SIEVING AND THE ERDŐS–KAC THEOREM

Andrew Granville
Université de Montréal

K. Soundararajan
University of Michigan

Abstract. We give a relatively easy proof of the Erdős-Kac theorem via computing moments. We show how this proof extends naturally in a sieve theory context, and how it leads to several related results in the literature.

Let $\omega(n)$ denote the number of distinct prime factors of the natural number n. The average value of $\omega(n)$ as n ranges over the integers below x is

$$\frac{1}{x}\sum_{n\le x}\omega(n)=\frac{1}{x}\sum_{p\le x}\sum_{\substack{n\le x\\ p|n}}1=\frac{1}{x}\sum_{p\le x}\left[\frac{x}{p}\right]=\frac{1}{x}\sum_{p\le x}\left(\frac{x}{p}+O(1)\right)=\log\log x+O(1).$$

It is natural to ask how $\omega(n)$ is distributed as one varies over the integers $n\le x$. A famous result of Hardy and Ramanujan (Hardy and Ramanujan, 1917) tells us that $\omega(n)\sim\log\log x$ for almost all $n\le x$; we say that $\omega(n)$ has *normal order* $\log\log n$. To avoid confusion let us state this precisely: given $\epsilon>0$ there exists x_ϵ such that if $x\ge x_\epsilon$ is sufficiently large, then $(1+\epsilon)\log\log x\ge\omega(n)\ge(1-\epsilon)\log\log x$ for all but at most ϵx integers $n\le x$. The functions $\log\log n$ and $\log\log x$ are interchangeable here since they are very close in value for all but the tiny integers $n\le x$.

Their proof revolves around the following wonderful inequality which they established by induction. Define $\pi_k(x)$ to be the number of integers $n\le x$ with $\omega(n)=k$. There exist constants $c_0,c_1>0$ such that for any $k\ge 0$ we have

$$\pi_k(x)<c_0\frac{x}{\log x}\frac{(\log\log x+c_1)^{k-1}}{(k-1)!},\tag{1}$$

for all $x\ge 2$. Hardy and Ramanujan exploited this by deducing that

$$\sum_{|k-\log\log x|\ge\epsilon\log\log x}\pi_k(x)\le c_0\frac{x}{\log x}\sum_{|k-\log\log x|\ge\epsilon\log\log x}\frac{(\log\log x\ +c_1)^{k-1}}{(k-1)!},$$

which is easily shown to be about $x/(\log x)^{\alpha}$ where $\alpha = \alpha_\epsilon = \epsilon^2/2 + O(\epsilon^3)$, far less than ϵx. In fact Hardy and Ramanujan squeezed a little more out of this idea, showing that if $\kappa(n) \to \infty$ as $n \to \infty$, no matter how slowly, then

$$|\omega(n) - \log\log n| \le \kappa(n)\sqrt{\log\log n} \tag{2}$$

for almost all integers $n \le x$.

Once we know that $\omega(n)$ has normal order $\log\log n$, we can ask finer questions about the distribution of $\omega(n)$. For instance how is $\omega(n) - \log\log n$ distributed? More specifically, how big is this typically in absolute value? Turán (Turán, 1934) found a very simple proof of the Hardy–Ramanujan result by showing that

$$\frac{1}{x}\sum_{n\le x}(\omega(n) - \log\log n)^2 = \{1 + o(1)\}\log\log x. \tag{3}$$

One deduces easily that $\omega(n)$ has *normal order* $\log\log n$: For, if there are $m_\epsilon(x)$ integers $\le x$ for which $|\omega(n) - \log\log n| \ge \epsilon\log\log x$ then by (3), $m_\epsilon(x) \le (1/\epsilon^2 + o(1))x/\log\log x$, which is $\le \epsilon x$ for sufficiently large x. Indeed the same argument also gives (2) for almost all $n \le x$.

