_x - \sum_{i\geqslant 2} \log Q_i$, $L_1 = 1 - \sum_{i\geqslant 2} L_i$. Since we have also assumed that $M_x = N_x$, we have

$$|\log P_{\mathtt{extra}} - L_1 \log x| \leqslant \log(x/N_x) + \sum_{i\geqslant 2} |\log Q_i - L_i \log x|.$$

Combining the two above displayed inequalities with (3.2), we conclude that

$$\sum_{i\geqslant 1} |\log P_i - V_i \log x| \leqslant \log(x/N_x) + 2\sum_{i\geqslant 2} |\log Q_i - L_i \log x| + 2\log s(N_x).$$

To complete the proof, recall that $\log Q_i = h(L_i \log x)$ and $r(t) = |h(t) - t|$, whence

$$\sum_{i \geqslant 2} |\log Q_i - L_i \log x| \leqslant \sum_{i \geqslant 2} r(L_i \log x) \leqslant \Theta_x.$$

This proves Lemma 2.1. □

## 4. Another realization of the GEM distribution

For the proof of Lemma 2.3 and Proposition 2.4, we will need to be more precise as to how the GEM process $\mathbf{L}$ is sampled in the coupling. The construction we present below is also the one used by Arratia [2].

**Definition 4.1** (The Poisson Process $\mathscr{R}$)**.**

(a) We denote by $\mathscr{R}$ the Poisson process on $\mathbb{R}^2_{>0}$ that has intensity measure $e^{-wy}\,\mathrm{d}w\,\mathrm{d}y$. Without loss of generality, we may assume that the $w$-coordinates of the points of $\mathscr{R}$ are always distinct.

(b) We index the points of $\mathscr{R} = \{(W_i, Y_i) : i \in \mathbb{Z}\}$ according to the following rules:
- $W_i < W_{i+1}$ for all $i \in \mathbb{Z}$;
- if we let $S_i := \sum_{\ell \geqslant i} Y_\ell$ for all $i \in \mathbb{Z}$, then we have $S_1 \leqslant \log x < S_0$.

*Remark.* By the Mapping Theorem (Proposition B.3), projecting $\mathscr{R}$ on the $w$-axis yields a Poisson Process with intensity $\frac{\mathrm{d}w}{w}$ and, similarly, projecting $\mathscr{R}$ on its $y$-axis yields a Poisson Process with intensity $\frac{\mathrm{d}y}{y}$. Therefore, the $w$-coordinates of the points in $\mathscr{R}$ have almost surely exactly one limit point at $0$ and they are almost surely unbounded. Hence, the indexing $(W_i, Y_i)$ in part (b) of the above definition is well-defined.

The following lemma describes the distribution of the point process $S_i$.

**Lemma 4.2** (Scale-invariant spacing lemma)**.** *The point process $\{S_i : i \in \mathbb{Z}\}$ is a Poisson process on the positive real line with intensity measure $\frac{\mathrm{d}s}{s}$.*

*Proof.* See [3, Lemma 7.1]. □

Using this lemma, we have the following description of the GEM distribution.

**Proposition 4.3** (Arratia, [2])**.** *The process $\left(1 - \frac{S_1}{\log x}, \frac{Y_1}{\log x}, \frac{Y_2}{\log x}, \dots\right)$ follows a GEM distribution.*

*Proof.* Applying the map $T_x(s) := \log_2 x - \log s$ to the points of $\{S_i : i \in \mathbb{Z}\}$ gives a homogeneous Poisson process on the real line with constant rate 1, by using the Mapping Theorem (Proposition B.3) with Lemma 4.2. Furthermore, we have $T_x(S_0) < 0 \leqslant T_x(S_1)$ and $(T_x(S_i))_{i \in \mathbb{Z}}$ increasing. Therefore, $T_x(S_1)$ and $T_x(S_{i+1}) - T_x(S_i)$ for all $i \geqslant 1$ are independent exponential random variables of parameter 1. If $X$ is a standard exponential random variable, then $1 - e^{-X}$ is a uniform random variable in $[0, 1]$. We conclude that $1 - \frac{S_1}{\log x}, \frac{Y_1}{S_1}, \frac{Y_2}{S_2}, \dots$ are independent uniform random variables in $[0, 1]$. The proposition follows by the characterization of the GEM distribution described in the introduction. □

Therefore, for the next sections, we will assume that the process $\mathbf{L} = (L_1, L_2, \dots)$ sampled for our coupling was determined by $\mathscr{R}$ by defining $L_1 := 1 - \frac{S_1}{\log x}$ and $L_j := \frac{Y_{j-1}}{\log x}$ for $j \geqslant 2$. In this setting, we now have that

$$J_x = \prod_{i \geqslant 1} e^{h(Y_i)}. \tag{4.1}$$

4.1. **Proof of Lemma 2.3.** We conclude this section by using the realization of the GEM described here to prove Lemma 2.3. Note that

$$\Theta_x = r(L_1 \log x) + \sum_{i \geqslant 1} r(Y_i) \leqslant \Theta_\infty + O(1) \tag{4.2}$$

with $\Theta_\infty := \sum_{i \in \mathbb{Z}} r(Y_i)$. With Campbell's Theorem (Proposition B.5), we directly compute that

$$\mathbb{E}[e^{\alpha \Theta_\infty}] = \exp\left( \int_0^\infty \frac{e^{\alpha r(y)} - 1}{y} \, \mathrm{d}y \right). \tag{4.3}$$

This integral is convergent for all fixed $\alpha > 0$ because $e^{\alpha r(y)} - 1 \ll_\alpha r(y) \ll \min\{y, y^{-2}\}$. Combining this fact with (4.2) proves that $\mathbb{E}[e^{\alpha \Theta_x}] \ll_\alpha 1$. We have thus established Lemma 2.3.

## 5. An integer-friendly version of $J_x$

Let $\Lambda$ be the von Mangoldt function, that is to say

$$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^k \text{ for some prime power } p^k, \\ 0 & \text{otherwise.} \end{cases}$$

Recall the Poisson Process $\mathscr{R} = (W_i, Y_i)_{i \in \mathbb{Z}}$ given in Definition 4.1. We then define

$$\mathscr{R}^* := \left\{ \left( W_i Y_i / h(Y_i), e^{h(Y_i)} \right) : i \in \mathbb{Z},\ Y_i > e^{-\gamma} \right\}.$$

Without loss of generality, we may assume that the quantities $W_i Y_i / h(Y_i)$ with $Y_i > e^{-\gamma}$ are all distinct. By the Mapping Theorem (Proposition B.3) applied to the map $(w, y) \to (wy/h(y), e^{h(y)})$, the random set $\mathscr{R}^*$ is a Poisson process on the space $\mathbb{R}_{>0} \times \{\text{prime powers}\}$ with mean measure $\mu^*$ satisfying

$$\mu^*(B \times \{q\}) = \int_B \frac{\Lambda(q)}{q^{1+t}} \, \mathrm{d}t$$

for any $B \subseteq \mathbb{R}_{>0}$ and any $q \in \mathbb{N}$. There is a unique way to relabel the points of $\mathscr{R}^*$ as $\{(T_i^*, Q_i^*) : i \in \mathbb{Z}\}$ such that:

- $T_i^* < T_{i+1}^*$ for all $i \in \mathbb{N}$;
- $\prod_{i=1}^\infty Q_i^* \leqslant x < \prod_{i=0}^\infty Q_i^*$.

We then define the random integer

$$J_x^* := \prod_{i=1}^\infty Q_i^*.$$

One advantage of introducing $J_x^*$ is that the computation for the distribution of $J_x^*$ can be done more easily and precisely than the distribution of $J_x$. We perform this calculation in this section. We will then see in the next section that the probability that $J_x$ is different from $J_x^*$ is small enough to be negligible in the calculation of the total variation distance in Section 7.

This random integer $J_x^*$ was used by Arratia in [2], and he estimated its distribution in his Lemma 2. For the sake of completeness, we repeat his proof in Lemma 5.2 below. But first, we need the following preliminary result on a particular sum of independent Poisson random variables.

**Lemma 5.1.** *Let $\lambda \in (0, 1)$, let $(X_k)_{k \geqslant 1}$ be a sequence of independent Poisson random variables such that $\mathbb{E}[X_k] = \frac{\lambda^k}{k}$, and let $Z := \sum_{k \geqslant 1} k \cdot X_k$. Then we have*

$$\mathbb{P}[Z = \ell] = (1 - \lambda) \cdot \lambda^\ell \qquad \text{for } \ell = 0, 1, 2 \ldots$$

*Proof using generating functions.* Let $s \in \mathbb{C}$ with $|s| < \lambda^{-1}$. We have that $\mathbb{E}[s^{X_k}] = \exp(\frac{\lambda^k}{k} \cdot (s-1))$. With the independence of the $X_k$'s, we have

$$\mathbb{E}[s^Z] = \prod_{k=1}^{\infty} \exp(\tfrac{\lambda^k}{k} \cdot (s^k - 1)) = \exp\big( -\log(1-\lambda s) + \log(1-\lambda)\big) = \sum_{j=0}^{\infty} (1-\lambda)\lambda^j \cdot s^j.$$

We finally recover the probability mass function of $Z$ using

$$\mathbb{P}[Z = \ell] = \frac{1}{2\pi i} \oint_{|s|=1} \frac{\mathbb{E}[s^Z]}{s^{\ell+1}} \, \mathrm{d}s. \qquad \square$$

*Proof using combinatorics.* For any $\ell \geqslant 1$, we have

$$\begin{aligned}\mathbb{P}[Z = \ell] &= \sum_{\substack{m_1,\dots,m_\ell \geqslant 0 \\ \sum j m_j = \ell}} \left( \prod_{j=1}^{\ell} \mathbb{P}[X_j = m_j] \prod_{j=\ell+1}^{\infty} \mathbb{P}[X_j = 0] \right) \\ &= \exp\left( -\sum_{j=1}^{\infty} \frac{\lambda^j}{j} \right) \cdot \lambda^\ell \cdot \sum_{\substack{m_1,\dots,m_\ell \geqslant 0 \\ \sum j m_j = \ell}} \prod_{j=1}^{\ell} \frac{(1/j)^{m_j}}{m_j!}.\end{aligned}$$

By a classical formula due to Cauchy [16, Proposition 1.3.2], the expression $\ell! \cdot \prod_{j=1}^{\ell} \frac{(1/j)^{m_j}}{m_j!}$ equals the number of permutations in $S_\ell$ with $m_j$ cycles of length $j$ for all $j$. In particular,

$$\sum_{\substack{m_1,\dots,m_\ell \geqslant 0 \\ \sum j m_j = \ell}} \prod_{j=1}^{\ell} \frac{(1/j)^{m_j}}{m_j!} = 1,$$

whence $\mathbb{P}[Z = \ell] = (1-\lambda) \cdot \lambda^\ell$, as claimed. Finally, we compute $\mathbb{P}[Z = 0]$ using the formula $\mathbb{P}[Z = 0] = 1 - \mathbb{P}[Z \geqslant 1]$. $\square$

**Lemma 5.2** (Arratia [2])**.** *For $x \geqslant 2$ and $1 \leqslant j \leqslant x$, we have*

$$\mathbb{P}[J_x^* = j] = \frac{1}{j \log x} \left( 1 + O\left( \frac{1}{\log x} \right) \right).$$

*Proof.* Let $t > 0$ and let $q$ be a prime power, and consider the random variable $N_q(t)$ that counts the number of points $(T_i^*, Q_i^*)$ with $T_i^* > t$ and $Q_i^* = q$. Moreover, let $I_t^* \coloneqq \prod q^{N_q(t)}$ with the product being over all prime powers $q$, and let $Z_p(t) \coloneqq \sum_{k=1}^{\infty} k N_{p^k}(t)$. Thus, we have

$$I_t^* = \prod_p p^{Z_p(t)}.$$

The family of random variables $(Z_p(t))_p$ is independent since $\mathscr{R}^*$ is a Poisson process. We also know that each random variable $N_{p^k}(t)$ follows a Poisson distribution with parameter $\frac{1}{kp^{k(1+t)}}$. Hence, $\mathbb{P}[Z_p(t) = \ell] = (1 - p^{-1-t}) p^{-\ell(1+t)}$ by Lemma 5.1 for all $\ell \geqslant 0$. Therefore,

$$\mathbb{P}[I_t^* = j] = \prod_p \mathbb{P}[Z_p(t) = \nu_p(j)] = \frac{j^{-1-t}}{\zeta(1+t)}.$$

All points of $\mathscr{R}^*$ have distinct $T^*$-coordinates with probability one. In this situation, we have that $J_x^* = j$ if and only if there exists exactly one point $(T^*, Q^*) \in \mathscr{R}^*$ such that $I_{T^*}^* = j$ and $Q^* > \frac{x}{j}$.

