Canad. Math. Bull. Vol. **00** (0), 2022 pp. 1–17
http://dx.doi.org/10.4153/xxxx


# Benford behavior and distribution in residue classes of large prime factors

Paul Pollack and Akash Singha Roy

*Abstract.* **We investigate the leading digit distribution of the $k$th largest prime factor of $n$ (for each fixed $k = 1, 2, 3, \dots$) as well as the the sum of all prime factors of $n$. In each case, we find that the leading digits are distributed according to Benford's law. Moreover, Benford behavior emerges simultaneously with equidistribution in arithmetic progressions uniformly to small moduli.**

## 1 Introduction

*Benford's law*, named for physicist Frank Benford (though discovered almost 60 years prior by Simon Newcomb), refers to the observation that in many naturally occurring data sets, the leading digits are far from uniformly distributed, with smaller digits more likely to occur. Let us make this precise. By the $N$ leading digits of the positive real number $x$, we mean the $N$ most significant digits. For example (working in base 10), 123.456 has first 4 leading digits 1234, and this is the same for 0.00123456. Now let $D$ and $b$ be integers with $b \geq 2$. We say a positive real number "begins with $D$ in base $b$" if its most significant digits in base $b$ are those of the base $b$ expansion of $D$. Then Benford's law, in base $b$, predicts that the proportion of terms in the data set beginning with $D$ should be approximately $\log(1 + D^{-1})/\log b$. For example, since $\frac{\log 2}{\log 10} = 0.3010\dots$, we expect to see a leading digit 1 in base 10 about 30% of the time.

For general background on Benford's law, see [5, 22]. In this paper, we are interested in data sets arising from positive-valued arithmetic functions. Let $f\colon \mathbb{N} \to \mathbb{R}_{>0}$. We say *$f$ obeys Benford's law in base $b$* (or that *$f$ is Benford in base $b$*) if, for each positive integer $D$, the asymptotic density of $n$ for which $f(n)$ begins with $D$ in base $b$ is $\log(1+D^{-1})/\log b$. Results on the "Benfordity" of particular arithmetic functions are scattered throughout the literature. For example, $f(n) = n!$ is Benford in every base $b$ (Diaconis [11]), as is the "primorial" $f(n) = \prod_{k=1}^{n} p_k$ (Massé–Schneider [21]). The classical partition function $p(n)$ is also Benford in every base (see Anderson–Rolen–Stoehr [2] or [21]). On the other hand, $f(n) = n$ is not Benford; the asymptotic density in question does not exist. This same obstruction to Benford's law persists if $f(n)$ is any positive-valued polynomial function of $n$. (See, for instance, the final section of [21]. It should be noted that these examples obey Benford's law in a weaker sense; namely Benford's law holds if asymptotic density is replaced with logarithmic density.)

When $f$ is multiplicative, whether or not $f$ is Benford in base $b$ can be interpreted as a problem in the theory of mean values of multiplicative functions. Namely, $f$ is

2020 Mathematics Subject Classification: Primary 11A63; Secondary 11N37, 11N64.
Keywords: Benford's law, smooth numbers, anatomy of integers.

Benford precisely when $f(n)^{2\pi i\ell/\log b}$ has mean value zero for each nonzero integer $\ell$. This criterion was noted by Aursukaree and Chandee [3] and used by them to show that the divisor function $d(n)$ is Benford in base 10. A more systematic study of the Benford behavior of multiplicative function, leveraging Halász's celebrated mean value theorem, was recently undertaken in [8]. For example, it is shown there that $\phi(n)$ is not Benford but that $|\tau(n)|$ is, where $\tau$ is Ramanujan's $\tau$-function.[1] All of the work in [8] is carried out in base 10, but both of the quoted results hold, by simple modifications of the proofs, in each fixed base $b \geq 2$.

Our concern in the present paper is with certain nonmultiplicative functions. Roughly speaking, we show that (for each fixed $k$) the $k$th largest prime factor of $n$ obeys Benford's law, as does the sum of all of the prime factors of $n$. (Both results hold for each base $b$.) In fact, our results are somewhat stronger than this.

We let $P_k(n)$ denote the $k$th largest prime factor of $n$; when $k = 1$, we write $P(n)$ in place of the more cumbersome $P_1(n)$. More precisely, if $n = p_1p_2p_3\cdots p_{\Omega(n)}$, with $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_{\Omega(n)}$, we set $P_k(n) = p_k$, with the convention that $P_k(n) = 0$ if $k > \Omega(n)$. Put

$$\Psi_k(x, y) := \#\{n \leq x : P_k(n) \leq y\}.$$

(When $k = 1$, it is usual to write $\Psi(x, y)$ in place of $\Psi_1(x, y)$.) Let $a \bmod q$ be a coprime residue class. For real $x, y \geq 2$, define

$$\Psi_k(x, y; b, D, q, a) := \#\{n \leq x : P_k(n) \leq y,\ P_k(n) \equiv a \pmod q,\ P_k(n) \text{ begins with } D \text{ in base } b\}.$$

***Theorem* 1.1** *Fix positive integers $k$, $b$, and $D$, with $b \geq 2$. Fix real numbers $U \geq 1$ and $\epsilon > 0$. Then*

$$\Psi_k(x, y; b, D, q, a) \sim \frac{1}{\phi(q)} \frac{\log(1 + D^{-1})}{\log b} \Psi_k(x, y),$$

*as $x, y \to \infty$, uniformly for $y \geq x^{1/U}$ and coprime residue classes $a \bmod q$ with $q \leq \frac{\log x}{(\log\log x)^{k-1+\epsilon}}$. In fact, if $k = 1$, we can take $q \leq (\log x)^A$ for any fixed $A$.*

To deduce that $P_k(n)$ is Benford, it suffices to take $q = 1$ and $y = x$. The additional generality of Theorem 1.1 seems of some interest. For example, Theorem 1.1 contains the result of Banks–Harman–Shparlinski [4] that $P(n)$, on integers $n \leq x$, is uniformly distributed in coprime residue classes mod $q$, for $q$ up to an arbitrary fixed power of $\log x$. Theorem 1.1 gives the corresponding result for $P_k(n)$, when $k > 1$, in the more restricted range $q \leq \log x/(\log\log x)^{k-1+\epsilon}$. This appears to be new; moreover, this range of $q$ is sharp up to the power of $\log\log x$, since $\gg x(\log\log x)^{k-2}/\log x$ values of $n \leq x$ have $P_k(n) = 2$.

Turning to the sum of the prime factors, we let $A(n) = \sum_{p^k \| n} kp$. That is, $A(n)$ is the sum of the prime factors of $n$, counting multiplicity. (The sum of the distinct prime factors of $n$ could be handled by similar arguments.) The function $A(n)$ was introduced by Alladi and first investigated by Alladi and Erdős [1].

[1] In this latter result, the notion of "asymptotic density" in the definition of a Benford function should be replaced with "asymptotic density relative to the set of $n$ with $\tau(n) \neq 0$".

Define

$$N(x,y;b,D,q,a) := \#\{n \le x : P(n) \le y, A(n) \equiv a \pmod q, \\ A(n) \text{ begins with } D \text{ in base } b\}.$$

***Theorem* 1.2** *Fix an integer $b \ge 2$, and a positive integer $D$. Fix real numbers $U \ge 1$ and $\epsilon > 0$. Then*

$$N(x,y;b,D,q,a) \sim \frac{1}{q}\frac{\log(1+D^{-1})}{\log b}\Psi(x,y),$$

*as $x, y \to \infty$, uniformly for $y \ge x^{1/U}$ and residue classes $a \bmod q$ with $q \le (\log x)^{\frac{1}{2}-\epsilon}$.*

As before, taking $y = x$ and $q = 1$ shows that $A(n)$ satisfies Benford's law. Again, the extra generality here seems interesting. For example, it is implicit in Theorem 1.2 that $A(n)$ is equidistributed mod $q$, uniformly for $q \le (\log x)^{\frac{1}{2}-\epsilon}$, a result which we have not seen explicitly stated in the literature before. (See Goldfeld [12] for the case of fixed $q$.) The same range of uniformity may follow from the method of Hall in [15] (who considered the distribution mod $q$ of $\sum_{p|n,\ p\nmid q} p$), but our proof exhibits the result as a simple consequence of quantitative mean value theorems.