We have now obtained some information about the distribution of $\omega(n)$, its average value, and the average difference between the value and the mean. Next we ask whether there is a distribution function for $\omega(n)$? In other words if, typically, the distance between $\omega(n)$ and $\log\log n$ is roughly of size $\sqrt{\log\log n}$ can we say anything about the distribution of

$$\frac{\omega(n) - \log\log n}{\sqrt{\log\log n}}? \tag{4}$$

In the late 1930s Mark Kac noticed that these developments bore more than a passing resemblance to developments in probability theory. He suggested that perhaps this distribution is *normal* and even conjectured certain number theory estimates which would imply that. Soon after describing this in a lecture, at which Paul Erdős was in the audience, Erdős and Kac were able to announce the result (Erdős and Kac, 1940): For any $\tau \in \mathbb{R}$, the proportion of the integers $n \le x$ for which $\omega(n) \le \log\log n + \tau\sqrt{\log\log n}$ tends to the limit

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\tau} \mathrm{e}^{-t^2/2}dt \tag{5}$$

as $x \to \infty$. In other words the quantity in (4) is distributed like a normal distribution with mean 0 and variance 1.

Erdős and Kac's original proof was based on the central limit theorem, and Brun's sieve. A different proof follows from the work of Selberg (Selberg, 1954) (extending and simplifying the work of (Sathe, 1953)) who obtained an asymptotic

formula for $\pi_k(x)$ uniformly in a wide range of k. Yet a third proof is provided by Halberstam (Halberstam, 1955) who showed how to compute the moments

$$\sum_{n\le x}(\omega(n)-\log\log x)^k, \tag{6}$$

for natural numbers k, and showed that these agreed with the moments of a normal distribution. Since the normal distribution is well-known to be determined by its moments, he deduced the Erdős-Kac theorem.Erdos@Erdős-Kac theorem

In this article, we give a simple method to compute the moments (6), and in fact we can obtain an asymptotic formula uniformly in a wide range of k. Then we discuss how such moments can be formulated for more general sequences assuming sieve type hypotheses.

THEOREM 1. *For any natural number k we let $C_k = \Gamma(k+1)/(2^{k/2}\Gamma(k/2+1))$. Uniformly for even natural numbers $k \le (\log\log x)^{1/3}$ we have*

$$\sum_{n\le x}(\omega(n)-\log\log x)^k = C_k x(\log\log x)^{k/2}\Big(1+O\Big(\frac{k^{3/2}}{\sqrt{\log\log x}}\Big)\Big),$$

and uniformly for odd natural numbers $k \le (\log\log x)^{1/3}$ we have

$$\sum_{n\le x}(\omega(n)-\log\log x)^k \ll C_k x(\log\log x)^{k/2}\frac{k^{3/2}}{\sqrt{\log\log x}}.$$

We will deduce this theorem from the following technical proposition.

PROPOSITION 2. *Define*

$$f_p(n) = \begin{cases} 1-\frac{1}{p} & \text{if } p \mid n \\ -\frac{1}{p} & \text{if } p \nmid n. \end{cases}$$

Let $z \ge 10^6$ be a real number. Uniformly for even natural numbers $k \le (\log\log z)^{\frac{1}{3}}$ we have

$$\sum_{n\le x}\Big(\sum_{p\le z} f_p(n)\Big)^k = C_k x(\log\log z)^{k/2}\Big(1+O\Big(\frac{k^3}{\log\log z}\Big)\Big) + O(2^k\pi(z)^k), \tag{7}$$

while, uniformly for odd natural numbers $k \le (\log\log z)^{1/3}$, we have

$$\sum_{n\le x}\Big(\sum_{p\le z} f_p(n)\Big)^k \ll C_k x(\log\log z)^{k/2}\frac{k^{3/2}}{\sqrt{\log\log z}} + 2^k\pi(z)^k. \tag{8}$$

Deduction of Theorem 1. We seek to evaluate $\sum_{n\le x}(\omega(n)-\log\log x)^k$ for natural numbers $k\le(\log\log x)^{1/3}$. Set $z=x^{1/k}$ and note that, for $n\le x$,

$$\omega(n)-\log\log x=\sum_{p\le z}f_p(n)+\sum_{\substack{p\mid n\\ p>z}}1+\Big(\sum_{p\le z}1/p-\log\log x\Big)=\sum_{p\le z}f_p(n)+O(k).$$

Thus for some positive constant c we obtain that

$$(\omega(n)-\log\log x)^k=\Big(\sum_{p\le z}f_p(n)\Big)^k+O\Big(\sum_{\ell=0}^{k-1}(ck)^{k-\ell}\binom{k}{\ell}\Big|\sum_{p\le z}f_p(n)\Big|^\ell\Big).$$