Thus, we have

$$\mathbb{1}_{J_x^*=j} = \sum_{(T^*,Q^*)\in\mathscr{R}^*} \mathbb{1}_{I_{T^*}^*=j \text{ and } Q^*>x/j}$$

almost surely. Taking expectations on both sides and using the Mecke equation (Proposition B.6), we get the distribution of $J_x^*$:

$$\mathbb{P}[J_x^* = j] = \int_0^\infty \left( \sum_{q>x/j} \frac{\Lambda(q)}{q^{1+t}} \right) \cdot \frac{j^{-1-t}}{\zeta(1+t)} \, \mathrm{d}t. \tag{5.1}$$

We want to estimate this integral. Let $S(u) \coloneqq \sum_{q\leqslant u} \frac{\Lambda(q)}{q}$. We have $S(u) = \log u + O(1)$ for $u \geqslant 1$ by Mertens' estimate [12, Theorem 3.4(a)]. We use partial summation to get

$$\sum_{q>x/j} \frac{\Lambda(q)}{q^{1+t}} = \int_{x/j}^\infty u^{-t} \, \mathrm{d}S(u) = \frac{(x/j)^{-t}}{t}(1+O(t)).$$

for all $t > 0$. By putting this estimate in (5.1), we have

$$\mathbb{P}[J_x^* = j] = \frac{1}{j} \int_0^\infty \frac{x^{-t}(1+O(t))}{t \cdot \zeta(1+t)} \, \mathrm{d}t.$$

Since $\zeta(1+t) \geqslant 1$ for all $t > 0$, the portion of the integral over $t \geqslant 1$ is $\ll 1/(x \log x)$. On the other hand, if $t \in (0,1]$, we have $1/\zeta(1+t) = t + O(t^2)$. We conclude that

$$\mathbb{P}[J_x^* = j] = \frac{1}{j} \int_0^1 x^{-t}(1+O(t)) \, \mathrm{d}t + O\left(\frac{1}{jx\log x}\right) = \frac{1}{j\log x}\left(1 + O\left(\frac{1}{\log x}\right)\right).$$

This concludes the proof. □

## 6. When $J_x$ and $J_x^*$ are different

To prove Proposition 2.4, we must get a hold of the distribution of $J_x$. Since we have a good approximation for the distribution of $J_x^*$, it will be enough, for our purposes, to show that the event $\{J_x \neq J_x^*\}$ occurs with low probability.

For any $t > 0$, consider the random variable

$$I_t \coloneqq \sum_{(W,Y)\in\mathscr{R}} Y \cdot \mathbb{1}_{W>t}.$$

We compute below the distribution of $I_t$.

**Proposition 6.1** (Arratia, [2])**.** *For any fixed $t > 0$, the random variable $I_t$ follows an exponential distribution of parameter $t$, i.e. $\mathbb{P}[I_t > y] = e^{-ty}$ for all $y > 0$.*

*Proof.* A direct application of Campbell's Theorem (Proposition B.5) implies that $\mathbb{E}[e^{sI_t}] = t/(t-s)$ for $\mathrm{Re}(s) < t$, which agrees with the moment generating function of an exponential distribution of parameter $t$. This completes the proof. □

Let $\eta$ be the smallest positive constant satisfying

$$\frac{a}{h(a)} \cdot \frac{h(b)}{b} \leqslant 1 + \frac{\eta}{\min\{a,b\}^2}.$$

for all $a, b > e^{-\gamma}$. Such a constant must exist by (2.4). In addition, let

$$r_0 \coloneqq \sup_{y>0} r(y).$$

Let us then define the following events:

- $\mathcal{E}_1 = \big\{S_1 < \log x - R_x - r_0,\ S_0 > \log x + R_x + 2r_0\big\}$, where $R_x \coloneqq \sum_{i\geqslant 2} r(Y_i)$.
- $\mathcal{E}_2$ is the event where $\frac{W_i}{W_0} > 1 + \frac{\eta}{\min\{Y_0,Y_i\}^2}$ for all $i \geqslant 1$.
- $\mathcal{E}_3$ is the event where $\frac{W_0}{W_i} > 1 + \frac{\eta}{\min\{Y_0,Y_i\}^2}$ for all $i \leqslant -1$.

The variable $R_x$ depends on the value of $x$ since the labeling of points in $\mathscr{R}$ change as $x$ grows, even if $\mathscr{R}$ stays fixed.

**Lemma 6.2.** *For $x > 1$, we have $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \subseteq \{J_x = J_x^*\}$. In particular,*

$$\mathbb{P}[J_x \neq J_x^*] \leqslant \mathbb{P}[\mathcal{E}_1^c] + \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2^c] + \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_3^c].$$

*Proof.* Recall that $J_x = \prod_{i\geqslant 1} e^{h(Y_j)}$ and assume that $\mathcal{E}_1$ occurs. Then

$$\log J_x = \sum_{i=1}^{\infty} h(Y_i) \leqslant \sum_{i\geqslant 1} \big(Y_i + r(Y_i)\big) \leqslant S_1 + R_x + r_0 < \log x, \tag{6.1}$$

since $r(Y_1) \leqslant r_0$. Similarly, we have

$$\log\big(J_x \cdot e^{h(Y_0)}\big) = \sum_{i=0}^{\infty} h(Y_i) \geqslant \sum_{i\geqslant 0} \big(Y_i - r(Y_i)\big) \geqslant S_0 - R_x - 2r_0 > \log x.$$

Therefore, $\mathcal{E}_1$ implies the inequalities

$$J_x < x < J_x e^{h(Y_0)}. \tag{6.2}$$

In particular, we must have $Y_0 > e^{-\gamma}$.

Assume now further that $\mathcal{E}_2$ and $\mathcal{E}_3$ also occur. We claim that this implies $J_x = J_x^*$.

Let $B_+$ be the set of integers $i \geqslant 1$ such that $Y_i > e^{-\gamma}$, and let $B_-$ be the set of integers $i \leqslant -1$ such that $Y_i > e^{-\gamma}$. By our assumption that $\mathcal{E}_2 \cap \mathcal{E}_3$ occurs and by the definition of $\eta$, we have

$$\frac{W_i Y_i}{h(Y_i)} > \frac{W_0 Y_0}{h(Y_0)} \qquad \text{if } i \in B_+ \tag{6.3}$$

and, similarly,

$$\frac{W_i Y_i}{h(Y_i)} < \frac{W_0 Y_0}{h(Y_0)} \qquad \text{if } i \in B_-. \tag{6.4}$$

Now, let $(T_j^*, Q_j^*)$ with $j \in \mathbb{Z}$ be the indexing of points of $\mathscr{R}^*$ given in Section 5, that is to say we have $T_j^* < T_{j+1}^*$ for all $j$, and such that $\prod_{j=0}^{\infty} Q_j^* > x \geqslant \prod_{j=1}^{\infty} Q_j^*$. Let $j_0 \in \mathbb{Z}$ be such that $Q_{j_0}^* = e^{h(Y_0)}$ (which exists because $Y_0 > e^{-\gamma}$). Using relations (6.3) and (6.4), we find that

$$\prod_{j>j_0} Q_j^* = \prod_{i\in B_+} e^{h(Y_i)} = \prod_{i\geqslant 1} e^{h(Y_i)} = J_x.$$

On the other hand, we have $\prod_{j\geqslant j_0} Q_j^* = e^{h(Y_0)} J_x$. Hence, using (6.2), we find that $\prod_{j>j_0} Q_j^* < x < \prod_{j\geqslant j_0} Q_j^*$. In particular, $j_0 = 0$ and thus $J_x^* = \prod_{j>j_0} Q_j^* = J_x$, as claimed. □

The following lemma requires a lot more care than in Arratia's paper [2] since he only needed $\mathbb{P}[J \neq J^*] \ll \frac{\log_2 x}{\log x}$ to be true for his coupling.

**Lemma 6.3.** *We have $\mathbb{P}[\mathcal{E}_1^c] \ll \frac{1}{\log x}$ for $x \geqslant 2$.*

*Proof.* We may assume that $x$ is large enough. We have $\mathbb{E}[e^{R_x}] \leqslant \mathbb{E}[e^{\Theta_\infty}] \ll 1$ by (4.3). Therefore, the event $R_x > \log_2 x$ occurs with probability $\ll \frac{1}{\log x}$. Using Lemma 4.2, we find that the probability that there exists $i \in \mathbb{Z}$ such that $|S_i - \log x| \leqslant 4r_0$ is $O(1/\log x)$. In conclusion,

$$\mathbb{P}[\mathcal{E}_1^c] \leqslant \sum_{k\in\{0,1\}} \mathbb{P}\Big[4r_0 < |S_k - \log x| \leqslant R_x + 2r_0,\ R_x \leqslant \log_2 x\Big] + O\bigg(\frac{1}{\log x}\bigg).$$

Note that if $W_k > (\log_2 x)^{-10}$ and $|S_k - \log x| \leqslant 2r_0 + \log_2 x$, then $|I_{(\log_2 x)^{-10}}| \geqslant \log x - \log_2 x - 2r_0$. So, using Proposition 6.1, we find that

$$\begin{aligned}\mathbb{P}\Big[|S_k - \log x| \leqslant \log_2 x + 2,\ W_k > (\log_2 x)^{-10}\Big] &\leqslant \mathbb{P}\big[I_{(\log_2 x)^{-10}} > \log x - \log_2 x - 2r_0\big]\\ &= \exp\bigg(-\frac{\log x - \log_2 x - 2r_0}{(\log_2 x)^{10}}\bigg)\end{aligned}$$

for $x$ large enough. Consequently,

$$\mathbb{P}[\mathcal{E}_1^c] \leqslant \sum_{k\in\{0,1\}} \mathbb{P}\Big[4r_0 < |S_k - \log x| \leqslant R_x + 2r_0,\ R_x \leqslant \log_2 x,\ W_k \leqslant (\log_2 x)^{-10}\Big] + O\bigg(\frac{1}{\log x}\bigg).$$

Let $m_0$ be the largest integer such that $2^{m-1}r_0 \leqslant \log_2 x$. Hence, if $2r_0 < R_x \leqslant \log_2 x$, then there exists a unique integer $m \in [2, m_0]$ such that $2^{m-1}r_0 < R_x \leqslant 2^m r_0$, in which case the condition $|S_k - \log x| \leqslant R_x + 2r_0$ implies that $|S_k - \log x| \leqslant 2^m r_0 + 2r_0 \leqslant 2^{m+1}r_0$. We conclude that

$$\mathbb{P}[\mathcal{E}_1^c] \leqslant \sum_{\substack{k\in\{0,1\}\\ 2\leqslant m\leqslant m_0}}\sum \mathbb{P}\Big[R > 2^{m-1}r_0,\ |S_k - \log x| \leqslant 2^{m+1}r_0,\ W_k \leqslant (\log_2 x)^{-10}\Big] + O\bigg(\frac{1}{\log x}\bigg).$$

Note that $\mathbb{1}_{R_x > 2^{m-1}r_0} \leqslant \frac{4^{1-m}}{r_0^2}\mathbb{1}_{R_x > 2r_0} R_x^2$. In addition, if $R_x > 2r_0$, then $\sum_{i\geqslant 2} r(Y_i)^2 \leqslant r_0 R_x < R_x^2/2$, whence $R_x^2 \leqslant 4\sum_{i>j\geqslant 2} r(Y_i)r(Y_j)$. We conclude that

$$\mathbb{P}[\mathcal{E}_1^c] \leqslant \sum_{\substack{k\in\{0,1\}\\ 2\leqslant m\leqslant m_0}}\sum \frac{4^{2-m}}{r_0^2}\mathbb{E}\bigg[\sum_{i>j>k} r(Y_i)r(Y_j)\cdot \mathbb{1}_{|S_k - \log x|\leqslant 2^{m+1}r_0}\cdot \mathbb{1}_{W_k\leqslant (\log_2 x)^{-10}}\bigg] + O\bigg(\frac{1}{\log x}\bigg),$$

Therefore, to complete the proof of the lemma, it is enough to show that

$$E(z) := \mathbb{E}\bigg[\sum_{k\in\{0,1\}}\sum_{i>j>k} r(Y_i)r(Y_j)\cdot \mathbb{1}_{|S_k-\log x|\leqslant z}\cdot \mathbb{1}_{W_k\leqslant(\log_2 x)^{-10}}\bigg] \ll \frac{z}{\log x} \tag{6.5}$$

uniformly for $z \in [0, 4\log_2 x]$.

For the rest of the proof, we fix $z \in [0, 4\log_2 x]$. Given $t'' > t' > t$, let

$$I_{t,t',t''} := \sum_{(W,Y)\in\mathscr{R},\ W\in\mathbb{R}_{>t}\setminus\{t',t''\}} Y.$$

Since the $W_i$'s are almost surely distinct, we have

$$E(z) \leqslant 2\cdot\mathbb{E}\bigg[\sum_{\substack{(W,Y),(W',Y'),(W'',Y'')\in\mathscr{R}\\ W''>W'>W}}\sum\sum r(Y')r(Y'')\cdot \mathbb{1}_{|Y+Y'+Y''+I_{W,W',W''}-\log x|\leqslant z}\cdot \mathbb{1}_{W\leqslant(\log_2 x)^{-10}}\bigg].$$

Hence, using the Mecke equation (Proposition B.6), we find that

$$E(z) \leqslant 2 \int\cdots\int_{\substack{0<w<w'<w''\\ w\leqslant(\log_2 x)^{-10}\\ y,y',y''\geqslant 0\\ y+y'+y''\leqslant \log x+z}} r(y')r(y'')\mathbb{P}\Big[\big|I_w+y+y'+y''-\log x\big|\leqslant z\Big]\frac{\mathrm{d}w\cdots\mathrm{d}y''}{e^{wy+w'y'+w''y''}},$$

where the integral is sixfold with variables $w, w', w'', y, y', y''$. Proposition 6.1 implies that

$$\mathbb{P}\Big[\big|I_w+y+y'+y''-\log x\big|\leqslant z\Big]\leqslant e^{-w(\log x-y-y'-y'')}(e^{wz}-e^{-wz})\ll wze^{-w(\log x-y-y')},$$

since $w\leqslant(\log_2 x)^{-10}$ and $z\leqslant 4\log_2 x$ here. Consequently,

$$\begin{aligned}E(z)&\ll \int\cdots\int_{\substack{0<w<w'<w''\\ y,y',y''\geqslant 0\\ y+y'+y''\leqslant \log x+z}} r(y')r(y'')zwe^{-w\log x-(w'-w)y'-(w''-w')(y'+y'')}\,\mathrm{d}w\cdots\mathrm{d}y''\\ &=\frac{z}{(\log x)^2}\iiiint_{\substack{t,y,y',y''\geqslant 0\\ y+y'+y''\leqslant \log x+z}}\frac{r(y')r(y'')}{y'(y'+y'')}te^{-t}\,\mathrm{d}t\,\mathrm{d}y\,\mathrm{d}y'\,\mathrm{d}y'',\end{aligned}$$

where we made the change of variables $t=w\log x$. Since $\int_0^\infty r(u)/u\,\mathrm{d}u\ll 1$ and $\log x+z\ll\log x$, relation (6.5) follows. This completes the proof of the lemma. □

**Lemma 6.4.** *We have* $\mathbb{P}[\mathcal{E}_1\cap\mathcal{E}_2^c]+\mathbb{P}[\mathcal{E}_1\cap\mathcal{E}_3^c]\ll 1/\log x$ *for* $x\geqslant 2$.