In addition to the already-mentioned references, the reader interested in number-theoretic investigations of Benford's law might also consult [20], [6], [9], [18], [7], and [24].

### Notation

Most of our notation is standard. Of note, we allow constants in $O$-symbols to depend on any parameter that has been declared as 'fixed'. When we refer to 'large' $x$, the threshold for large enough may also depend on these parameters. We write $A \gtrsim B$ as an abbreviation for $A \ge (1+o(1))B$.

## 2 Benford's law for $P_k(n)$: Proof of Theorem 1.1

We make crucial use of both the results and methods of Knuth and Trabb Pardo [19], who were the first to seriously investigate $P_k(n)$ when $k > 1$. We define functions $\rho_k(\alpha)$, for integers $k \ge 0$ and real $\alpha$, as follows:

$$\rho_k(\alpha) = 0 \quad \text{if} \quad \alpha \le 0 \text{ or } k = 0,$$

$$\rho_k(\alpha) = 1 \quad \text{for} \quad 0 < \alpha \le 1 \text{ and } k \ge 1,$$

$$\rho_k(\alpha) = 1 - \int_1^\alpha (\rho_k(t-1) - \rho_{k-1}(t-1))\frac{\mathrm{d}t}{t}, \quad \text{for} \quad \alpha > 1 \text{ and } k \ge 1. \tag{2.1}$$

Much is known about the asymptotic behavior of $\rho_k(\alpha)$ as $\alpha \to \infty$; for $k = 1$, see for instance [10], while for $k \ge 2$, see equations (6.4) and (6.15) in [19]. For our purposes, much weaker information suffices. We assume as known that each $\rho_k$ ($k = 1, 2, 3, \dots$) is positive-valued and weakly decreasing on $(0,\infty)$, and that $\lim_{\alpha\to\infty}\rho_k(\alpha) = 0$.

The following result, which connects the $\rho_k$ with the distribution of $P_k(n)$, appears as eq. (4.7) in [19] (and is a consequence of the stronger assertion (4.8) shown there).

***Proposition* 2.1** *Fix a positive integer $k$ and a real number $U \geq 1$. For all $x, y \geq 2$,*

$$\Psi_k(x,y) = \rho_k(u)x + O(x/\log x), \tag{2.2}$$

*uniformly for $y \geq x^{1/U}$, where $u := \frac{\log x}{\log y}$. In particular, $\Psi_k(x,y) \sim \rho_k(u)x$ as $x \to \infty$, uniformly for $y \geq x^{1/U}$.*

(In [19], it is assumed that the ratio $\frac{\log x}{\log y}$ is fixed, rather than merely bounded. However, the proof given actually establishes (2.2) in the full range of Proposition 2.1.)

The next result is a variant of Theorem 1.1 where we require that $P_k(n)$ be bounded below by a fixed power of $x$.

***Proposition* 2.2** *Fix positive integers $k$, $b$, and $D$ with $b \geq 2$. Fix real numbers $A \geq 1$, $U \geq 1$, and fix a real number $U' > U$. The number of $n \leq x$ for which $P_k(n) \equiv a \pmod q$, $P_k(n)$ begins with the digits of $D$ in base $b$, and $P_k(n) \in (x^{1/U'}, y]$, is*

$$\frac{1}{\phi(q)} \frac{\log(1+D^{-1})}{\log b} (\rho_k(u) - \rho_k(U'))x + o(x/\phi(q)),$$

*where $u := \frac{\log x}{\log y}$, where $x, y \to \infty$ with $y \geq x^{1/U}$, and where $a \bmod q$ is a coprime residue class with $q \leq (\log x)^A$.*

The proof of Proposition 2.2 requires two classical results from the theory of primes in arithmetic progressions. Let $\pi(x;q,a)$ denote the count of primes $p \leq x$ with $p \equiv a \pmod q$.

***Proposition* 2.3 (*Brun–Titchmarsh*)** *If $a$ and $q$ are coprime integers with $0 < 2q \leq x$, then*

$$\pi(x;q,a) \ll \frac{1}{\phi(q)} \frac{x}{\log(x/q)}.$$

*Here the implied constant is absolute.*

***Proposition* 2.4 (*Siegel–Walfisz*)** *Fix a real number $A > 0$. If $a, q$ are coprime integers with $1 \leq q \leq (\log x)^A$, and $x \geq 3$, then*

$$\pi(x;q,a) = \frac{1}{\phi(q)} \int_2^x \frac{1}{\log t}\, \mathrm{d}t + O_A(x \exp(-C\sqrt{\log x})).$$

*Here $C$ is a certain absolute constant.*

For proofs of these results, see [23, Theorem 3.9, p. 90] and [23, Corollary 11.21, p. 382].

**Proof** First note that we can (and will) always assume that $y \leq x$, since the cases when $y > x$ are covered by the case $y = x$.

By a standard compactness argument, when proving Proposition 2.2 we may assume that $u = \frac{\log x}{\log y}$ is fixed. To see this, suppose Proposition 2.2 holds when $u$ is fixed but

does not hold in general. Then for some $\epsilon > 0$, there are choices of $x, y, a$ and $q$ with $x$ arbitrarily large, $x \geq y \geq x^{1/U}$, and $q \leq (\log x)^A$ for which our count exceeds

$$\frac{1}{\phi(q)} \frac{\log(1+D^{-1})}{\log b} (\rho_k(u) - \rho_k(U') + \epsilon)x, \tag{2.3}$$

or there are such choices of $x, y, a$ and $q$ for which our count falls below

$$\frac{1}{\phi(q)} \frac{\log(1+D^{-1})}{\log b} (\rho_k(u) - \rho_k(U') - \epsilon)x.$$

We will assume we are in the former case; the latter can be handled analogously. By compactness, we may choose $x, y, a, q$ so that $u \to u_0$, for some $u_0 \in [1, U]$.

We first rule out $u_0 = 1$. As $y \leq x$, the condition $P_k(n) \leq y$ is always at least as strict as the condition $P_k(n) \leq x$ (which holds vacuously, as we are counting numbers $n \leq x$). Moreover, the $u = 1$ case of Proposition 2.2 is true by hypothesis. Putting these observations together, we see that the count of $n$ corresponding to $x, y, a, q$ is at most

$$\frac{1}{\phi(q)} \frac{\log(1+D^{-1})}{\log b} (\rho_k(1) - \rho_k(U') + o(1))x.$$

But if $u \to 1$, then $\rho_k(u) \to \rho_k(1)$, and this estimate is eventually incompatible with (2.3).

Thus, it must be that $u_0 > 1$. Here we may obtain a contradiction by a slightly tweaked argument. For any fixed $\delta > 0$, we eventually have $u > u_0 - \delta$. So the condition $P_k(n) \leq y$ is eventually stricter than the condition $P_k(n) \leq x^{1/(u_0-\delta)}$. If $\delta$ is fixed sufficiently small (in terms of $\epsilon$), the $u = u_0 - \delta$ case of Proposition 2.2 gives an estimate contradicting (2.3).

We thus turn to proving the modified statement with the extra condition that $u$ is fixed.