When we sum this up over all integers $n\le x$ the first term above is handled through (7, 8). To handle the remainder terms we estimate $\sum_{n\le x}\big|\sum_{p\le z}f_p(n)\big|^\ell$ for $\ell\le k-1$. When ℓ is even this is once again available through (7). Suppose ℓ is odd. By Cauchy–Schwarz we get that

$$\sum_{n\le x}\Big|\sum_{p\le z}f_p(n)\Big|^\ell\le\Big(\sum_{n\le x}\Big(\sum_{p\le z}f_p(n)\Big)^{\ell-1}\Big)^{1/2}\Big(\sum_{n\le x}\Big(\sum_{p\le z}f_p(n)\Big)^{\ell+1}\Big)^{1/2},$$

and using (7) we deduce that this is

$$\ll\sqrt{C_{\ell-1}C_{\ell+1}}x(\log\log z)^{\ell/2}.$$

Proof of Proposition 2. If $r=\prod_i p_i^{\alpha_i}$ is the prime factorization of r we put $f_r(n)=\prod_i f_{p_i}(n)^{\alpha_i}$. Then we may write

$$\sum_{n\le x}\Big(\sum_{p\le z}f_p(n)\Big)^k=\sum_{p_1,\dots,p_k\le z}\sum_{n\le x}f_{p_1\cdots p_k}(n).$$

To proceed further, let us consider more generally $\sum_{n\le x}f_r(n)$.

Suppose $r=\prod_{i=1}^s q_i^{\alpha_i}$ where the q_i are distinct primes and $\alpha_i\ge1$. Set $R=\prod_{i=1}^s q_i$ and observe that if $d=(n,R)$ then $f_r(n)=f_r(d)$. Therefore, with τ denoting the divisor function,

$$\begin{aligned}\sum_{n\le x}f_r(n)&=\sum_{d\mid R}f_r(d)\sum_{\substack{n\le x\\(n,R)=d}}1=\sum_{d\mid R}f_r(d)\Big(\frac{x}{d}\frac{\varphi(R/d)}{R/d}+O(\tau(R/d))\Big)\\&=\frac{x}{R}\sum_{d\mid R}f_r(d)\varphi(R/d)+O(\tau(R)).\end{aligned}$$

Thus seting

$$G(r):=\frac{1}{R}\sum_{d\mid R}f_r(d)\varphi(R/d)=\prod_{q^\alpha\|r}\Big(\frac{1}{q}\Big(1-\frac{1}{q}\Big)^\alpha+\Big(\frac{-1}{q}\Big)^\alpha\Big(1-\frac{1}{q}\Big)\Big),$$

we conclude that

$$\sum_{n\le x} f_r(n) = G(r)x + O(\tau(R)).$$

Observe that $G(r) = 0$ unless r is square-full and so

$$\sum_{n\le x}\Big(\sum_{p\le z} f_p(n)\Big)^k = x \sum_{\substack{p_1,\dots,p_k\le z\\ p_1\cdots p_k \text{ square-full}}} G(p_1\cdots p_k) + O(2^k\pi(z)^k). \tag{9}$$

Suppose $q_1 < q_2 < \ldots < q_s$ are the distinct primes in $p_1\cdots p_k$. Note that since $p_1\cdots p_k$ is square-full we have $s \le k/2$. Thus our main term above is

$$\sum_{s\le k/2}\sum_{q_1<q_2<\ldots<q_s\le z}\sum_{\substack{\alpha_1,\dots,\alpha_s\ge 2\\ \sum_i \alpha_i = k}} \frac{k!}{\alpha_1!\cdots\alpha_s!} G(q_1^{\alpha_1}\cdots q_s^{\alpha_s}).$$

When k is even there is a term $s = k/2$ (and all $\alpha_i = 2$) which gives rise to the Gaussian moments. This term contributes

$$\frac{k!}{2^{k/2}(k/2)!}\sum_{\substack{q_1,\dots,q_{k/2}\le z\\ q_i \text{ distinct}}}\prod_{i=1}^{k/2}\frac{1}{q_i}\Big(1-\frac{1}{q_i}\Big).$$