*Proof.* Recall that $\mathcal{E}_2$ failing means that there exists $i\geqslant 1$ such that $W_i/W_0\leqslant 1+\eta/\min\{Y_0,Y_i\}^2$. In addition, recall that the event $\mathcal{E}_1$ implies that $Y_0>e^{-\gamma}$ (this was explained in the beginning of the proof of Lemma 6.2). Hence, we have that

$$\mathbb{P}[\mathcal{E}_1\cap\mathcal{E}_2^c]\leqslant\mathbb{E}\Bigg[\sum\sum_{(W,Y),(W',Y')\in\mathscr{R}}\mathbb{1}_{1<W'/W\leqslant 1+\eta/\min\{Y,Y'\}^2}\cdot\mathbb{1}_{Y',Y>e^{-\gamma}}\cdot\mathbb{1}_{I_W\in(\log x-Y,\log x]}\Bigg].$$

We use Mecke's equation as in the proof of (6.5) to get that

$$\mathbb{P}[\mathcal{E}_1\cap\mathcal{E}_2^c]\leqslant\iiiint_{\substack{0<w<w'<w(1+\eta/\min\{y,y'\}^2)\\ y,y'>e^{-\gamma},\ y'\leqslant\log x}}\mathbb{P}\big[\log x-y<I_w+y'\leqslant\log x\big]\cdot e^{-wy-w'y'}\,\mathrm{d}w\,\mathrm{d}w'\,\mathrm{d}y\,\mathrm{d}y'.$$

We have $e^{-w'y'}\leqslant e^{-wy'}$, thus

$$\mathbb{P}[\mathcal{E}_1\cap\mathcal{E}_2^c]\ll\iiint_{\substack{w>0,\ y,y'>e^{-\gamma}\\ y'\leqslant\log x}}\mathbb{P}\big[\log x-y<I_w+y'\leqslant\log x\big]\cdot\frac{we^{-w(y+y')}}{\min\{y,y'\}^2}\,\mathrm{d}w\,\mathrm{d}y\,\mathrm{d}y'.$$

By Proposition 6.1, we have

$$\mathbb{P}\big[\log x-y<I_w+y'\leqslant\log x\big]\leqslant\begin{cases}e^{-w(\log x-y-y')} & \text{if } y\leqslant\log x,\\ 1 & \text{if } y>\log x.\end{cases}$$

Therefore,

$$\begin{aligned}\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2^c] &\ll \iiint\limits_{\substack{w>0,\ y,y'>e^{-\gamma}\\ y,y'\leqslant \log x}} \frac{we^{-w\log x}}{\min\{y,y'\}^2}\,\mathrm{d}w\,\mathrm{d}y\,\mathrm{d}y' + \iiint\limits_{\substack{w>0,\ y,y'>e^{-\gamma}\\ y'\leqslant \log x<y'}} \frac{we^{-w(y+y')}}{(y')^2}\,\mathrm{d}w\,\mathrm{d}y\,\mathrm{d}y'\\ &\leqslant \iint\limits_{e^{-\gamma}<y,y'\leqslant \log x} \frac{1}{\min\{y,y'\}^2(\log x)^2}\,\mathrm{d}y\,\mathrm{d}y' + \iint\limits_{\substack{y,y'>e^{-\gamma}\\ y'\leqslant \log x<y'}} \frac{1}{(yy')^2}\,\mathrm{d}y\,\mathrm{d}y'\\ &\ll 1/\log x.\end{aligned}$$

This completes the proof of the claimed bound on $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2^c]$.

Finally, we bound $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_3^c]$ using a very similar argument. We have

$$\begin{aligned}\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_3^c] &\leqslant \mathbb{E}\Bigg[\sum\sum_{(W,Y),(W',Y')\in\mathscr{R}} \mathbb{1}_{1<W/W'\leqslant 1+\eta/\min\{Y,Y'\}^2}\cdot \mathbb{1}_{Y',Y>e^{-\gamma}}\cdot \mathbb{1}_{I_W\in(\log x-Y,\log x]}\Bigg]\\ &= \iiiint\limits_{\substack{0<w'<w<w'(1+\eta/\min\{y,y'\}^2)\\ y,y'>e^{-\gamma}}} \mathbb{P}\big[\log x - y < I_w \leqslant \log x\big]\cdot e^{-wy-w'y'}\,\mathrm{d}w\,\mathrm{d}w'\,\mathrm{d}y\,\mathrm{d}y'\\ &\leqslant J_1 + J_2,\end{aligned}$$

where

$$J_1 \coloneqq \iiiint\limits_{\substack{0<w'<w<w'(1+\eta/\min\{y,y'\}^2)\\ y,y'>e^{-\gamma},\ y\leqslant \log x}} e^{-w\log x-w'y'}\,\mathrm{d}w\,\mathrm{d}w'\,\mathrm{d}y\,\mathrm{d}y'$$

and

$$J_2 \coloneqq \iiiint\limits_{\substack{0<w'<w<w'(1+\eta/\min\{y,y'\}^2)\\ y>\log x,\ y'>e^{-\gamma}}} e^{-wy-w'y'}\,\mathrm{d}w\,\mathrm{d}w'\,\mathrm{d}y\,\mathrm{d}y'.$$

Using $e^{-w\log x}\leqslant e^{-w'\log x}$, we find that

$$\begin{aligned}J_1 &\ll \iiint\limits_{\substack{w'>0,\ y,y'>e^{-\gamma}\\ y\leqslant \log x}} \frac{w'e^{-w'(y'+\log x)}}{\min\{y,y'\}^2}\,\mathrm{d}w'\,\mathrm{d}y\,\mathrm{d}y'\\ &= \iint\limits_{\substack{y,y'>e^{-\gamma}\\ y\leqslant \log x}} \frac{1}{\min\{y,y'\}^2(y'+\log x)^2}\,\mathrm{d}y\,\mathrm{d}y'\\ &\ll 1/\log x.\end{aligned}$$

Similarly, we have

$$\begin{aligned} J_2 &\ll \iiint\limits_{\substack{w'>0,\ y'>e^{-\gamma}\\ y>\log x}} \frac{w' e^{-w'(y+y')}}{\min\{y,y'\}^2}\,\mathrm{d}w'\,\mathrm{d}y\,\mathrm{d}y' \\ &= \iint\limits_{\substack{y'>e^{-\gamma}\\ y>\log x}} \frac{1}{\min\{y,y'\}^2(y+y')^2}\,\mathrm{d}y\,\mathrm{d}y' \\ &\ll 1/\log x. \end{aligned}$$

This implies that $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_3^c] \ll 1/\log x$, thus completing the proof of the lemma. □

As an immediate corollary of Lemmas 6.2, 6.3 and 6.4, we have:

**Proposition 6.5.** *For $x \geqslant 2$, we have $\mathbb{P}[J_x \neq J_x^*] \ll 1/\log x$.*

## 7. Proof of Proposition 2.4

With a good estimation of the distribution of $J_x^*$ and with the fact that $J_x = J_x^*$ with a quantifiably high probability, we are able to give an upper bound on the total variation distance between the distribution of $M_x$ and $N_x$.

*Proof of Proposition 2.4.* Recall that we constructed $N_x$ in the coupling such that $\mathbb{P}[M_x \neq N_x] = d_{\mathrm{TV}}(\mu_x, \nu_x)$ with $\mu_x$ and $\nu_x$ being the distribution of $M_x$ and $N_x$ respectively. Note that $M_x > x$ must imply that $J_x > x$, which means that $\mathcal{E}_1$ cannot happen by (6.1). Hence, using Lemma 6.3, we see that it is enough to show that

$$\mathbb{P}[M_x \in A] = \frac{\#A}{\lfloor x \rfloor} + O\Big(\frac{1}{\log x}\Big)$$

uniformly over all $A \subseteq \mathbb{Z} \cap [1, x]$.

For each $j \leqslant x$, we define the sets

$$A_j := \bigcup_{p:\ pj \in A} \left( \frac{\theta(p-1)}{\theta(x/j)}, \frac{\theta(p)}{\theta(x/j)} \right].$$

Note that $M_x \in A$ and $J_x \leqslant x/2$ if, and only if, $J_x = j$ for some $j \leqslant x/2$ and $U_1' \in A_j$. With this in mind, we define the following events:

- $\mathcal{B}_1 = \{J_x = J_x^* \leqslant x/2\}$;
- $\mathcal{B}_2 = \{M_x \in A\}$;
- $\mathcal{B}_3 = \{J_x^* = j \text{ and } U_1' \in A_j \text{ for some } j \leqslant x/2\}$.

Since $\mathcal{B}_1 \cap \mathcal{B}_2 = \mathcal{B}_1 \cap \mathcal{B}_3$, we have

$$\Big|\mathbb{P}[\mathcal{B}_2] - \mathbb{P}[\mathcal{B}_3]\Big| \leqslant \mathbb{P}[\mathcal{B}_1^c] \leqslant \mathbb{P}\big[J_x \neq J_x^*\big] + \mathbb{P}\big[x/2 < J_x^* \leqslant x\big] \ll \frac{1}{\log x}.$$

by Lemma 5.2 and by Proposition 6.5. Therefore, we have

$$\begin{aligned}\mathbb{P}\big[M_x \in A\big] &= \sum_{j\leqslant x/2} \mathbb{P}\big[J_x^* = j,\ U_1' \in A_j\big] + O\bigg(\frac{1}{\log x}\bigg)\\ &= \sum_{p,j:\ pj\in A} \sum \frac{\log p}{\theta(x/j)} \cdot \mathbb{P}[J_x^* = j] + O\bigg(\frac{1}{\log x}\bigg). \end{aligned} \tag{7.1}$$

Since $\theta(t)/t = 1 + O\big((\log t)^{-2}\big)$ for all $t \geqslant 2$ by the Prime Number Theorem [12, Theorem 8.1], we have

$$\frac{\mathbb{P}[J_x^* = j]}{\theta(x/j)} = \frac{1}{\lfloor x\rfloor \log x}\bigg(1 + O\bigg(\frac{1}{(\log(x/j))^2} + \frac{1}{\log x}\bigg)\bigg) \tag{7.2}$$

for $j \leqslant x/2$. In addition, note that

$$\sum_{p,j:\ pj\in A} \sum \log p \cdot \bigg(\frac{1}{(\log(x/j))^2} + \frac{1}{\log x}\bigg) \ll \sum_{j\leqslant x/2} \bigg(\frac{x}{j(\log(x/j))^2} + \frac{x}{j\log x}\bigg) \ll x. \tag{7.3}$$

By combining (7.1), (7.2) and (7.3), we reduce the problem to showing the following:

$$\sum_{p,j:\ pj\in A} \sum \frac{\log p}{\log x} = \#A + O\bigg(\frac{x}{\log x}\bigg). \tag{7.4}$$

If we set $L(a) = \sum_{p|a} \log p$, then we have

$$\sum_{p,j:\ pj\in A} \sum \frac{\log p}{\log x} = \sum_{a\in A} \frac{L(a)}{\log x} = \#A - \sum_{a\in A} \frac{\log(x/a)}{\log x} - \sum_{a\in A} \frac{\log a - L(a)}{\log x}.$$

Since $\log(x/a) \geqslant 0$ for $a \leqslant x$, and we also have $L(a) \leqslant \log a$ for all $a$, relation (7.4) holds uniformly over all choices of $A \subseteq \mathbb{Z}\cap[1,x]$ if, and only if,

$$\sum_{a\leqslant x} \log(x/a) \ll x \quad \text{and} \quad \sum_{a\leqslant x} (\log a - L(a)) \ll x. \tag{7.5}$$

The first estimate follows readily by partial summation (or by Stirling's formula). The second one is a simple consequence of the identity $\log a - L(a) = \sum_{p^\nu \| a,\, \nu\geqslant 2} (\nu-1)\log p$. This establishes (7.5), and hence it completes the proof of Proposition 2.4. □

# PART II. FACTORIZATION INTO $k$ PARTS

## 8. THEOREM 2 FOR THE PROBABILISTIC MODEL

With the following probabilistic version of Theorem 2 given by Donnelly and Tavaré in 1987 [8, Section 3], we see how a Poisson–Dirichlet process relates to the Dirichlet distribution. We repeat their proof since they use a Poisson process that will come in handy later.