For each nonnegative integer $j$, let $\mathcal{I}_j$ denote the interval

$$\mathcal{I}_j := [u_j, v_j), \quad \text{where} \quad u_j := Db^j, \ v_j := (D+1)b^j. \tag{2.4}$$

Then our count of $n$ is given by

$$\sum_{j \geq 0} \sum_{\substack{p \in \mathcal{I}_j \cap (x^{1/U'}, y] \\ p \equiv a \pmod q}} \sum_{\substack{n \leq x \\ P_k(n) = p}} 1. \tag{2.5}$$

Let $\mathcal{J}$ be the collection of nonnegative integers $j$ with $\mathcal{I}_j \subset (x^{1/U'}, y/\exp(\sqrt{\log x}))$. Then at the cost of another error of size $o(x/\phi(q))$, we can restrict the triple sum in (2.5) to $j \in \mathcal{J}$. Indeed, the $n$ counted by the triple sum above that are excluded by this restriction have either a prime divisor in $P := (x^{1/U'}, bx^{1/U'}]$ or in $P' := [y/b \exp(\sqrt{\log x}), y]$, and the number of such $n \leq x$ is at most

$$x \sum_{\substack{p \in P \cup P' \\ p \equiv a \pmod q}} 1/p = o(x/\phi(q)),$$

by partial summation and the Brun–Titchmarsh theorem (Proposition 2.3). We proceed to estimate, for each $j \in \mathcal{J}$, the corresponding inner sums in (2.5) over $p$ and $n$.

If $p$ is prime and $P_k(n) = p$, then $n = mp$ where $m \le x/p$, $P_k(m) \le p$, and $P_{k-1}(m) \ge p$. The converse also holds. Thus, if $j \in \mathcal{J}$ and $p \in \mathcal{I}_j$,

$$\sum_{\substack{n \le x \\ P_k(n)=p}} 1 = \Psi_k(x/p, p) - \Psi_{k-1}(x/p, p-\epsilon)$$

for (say) $\epsilon = \frac{1}{2}$. Hence,

$$\sum_{\substack{p \in \mathcal{I}_j \\ p \equiv a \pmod q}} \sum_{\substack{n \le x \\ P_k(n)=p}} 1 = \sum_{\substack{p \in \mathcal{I}_j \\ p \equiv a \pmod q}} \Psi_k(x/p, p) - \sum_{\substack{p \in \mathcal{I}_j \\ p \equiv a \pmod q}} \Psi_{k-1}(x/p, p-\epsilon).$$

To continue, observe that for $j \in \mathcal{J}$,

$$\begin{aligned}
&\sum_{\substack{p \in \mathcal{I}_j \\ p \equiv a \pmod q}} \Psi_k(x/p, p) - \frac{1}{\phi(q)} \int_{\mathcal{I}_j} \Psi_k(x/t, t) \frac{\mathrm{d}t}{\log t} \\
&\quad = \sum_{\substack{u_j \le p < v_j \\ p \equiv a \pmod q}} \sum_{\substack{n \le x/p \\ P_k(n) \le p}} 1 - \frac{1}{\phi(q)} \int_{u_j}^{v_j} \sum_{\substack{n \le x/t \\ P_k(n) \le t}} \frac{1}{\log t} \, \mathrm{d}t \\
&\quad = \sum_{n \le x/u_j} \left( \sum_{\substack{m < p \le M \\ p \equiv a \pmod q}} 1 - \frac{1}{\phi(q)} \int_m^M \frac{\mathrm{d}t}{\log t} + O(1) \right),
\end{aligned}$$

where $m$ and $M$ are defined by

$$m := \max\{u_j, P_k(n)\}, \quad M := \min\{x/n, v_j\},$$

and where the last displayed sum on $n$ is understood to be extended only over those $n \le x/u_j$ for which $m \le M$. By the Siegel–Walfisz theorem (Proposition 2.4),

$$\sum_{\substack{m < p \le M \\ p \equiv a \pmod q}} 1 - \frac{1}{\phi(q)} \int_m^M \frac{\mathrm{d}t}{\log t} \ll M \exp(-C\sqrt{\log M}) \ll \frac{x}{n} \exp(-C'\sqrt{\log x}).$$

where $C$ is an absolute positive constant and $C' = C/\sqrt{U'}$. (This use of the Siegel–Walfisz theorem explains the restriction $q \le (\log x)^A$ in the statement of Proposition 2.2.) Putting this back in above and summing on $n$, we find that (for large $x$)

$$\sum_{\substack{p \in \mathcal{I}_j \\ p \equiv a \pmod q}} \Psi_k(x/p, p) - \frac{1}{\phi(q)} \int_{\mathcal{I}_j} \Psi_k(x/t, t) \frac{\mathrm{d}t}{\log t} \ll x \log x \cdot \exp(-C'\sqrt{\log x}) + \frac{x}{u_j}. \tag{2.6}$$

A nearly identical calculation gives the same bound for the difference

$$\sum_{\substack{p \in \mathcal{I}_j \\ p \equiv a \pmod q}} \Psi_{k-1}(x/p, p-\epsilon) - \frac{1}{\phi(q)} \int_{\mathcal{I}_j} \Psi_{k-1}(x/t, t) \frac{\mathrm{d}t}{\log t}.$$

Since $u_{j+1}/u_j \geq 2$ and the smallest $j \in \mathcal{J}$ has $u_j \geq x^{1/U'}$, the expression on the right-hand side of (2.6), when summed on $j \in \mathcal{J}$, is $\ll x(\log x)^2 \exp(-C'\sqrt{\log x}) + x^{1-1/U'}$, and this is certainly $o(x/\phi(q))$. As a consequence, instead of our original triple sum (2.5), it is enough to estimate

$$\frac{x}{\phi(q)} \sum_{j\in\mathcal{J}} \frac{1}{x} \int_{\mathcal{I}_j} (\Psi_k(x/t,t) - \Psi_{k-1}(x/t,t)) \frac{\mathrm{d}t}{\log t}. \tag{2.7}$$

We now apply Proposition 2.1, noting that for each $t \in \mathcal{I}_j$, we have $\frac{\log(x/t)}{\log t} = \frac{\log x}{\log t} - 1 \leq U' - 1$ as well as $\log(x/t) \geq \log(y/t) \geq \sqrt{\log x}$. We find that

$$\begin{aligned}\frac{1}{x} \int_{\mathcal{I}_j} &(\Psi_k(x/t,t) - \Psi_{k-1}(x/t,t)) \frac{\mathrm{d}t}{\log t} \\ &= \int_{\mathcal{I}_j} \frac{1}{t} \left(\rho_k\left(\frac{\log x}{\log t} - 1\right) - \rho_{k-1}\left(\frac{\log x}{\log t} - 1\right)\right) \frac{\mathrm{d}t}{\log t} + O\left(\int_{\mathcal{I}_j} \frac{1}{t\sqrt{\log x}} \frac{\mathrm{d}t}{\log t}\right).\end{aligned}$$

The error term, when summed on $j \in \mathcal{J}$, is $\ll \frac{1}{\sqrt{\log x}} \int_2^x \frac{\mathrm{d}t}{t\log t} \ll \log\log x/\sqrt{\log x}$, and so is $o(1)$; inserted back into (2.7), we see this gives rise to a final error of size $o(x/\phi(q))$ in our count, which is acceptable. To deal with the remaining integrals, we write $u_j = x^{\mu_j}$ and $v_j = x^{\nu_j}$ and make the change of variables $\alpha = \frac{\log x}{\log t}$. Then $\mathrm{d}\alpha = -\frac{\log x}{t(\log t)^2}\,\mathrm{d}t$, so that $\frac{\mathrm{d}t}{t\log t} = -\frac{\mathrm{d}\alpha}{\alpha}$ and

$$\begin{aligned}\sum_{j\in\mathcal{J}} \int_{\mathcal{I}_j} \frac{1}{t} &\left(\rho_k\left(\frac{\log x}{\log t} - 1\right) - \rho_{k-1}\left(\frac{\log x}{\log t} - 1\right)\right) \frac{\mathrm{d}t}{\log t} \\ &= \sum_{j\in\mathcal{J}} \int_{1/\mu_j}^{1/\nu_j} -\frac{\rho_k(\alpha-1) - \rho_{k-1}(\alpha-1)}{\alpha}\,\mathrm{d}\alpha.\end{aligned}$$