By ignoring the distinctness condition, we see that the sum over q's is bounded above by $(\sum_{p\le z}(1-1/p)/p)^{k/2}$. On the other hand, if we consider $q_1,\dots,q_{k/2-1}$ as given then the sum over $q_{k/2}$ is plainly at least $\sum_{\pi_{k/2}\le p\le z}(1-1/p)/p$ where we let π_n denote the nth smallest prime. Repeating this argument, the sum over the q's is bounded below by $(\sum_{\pi_{k/2}\le p\le z}(1-1/p)/p)^{k/2}$. Therefore the term with $s = k/2$ contributes

$$\frac{k!}{(k/2)!2^{k/2}}(\log\log z + O(1+\log\log k))^{k/2}. \tag{10}$$

To estimate the terms $s < k/2$ we use that $0 \le G(q_1^{\alpha_1}\cdots q_s^{\alpha_s}) \le 1/(q_1\cdots q_s)$ and so these terms contribute

$$\le \sum_{s<k/2}\frac{k!}{s!}\Big(\sum_{q\le z}\frac{1}{q}\Big)^s \sum_{\substack{\alpha_1,\dots,\alpha_s\ge 2\\ \sum_i\alpha_i = k}}\frac{1}{\alpha_1!\cdots\alpha_s!}.$$

The number of ways of writing $k = \alpha_1 + \ldots + \alpha_s$ with each $\alpha_i \ge 2$ equals the number of ways of writing $k-s = \alpha_1' + \ldots + \alpha_s'$ where each $\alpha_i' \ge 1$ and is therefore $\binom{k-s}{s}$. Thus these remainder terms contribute

$$\le \sum_{s<k/2}\frac{k!}{s!2^s}\binom{k-s}{s}(\log\log z + O(1))^s. \tag{11}$$

Proposition 2 follows upon combining (9), (10), and (11).

The main novelty in our proof above is the introduction of the function $f_r(n)$ whose expectation over integers n below x is small unless r is square-full. This leads easily to a recognition of the main term in the asymptotics of the moments. Previous approaches expanded out $(\omega(n)-\log\log x)^k$ using the binomial theorem, and then there are several main terms which must be carefully cancelled out before the desired asymptotic emerges. Our use of this simpler technique was inspired by (Montgomery and Soundararajan, 2004). Recently Rizwanur Khan (Khan, 2006) builds on this idea to prove that the spacings between normal numbers obey a Poisson distribution law.

This technique extends readily to the study of $\omega(n)$ in many other sequences. We formulate this in a sieve like setting:

Let $\mathcal{A} = \{a_1, \dots, a_x\}$ be a (multi)-set of x (not necessarily distinct) natural numbers. Let $\mathcal{A}_d = \#\{n \le x : d \mid a_n\}$. We suppose that there is a real valued, non-negative multiplicative function $h(d)$ such that for square-free d we may write

$$\mathcal{A}_d = \frac{h(d)}{d}x + r_d.$$

It is natural to suppose that $0 \le h(d) \le d$ for all square-free d, and we do so below. Here r_d denotes a remainder term which we expect to be small: either small for all d, or maybe just small on average over d.

Let $\mathcal{P}$ be any set of primes. In sieve theory one attempts to estimate $\#\{n \le x : (a_n, m) = 1\}$ for $m = \prod_{p\in\mathcal{P}} p$, in terms of the function h and the error terms r_d. Here we want to understand the distribution of values of $\omega_{\mathcal{P}}(a)$, as we vary through elements a of $\mathcal{A}$, where $\omega_{\mathcal{P}}(a)$ is defined to be the number of primes $p \in \mathcal{P}$ which divide a. We expect that the distribution of $\omega_{\mathcal{P}}(a)$ is normal with "mean" and "variance" given by

$$\mu_{\mathcal{P}} := \sum_{p\in\mathcal{P}} \frac{h(p)}{p} \quad \text{and} \quad \sigma^2_{\mathcal{P}} := \sum_{p\in\mathcal{P}} \frac{h(p)}{p}\left(1-\frac{h(p)}{p}\right),$$

and wish to find conditions under which this is true. There is a simple heuristic which explains why this should usually be true: Suppose that for each prime p we have a sequence of independent random variables $b_{1,p}, \dots, b_{x,p}$ each of which is 1 with probability $h(p)/p$ and 0 otherwise; and we let b_j be the product of the primes p for which $b_{j,p} = 1$. The b_j form a probabilistic model for the a_j satisfying our sieve hypotheses, the key point being that, in the model, whether or not b_j is divisible by different primes is independent. One can use the central limit theorem to show that, as $x \to \infty$, the distribution of $\omega_{\mathcal{P}}(b)$ becomes normal with mean $\mu_{\mathcal{P}}$ and variance $\sigma^2_{\mathcal{P}}$.