**Proposition 8.1** (Donnelly–Tavaré [8])**.** *Let* $\mathbf{V} = (V_1, V_2, \ldots)$ *be a Poisson–Dirichlet process and let* $(C_i)_{i\geqslant 1}$ *be a sequence of i.i.d. random variables (also independent of* $\mathbf{V}$*) with* $\mathbb{P}[C_i = j] = \alpha_j$ *for all* $i \geqslant 1$*. Then the random vector*

$$\left( \sum_{i:\ C_i=1} V_i, \ldots, \sum_{i:\ C_i=k} V_i \right)$$

*is distributed according to* $\mathrm{Dir}(\boldsymbol{\alpha})$.

*Proof.* Let $X_1 > X_2 > \cdots$ be the points of a Poisson process on $\mathbb{R}_{>0}$ with intensity $\frac{e^{-x}}{x}\,\mathrm{d}x$, and let $S \coloneqq \sum_{i\geqslant 1} X_i$. For each $i \geqslant 1$, Let $V_i \coloneqq \frac{X_i}{S}$. We then know that $(V_1, V_2, \ldots)$ follows the Poisson–Dirichlet distribution [9, Theorem 2.2]. With the Colouring Theorem (Proposition B.4), the point processes

$$\Pi_m = \{X_i : C_i = m\}$$

form independent Poisson processes of intensity $\alpha_m \cdot \frac{e^{-x}}{x}\,\mathrm{d}x$ for all $m = 1, \ldots, k$. We then let

$$S_m \coloneqq \sum_{X\in\Pi_m} X$$

for $m = 1, 2, \ldots, k$. With Campbell's Theorem (Proposition B.5), we compute the moment generating function

$$\log \mathbb{E}[e^{sS_m}] = \alpha_m \cdot \int_0^\infty \frac{e^{sx} - 1}{x} \cdot e^{-x}\,\mathrm{d}x = -\alpha_m \log(1 - s)$$

for $\mathrm{Re}(s) < 1$. This coincides with the moment generating function of the distribution $\mathrm{Gamma}(\alpha_m, 1)$. We thus deduce that the vector

$$\left( \sum_{i:\ C_i=1} V_i, \ldots, \sum_{i:\ C_i=k} V_i \right) = \left( \frac{S_1}{\sum_{m=1}^k S_m}, \ldots, \frac{S_k}{\sum_{m=1}^k S_m} \right)$$

follows the distribution $\mathrm{Dir}(\boldsymbol{\alpha})$ (see [11, Chapter 49, pp. 485-487] for a proof). □

## 9. A COUPLING FOR $\mathbf{D}_x$

Fix $k \geqslant 2$, $\boldsymbol{\alpha} \in \Delta^{k-1}$ with $\alpha_i > 0$ for all $i$, and $f \in \mathcal{F}_k(\boldsymbol{\alpha})$. In addition, let $\mathbf{D}_x$ be a random variable satisfying (1.5), and let $\mathbf{Z}$ be a $\Delta^{k-1}$-valued random variable distributed according to $\mathrm{Dir}(\boldsymbol{\alpha})$. We will construct an appropriate coupling between $\mathbf{D}_x$ and $\mathbf{Z}$. This will build on the coupling given in Section 2. To state our result, we also introduce the $\Delta^{k-1}$-valued random variable

$$\boldsymbol{\delta}_x \coloneqq \left( \frac{\log D_{x,1}}{\log N_x}, \ldots, \frac{\log D_{x,k}}{\log N_x} \right) \quad \text{and} \quad \boldsymbol{\delta}_x^* \coloneqq \left( \frac{\log D_{x,1}}{\log x}, \ldots, \frac{\log D_{x,k}}{\log x} \right), \tag{9.1}$$

which are more convenient to work with than $\mathbf{D}_x$ because we can compare them to $\mathbf{Z}$. To this end, we write $\|\cdot\|_\infty$ for the supremum norm in $\mathbb{R}^k$. Note that

$$\|\boldsymbol{\delta}_x - \boldsymbol{\delta}_x^*\|_\infty = \frac{\log(x/N_x)}{(\log x)(\log N_x)} \cdot \max_{1\leqslant i\leqslant k} \log D_{x,i} \leqslant \frac{\log(x/N_x)}{\log x}. \tag{9.2}$$

In addition, we have the following lemma:

**Lemma 9.1.** *Let $x \geqslant 2$ and assume the above notation. There exists a probability space $\Omega'$ containing copies of the random variables $N_x$, $\boldsymbol{\delta}_x$, $\boldsymbol{\delta}_x^*$, $\mathbf{V}$ and $\mathbf{Z}$, and an event $\mathcal{E}$ such that:*

*(a) $\mathbb{P}[\mathcal{E}^c] \ll 1/\log x$;*
*(b) If $\mathcal{E}$ occurs, then*

$$\|\boldsymbol{\delta}_x - \mathbf{Z}\|_\infty \leqslant \frac{2\cdot\log(x/N_x) + 3\cdot\log s(N_x) + 2\cdot\Theta_x}{\log x}.$$

*Proof.* Our starting point is the coupling of $N_x$ and $\mathbf{V}$ described in Section 2. We equip this space with some additional random variables that are all independent of each other and of $N_x$ and $\mathbf{V}$, and whose role will become apparent later:

- a sequence $(C_i')_{i\geqslant 1}$ of random variables such that $\mathbb{P}[C_i' = \ell] = \alpha_\ell$ for all $i \geqslant 1$ and all $\ell = 1, 2, \dots, k$;
- for each natural number $n$ and each prime $p$, let $\big(X_p^{(1)}(n), \dots, X_p^{(k)}(n)\big)$ be independent random vectors

$$\mathbb{P}\big[X_p^{(i)}(n) = e_i\ \forall i \leqslant k\big] = f(p^{e_1}, \dots, p^{e_k}),$$

for every non-negative integer solutions to $e_1 + \cdots + e_k = \nu_p(n)$.

Finally, we take $\mathcal{E} = \{M_x = N_x\}$, which satisfies the required property (a) by Proposition 2.4.

We will now show that the space we constructed above contains copies of $\boldsymbol{\delta}_x$, $\boldsymbol{\delta}_x^*$ and of $\mathbf{Z}$ satisfying part (b). The construction of $\boldsymbol{\delta}_x$ and $\boldsymbol{\delta}_x^*$ is rather straightforward. Indeed, using the fact that $f \in \mathcal{F}_k(\boldsymbol{\alpha})$ (property (c) of Definition 1.1), we may easily check that the random $k$-tuple $(\prod_p p^{X_p^{(i)}(N_x)})_{i=1}^k$ satisfies (1.5). Thus, we may take

$$D_{x,i} \coloneqq \prod_p p^{X_p^{(i)}(N_x)} \quad \text{for } i = 1, 2, \dots, k,$$

so that the distribution of $\mathbf{D}_x = (D_{x,1}, \dots, D_{x,k})$ is indeed in accordance with (1.5). We then define $\boldsymbol{\delta}_x$ and $\boldsymbol{\delta}_x^*$ as in (9.1).

Next, we define $\mathbf{Z}$. Firstly, we introduce a new sequence of random variables $(C_i)_{i\geqslant 1}$. To define them, fix $n \in \mathbb{Z}\cap[1,x]$ and, for every prime $p|n^\flat = n/s(n)$, fix a number $\ell(p) \in \{1,\dots,k\}$. Moreover, fix choices of $\ell_1', \ell_2', \dots \in \{1, 2, \dots, k\}$. Recall that $N_x = P_1P_2\cdots$ with the $P_i$'s forming a non-increasing sequence of primes or ones. Assume that the events $N_x = n$, $X_p^{(j)} = \mathbb{1}_{j=\ell(p)}$ for $j = 1, \dots, k$, and $C_i' = \ell_i'$ occur.

- If $i \geqslant 1$ is such that $P_i|n^\flat$, then we let $C_i = \ell(P_i)$.
- If $i \geqslant 1$ is such that $P_i|s(n)$, then we let $C_i = \ell_i'$.

Since $f \in \mathcal{F}_k(\boldsymbol{\alpha})$ satisfies property (b) of Definition 1.1, a straightforward computation reveals that the random variables $(C_i)_{i\geqslant 1}$ are independent of each other, of $N_x$ and of $\mathbf{V}$, and they satisfy

$$\mathbb{P}[C_i = \ell] = \alpha_\ell \quad \text{for } i = 1, 2, \dots \text{ and } \ell = 1, 2, \dots, k.$$

Finally, equipped with these random variables and motivated by the proof of Donnelly and Tavaré in Section 8, we take

$$\mathbf{Z} \coloneqq \left( \sum_{i \geqslant 1:\ C_i = 1} V_1, \dots, \sum_{i \geqslant 1:\ C_i = k} V_k \right),$$

which is distributed according to $\mathrm{Dir}(\boldsymbol{\alpha})$ by the discussion in Section 8.

In order to complete the proof of the lemma, it remains to show that (b) holds when $\mathcal{E} = \{M_x = N_x\}$. Define the auxiliary random variable

$$\boldsymbol{\rho}_x \coloneqq \left( \sum_{i \geqslant 1:\ C_i = 1} \frac{\log P_i}{\log x}, \dots, \sum_{i \geqslant 1:\ C_i = k} \frac{\log P_i}{\log x} \right).$$

When $\mathcal{E}$ holds, then Lemma 2.1 implies that

$$\sum_{i \geqslant 1} |\log P_i - V_i \log x| \leqslant \log(x/N_x) + 2 \cdot \log s(N_x) + 2 \cdot \Theta_x.$$

Thus, we readily find that

$$\|\boldsymbol{\rho}_x - \mathbf{Z}\|_\infty \leqslant \frac{\log(x/N_x) + 2 \cdot \log s(N_x) + 2 \cdot \Theta_x}{\log x}. \tag{9.3}$$

In addition, by the construction of the random variables $C_i$, we claim that

$$\|\boldsymbol{\delta}^*_x - \boldsymbol{\rho}_x\|_\infty \leqslant \frac{\log s(N_x)}{\log x}. \tag{9.4}$$

Indeed, if $N_x = n$, then we have

$$\begin{aligned} \log D_{x,\ell} &= \sum_{p \mid n^\flat} X_p^{(\ell)}(n) \log p + \sum_{p \mid s(n)} X_p^{(\ell)}(n) \log p \\ &= \sum_{i \geqslant 1:\ P_i \mid n^\flat} \mathbb{1}_{X_{P_i}^{(\ell)}(n) = 1} \log P_i + \sum_{p \mid s(n)} X_p^{(\ell)}(n) \log p \\ &= \sum_{i \geqslant 1:\ P_i \mid n^\flat} \mathbb{1}_{C_i = \ell} \log P_i + \sum_{p \mid s(n)} X_p^{(\ell)}(n) \log p, \end{aligned}$$

whence (9.4) follows readily.

Combining (9.2)-(9.4) completes the proof of part (b), and hence of the lemma. □

In order to make use of Lemma 9.1, we need to introduce some further notation. Let

$$R_{\mathrm{NT}} \coloneqq 3 \cdot \frac{\log(x/N_x) + \log s(N_x)}{\log x} \quad \text{and} \quad R_{\mathrm{PD}} \coloneqq 2 \cdot \frac{\Theta_x}{\log x}.$$

We write $B(\mathbf{x}, r)$ for the closed ball of radius $r \geqslant 0$ centered at $\mathbf{x}$ in the normed vector space $(\mathbb{R}^k, \|\cdot\|_\infty)$. We recall that $\Delta^{k-1}$ is the standard $(k-1)$-dimensional simplex. For any subset $A \subseteq \Delta^{k-1}$, we write $\partial A$ for the boundary of $A$ in the relative topology of $\Delta^{k-1}$ (cf. Definition C.1(c)).

In addition, we define the following events:

- $\mathcal{E}_{\mathrm{NT}}(A) \coloneqq \big\{ B(\boldsymbol{\delta}^*_x, R_{\mathrm{NT}}) \cap \partial A = \emptyset \big\}$;
- $\mathcal{E}_{\mathrm{PD}}(A) \coloneqq \big\{ B(\mathbf{Z}, R_{\mathrm{PD}}) \cap \partial A = \emptyset \big\}$.

**Lemma 9.2.** *For any measurable set $A \subseteq \Delta^{k-1}$, we have*

$$\mathbb{P}\big[\boldsymbol{\delta}_x \in A\big] = \mathbb{P}\big[\mathbf{Z} \in A\big] + O\left(\mathbb{P}\big[\mathcal{E}_{\mathrm{NT}}(A)^c\big] + \mathbb{P}\big[\mathcal{E}_{\mathrm{PD}}(A)^c\big] + \frac{1}{\log x}\right).$$

*Proof.* Let $\mathcal{E}$ be as in Lemma 9.1. Hence, it suffices to show that

$$\Big|\mathbb{P}\big[\mathcal{E} \cap \{\boldsymbol{\delta}_x \in A\}\big] - \mathbb{P}\big[\mathcal{E} \cap \{\mathbf{Z} \in A\}\big]\Big| \leqslant \mathbb{P}\big[\mathcal{E}_{\mathrm{NT}}(A)^c\big] + \mathbb{P}\big[\mathcal{E}_{\mathrm{PD}}(A)^c\big]. \tag{9.5}$$

We shall prove the following stronger statement: if $\mathcal{E} \cap \mathcal{E}_{\mathrm{NT}}(A) \cap \mathcal{E}_{\mathrm{PD}}(A)$ occurs, then $\boldsymbol{\delta}_x \in A$ if, and only if, $\mathbf{Z} \in A$.