From (2.1), $-\frac{\rho_k(\alpha-1)-\rho_{k-1}(\alpha-1)}{\alpha} = \rho_k'(\alpha)$, so that this last sum on $j$ simplifies to $\sum_{j\in\mathcal{J}}(\rho_k(1/\nu_j) - \rho_k(1/\mu_j))$. Now, following [19], we introduce the notation $F_k(\beta) = \rho_k(1/\beta)$. By the mean value theorem,

$$\begin{aligned}\rho_k(1/\nu_j) - \rho_k(1/\mu_j) &= F_k(\nu_j) - F_k(\mu_j) \\ &= (\nu_j - \mu_j)F_k'(t_j) = \frac{\log(1+D^{-1})}{\log x} F_k'(t_j)\end{aligned}$$

for some $t_j \in (\mu_j, \nu_j)$. Thus,

$$\begin{aligned}\sum_{j\in\mathcal{J}} (\rho_k(1/\nu_j) - \rho_k(1/\mu_j)) &= \frac{\log(1+D^{-1})}{\log b} \sum_{j\in\mathcal{J}} F_k'(t_j) \cdot \frac{\log b}{\log x} \\ &= \frac{\log(1+D^{-1})}{\log b} \sum_{j\in\mathcal{J}} F_k'(t_j) \cdot (\mu_{j+1} - \mu_j).\end{aligned}$$

Since each $t_j \in (\mu_j, \nu_j) \subset (\mu_j, \mu_{j+1})$, the final sum on $j$ is essentially a Riemann sum. To make this precise, let $j_0 = \min \mathcal{J}$ and $j_1 = \max \mathcal{J}$. Then

$$F_k'(1/U')\left(\mu_{j_0} - \frac{1}{U'}\right) + \sum_{j\in\mathcal{J}} F_k'(t_j)(\mu_{j+1} - \mu_j) + F_k'(1/u)\left(\frac{1}{u} - \mu_{j_1+1}\right)$$

is a genuine Riemann sum for $\int_{1/U'}^{1/u} F_k'(t)\,\mathrm{d}t$, whose mesh size goes to 0 as $x \to \infty$. But the terms we have added to the sum on $j \in \mathcal{J}$ contribute $o(1)$, as $x \to \infty$. It follows that $\sum_{j\in\mathcal{J}} F_k'(t_j)(\mu_{j+1} - \mu_j) \to \int_{1/U'}^{1/u} F_k'(t)\,\mathrm{d}t = F_k(1/u) - F_k(1/U') = \rho_k(u) - \rho_k(U')$. Collecting estimates completes the proof of the proposition in the case when $u$ is fixed. ∎

To deduce Theorem 1.1, it remains to handle the contribution from $n$ with $P_k(n) \le x^{1/U'}$.

The following lemma bounds the number of integers with a large smooth divisor. A proof is sketched in Exercise 293 on p. 554 of [26], with a solution in [25, pp. 305–306]. By the *y-smooth part* of a number $n$, we mean $\prod_{\substack{p^e \| n \\ p \le y}} p^e$.

***Lemma* 2.5** *For all $x \ge z \ge y \ge 2$, the number of $n \le x$ whose $y$-smooth part exceeds $z$ is $O\left(x \exp\left(-\frac{1}{2}\frac{\log z}{\log y}\right)\right)$.*

***Lemma* 2.6** *Fix a positive integer $k$ and a real number $B \ge 1$.*

- *When $k = 1$, the number of $n \le x$ with $P_k(n) \le y$ and $P_k(n) \equiv a \pmod q$ is*

$$\ll \frac{x}{\phi(q)} \exp\left(-\frac{1}{8}u\right) + x\left(\frac{\log(3q)}{\log x}\right)^B \cdot \exp\left(-\frac{1}{8}u\right),$$

  *uniformly for $x \ge y \ge 3$ with $y \le x^{1/4}$, and $a \bmod q$ any coprime residue class with $q \le x^{1/8}$. As usual, $u = \frac{\log x}{\log y}$.*
- *When $k \ge 2$, the number of $n \le x$ with $P_k(n) \le y$ and $P_k(n) \equiv a \pmod q$ is*

$$\ll \frac{x}{\log x}(\log\log x)^{k-2}\log(3q) + \frac{x}{\phi(q)}\frac{(\log u)^{k-2}}{u},$$

  *uniformly in the same range of $x$, $y$, and $q$.*

**Proof** We will restrict attention to $n > x^{3/4}$; this is permissible, since $x^{3/4}$ is dwarfed by either of our target upper bounds. We let $p = P_k(n)$ and write $n = p_1 \cdots p_{k-1} ps$, where $p_1 \ge p_2 \ge \cdots \ge p_{k-1} \ge p$ and $P(s) \le p$.

We first show we can assume $s \le x^{1/2}$. Indeed, suppose $s > x^{1/2}$. Then, with $m = n/p$, we have that $m \le x/p$ and that the $p$-smooth part of $m$ exceeds $x^{1/2}$. Applying Lemma

2.5, we see that for every $p \le y$, the number of corresponding $m$ is

$$\begin{aligned} \ll \frac{x}{p}\exp(-\frac{1}{4}\frac{\log x}{\log p}) &\ll \frac{x}{p}\exp(-\frac{1}{8}\frac{\log x}{\log p})\cdot\exp(-\frac{1}{8}\frac{\log x}{\log p}) \\ &\ll \frac{x}{(\log x)^B}\frac{(\log p)^B}{p}\exp(-\frac{1}{8}\frac{\log x}{\log p}) \\ &\ll \frac{x}{(\log x)^B}\frac{(\log p)^B}{p}\exp(-\frac{1}{8}u). \end{aligned}$$

Now we sum on $p \le y$ with $p \equiv a \pmod q$. We split the sum at $3q^2$, using Mertens' theorem to bound the first half and the Brun–Titchmarsh theorem (with partial summation) for the second; this gives

$$\begin{aligned} \sum_{\substack{p\le y\\ p\equiv a \pmod q}} \frac{(\log p)^B}{p} &\le \sum_{p\le 3q^2}\frac{(\log p)^B}{p} + \sum_{\substack{3q^2<p\le y\\ p\equiv a \pmod q}}\frac{(\log p)^B}{p} \\ &\ll (\log(3q))^{B-1}\sum_{p\le 3q^2}\frac{\log p}{p} + \frac{1}{\phi(q)}(\log y)^B \\ &\ll (\log(3q))^B + \frac{(\log y)^B}{\phi(q)}. \end{aligned}$$

Substituting this estimate into the previous display, we conclude that the $n$ with $s > x^{1/2}$ contribute

$$\begin{aligned} &\ll \frac{x}{u^B\phi(q)}\exp(-\frac{1}{8}u) + x\left(\frac{\log(3q)}{\log x}\right)^B\cdot\exp(-\frac{1}{8}u) \\ &\ll \frac{x}{\phi(q)}\exp(-\frac{1}{8}u) + x\left(\frac{\log(3q)}{\log x}\right)^B\cdot\exp(-\frac{1}{8}u). \end{aligned} \tag{2.8}$$

This is already enough to settle the $k = 1$ case of Lemma 2.6. Indeed, in that case $n = ps$, where $p = P(n)$, and $s = n/P(n) \ge n/y > x^{3/4}/y \ge x^{1/2}$.