PROPOSITION 3. *Uniformly for all natural numbers $k \le \sigma_{\mathcal{P}}^{2/3}$ we have*

$$\sum_{a\in\mathcal{A}} (\omega_{\mathcal{P}}(a)-\mu_{\mathcal{P}})^k = C_k x\sigma_{\mathcal{P}}^k\Big(1+O\Big(\frac{k^3}{\sigma_{\mathcal{P}}^2}\Big)\Big) + O\Big(\mu_{\mathcal{P}}^k \sum_{d\in D_k(\mathcal{P})} |r_d|\Big),$$

if k is even, and

$$\sum_{a\in\mathcal{A}} (\omega_{\mathcal{P}}(a)-\mu_{\mathcal{P}})^k \ll C_k x\sigma_{\mathcal{P}}^k \frac{k^{\frac{3}{2}}}{\sigma_{\mathcal{P}}} + \mu_{\mathcal{P}}^k \sum_{d\in D_k(\mathcal{P})} |r_d|,$$

if k is odd. Here $D_k(\mathcal{P})$ denotes the set of squarefree integers which are the product of at most k primes all from the set $\mathcal{P}$.

Proof. The proof is similar to that of Proposition 2, and so we record only the main points. We define $f_p(a) = 1 - h(p)/p$ if $p \mid a$ and $-h(p)/p$ if $p \nmid a$. If $r = \prod_i p_i^{\alpha_i}$ is the prime factorization of r we put $f_r(a) = \prod_i f_{p_i}(a)^{\alpha_i}$. Note that $\omega_{\mathcal{P}}(a) - \mu_{\mathcal{P}} = \sum_{p\in\mathcal{P}} f_p(a)$, and so

$$\sum_{a\in\mathcal{A}} (\omega_{\mathcal{P}}(a)-\mu_{\mathcal{P}})^k = \sum_{p_1,\dots,p_k\in\mathcal{P}} \sum_{a\in\mathcal{A}} f_{p_1\cdots p_k}(a). \tag{12}$$

As in Proposition 2, consider more generally $\sum_{a\in\mathcal{A}} f_r(a)$. Suppose $r = \prod_{i=1}^s q_i^{\alpha_i}$ where the q_i are distinct primes and each $\alpha_i \ge 1$. Set $R = \prod_{i=1}^s q_i$ and observe that if $d = (a, R)$ then $f_r(a) = f_r(d)$. Note that

$$\begin{aligned}\sum_{\substack{a\in\mathcal{A}\\(a,R)=d}} 1 &= \sum_{a\in\mathcal{A}} \sum_{\substack{e\mid(R/d)\\ de\mid n}} \mu(e) = \sum_{e\mid R/d} \mu(e)\mathcal{A}_{de}\\ &= x\,\frac{h(d)}{d} \prod_{p\mid(R/d)} \Big(1-\frac{h(p)}{p}\Big) + \sum_{e\mid(R/d)} \mu(e) r_{de}.\end{aligned}$$

Therefore

$$\begin{aligned}\sum_{a\in\mathcal{A}} f_r(a) &= \sum_{d\mid R} f_r(d) \sum_{\substack{a\in\mathcal{A}\\(a,R)=d}} 1\\ &= x\sum_{d\mid R} f_r(d)\,\frac{h(d)}{d} \prod_{p\mid(R/d)} \Big(1-\frac{h(p)}{p}\Big) + \sum_{d\mid R} f_r(d) \sum_{e\mid(R/d)} \mu(e) r_{de}\\ &= G(r)x + \sum_{m\mid R} r_m E(r,m),\end{aligned} \tag{13}$$

where

$$G(r) = \prod_{q^\alpha \| r} \Big(\frac{h(q)}{q}\Big(1-\frac{h(q)}{q}\Big)^\alpha + \Big(\frac{-h(q)}{q}\Big)^\alpha \Big(1-\frac{h(q)}{q}\Big)\Big), \tag{14}$$

and

$$E(r,m)=\prod_{\substack{q^\alpha\|r\\ q|m}}\left(\Big(1-\frac{h(q)}{q}\Big)^\alpha-\Big(\frac{-h(q)}{q}\Big)^\alpha\right)\prod_{\substack{q^\alpha\|r\\ q|(R/m)}}\Big(\frac{-h(q)}{q}\Big)^\alpha. \tag{15}$$