Indeed, assume that $\mathcal{E} \cap \mathcal{E}_{\mathrm{NT}}(A) \cap \mathcal{E}_{\mathrm{PD}}(A)$ occurs, and we also have $\boldsymbol{\delta}_x \in A$. We must show that $\mathbf{Z} \in A$. Assume for contradiction that $\mathbf{Z} \notin A$.

Let $R'_{\mathrm{NT}} := (\log x)^{-1}(2 \cdot \log(x/N_x) + 3 \cdot \log s(N_x))$. Note that $B(\boldsymbol{\delta}_x, R'_{\mathrm{NT}}) \subseteq B(\boldsymbol{\delta}^*_x, R_{\mathrm{NT}})$ by (9.2). By assumption of $\mathcal{E}_{\mathrm{NT}}(A)$, we know that $B(\boldsymbol{\delta}_x, R'_{\mathrm{NT}}) \cap \partial A = \emptyset$. Since $\boldsymbol{\delta}_x \in A$, we use Lemma C.2 to find

$$B(\boldsymbol{\delta}_x, R'_{\mathrm{NT}}) \cap \Delta^{k-1} \subseteq A.$$

On the other hand, since $\mathbf{Z} \in \Delta^{k-1} \setminus A$ and $B(\mathbf{Z}, R_{\mathrm{PD}}) \cap \partial A = \emptyset$, we must have

$$B(\mathbf{Z}, R_{\mathrm{PD}}) \cap \Delta^{k-1} \subseteq \Delta^{k-1} \setminus A.$$

In particular, we find that $B(\boldsymbol{\delta}_x, R'_{\mathrm{NT}}) \cap B(\mathbf{Z}, R_{\mathrm{PD}}) \cap \Delta^{k-1} = \emptyset$. On the other hand, we know that $\|\boldsymbol{\delta}_x - \mathbf{Z}\|_\infty \leqslant R'_{\mathrm{NT}} + R_{\mathrm{PD}}$ when $\mathcal{E}$ occurs by Lemma 9.1(b). In particular, there exists a point $\mathbf{y}$ on the line segment connecting $\boldsymbol{\delta}_x$ and $\mathbf{Z}$ such that $\|\boldsymbol{\delta}_x - \mathbf{y}\|_\infty \leqslant R'_{\mathrm{NT}}$ and $\|\mathbf{y} - \mathbf{Z}\|_\infty \leqslant R_{\mathrm{PD}}$. Since the sets $\Delta^{k-1}$, $B(\boldsymbol{\delta}_x, R'_{\mathrm{NT}})$ and $B(\mathbf{Z}, R_{\mathrm{PD}})$ are convex, we conclude that $\mathbf{y}$ lies in their intersection. But we had seen before that this intersection is the empty set. We have thus arrived at a contradiction. This completes the proof that if $\mathcal{E} \cap \mathcal{E}_{\mathrm{NT}}(A) \cap \mathcal{E}_{\mathrm{PD}}(A)$ occurs, and we also know that $\boldsymbol{\delta}_x \in A$, then we must also have that $\mathbf{Z} \in A$.

Conversely, we may show by a simple variation of the above argument that if $\mathcal{E} \cap \mathcal{E}_{\mathrm{NT}}(A) \cap \mathcal{E}_{\mathrm{PD}}(A)$ occurs, and we also know that $\mathbf{Z} \in A$, then we must also have that $\boldsymbol{\delta}_x \in A$. This completes the proof of the lemma. □

To prove Theorem 2, we need to bound $\mathbb{P}[\mathcal{E}_{\mathrm{NT}}(A)^c]$ and $\mathbb{P}[\mathcal{E}_{\mathrm{PD}}(A)^c]$ when $A$ equals the set

$$\Delta^{k-1}_{\mathbf{u}} := \{\mathbf{x} \in \Delta^{k-1} : x_i \leqslant u_i \ \forall i < k\}.$$

with $\mathbf{u} \in (0,1]^{k-1}$. Note that

$$\partial\big(\Delta^{k-1}_{\mathbf{u}}\big) \subseteq \{\mathbf{x} \in \Delta^{k-1} : \exists i < k \text{ such that } x_i = u_i < 1\},$$

(here it is important that the boundary of $\Delta^{k-1}_{\mathbf{u}}$ is defined with respect to the topology of $\Delta^{k-1}$). Therefore,

$$\mathcal{E}_{\mathrm{NT}}(\Delta^{k-1}_{\mathbf{u}})^c \subseteq \bigcup_{\substack{i<k \\ u_i \neq 1}} \mathcal{B}^{(i)}_{\mathrm{NT}}(u_i), \quad \text{where} \quad \mathcal{B}^{(i)}_{\mathrm{NT}}(u) := \Big\{\big|\log D_{x,i} - u\log x\big| \leqslant 3\log(x \cdot s(N_x)/N_x)\Big\},$$

as well as

$$\mathcal{E}_{\mathrm{PD}}(\Delta^{k-1}_{\mathbf{u}})^c \subseteq \bigcup_{\substack{i<k \\ u_i \neq 1}} \mathcal{B}^{(i)}_{\mathrm{PD}}(u_i), \quad \text{where} \quad \mathcal{B}^{(i)}_{\mathrm{PD}}(u) := \left\{\big|Z_i - u\big| \leqslant \frac{2\Theta_x}{\log x}\right\},$$

We then have the following two crucial estimates:

**Lemma 9.3.** *For $i \in \{1, 2, \dots, k-1\}$ and $u \in (0,1)$, we have the uniform estimate*

$$\mathbb{P}\big[\mathcal{B}_{\mathrm{NT}}^{(i)}(u)\big] \ll \frac{1}{(1+u\log x)^{1-\alpha_i}(1+(1-u)\log x)^{\alpha_i}}.$$

**Lemma 9.4.** *For $i \in \{1, 2, \dots, k-1\}$ and $u \in (0,1)$, we have that*

$$\mathbb{P}\big[\mathcal{B}_{\mathrm{PD}}^{(i)}(u)\big] \ll \frac{1}{(1+u\log x)^{1-\alpha_i}(1+(1-u)\log x)^{\alpha_i}}.$$

Using Lemmas 9.2-9.4 together with Proposition 8.1 yields immediately Theorem 2. Thus, it remains to prove Lemmas 9.3 and 9.4, which we do in the next two sections.

*Remarks.* (a) The proofs of Lemmas 9.3 and 9.4 are rather involved. However, it is possible to obtain slightly weaker versions of them in a rather easy manner, which will then lead to a correspondingly weaker version of Theorem 2.

First, we prove a variation of Lemma 9.2. Let $\mathcal{E}^*(A)$ be the event that $B(\mathbf{Z}, R^*) \cap \partial A = \emptyset$ with

$$R^* := \frac{2\cdot\log(x/N_x) + 3\cdot\log s(N_x) + 2\cdot\Theta_x}{\log x}.$$

A straightforward modification of the proof of Lemma 9.2 implies that

$$\mathbb{P}[\mathbf{D}_x \in A] = \mathbb{P}[\mathbf{Z} \in A] + O\left(\mathbb{P}\big[\mathcal{E}^*(A)^c\big] + \frac{1}{\log x}\right) \tag{9.6}$$

for any measurable set $A$. With Chernoff's bound and Lemmas 2.2-2.3, the probability that $R^* > \log_2 x$ is $\ll 1/\log x$. Moreover, for any $\delta \in (0, 1/4]$, we can show by a direct computation with the Dirichlet distribution the uniform bound

$$\mathbb{P}\Big[\exists j < k \text{ such that } |Z_j - u_j| \leqslant \delta\Big] \ll \sum_{1\leqslant j<k} \frac{\delta}{(u_j+\delta)^{1-\alpha_j}(1-u_j+\delta)^{\alpha_j}}$$

for all $\mathbf{u} \in [0,1]^{k-1}$ (see Lemma 11.1 below for a proof of this claim). Taking $\delta = \frac{\log_2 x}{\log x}$ proves that

$$\mathbb{P}\big[D_{x,i} \leqslant N_x^{u_i} \ \forall i < k\big] = F_{\boldsymbol{\alpha}}(\mathbf{u}) + O\left(\sum_{\substack{1\leqslant i<k\\ u_i\neq 1}} \frac{1}{(1+u_i\frac{\log x}{\log_2 x})^{1-\alpha_i}(1+(1-u_i)\frac{\log x}{\log_2 x})^{\alpha_i}}\right).$$

(b) As a matter of fact, the above proof could have worked with Arratia's original coupling from [2]. Under this coupling, we have $\mathbb{P}[M_x \neq N_x] \ll \frac{\log_2 x}{\log x}$ (see Remark (b) at the end of Section 2), so that Lemma 9.1 would hold with part (a) replaced by the weaker bound $\mathbb{P}[\mathcal{E}^c] \ll \frac{\log_2 x}{\log x}$, and thus (9.6) would hold with $\frac{\log_2 x}{\log x}$ in place of $\frac{1}{\log x}$.

## 10. Proof of Lemma 9.3 using elementary number-theoretic techniques

Note that

$$\mathbb{P}\big[N_x^\flat \leqslant x/(\log x)^3\big] \leqslant \frac{1}{\log x}\mathbb{E}\big[(x/N_x^\flat)^{1/3}\big] \ll \frac{1}{\log x} \tag{10.1}$$

by Lemma 2.2 applied with $\alpha = \beta = 1/3$.

Fix now $i \in \{1, \dots, k-1\}$ and $u \in (0,1)$. For simplicity, let us write

$$\alpha = \alpha_i$$

for the remainder of this section. Given $z$, let us define

$$S(z) \coloneqq \sum_{\substack{n\leqslant x\\ n^\flat\in(\frac{x}{z},\frac{2x}{z}]}} \sum_{\substack{d_1\cdots d_k=n\\ d_i\in[x^uz^{-3},x^uz^3]}} f(d_1,\dots,d_k).$$

Using (10.1) and (1.5), we find that

$$\mathbb{P}\big[\mathcal{B}_{\mathrm{NT}}^{(i)}(u)\big] \leqslant \frac{1}{\lfloor x\rfloor} \sum_{1\leqslant m\leqslant 5\log_2 x} S(2^m) + O\bigg(\frac{1}{\log x}\bigg). \tag{10.2}$$

Hence Lemma 9.3 follows readily from the estimate below:

**Lemma 10.1.** *There exists a universal constant $x_0\geqslant 2$ such that if $x\geqslant x_0$ and $z\in[2,(\log x)^4]$, then we have the uniform estimate*

$$S(z) \ll \frac{\log z}{\sqrt{z}}\cdot\frac{x}{(1+u\log x)^{1-\alpha}(1+(1-u)\log x)^{\alpha}}.$$

*Proof.* If $d_1\cdots d_k=n$, then we may uniquely write $d_j=d_j'd_j''$ with $d_j'|n^\flat$ and $d_j''|s(n)$. In particular, $d_j''\leqslant s(n)\leqslant x/n^\flat\leqslant z$ for all $j$. Hence, if $d_i\in[x^uz^{-3},x^uz^3]$, we must also have that $d_i'\in[x^uz^{-4},x^uz^3]$. Hence, using property (c) of Definition 1.1, we deduce that

$$\sum_{\substack{d_1\cdots d_k=n\\ d_i\in[x^uz^{-3},x^uz^3]}} f(d_1,\dots,d_k) \leqslant \sum_{\substack{d_1'\cdots d_k'=n^\flat\\ d_i'\in[x^uz^{-4},x^uz^3]}} f(d_1',\dots,d_k') \sum_{d_1''\cdots d_k''=s(n)} f(d_1'',\dots,d_k'').$$

Relation (1.4) implies that $f(d_1',\dots,d_k')=\prod_{j=1}^k\alpha_j^{\omega(d_j')}$. Moreover, using property (a) of Definition 1.1, we find that the sum over $d_j''$ equals 1. In conclusion,

$$\sum_{\substack{d_1\cdots d_k=n\\ d_i\in[x^uz^{-3},x^uz^3]}} f(d_1,\dots,d_k) \leqslant \sum_{\substack{d_1'\cdots d_k'=n^\flat\\ d_i'\in[x^uz^{-4},x^uz^3]}} \prod_{j=1}^k \alpha_j^{\omega(d_j')} = \sum_{\substack{d|n^\flat\\ d\in[x^uz^{-4},x^uz^3]}} \alpha^{\omega(d)}(1-\alpha)^{\omega(n^\flat/d)}.$$

Let $n^\flat=dm$ and $s(n)=b$. If $n^\flat\in[\frac{x}{z},\frac{2x}{z}]$ and $d\in[x^uz^{-4},x^uz^3]$, then $m\in[x^{1-u}z^{-2},2x^{1-u}z^3]$. Hence, we find that

$$\begin{aligned} S(z) &\leqslant \sum_{\substack{dm\in[\frac{x}{z},\frac{2x}{z}]\\ d\in[x^uz^{-4},x^uz^3]\\ m\in[x^{1-u}z^{-2},2x^{1-u}z^3]}}\sum \alpha^{\omega(d)}(1-\alpha)^{\omega(m)} \sum_{\substack{b\leqslant z\\ b \text{ square-full}}} 1\\ &\ll \sqrt{z} \sum_{\substack{dm\in[\frac{x}{z},\frac{2x}{z}]\\ d\in[x^uz^{-4},x^uz^3]\\ m\in[x^{1-u}z^{-2},2x^{1-u}z^3]}}\sum \alpha^{\omega(d)}(1-\alpha)^{\omega(m)} = \sqrt{z}\cdot(S_1+S_2), \end{aligned} \tag{10.3}$$

where $S_1$ denotes the double sum over $d$ and $m$ with the additional constraint $d\leqslant\sqrt{2x/z}$, and $S_2$ is the corresponding sum over pairs $(d,m)$ with $d>\sqrt{2x/z}$ and $m\leqslant\sqrt{2x/z}$.