Now suppose that $k \ge 2$ and that $s \le x^{1/2}$. Then

$$p_1^k \ge p_1\cdots p_{k-1}p = n/s > x^{3/4}/x^{1/2} = x^{1/4},$$

so that $p_1 \ge x^{1/4k}$. Hence, given $p_2, \dots, p_{k-1}$, $p$, and $s$, the number of possibilities for $p_1$ (and thus also for $n$) is $\ll \pi(x/p_2\cdots p_{k-1}ps) \ll x/p_2\cdots p_{k-1}ps\log x$. Observe that $s$ is $p$-smooth, while each $p_i \in [p, x]$. We have that $\sum_{s\ p\text{-smooth}} 1/s = \prod_{\text{prime } \ell\le p}(1-1/\ell)^{-1} \ll \log p$. Also (when $p \le y$), $\sum_{p\le p_i\le x} 1/p_i \ll \log\frac{\log x}{\log p}$. Hence, the number of possibilities for $n$ given $p$ is

$$\ll \frac{x}{\log x}\left(\log\frac{\log x}{\log p}\right)^{k-2}\frac{\log p}{p}.$$

We now sum on $p \le y$ with $p \equiv a \pmod q$. Estimating crudely, we see that the $p \le 3q^2$ contribute

$$\ll \frac{x}{\log x}(\log\log x)^{k-2}\log(3q).$$

To handle the remaining contribution in the case when $y > 3q^2$, we apply partial summation; by Brun–Titchmarsh,

$$\sum_{\substack{3q^2<p\le y\\ p\equiv a \pmod q}} \left(\log\frac{\log x}{\log p}\right)^{k-2}\frac{\log p}{p} \ll \frac{1}{\phi(q)}(\log u)^{k-2} - \int_{3q^2}^{y} \pi(t;q,a)\,\mathrm{d}\left(\left(\log\frac{\log x}{\log t}\right)^{k-2}\frac{\log t}{t}\right).$$

Since $\left(\log\frac{\log x}{\log t}\right)^{k-2}\frac{\log t}{t}$ is a decreasing function of $t$ on $[3q^2, y]$, the bound $\pi(t;q,a) \ll t/\phi(q)\log t$ implies that

$$-\int_{3q^2}^{y} \pi(t;q,a)\,\mathrm{d}\left(\left(\log\frac{\log x}{\log t}\right)^{k-2}\frac{\log t}{t}\right) \ll -\frac{1}{\phi(q)}\int_{3q^2}^{y}\frac{t}{\log t}\,\mathrm{d}\left(\left(\log\frac{\log x}{\log t}\right)^{k-2}\frac{\log t}{t}\right).$$

Integrating by parts again,

$$\begin{aligned}\int_{3q^2}^{y}\frac{t}{\log t}\,\mathrm{d}\left(\left(\log\frac{\log x}{\log t}\right)^{k-2}\frac{\log t}{t}\right) &= -\int_{3q^2}^{y}\left(\log\frac{\log x}{\log t}\right)^{k-2}\frac{\log t}{t}\,\mathrm{d}\left(\frac{t}{\log t}\right) + O((\log\log x)^{k-2})\\ &\ll \int_{3q^2}^{y}\left(\log\frac{\log x}{\log t}\right)^{k-2}\frac{\mathrm{d}t}{t} + O((\log\log x)^{k-2}).\end{aligned}$$

Making the change of variables $\alpha = \frac{\log t}{\log x}$,

$$\int_{3q^2}^{y}\left(\log\frac{\log x}{\log t}\right)^{k-2}\frac{\mathrm{d}t}{t} \le \log x\int_0^{1/u}(\log(1/\alpha))^{k-2}\,\mathrm{d}\alpha \ll \log x\cdot\frac{1}{u}(\log u)^{k-2}.$$

(In the last step, we use that $\int_0^z(\log(1/\alpha))^{k-2}\,\mathrm{d}\alpha$ has the form $z\cdot Q(\log(1/z))$, where $Q$ is a monic polynomial with degree $k-2$.) Collecting estimates, we conclude that when $k \ge 2$, the $n$ with $s \le x^{1/2}$ make a contribution

$$\ll \frac{x}{\log x}(\log\log x)^{k-2}\log(3q) + \frac{x}{\phi(q)}\frac{(\log u)^{k-2}}{u}.$$

Since this upper bound dominates the contribution (2.8) from $n$ with $s > x^{1/2}$, the $k \ge 2$ cases of Lemma 2.6 follow. ∎

**Proof** Fix $\eta > 0$. We will show that the count of $n$ in question is eventually[2] larger than $\frac{1}{\phi(q)} \frac{\log(1+D^{-1})}{\log b} (\rho_k(u) - \eta) x$ and eventually smaller than $\frac{1}{\phi(q)} \frac{\log(1+D^{-1})}{\log b} (\rho_k(u) + \eta) x$, and hence is $\sim \frac{1}{\phi(q)} \frac{\log(1+D^{-1})}{\log b} \rho_k(u) x$. Since $\Psi_k(x, y) \sim \rho_k(u) x$, Theorem 1.1 then follows.

The required lower bound is immediate from Proposition 2.2: It suffices to apply that Proposition with $U'$ fixed large enough that $\rho_k(U') < \eta$.

We turn now to the upper bound. Apply Lemma 2.6, taking $B = A + 1$ in the case $k = 1$. That lemma implies the existence of a constant $C$, depending only on $k$ (and on $A$, if $k = 1$) such that the following holds: For any fixed $U' \geq 4$, the number of $n \leq x$ with $P_k(n) \equiv a \pmod q$ and $P_k(n) \leq x^{1/U'}$ is eventually at most $C \frac{x}{\phi(q)} \frac{(\log U')^{k-2}}{U'}$. If we choose $U' > U$ so large that $C \frac{(\log U')^{k-2}}{U'} < \eta \frac{\log(1+D^{-1})}{\log b}$, the desired upper bound then follows from Proposition 2.2. ∎

## 3 Benford's law for the sum of the prime factors: Proof of Theorem 1.2

For multiplicative functions $F, G$ taking values on or inside the complex unit circle, we define (following Granville and Soundararajan [13]) the *distance between $F$ and $G$, up to $x$*, by

$$\mathbb{D}(F, G; x) = \sqrt{\sum_{p \leq x} \frac{1 - \mathrm{Re}(F(p)\overline{G(p)})}{p}}.$$

The following statement (Corollary 4.12 on p. 494 of [26]), due to Montgomery and Tenenbaum, makes quantitatively precise a result of Halász [14] that $F$ has mean value 0 unless $F$ "pretends" to be $n^{it}$ for some $t$.

***Proposition* 3.1** *Let $F$ be a multiplicative function with $|F(n)| \leq 1$ for all $n$. For $x \geq 2$ and $T \geq 2$, let*

$$m(x, T) = \min_{|t| \leq T} \mathbb{D}(F, n^{it}; x)^2.$$

*Then*

$$\sum_{n \leq x} F(n) \ll x \frac{1 + m(x, T)}{\mathrm{e}^{m(x,T)}} + \frac{x}{T}.$$

*Here the implied constant is absolute.*

When $F$ is real-valued, the following (slightly weakened version of a) theorem of Hall and Tenenbaum [16] allows us to consider only $\mathbb{D}(F, 1; x)$.