First, we bound $S_1$. Note that for the conditions on $d$ to be compatible we must have that $u\leqslant 2/3$, provided that $x$ is large enough. Assuming this is the case, we apply twice Proposition

A.3 to find that

$$\begin{aligned} S_1 &\leqslant \sum_{\substack{d\leqslant\sqrt{2x/z}\\ d\in[x^u z^{-4},x^u z^3]}} \alpha^{\omega(d)} \sum_{m\leqslant\frac{2x}{dz}} (1-\alpha)^{\omega(m)} \\ &\ll \sum_{\substack{d\leqslant\sqrt{2x/z}\\ d\in[x^u z^{-4},x^u z^3]}} \alpha^{\omega(d)} \cdot \frac{x(\log x)^{-\alpha}}{dz} \\ &\ll \frac{x(\log z)(1+u\log x)^{\alpha-1}(\log x)^{-\alpha}}{z}. \end{aligned} \tag{10.4}$$

We bound $S_2$ in a similar manner. For this sum to be non-empty, we have $u \geqslant 1/3$. If this is the case, then we have

$$\begin{aligned} S_2 &\leqslant \sum_{\substack{m\leqslant\sqrt{2x/z}\\ m\in[x^{1-u}z^{-3},2x^{1-u}z^2]}} (1-\alpha)^{\omega(m)} \sum_{d\leqslant\frac{2x}{mz}} \alpha^{\omega(d)} \\ &\ll \sum_{\substack{m\leqslant\sqrt{2x/z}\\ m\in[x^{1-u}z^{-3},2x^{1-u}z^2]}} (1-\alpha)^{\omega(m)} \cdot \frac{x(\log x)^{\alpha-1}}{mz} \\ &\ll \frac{x(\log z)(1+(1-u)\log x)^{-\alpha}(\log x)^{\alpha-1}}{z}. \end{aligned} \tag{10.5}$$

Combining (10.3), (10.4) and (10.5), while keeping in mind that $S_1 = 0$ if $u > 2/3$ and that $S_2 = 0$ if $u < 1/3$, completes the proof of the lemma, □

## 11. Proof of Lemma 9.4 using probabilistic techniques

Since $\mathbf{Z}$ follows the distribution $\mathrm{Dir}(\boldsymbol{\alpha})$, its $i$-th component $Z_i$ follows the $\mathrm{Beta}(\alpha_i, 1-\alpha_i)$ distribution, meaning that if $[a,b] \subseteq [0,1]$, then

$$\mathbb{P}\big[Z_i \in [a,b]\big] = \frac{1}{\Gamma(\alpha_i)\Gamma(1-\alpha_i)} \int_a^b \frac{\mathrm{d}t}{t^{1-\alpha_i}(1-t)^{\alpha_i}} = \frac{\sin(\pi\alpha_i)}{\pi} \int_a^b \frac{\mathrm{d}t}{t^{1-\alpha_i}(1-t)^{\alpha_i}}, \tag{11.1}$$

by Euler's reflection formula. For this reason, we need the following preliminary estimate.

**Lemma 11.1.** *Uniformly for $u, \delta, \alpha \in [0,1]$, we have*

$$\sin(\pi\alpha) \int_{[u-\delta,u+\delta]\cap[0,1]} \frac{\mathrm{d}t}{t^{1-\alpha}(1-t)^{\alpha}} \ll \frac{\delta}{(u+\delta)^{1-\alpha}(1-u+\delta)^{\alpha}}.$$

*In particular, if $\delta \geqslant 1/\log x$, then*

$$\sin(\pi\alpha) \int_{[u-\delta,u+\delta]\cap[0,1]} \frac{\mathrm{d}t}{t^{1-\alpha}(1-t)^{\alpha}} \ll \frac{\delta\log x}{(1+u\log x)^{1-\alpha}(1+(1-u)\log x)^{\alpha}}.$$

*Proof.* The lemma holds trivially when $\alpha \in \{0,1\}$, so let us assume that $\alpha \in (0,1)$. Note that both sides of the claimed inequality are invariant under the change of variables $(\alpha, u) \to (1-\alpha, 1-u)$. Hence, we may assume without loss of generality that $u \in [0, 1/2]$. In addition, observe that

$$\sin(\pi\alpha) \int_0^1 \frac{\mathrm{d}t}{t^{1-\alpha}(1-t)^{\alpha}} = \frac{\sin(\pi\alpha)}{\Gamma(\alpha)\Gamma(1-\alpha)} = \pi.$$

Hence the lemma is trivially true if $\delta \in [1/4, 1]$.

We have thus reduced the proof to the case when $\alpha \in (0,1)$, $u \in [0, 1/2]$ and $\delta \in [0, 1/4]$. In particular, $u + \delta \leqslant 3/4$, so that $1 - t \in [1/4, 1]$ for all $t \in [0, u + \delta]$. It thus suffices to prove that

$$\sin(\pi\alpha) \int_{[u-\delta,u+\delta]\cap[0,1]} \frac{\mathrm{d}t}{t^{1-\alpha}} \ll \frac{\delta}{(u+\delta)^{1-\alpha}}. \tag{11.2}$$

Assume first that $\delta \leqslant u/2$. We then have $t \geqslant u/2$ whenever $t \in [u - \delta, u + \delta]$. Using also the trivial bound $\sin(\pi\alpha) \leqslant 1$, we conclude that

$$\sin(\pi\alpha) \int_{[u-\delta,u+\delta]\cap[0,1]} \frac{\mathrm{d}t}{t^{1-\alpha}} \leqslant \int_{[u-\delta,u+\delta]\cap[0,1]} \frac{\mathrm{d}t}{(u/2)^{1-\alpha}} \ll \frac{\delta}{(u+\delta)^{1-\alpha}}.$$

This proves the lemma in this case.

Finally, assume that $\delta \geqslant u/2$. We then use that

$$\sin(\pi\alpha) \int_{[u-\delta,u+\delta]\cap[0,1]} \frac{\mathrm{d}t}{t^{1-\alpha}} \leqslant \sin(\pi\alpha) \int_0^{3\delta} \frac{\mathrm{d}t}{t^{1-\alpha}} = \frac{\sin(\pi\alpha)}{\alpha} \cdot (3\delta)^\alpha \ll \delta^\alpha \asymp \frac{\delta}{(u+\delta)^{1-\alpha}}.$$

This completes the proof of the lemma in all cases. □

Let us now show Lemma 9.4. For the simplicity of notation, let us fix $i \in \{1, 2, \dots, k\}$ and $u \in (0,1)$, and let us set $\alpha = \alpha_i$ and

$$\Delta = \frac{1}{(1 + u\log x)^{1-\alpha}(1 + (1-u)\log x)^\alpha}.$$

Thus, our goal is to show that

$$\mathbb{P}\left[\big|Z_i - u\big| \leqslant \frac{2\Theta_x}{\log x}\right] \ll \Delta. \tag{11.3}$$

Recall that

$$\Theta_x = \sum_{j\geqslant 1} r(V_j \log x),$$

and that there exists an absolute constant $c \geqslant 1$ such that $r(y) \leqslant c\min\{y, y^{-2}\}$ for all $y > 0$ (see (2.4)). Since there are at most 10 indices $j$ such that $V_j \geqslant 0.1$, we find that

$$\Theta_x \leqslant \Theta'_x + 10c, \quad \text{where} \quad \Theta'_x := \sum_{j\geqslant 1:\ V_j<0.1} r(V_j \log x).$$

Now, using Lemma 11.1 and relation (11.1), we have that

$$\mathbb{P}\left[\big|Z_i - u\big| \leqslant \frac{200c}{\log x}\right] \ll \Delta.$$

On the other hand, if $\frac{200c}{\log x} < |Z_i - u| \leqslant \frac{2\Theta_x}{\log x} \leqslant \frac{20c + 2\Theta'_x}{\log x}$, then we must have $\Theta'_x > 90c$. In particular, there exists $m \in \mathbb{Z}$ such that $2^m > 40c$ and $\Theta'_x \in (2^m, 2^{m+1}]$, whence $|Z_i - u| \leqslant 2^{m+2}/\log x$. We thus conclude that

$$\mathbb{P}\left[\big|Z_i - u\big| \leqslant \frac{20c + 2\Theta'_x}{\log x}\right] \leqslant \sum_{m\in\mathbb{Z}:\ 2^m>40c} \mathbb{P}\left[\big|Z_i - u_i\big| \leqslant \frac{20c + 2^{m+2}}{\log x},\ \Theta_x > 2^m\right] + O(\Delta).$$

Therefore Lemma 9.4 will follow if we can prove that

$$\mathbb{P}\left[\big|Z_i - u\big| \leqslant \frac{20c + 4\kappa}{\log x},\ \Theta'_x > \kappa\right] \ll \frac{\Delta}{\kappa} \qquad \text{for all } \kappa \geqslant 40c. \tag{11.4}$$

We shall make a further reduction. If we set

$$G(\lambda) \coloneqq \#\big\{j \geqslant 1 : V_j \log x \in (\lambda, 2\lambda]\big\},$$

then we have

$$\begin{aligned}\Theta'_x = \sum_{j\geqslant 1:\ V_j<0.1} r(V_j \log x) &\leqslant c \sum_{j\geqslant 1:\ V_j<0.1} \min\big\{V_i \log x, (V_i \log x)^{-2}\big\} \\ &\leqslant c \sum_{\substack{m\geqslant 0\\ 2^m\leqslant 0.1\log x}} \frac{G(2^m)}{4^m} + c\sum_{m<0} 2^{m+1} G(2^m) \\ &\leqslant 5c \max_{\substack{m\geqslant 0\\ 2^m\leqslant 0.1\log x}} \left(\frac{G(2^m)}{2^{3m/2}}\right) + 5c \max_{m<0}\big(2^{m/2} G(2^m)\big),\end{aligned}$$

since $\sum_{m\geqslant 0} 2^{-m/2} < 5$ and $\sum_{m<0} 2^{1+m/2} < 5$. We thus find that

$$\begin{aligned}\mathbb{P}\bigg[\big|Z_i - u\big| \leqslant \frac{20c+4\kappa}{\log x},\ \Theta'_x > \kappa\bigg] \leqslant &\sum_{\substack{m\geqslant 0\\ 2^m\leqslant 0.1\log x}} \mathbb{P}\bigg[\big|Z_i - u\big| \leqslant \frac{20c+4\kappa}{\log x},\ G(2^m) > \frac{2^{3m/2}\kappa}{10c}\bigg] \\ &+ \sum_{m<0} \mathbb{P}\bigg[\big|Z_i - u\big| \leqslant \frac{20c+4\kappa}{\log x},\ G(2^m) > \frac{\kappa 2^{-m/2}}{10c}\bigg].\end{aligned}$$

Note that in both sums, we have $G(2^m) \geqslant 4$, since $\kappa \geqslant 40c$. Hence, Markov's inequality implies

$$\begin{aligned}\mathbb{P}\bigg[\big|Z_i - u\big| \leqslant \frac{20c+4\kappa}{\log x},\ \Theta'_x > \kappa\bigg] \leqslant &\sum_{\substack{m\geqslant 0\\ 2^m\leqslant 0.1\log x}} \frac{250c^2}{\kappa^2 2^{3m}} \mathbb{E}\Big[\mathbb{1}_{|Z_i-u|\leqslant \frac{20c+4\kappa}{\log x}} \cdot \mathbb{1}_{G(2^m)\geqslant 4} \cdot G(2^m)^2\Big] \\ &+ \sum_{m<0} \frac{250c^2 2^m}{\kappa^2} \mathbb{E}\Big[\mathbb{1}_{|Z_i-u|\leqslant \frac{20c+4\kappa}{\log x}} \cdot \mathbb{1}_{G(2^m)\geqslant 4} \cdot G(2^m)^2\Big].\end{aligned}$$

This reduces (11.4), and thus Lemma 9.4, to proving the following estimate:

**Lemma 11.2.** *Uniformly for $\mu \geqslant 1$ and $\lambda \in (0, 0.1\log x]$, we have*

$$\mathbb{E}\Big[\mathbb{1}_{|Z_i-u|\leqslant \frac{\mu}{\log x}} \cdot \mathbb{1}_{G(\lambda)\geqslant 4} \cdot G(\lambda)^2\Big] \ll (\lambda+\mu)\cdot\Delta.$$

*Proof.* Let us call $E$ the quantity we seek to bound from above. Consider two independent Poisson processes $(A_j)_{j\geqslant 1}$ with intensity $\alpha\frac{e^{-x}}{x}\,\mathrm{d}x$ and $(B_j)_{j\geqslant 1}$ with intensity $(1-\alpha)\frac{e^{-x}}{x}\,\mathrm{d}x$. Hence, the union of the two processes is a new Poisson process of intensity $\frac{e^{-x}}{x}\,\mathrm{d}x$. The sum $S_A \coloneqq \sum_{j\geqslant 1} A_j$ has distribution Gamma$(\alpha, 1)$ and the sum $S_B \coloneqq \sum_{j\geqslant 1} B_j$ has distribution Gamma$(1-\alpha, 1)$. If we also set $S \coloneqq S_A + S_B$, then, as we saw in the proof of Proposition 8.1, we have

$$E = \mathbb{E}\Big[\mathbb{1}_{|S_A/S-u|\leqslant \frac{\mu}{\log x}} \cdot \mathbb{1}_{G_A(\lambda)+G_B(\lambda)\geqslant 4} \cdot \Big(G_A(\lambda)+G_B(\lambda)\Big)^2\Big],$$

where $G_A(\lambda) = \sum_{j\geqslant 1} \mathbb{1}_{A_j \log x\in(\lambda S, 2\lambda S]}$ and $G_B(\lambda) = \sum_{j\geqslant 1} \mathbb{1}_{B_j \log x\in(\lambda S, 2\lambda S]}$. We know $G_A(\lambda) + G_B(\lambda) \geqslant 4$. Hence, if $G_A(\lambda) \geqslant G_B(\lambda)$, then we must have $G_A(\lambda) \geqslant 2$, whence $G_A(\lambda) \leqslant$