[2] Here and later in this proof, "eventually" refers to the limit as taken in Theorem 1.1. That is, a statement holds eventually if there is a real number $T$ such that the statement is true whenever $x, y \geq T$, with $y \geq x^{1/U}$, and with $a \bmod q$ a coprime residue class modulo $q \leq \frac{\log x}{(\log\log x)^{k-1+\epsilon}}$ or, when $k = 1$, modulo $q \leq (\log x)^A$.

***Proposition* 3.2** *Let $F$ be a real-valued multiplicative function with $|F(n)| \le 1$ for all $n$. Then*

$$\sum_{n\le x} F(n) \ll x \exp(-0.3 \cdot \mathbb{D}(F,1;x)^2).$$

***Lemma* 3.3** *Fix $\delta > 0$ and fix $U \ge 1$. For all large $x$, the number of $n \le x$ with $P(n) \le y$ and $A(n) \equiv a \pmod q$ is*

$$\frac{\Psi(x,y)}{q} + O(x/(\log x)^{\frac12-\delta}),$$

*for all $x \ge y \ge x^{1/U}$ and residue classes $a \bmod q$ with $q \le \log x$.*

**Proof** By the orthogonality relations for additive characters,

$$\begin{aligned}\sum_{\substack{n\le x\\ P(n)\le y\\ A(n)\equiv a \pmod q}} 1 &= \sum_{n\le x} \mathbb{1}_{P(n)\le y}\mathbb{1}_{A(n)\equiv a \pmod q}\\ &= \sum_{n\le x}\mathbb{1}_{P(n)\le y}\left(\frac1q \sum_{r \bmod q} e^{2\pi i r(A(n)-a)/q}\right)\\ &= \frac{\Psi(x,y)}{q} + \frac1q \sum_{\substack{r \bmod q\\ r\not\equiv 0 \pmod q}} e^{-2\pi i a r/q}\sum_{n\le x}\mathbb{1}_{P(n)\le y}e^{2\pi i r A(n)/q}.\end{aligned}$$

Hence, it suffices to show that

$$\sum_{n\le x}\mathbb{1}_{P(n)\le y}e^{2\pi i r A(n)/q} = O(x/(\log x)^{1/2-\delta}) \tag{3.1}$$

for each nonzero residue class $r \bmod q$.

Write $r/q = r'/q'$ in lowest terms, so that $q' > 1$. If $q' = 2$, then $r' = 1$, and $F(n) := \mathbb{1}_{P(n)\le y}e^{2\pi i r A(n)/q} = \mathbb{1}_{P(n)\le y}(-1)^{A(n)}$ is a real-valued multiplicative function of modulus at most 1. Moreover, $\mathbb{D}(F,1;x)^2 \ge \sum_{2<p\le y} 2/p = 2\log\log x + O(1)$. By Proposition 3.2, the left-hand side of (3.1) is $O(x/(\log x)^{0.6})$, which is more than we need. So we may assume $q' > 2$.

When $q' > 2$, we apply Proposition 3.1 taking $T = \log x$. Let $t$ be any real number with $|t| \le T$. We set $z = \exp((\log x)^\delta)$ and start from the lower bound

$$\mathbb{D}(F, n^{it};x)^2 \ge \sum_{z<p\le y} \frac{1-\mathrm{Re}(e^{2\pi i r' p/q'}p^{-it})}{p}. \tag{3.2}$$

To estimate the right-hand sum, we split the range of summation into blocks on which $p^{-it}$ is essentially constant.

Cover $(z,y]$ with intervals $\mathcal{I} = (u, u(1+1/(\log x)^2)]$, allowing the rightmost interval to jut out slightly past $y$ but no further than $y + y/(\log x)^2$. On each interval $\mathcal{I}$, every

$p \in \mathcal{I}$ satisfies $|t \log p - t \log u| \le |t|/(\log x)^2 \le 1/\log x$, so that

$$|p^{-it} - u^{-it}| = \left| \int_{t \log u}^{t \log p} \exp(-i\theta)\, d\theta \right| \le 1/\log x,$$

and

$$\sum_{p \in \mathcal{I}} \frac{1 - \mathrm{Re}(\mathrm{e}^{2\pi i r' p/q'} p^{-it})}{p} = \sum_{p \in \mathcal{I}} \frac{1 - \mathrm{Re}(\mathrm{e}^{2\pi i r' p/q'} u^{-it})}{p} + O\left( \frac{1}{\log x} \sum_{p \in \mathcal{I}} \frac{1}{p} \right). \quad (3.3)$$

The error term when summed over all intervals $\mathcal{I}$ will be $O(\log\log x/\log x)$, which is negligible for us. So we focus on the main term. Observe that $p = (1 + o(1))u$ for every $p \in \mathcal{I}$. (Here and below, asymptotic notation refers to the behavior as $x \to \infty$.) Thus,

$$\begin{aligned} \sum_{p \in \mathcal{I}} \frac{1 - \mathrm{Re}(\mathrm{e}^{2\pi i r' p/q'} u^{-it})}{p} &\gtrsim \frac{1}{u} \sum_{p \in \mathcal{I}} (1 - \mathrm{Re}(\mathrm{e}^{2\pi i r' p/q'} u^{-it})) \\ &\gtrsim \frac{1}{u} \sum_{\substack{a' \bmod q' \\ \gcd(a', q') = 1}} (1 - \mathrm{Re}(\mathrm{e}^{2\pi i r' a'/q'} u^{-it})) \pi(\mathcal{I}; q', a'), \end{aligned}$$

where $\pi(\mathcal{I}; q', a')$ denotes the number of primes $p \in \mathcal{I}$ with $p \equiv a' \pmod{q'}$. By the Siegel–Walfisz theorem [26, Theorem 8.17, p. 376], $\pi(\mathcal{I}; q', a') \sim \frac{1}{\phi(q')} \pi(\mathcal{I})$, where $\pi(\mathcal{I})$ is the total count of primes belonging to $\mathcal{I}$. Thus, the above right-hand side is

$$\begin{aligned} \gtrsim \frac{\pi(\mathcal{I})}{\phi(q')u} \sum_{\substack{a' \bmod q' \\ \gcd(a', q') = 1}} (1 - \mathrm{Re}(\mathrm{e}^{2\pi i r' a'/q'} u^{-it})) &= \frac{\pi(\mathcal{I})}{\phi(q')u} (\phi(q') - \mathrm{Re}(\mu(q') u^{-it})) \\ &\ge \frac{1}{2} \pi(\mathcal{I})/u \gtrsim \frac{1}{2} \sum_{p \in \mathcal{I}} \frac{1}{p}; \end{aligned} \quad (3.4)$$

here we use that $\sum_{a' \pmod{q'},\ \gcd(a', q') = 1} \mathrm{e}^{2\pi i a' r'/q'} = \mu(q')$ (see, for example, [17, Theorem 272, p. 309]) and that $\phi(q') - \mathrm{Re}(\mu(q') u^{-it}) \ge \phi(q') - 1 \ge \frac{1}{2}\phi(q')$, as $q' > 2$. Combining the last two displays and summing on $\mathcal{I}$,

$$\sum_{\mathcal{I}} \sum_{p \in \mathcal{I}} \frac{1 - \mathrm{Re}(\mathrm{e}^{2\pi i r' p/q'} u^{-it})}{p} \gtrsim \frac{1}{2} \sum_{\mathcal{I}} \sum_{p \in \mathcal{I}} \frac{1}{p} \ge \frac{1}{2} \sum_{z < p \le y} \frac{1}{p} \gtrsim \frac{1}{2}(1 - \delta) \log\log x.$$

From (3.3) (and the immediately following remark about the error term), the same lower bound holds for $\sum_{\mathcal{I}} \sum_{p \in \mathcal{I}} \frac{1 - \mathrm{Re}(\mathrm{e}^{2\pi i r' p/q'} p^{-it})}{p}$. This double sum essentially coincides with the right-hand side of (3.2), except for possibly including contributions from a few values of $p > y$. But those contributions are $O(1)$, in fact $\ll \sum_{y < p < y + y/(\log x)^2} 1/p \ll 1/(\log x)^2$. Thus, $\mathbb{D}(F, n^{it}; x)^2 \gtrsim \frac{1}{2}(1 - \delta) \log\log x$. In particular, $\mathbb{D}(F, n^{it}; x)^2 \ge (\frac{1}{2} - \frac{9}{10}\delta) \log\log x$ once $x$ is sufficiently large (in terms of $\delta$ and $U$). Since this lower bound holds uniformly in $t$ with $|t| \le T$, the desired inequality (3.1) follows from Proposition 3.1. ∎

Using Lemma 3.3, we can establish the following $A(n)$-analogue of Proposition 2.2.

***Proposition 3.4*** *Fix positive integers $k, D$, and $b$ with $b \geq 2$. Fix real numbers $U' > U \geq 1$, and fix $\epsilon > 0$. The number of $n \leq x$ for which $A(n) \equiv a \pmod q$, $P(n)$ begins with the digits of $D$ in base $b$, and $P(n) \in (x^{1/U'}, y]$ is*

$$\frac{1}{q} \frac{\log(1 + D^{-1})}{\log b} \left(\rho(u) - \rho(U')\right) x + o(x/q),$$

*where $u := \frac{\log x}{\log y}$, where $x, y \to \infty$ with $y \geq x^{1/U}$, and where $a \bmod q$ is any residue class with $q \leq (\log x)^{\frac{1}{2} - \epsilon}$.*

**Proof** The proof is similar to the case $k = 1$ of Proposition 2.2, with the needed input on $\Psi(x, y)$ replaced by appeals to Lemma 3.3. We may assume $y = x^{1/u}$ where $u \geq 1$ is fixed. With the intervals $\mathcal{I}_j$ defined as in (2.4), the desired count of $n$ is given by the triple sum

$$\sum_{j \geq 0} \sum_{p \in \mathcal{I}_j \cap (x^{1/U'}, y]} \sum_{\substack{n \leq x \\ P(n) = p \\ A(n) \equiv a \pmod q}} 1. \tag{3.5}$$

At the cost of a negligible error, we may restrict the outer sum to $j \in \mathcal{J}$, where $\mathcal{J}$ is the collection of nonnegative integers $j$ with $\mathcal{I}_j \subset (x^{1/U'}, y/\exp(\sqrt{\log x}))$; indeed, defining (as before) $P := (x^{1/U'}, bx^{1/U'}]$ and $P' := [y/b \exp(\sqrt{\log x}), y]$, the incurred error is of size

$$\ll x \sum_{p \in P \cup P'} 1/p \ll x/(\log x)^{1/2},$$

which is $o(x/q)$. Now suppose $j \in \mathcal{J}$ and $p \in \mathcal{I}_j$; then by Lemma 3.3,

$$\sum_{\substack{n \leq x \\ P(n) = p \\ A(n) \equiv a \pmod q}} 1 = \sum_{\substack{m \leq x/p \\ P(m) \leq p \\ A(m) \equiv a - p \pmod q}} 1 = \frac{1}{q} \Psi(x/p, p) + O\left( \frac{x}{p(\log (x/p))^{\frac{1}{2}(1-\epsilon)}} \right).$$

Summing on all $j \in \mathcal{J}$ and all $p \in \mathcal{I}_j$, the contribution from $O$-terms is

$$\ll x \sum_{x^{1/U'} < p \leq x/2} \frac{1}{p(\log (x/p))^{\frac{1}{2}(1-\epsilon)}} \ll \frac{x}{(\log x)^{\frac{1}{2}(1-\epsilon)}},$$

which is $o(x/q)$. (Perhaps the simplest way to estimate this last sum on $p$ is to consider, for each $j$, the contribution from $p$ with $x/p \in (e^j, e^{j+1}]$.) On the other hand, the calculations from the proof of Proposition 2.2 (with $k = 1, q = 1$) already show that

$$\sum_{j \in \mathcal{J}} \sum_{p \in \mathcal{I}_j} \Psi(x/p, p) = \frac{\log(1 + D^{-1})}{\log b} (\rho(u) - \rho(U') + o(1)) x.$$

Collecting estimates, we deduce that (3.5) is $\frac{1}{q} \frac{\log(1+D^{-1})}{\log b} \left(\rho(u) - \rho(U')\right) x + o(x/q)$, as desired. ∎

Proposition 3.4 implies the following variant of Theorem 1.2, with the leading digits of $P(n)$ prescribed (instead of those of $A(n)$).

***Proposition* 3.5** *Fix positive integers $k$, $D$, and $b$ with $b \geq 2$. Fix a real number $U \geq 1$, and fix $\epsilon > 0$. The number of $n \leq x$ for which $A(n) \equiv a \pmod q$, $P(n)$ begins with the digits of $D$ in base $b$, and $P(n) \leq y$ is*

$$\sim \frac{1}{q} \frac{\log(1+D^{-1})}{\log b} \Psi(x, y),$$

*where $x, y \to \infty$ with $y \geq x^{1/U}$, and where $a \bmod q$ is any residue class with $q \leq (\log x)^{\frac{1}{2}-\epsilon}$.*

**Proof** The proof parallels that of Theorem 1.1. It suffices to show that the count of $n$ in question is eventually larger than $\frac{1}{q}\frac{\log(1+D^{-1})}{\log b}(\rho(u)-\eta)x$ and eventually smaller than $\frac{1}{q}\frac{\log(1+D^{-1})}{\log b}(\rho(u)+\eta)x$. The lower bound follows from Proposition 3.4, fixing $U'$ large enough that $\rho(U') < \eta$. For the upper bound, we fix $U'$ large enough that $\rho(U') < \eta\frac{\log(1+D^{-1})}{\log b}$; the upper bound inequality then follows from Lemma 3.3 and Proposition 3.4. ∎

To finish the proof of Theorem 1.2, we show that $P(n)$ and $A(n)$ usually have the same leading digits. We begin by observing that $P(n)$ and $A(n)$ are usually close.

***Lemma* 3.6** *Fix $\delta > 0$. For large $x$, the number of $n \leq x$ for which $A(n) > (1+\delta)P(n)$ is $O(x(\log\log x)^2/\log x)$.*

**Proof** Put $y := x^{1/2\log\log x}$. We may suppose that $P(n) > y$, since by standard results on the distribution of smooth numbers (e.g., Theorem 5.1 on p. 512 of [26]) this condition excludes only $O(x/\log x)$ integers $n \leq x$. If $A(n) > (1+\delta)P(n)$ for one of these remaining $n$, then $\delta P(n) < \sum_{k>1} P_k(n) \leq \Omega(n)P_2(n) \leq 2P_2(n)\log x$. Hence, $n$ is divisible by $pp'$ for primes $p, p'$ with $p > y$ and $p' \in (\frac{\delta}{2}p/\log x, p]$. The number of such $n \leq x$ is

$$x \sum_{y<p\leq x} \sum_{\frac{\delta}{2}\frac{p}{\log x}<p'\leq p} \frac{1}{pp'} \ll x \sum_{y<p\leq x} \frac{1}{p}\frac{\log\log x}{\log p} \ll x\frac{\log\log x}{\log y} \ll x\frac{(\log\log x)^2}{\log x}.$$

Here the sum on $p'$ has been estimated using Mertens' theorem with the usual $1/\log$ error term [26, Theorem 1.10, p. 18]. ∎

***Lemma* 3.7** *Fix positive integers $N$ and $b$, with $b \geq 2$, and fix a real number $\epsilon > 0$. Among all $n \leq x$ with $A(n) \equiv a \pmod q$, the number of $n$ for which the $N$ leading base $b$ digits of $P(n)$* do not *coincide with those of $A(n)$ is $o(x/q)$, as $x \to \infty$, uniformly in residue classes $a \bmod q$ with $q \leq (\log x)^{\frac{1}{2}-\epsilon}$.*

**Proof** Since $b$ and $N$ are fixed, it is enough to prove the estimate of the lemma under the assumption that the $N$ leading digits in the base $b$ expansion of $P(n)$ are fixed, say as the digits of the positive integer $D$.