$G_A(\lambda)^2/2$. So, we find that

$$\Big(G_A(\lambda)+G_B(\lambda)\Big)^2 \leqslant 4G_A(\lambda)^2 = 4G_A(\lambda) + 8\sum_{k>j\geqslant 1}\sum \mathbb{1}_{A_k\log x,\, A_j\log x\in(\lambda S,2\lambda S]}$$
$$\leqslant 16\sum_{k>j\geqslant 1}\sum \mathbb{1}_{A_k\log x,\, A_j\log x\in(\lambda S,2\lambda S]}.$$

The analogous inequality also holds when $G_B(\lambda)\geqslant G_A(\lambda)$, with the roles of $A$ and $B$ reversed. We conclude that

$$E\leqslant 16(E_A+E_B),$$

where

$$E_A := \mathbb{E}\bigg[\mathbb{1}_{|S_A/S-u|\leqslant \frac{\mu}{\log x}}\sum_{k>j\geqslant 1}\sum \mathbb{1}_{A_k\log x,\, A_j\log x\in(\lambda S,2\lambda S]}\bigg]$$

and $E_B$ is defined analogously. Using Mecke's equation (cf. Proposition B.6), we find that

$$E_A = \iint\limits_{2a_1>a_2>a_1>0} \mathbb{E}\bigg[\mathbb{1}_{|\frac{a_1+a_2+S_A}{a_1+a_2+S}-u|\leqslant\frac{\mu}{\log x}}\prod_{j=1}^{2}\mathbb{1}_{a_j\log x\in(\lambda(a_1+a_2+S),2\lambda(a_1+a_2+S)]}\bigg]\frac{e^{-a_1-a_2}}{a_1a_2}\,\mathrm{d}a_1\,\mathrm{d}a_2$$
$$= \frac{1}{\Gamma(\alpha)\Gamma(1-\alpha)}\iiiint\limits_{\substack{2a_1>a_2>a_1>0,\ s_1,s_2>0\\ |\frac{a_1+a_2+s_1}{a_1+a_2+s_1+s_2}-u|\leqslant\frac{\mu}{\log x}\\ a_j\log x\in(\lambda(a_1+a_2+s_1+s_2),2\lambda(a_1+a_2+s_1+s_2)]\ (j=1,2)}} \frac{e^{-a_1-a_2-s_1-s_2}}{a_1a_2s_1^{1-\alpha}s_2^{\alpha}}\,\mathrm{d}a_1\,\mathrm{d}a_2\,\mathrm{d}s_1\,\mathrm{d}s_2,$$

where we used that the $a_j$'s lie in the same dyadic interval to deduce that $a_2<2a_1$. We make the change of variables $t=s_1/(s_1+s_2)$ and $s=s_1+s_2$. Since $\lambda/\log x\leqslant 0.1$ and $a_1<a_2<2a_1$, the conditions $a_j\log x\in\lambda(a_1+a_2+s),2\lambda(a_1+a_2+s)]$ for $j=1,2$ imply that $a_j\log x\in(\lambda s,5\lambda s]$. Knowing also that $|\frac{a_1+a_2+s_1}{a_1+a_2+s}-u|\leqslant\frac{\mu}{\log x}$, we find $|t-u|=|\frac{s_1}{s_1+s_2}-u|\leqslant\frac{10\lambda+\mu}{\log x}$. Finally, we use Euler's reflection formula to write $\Gamma(\alpha)\Gamma(1-\alpha)=\pi/\sin(\pi\alpha)$. We conclude that

$$E_A\leqslant\frac{\sin(\pi\alpha)}{\pi}\iiiint\limits_{\substack{a_2>a_1>0,\ s>0,\ t\in[0,1]\\ |t-u|\leqslant(10\lambda+\mu)/\log x\\ a_j\log x\in(\lambda s,5\lambda s]\ (j=1,2)}}\frac{e^{-a_1-a_2-s}}{a_1a_2t^{1-\alpha}(1-t)^{\alpha}}\,\mathrm{d}a_1\,\mathrm{d}a_2\,\mathrm{d}s\,\mathrm{d}t.$$

For every fixed value of $s$, the integral over $a_1$ and $a_2$ is $\leqslant(\log 5)^2$, since $\int_w^{5w}\frac{e^{-a}}{a}\,\mathrm{d}\leqslant\log 5$ for any $w>0$. We also have $\int_0^{\infty}e^{-s}\,\mathrm{d}s=1$. We thus conclude that

$$E_A\leqslant\frac{(\log 5)^2\sin(\pi\alpha)}{\pi}\int\limits_{\substack{t\in[0,1]\\ |t-u|\leqslant(10\lambda+\mu)/\log x}}\frac{\mathrm{d}t}{t^{1-\alpha}(1-t)^{\alpha}}.$$

Using Lemma 11.1 shows that $E_A\ll(\lambda+\mu)\Delta$. The same estimate holds for $E_B$ too, thus completing the proof of the lemma. □

# PART III. APPENDICES

Because this paper lies in the intersection of number theory and probability theory, readers coming from one of these fields might not be familiar with standard results of the other one. For this reason, we gather here some key results from both fields. We shall also need a standard fact from topology. We present these results in the following three sections.

## APPENDIX A. TOOLS FROM NUMBER THEORY

Let $\theta(x) \coloneqq \sum_{p \leqslant x} \log p$ and let $\psi(x) \coloneqq \sum_{n \leqslant x} \Lambda(n)$ where

$$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^k \text{ for some prime power } p^k, \\ 0 & \text{otherwise.} \end{cases}$$

Understanding these functions gives information about the distribution of primes. In 1896, Charles de la Vallée Poussin and Jacques Hadamard proved the Prime Number Theorem, which gives an approximation for $\psi(x)$. The formulation we shall use in this paper is the following weaker version of their result:

**Proposition A.1** (The Prime Number Theorem)**.** *For $x \geqslant 2$, we have*

$$\theta(x) = x + O\left(\frac{x}{(\log x)^3}\right)$$

*and*

$$\psi(x) = x + O\left(\frac{x}{(\log x)^3}\right).$$

For a proof, see [12, Chapter 8].

**Proposition A.2** (Strong Mertens' estimate)**.** *For $x \geqslant 2$, we have*

$$\sum_{p^k \leqslant x} \frac{1}{kp^k} = \log_2 x + \gamma + O\left(\frac{1}{(\log x)^3}\right).$$

*Proof.* By partial summation and the Prime Number Theorem (Proposition A.1), we directly have

$$\sum_{p^k \leqslant x} \frac{1}{kp^k} = \int_{2^-}^{x} \frac{\mathrm{d}\psi(t)}{t \log t} = \log\log x + c + O\left(\frac{1}{(\log x)^3}\right)$$

for some real constant $c$. Note that

$$\sum_{p^k \leqslant x} \frac{1}{kp^k} + \log \prod_{p \leqslant x} \left(1 - \frac{1}{p}\right) = -\sum_{\substack{p^k > x \\ p \leqslant x}} \frac{1}{kp^k}.$$

Since the right-hand side above tends to $0$, and Mertens' estimate says that

$$\prod_{p \leqslant x} \left(1 - \frac{1}{p}\right) \sim \frac{e^{-\gamma}}{\log x}$$

(see [12, Theorem 3.4(c)] for a proof), then we have $c = \gamma$. □

In the proof of Theorem 2, we need the following estimate:

**Proposition A.3.** *Uniformly for $\alpha \in [0,1]$ and $x, y \geqslant 2$, we have*

$$\sum_{n\leqslant x} \alpha^{\omega(n)} \ll x(\log x)^{\alpha-1} \quad \text{and} \quad \sum_{x/y<n\leqslant xy} \frac{\alpha^{\omega(n)}}{n} \ll (\log y)(\log(xy))^{\alpha-1}.$$

*Proof.* The first bound follows readily by Theorem 14.2 in [12]. Let us now prove the second one. When $y \geqslant \sqrt{x}$, we have

$$\sum_{x/y<n\leqslant xy} \frac{\alpha^{\omega(n)}}{n} \leqslant \sum_{p|n \,\Rightarrow p\leqslant y^3} \frac{\alpha^{\omega(n)}}{n} = \prod_{p\leqslant y^3} \left(1 + \frac{\alpha}{p-1}\right) \ll (\log y)^{\alpha}$$

by the inequality $1 + t \leqslant e^t$ and Mertens' theorem (Proposition A.2). This completes the proof in this case. Finally, assume that $y \leqslant \sqrt{x}$. We then have

$$\sum_{x/y<n\leqslant xy} \frac{\alpha^{\omega(n)}}{n} \leqslant \sum_{\substack{m\in\mathbb{Z}\\ e^m\in[x/y,exy]}} \sum_{n\in(e^{m-1},e^m]} \frac{\alpha^{\omega(n)}}{n} \leqslant \sum_{\substack{m\in\mathbb{Z}\\ e^m\in[x/y,exy]}} e^{1-m} \sum_{n\leqslant e^m} \alpha^{\omega(n)}.$$

For every $m$ as above, the innermost sum is $\ll e^m m^{\alpha-1} \asymp e^m(\log(xy))^{\alpha-1}$ by the first part of the lemma and our assumption that $y \leqslant \sqrt{x}$. Since there are $\leqslant 1 + \log y \ll \log y$ choices for $m$, the needed estimate follows in this last case too. □

## Appendix B. Tools from Probability Theory

B.1. **The Total Variation Distance.** The *total variation distance* is a metric between two probability distributions. Let $\mu$ and $\nu$ be two probability measures on $\mathbb{C}$. Then

$$d_{\mathrm{TV}}(\mu,\nu) := \sup |\mu(A) - \nu(A)|, \tag{B.1}$$

where the supremum is taken over all Lebesgue-measurable subsets of $\mathbb{C}$. For any real number $a$, let

$$a^+ := \max\{a, 0\} \quad \text{and} \quad a^- := \max\{-a, 0\}.$$

When $\mu$ and $\nu$ are supported on $\mathbb{Z}_{\geqslant 1}$, here are some alternative definitions of the total variation distance:

**Lemma B.1.** *Let $\mu$ and $\nu$ be two probability measure supported on $\mathbb{N}$. Then*

$$d_{\mathrm{TV}}(\mu,\nu) = \sum_{i\geqslant 1} (\mu(i) - \nu(i))^+ = \sum_{i\geqslant 1} (\mu(i) - \nu(i))^-$$

*Proof.* Let $E := \{i \in \mathbb{N} : \mu(i) > \nu(i)\}$ and let $\xi := \mu - \nu$. Note that for any $B \subseteq \mathbb{N}$, we have $\xi(B \cap E) \geqslant 0 \geqslant \xi(B \cap E^c)$. Therefore, for any $A \subseteq \mathbb{N}$, we have

$$\xi(A) = \xi(E) + \xi(A \cap E^c) - \xi(A^c \cap E) \leqslant \xi(E).$$

and

$$\xi(A) = \xi(E^c) + \xi(A \cap E) - \xi(A^c \cap E^c) \geqslant \xi(E^c)$$

Therefore, $d_{\mathrm{TV}}(\mu,\nu) = \sup_{A\subseteq\mathbb{N}} |\xi(A)| = \max\{\xi(E), -\xi(E^c)\}$. Note that $\xi(E) = \sum_{i\geqslant 1}(\mu(i) - \nu(i))^+$, that $-\xi(E^c) = \sum_{i\geqslant 1}(\mu(i) - \nu(i))^-$ and that

$$\sum_{i\geqslant 1} (\mu(i) - \nu(i))^+ = \sum_{i\geqslant 1} \left[(\mu(i) - \nu(i)) + (\mu(i) - \nu(i))^-\right] = \sum_{i\geqslant 1} (\mu(i) - \nu(i))^-.$$

The lemma follows. □

The total variation distance will be especially useful in this paper because of the following proposition. In [2, Section 3.8], Arratia proved that for any two random variables $X$ and $Y$ returning positive integers, we can always construct $X'$ and $Y'$ within the same probability space such that:

- $X'$ and $Y'$ have the same distribution as $X$ and $Y$, respectively;
- $Y'$ is a function of $X', U, V$ where $(U, V)$ is a point uniformly chosen in the unit square independent of $X'$;
- $\mathbb{P}[X' \neq Y'] = d_{\mathrm{TV}}(X, Y)$.