For $M$ a (fixed) positive integer to be specified momentarily, we let $D'$ be the integer obtained by tacking $M$ copies of the digit '$b-1$' on to the end of the $b$-ary expansion of $D$. Thus, $D' = b^M D + (b^M - 1)$.

Suppose $P(n)$ begins with $D$ in base $b$ but $A(n)$ does not. We take two cases. First, it may be that $P(n)$ begins with $D$ but not $D'$; in that case, for $A(n)$ to not begin with $D$ we must have $A(n)/P(n) > 1 + 1/D'$. By Lemma 3.6, the number of such $n \le x$ is $O(x(\log\log x)^2/\log x)$, which is $o(x/q)$. On the other hand, if $P(n)$ begins with $D'$, we apply Proposition 3.5. Taking $y = x$ there, we see that the number of $n \le x$ for which $P(n)$ begins with $D'$ and $A(n) \equiv a \pmod q$ is $\sim \frac{\log(1+1/D')}{\log b}\frac{x}{q}$. Since the coefficient $\frac{\log(1+1/D')}{\log b}$ of $\frac{x}{q}$ in this estimate can be made as small as we like by fixing $M$ large enough, we obtain the lemma. ∎

Theorem 1.2 follows from combining Proposition 3.5 with Lemma 3.7.

***Remark*** The range of uniformity in $q$ can be widened under the assumption that $q$ is supported on sufficiently large primes. More precisely, for any fixed $Q \ge 2$, the result of Theorem 1.2 holds uniformly for $q \le (\log x)^{1-1/Q-\epsilon}$, provided the least prime $P^-(q)$ dividing $q$ is at least $Q+1$. The key observation is that, in the notation of Lemma 3.3, such $q$ have $\phi(q') \ge P^-(q) - 1 \ge Q$, which shows that

$$\frac{\pi(\mathcal{I})}{\phi(q')u}(\phi(q') - \mathrm{Re}(\mu(q')u^{-it})) \ge \left(1 - \frac{1}{Q}\right)\frac{\pi(\mathcal{I})}{u}$$

in the display (3.4). The remainder of the proof requires only minor modifications.

## Acknowledgements

We thank the referees for their careful reading of the manuscript. P.P. is supported by the National Science Foundation under award DMS-2001581.

## References

[1] K. Alladi and P. Erdős, *On an additive arithmetic function*, Pacific J. Math. **71** (1977), no. 2, 275–294.
[2] T. C. Anderson, L. Rolen, and R. Stoehr, *Benford's law for coefficients of modular forms and partition functions*, Proc. Amer. Math. Soc. **139** (2011), 1533–1541.
[3] S. Aursukaree and V. Chandee, *Equidistribution of* $\log(d(n))$, Proceedings of Annual Pure and Applied Mathematics Conference, Chulalongkorn University, Thailand, May 2016, pp. 399–410.
[4] W. D. Banks, G. Harman, and I. E. Shparlinski, *Distributional properties of the largest prime factor*, Michigan Math. J. **53** (2005), 665–681.
[5] A. Berger and T. P. Hill, *An introduction to Benford's law*, Princeton University Press, Princeton, NJ, 2015.
[6] A. Best, P. Dynes, X. Edelsbrunner, B. McDonald, S. J. Miller, K. Tor, C. Turnage-Butterbaugh, and M. Weinstein, *Benford behavior of Zeckendorf decompositions*, Fibonacci Quart. **52** (2014), no. 5, 35–46.
[7] ———, *Benford behavior of generalized Zeckendorf decompositions*, Combinatorial and additive number theory. II, Springer Proc. Math. Stat., vol. 220, Springer, Cham, 2017, pp. 25–37.
[8] V. Chandee, X. Li, P. Pollack, and A. Singha Roy, *Benford's law for multiplicative functions*, submitted; https://arxiv.org/abs/2203.13117.
[9] E. Chen, P. S. Park, and A. A. Swaminathan, *On logarithmically Benford sequences*, Proc. Amer. Math. Soc. **144** (2016), 4599–4608.
[10] N. G. de Bruijn, *The asymptotic behaviour of a function occurring in the theory of primes*, J. Indian Math. Soc. (N.S.) **15** (1951), 25–32.

[11] P. Diaconis, *The distribution of leading digits and uniform distribution* mod 1, Ann. Probability **5** (1977), 72–81.
[12] D. Goldfeld, *On an additive prime divisor function of Alladi and Erdős*, Analytic number theory, modular forms and $q$-hypergeometric series, Springer Proc. Math. Stat., vol. 221, Springer, Cham, 2017, pp. 297–309.
[13] A. Granville and K. Soundararajan, *Pretentious multiplicative functions and an inequality for the zeta-function*, Anatomy of integers, CRM Proc. Lecture Notes, vol. 46, Amer. Math. Soc., Providence, RI, 2008, pp. 191–197.
[14] G. Halász, *Über die Mittelwerte multiplikativer zahlentheoretischer Funktionen*, Acta Math. Acad. Sci. Hungar. **19** (1968), 365–403.
[15] R. R. Hall, *On the probability that $n$ and $f(n)$ are relatively prime. III*, Acta Arith. **20** (1972), 267–289.
[16] R. R. Hall and G. Tenenbaum, *Effective mean value estimates for complex multiplicative functions*, Math. Proc. Cambridge Philos. Soc. **110** (1991), 337–351.
[17] G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers*, sixth ed., Oxford University Press, Oxford, 2008.
[18] M. Jameson, J. Thorner, and L. Ye, *Benford's law for coefficients of newforms*, Int. J. Number Theory **12** (2016), 483–494.
[19] D. E. Knuth and L. Trabb Pardo, *Analysis of a simple factorization algorithm*, Theoret. Comput. Sci. **3** (1976/77), 321–348.
[20] A. V. Kontorovich and S. J. Miller, *Benford's law, values of $L$-functions and the $3x+1$ problem*, Acta Arith. **120** (2005), 269–297.
[21] B. Massé and D. Schneider, *Fast growing sequences of numbers and the first digit phenomenon*, Int. J. Number Theory **11** (2015), 705–719.
[22] S. J. Miller (ed.), *Benford's Law: theory and applications*, Princeton University Press, Princeton, NJ, 2015.
[23] H. L. Montgomery and R. C. Vaughan, *Multiplicative number theory. I. Classical theory*, Cambridge Studies in Advanced Mathematics, vol. 97, Cambridge University Press, Cambridge, 2007.
[24] P. Pollack and A. Singha Roy, *Dirichlet, Sierpiński, and Benford*, J. Number Theory **239** (2022), 352–364.
[25] G. Tenenbaum, *Théorie analytique et probabiliste des nombres: 307 exercices corrigés*, Échelles collection, Belin, 2014, prepared with the collaboration of Jie Wu.
[26] ———, *Introduction to analytic and probabilistic number theory*, third ed., Graduate Studies in Mathematics, vol. 163, American Mathematical Society, Providence, RI, 2015.

*Department of Mathematics, University of Georgia, Boyd Research and Education Center, Athens, GA, 30602*
*e-mail:* pollack@uga.edu, akash01s.roy@gmail.com.