We repeat his proof here. Let $\mu$ and $\nu$ be two probability measures supported on $\mathbb{N}$. Let $z_j \coloneqq \sum_{i \leqslant j} \frac{(\mu(i)-\nu(i))^-}{d_{\mathrm{TV}}(\mu,\nu)}$ (with $z_0 \coloneqq 0$). We consider the function $f_{\mu,\nu} \colon \mathbb{N} \times (0,1)^2 \to \mathbb{N}$ defined as

$$f_{\mu,\nu}(m; a, b) \coloneqq \begin{cases} m & \text{if } a \cdot \mu(m) \leqslant \nu(m), \\ \sum_{i \geqslant 1} i \cdot \mathbb{1}_{z_{i-1} < b \leqslant z_i} & \text{otherwise.} \end{cases}$$

This is the function used in Section 2 for the extraction of $N_x$. Note that if $a, b \in (0,1)$ and $m \in \mathbb{N}$, then we have the equivalency

$$f_{\mu,\nu}(m; a, b) \neq m \quad \Longleftrightarrow \quad a \cdot \mu(m) > \nu(m). \tag{B.2}$$

Indeed, the direction "⇒" is obvious. To see the converse direction, note that if $a \cdot \mu(m) > \nu(m)$, then $(\mu(m) - \nu(m))^- = 0$, and hence the interval $(z_{m-1}, z_m]$ is empty.

**Lemma B.2** (Arratia, [2])**.** *Let $\mu$ and $\nu$ be two probability measures supported on $\mathbb{N}$, let $X$ be a random variable with law $\mu$, and let $U$ and $U'$ be two uniform random variables in $(0,1)$ such that $X, U, U'$ are independent. Let $Y \coloneqq f_{\mu,\nu}(X; U, U')$ with $f_{\mu,\nu}$ defined as above. Then $\mathbb{P}[X \neq Y] = d_{\mathrm{TV}}(\mu, \nu)$ and $\mathbb{P}[Y \in A] = \nu(A)$.*

*Proof.* Using (B.2), we find that

$$\{X \neq Y\} = \{U \cdot \mu(X) > \nu(X)\}. \tag{B.3}$$

Furthermore, we directly compute that

$$\mathbb{P}\big[U \cdot \mu(m) > \nu(m), X = m\big] = (\mu(m) - \nu(m))^+. \tag{B.4}$$

Therefore,

$$\mathbb{P}[X \neq Y] = \mathbb{P}[U \cdot \mu(X) > \nu(X)] = \sum_{m \geqslant 1} (\mu(m) - \nu(m))^+ = d_{\mathrm{TV}}(\mu, \nu) \tag{B.5}$$

with Lemma B.1.

Next, we prove that $\mathbb{P}[Y = n] = \nu(n)$ for any $n \in \mathbb{N}$. Note that $Y = n$ if, and only if, one of two disjoint events happen: either we have $U \cdot \mu(n) \leqslant \nu(n)$ with $X = n$, or we have $U \cdot \mu(X) > \nu(X)$ with $z_{n-1} < U' \leqslant z_n$. Therefore, with (B.4) and (B.5), we have

$$\begin{aligned} \mathbb{P}[Y = n] &= \mathbb{P}\big[U \cdot \mu(n) \leqslant \nu(n), X = n\big] + \mathbb{P}\big[U \cdot \mu(X) > \nu(X)\big] \cdot \mathbb{P}\big[z_{n-1} < U' \leqslant z_n\big] \\ &= \Big(\mu(n) - \big(\mu(n) - \nu(n)\big)^+\Big) + d_{\mathrm{TV}}(\mu, \nu) \cdot \frac{(\mu(n) - \nu(n))^-}{d_{\mathrm{TV}}(\mu, \nu)} = \nu(n), \end{aligned}$$

where we used (B.3) and (B.5) to show that $\mathbb{P}\big[U \cdot \mu(X) > \nu(X)\big] = d_{\mathrm{TV}}(\mu, \nu)$. This concludes the proof of the lemma. □

B.2. **Poisson Processes.** The following definition and propositions are borrowed from Kingman's book on Poisson processes [10]. Let $(S, \mathcal{S})$ be a measurable space with $S$ being a subset of $\mathbb{R}^d$ for some $d \geqslant 1$. A *Poisson process* on a state space $S$ with mean measure $\mu$ is a random countable subset $\Pi \subseteq S$ such that:

- for any disjoint measurable subsets $A_1, \dots, A_n$ of $S$, the random variables $\#(\Pi \cap A_1), \dots, \#(\Pi \cap A_n)$ are independent;
- the random variable $\#(\Pi \cap A)$ is a Poisson random variable of parameter $\mu(A)$ for any $A \subset S$ measurable.

We can see $\Pi$ as an element of the measurable space $(\Omega_S, \mathcal{F}_S)$ where $\Omega_S$ is the set of countable subsets of $S$ and $\mathcal{F}_S$ is the smallest $\sigma$-algebra for which the map $\Pi \mapsto \#\{\Pi \cap B\}$ is measurable for all $B \in \mathcal{S}$. If $\mu$ has no atoms, meaning no singleton with positive probability, and is $\sigma$-finite, meaning that $S$ is a countable union of measurable sets with finite measure, then a Poisson process with mean measure $\mu$ always exists (see [10], the Existence Theorem in section 2.5 for a proof). If $\mu$ is absolutely continuous with respect to the Lebesgue measure, then the function $\lambda \colon S \to \mathbb{R}_{\geqslant 0}$ such that $\mu(A) = \int_A \lambda(x)\,\mathrm{d}x$ for all measurable subsets $A \subseteq S$ is called the *intensity* of the Poisson process. Here are a few important propositions about Poisson processes that we will use in the paper and proved in [10]:

**Proposition B.3** (Mapping Theorem, [10] section 2.3)**.** *If $\Pi$ is a Poisson process with mean measure $\mu$ on $S$, and $f : S \to T$ is a measurable function such that $\mu^*(B) \coloneqq \mu(f^{-1}(B))$ has no atoms, then $f(\Pi)$ is a Poisson process on $T$ with measure $\mu^*$.*

**Proposition B.4** (Colouring Theorem, [10] section 5.1)**.** *If $\Pi$ is a Poisson process with mean measure $\mu$ on $S$, and the points are randomly coloured with $k$ colours such that the probability of a point receiving the colour $i$ is $p_i$, and such that the colour of a point is independent of different points and of the position of the point. Let $\Pi_i$ be the subset of $\Pi$ with colour $i$. Then all the $\Pi_i$ are independent Poisson processes with mean measures $\mu_i = p_i\mu$.*

**Proposition B.5** (Campbell's Theorem, [10] section 3.2)**.** *Let $\Pi$ be any Poisson process on $S$ with mean measure $\mu$. Let $f \colon S \to \mathbb{R}$ be a measurable function. Then*

$$\Sigma = \sum_{X \in \Pi} f(X)$$

*is absolutely convergent almost surely if and only if*

$$\int_S \min\{f, 1\}\,\mathrm{d}\mu < \infty.$$

*If this condition holds, then*

$$\mathbb{E}[e^{s\Sigma}] = \exp\left(\int_S (e^{sf} - 1)\,\mathrm{d}\mu\right)$$

*for any complex $s$ for which the integral converges. Moreover,*

$$\mathbb{E}[\Sigma] = \int_S f\,\mathrm{d}\mu$$

*if the integral converges. In the case where it converges, we also have*

$$\mathrm{Var}[\Sigma] = \int_S f^2\,\mathrm{d}\mu.$$

Many probabilities or expectations that involve Poisson processes in this paper can be reformulated as

$$\mathbb{E}\sum_{X\in\Pi} f(\Pi\setminus\{X\}, X).$$

In these cases, there is a generalization of the formula for $\mathbb{E}[\Sigma]$ in Campbell's Theorem, called the Mecke equation, allowing us to compute these objects:

**Proposition B.6** (Mecke equation, [13] Theorem 4.5)**.** *Let $\Pi$ be a Poisson process on $S$ with a $\sigma$-finite mean measure $\mu$, and let $f : \Omega_S \times S^k \to [0,\infty)$ be measurable. Then we have*

$$\mathbb{E}\sum_{\substack{X_1,\ldots,X_k\\ \textit{all distinct}}} f(\Pi\setminus\{X_1,\ldots,X_k\}; X_1,\ldots,X_k) = \int_S\cdots\int_S \mathbb{E}[f(\Pi; x_1,\ldots,x_k)]\,\mathrm{d}\mu(x_1)\cdots\mathrm{d}\mu(x_k).$$

## APPENDIX C. TOOLS FROM TOPOLOGY

Recall the following basic definitions.

**Definition C.1.** Let $(X,\mathcal{O})$ be a topological space, where $\mathcal{O}$ is the set of open sets of $X$, and let $A\subseteq X$.

(a) We define the *interior* of $A$ to equal $\mathrm{int}(A) \coloneqq \{x\in X : \exists O\in\mathcal{O}$ such that $x\in O\subseteq A\}$.
(b) We define the *closure* of $A$ to equal $\overline{A} \coloneqq X\setminus\mathrm{int}(X\setminus A)$.
(c) We define the *boundary* of $A$ to equal $\partial A \coloneqq \overline{A}\setminus\mathrm{int}(X\setminus A) = \overline{A}\cap\overline{X\setminus A}$.
(d) We say that $A$ is *disconnected* if there exist two open sets $O_1, O_2$ such that $A = (A\cap O_1)\sqcup(A\cap O_2)$. We say that $A$ is *connected* if it is not disconnected.

**Lemma C.2.** *Let $X$ be a topological space, let $A\subseteq X$, and let $B$ be a connected subset of $\mathbb{R}^k$ such that $B\cap\partial A = \emptyset$. Then either $B\subseteq A$ or $B\subseteq X\setminus A$.*

*Proof.* Since $\partial A = \overline{A}\cap\overline{A^c}$, we have that $\partial A$, $\mathrm{int}(A)$ and $\mathrm{int}(X\setminus A)$ are disjoint sets partitioning $X$. By our assumption that $B\cap\partial A = \emptyset$, we must thus have $B\subseteq\mathrm{int}(A)\sqcup\mathrm{int}(X\setminus A)$. By our assumption that $B$ is connected, we must then have that either $B\subseteq\mathrm{int}(A)\subseteq A$ or that $B\subseteq\mathrm{int}(X\setminus A)\subseteq X\setminus A$. This completes the proof. □

## REFERENCES


[1] R. Arratia. *On the central role of scale invariant Poisson processes on* $(0,\infty)$. Microsurveys in discrete probability (Princeton, NJ, 1997), 21-–41. DIMACS Ser. Discrete Math. Theoret. Comput. Sci., 41. American Mathematical Society, Providence, RI, 1998.

[2] R. Arratia. *On the amount of dependence in the prime factorization of a random integer.* Contemporary Combinatorics, 29–91. Bolyai Soc. Math. Stud., 10. János Bolyai Math. Soc., Budapest, 2002.

[3] R. Arratia, A. D. Barbour and S. Tavaré. *A tale of three couplings: Poisson-Dirichlet and GEM approximations for random permutations.* Combin. Probab. Comput. **15** (2006), no. 1–2, 31–62.

[4] G. Bareikis and E. Manstavičius, *On the DDT Theorem.* Acta Arith. **126** (2007), no. 2, 155–168.

[5] P. Billingsley. *On the distribution of large prime divisors.* Period. Math. Hungar. **2** (1972), 283–289.

[6] R. de la Bretèche and G. Tenenbaum. *Sur les processus arithmétiques liés aux diviseurs.* Adv. in Appl. Probab. **48** (2016), no. A, 63—76.

[7] J.-M. Deshouillers, F. Dress and G. Tenenbaum. *Lois de répartition des diviseurs. I.* Acta Arith. **34** (1979), no. 4, 273–285 (loose errata).

[8] P. Donnelly and S. Tavaré. *The population genealogy of the infinitely-many neutral alleles model.* J. Math. Biol. **25** (1987), no. 4, 381–391.

[9] S. Feng. *The Poisson-Dirichlet distribution and related topics.* Models and asymptotic behaviors. Probab. Appl. (New York). Springer, Heidelberg, 2010.

[10] J. F. C. Kingman. *Poisson processes.* Oxford Stud. Probab., 3. Oxford Sci. Publ. The Clarendon Press, Oxford University Press, New York, 1993.
[11] S. Kotz, N. Balakrishnan and N. L. Johnson. *Continuous multivariate distributions. Vol. 1.* Models and applications. Second edition. Wiley Ser. Probab. Statist. Appl. Probab. Statist. Wiley-Interscience, New York, 2000.
[12] D. Koukoulopoulos. *The distirbution of prime numbers.* Grad. Stud. Math., 203. American Mathematical Society, Providence, RI, 2019.
[13] G. Last and M. Penrose. *Lectures on the Poisson process.* IMS Textb., 7. Cambridge University Press, Cambridge, 2018.
[14] S.-K. Leung. *Dirichlet law for factorisation of integers, polynomials and permutations.* Math. Proc. Cambridge Philos. Soc. **175** (2023), no. 3, 649—676.
[15] S. Nyandwi and A. Smati. *Distribution laws of pairs of divisors.* Integers. **13** (2013). Paper No. A13, 13.
[16] R. P. Stanley. *Enumerative Combinatorics. Volume 1.* Second edition Cambridge Stud. Adv. Math., 49. Cambridge University Press, Cambridge, 2012.
[17] G. Tenenbaum. *A rate estimate in Billingsley's theorem for the size distribution of large prime factors.* Q. J. Math. **51** (2000), no. 3, 385–403.

DÉPARTEMENT DE MATHÉMATIQUES ET DE STATISTIQUE, UNIVERSITÉ DE MONTRÉAL, CP 6128 SUCC. CENTRE-VILLE, MONTRÉAL, QC H3C 3J7, CANADA
*Email address*: tony.haddad@umontreal.ca

DÉPARTEMENT DE MATHÉMATIQUES ET DE STATISTIQUE, UNIVERSITÉ DE MONTRÉAL, CP 6128 SUCC. CENTRE-VILLE, MONTRÉAL, QC H3C 3J7, CANADA
*Email address*: dimitris.koukoulopoulos@umontreal.ca