We input the above analysis in (12). Consider first the main terms that arise. Notice that $G(r)=0$ unless r is square-full, and so the main terms look exactly like the corresponding main terms in Proposition 2. We record the only small difference from the analysis there. When k is even there is a leading contribution from the terms with $s=k/2$ and all $\alpha_i=2$ (in notation analogous to Proposition 2); this term contributes

$$\frac{k!}{2^{k/2}(k/2)!}\sum_{\substack{q_1,\dots,q_{k/2}\in\mathcal{P}\\ q_i \text{ distinct}}}\prod_{i=1}^{k/2}\frac{h(q_i)}{q_i}\Big(1-\frac{h(q_i)}{q_i}\Big).$$

The sum over q's is bounded above by $\sigma_{\mathcal{P}}^k$, and is bounded below by

$$\Bigg(\sum_{\substack{p\in\mathcal{P}\\ p\ge\pi_{k/2}(\mathcal{P})}}\frac{h(p)}{p}\Big(1-\frac{h(p)}{p}\Big)\Bigg)^{k/2}\ge(\sigma_{\mathcal{P}}^2-k/8)^{k/2},$$

where we let $\pi_n(\mathcal{P})$ denote the n-th smallest prime in $\mathcal{P}$ and made use of the fact that $0\le(h(p)/p)(1-h(p)/p)\le 1/4$. The remainder of the argument is exactly the same as in Proposition 2.

Finally we need to deal with the "error" term contribution to (12). To estimate the error terms that arise in (12), we use that $|E(p_1\cdots p_k,m)|\le\prod_{p_i\nmid m}h(p_i)/p_i$. Thus the error term is

$$\le\sum_{\ell=1}^{k}\sum_{\substack{m=q_1\dots q_\ell\ge 1\\ q_1<q_2<\cdots<q_\ell\in\mathcal{P}}}|r_m|\sum_{\substack{p_1,\dots,p_k\in\mathcal{P}\\ m|p_1\cdots p_k}}\prod_{p_i\nmid m}\frac{h(p_i)}{p_i}.$$

Fix m and let $e_j=\#\{i:\ p_i=q_j\}$ for each j, $1\le j\le\ell$. Then there are $e_0:=k-(e_1+\cdots+e_\ell)\le k-\ell$ primes p_i which are not equal to any q_j, and so their contribution to the final sum is $\le\mu_{\mathcal{P}}^{e_0}$. Therefore the final sum is

$$\le\sum_{0\le e_0\le k-\ell}\binom{k}{e_0}\mu_{\mathcal{P}}^{e_0}\sum_{\substack{e_1+\cdots+e_\ell=k-e_0\\ \text{each } e_i\ge 1}}\frac{(k-e_0)!}{e_1!\cdots e_\ell!}$$

$$\le\sum_{0\le e_0\le k-1}\binom{k}{e_0}\mu_{\mathcal{P}}^{e_0}\ell^{k-e_0}\le(\mu_{\mathcal{P}}+\ell)^k\ll 2\mu_{\mathcal{P}}^k,$$

since $k^3 \le \sigma_{\mathcal{P}}^2 \le \mu_{\mathcal{P}}$. This completes the proof of the proposition.

One way of using Proposition 3 is to take $\mathcal{P}$ to be the set of primes below z where z is suitably small so that the error term arising from the $|r_d|$'s is negligible. If the numbers a in $\mathcal{A}$ are not too large, then there cannot be too many primes larger than z that divide a, and so Proposition 3 furnishes information about $\omega(a)$. Note that we used precisely such an argument in deducing Theorem 1 from Proposition 2.

In this manner, Proposition 3 may be used to prove the Erdős-Kac theorem for many interesting sequences of integers. For example, Halberstam (Halberstam, 1956) showed such a result for the shifted primes $p-1$, which the reader can now deduce from Proposition 3 and the Bombieri–Vinogradov theorem.

Similarly, one can take $\mathcal{A} = \{f(n) : n \le x\}$ for $f(t) \in \mathbb{Z}[t]$. In this case $h(p)$ is bounded by the degree of f except at finitely many primes, and the prime ideal theorem implies that $\mu_{\mathcal{P}}, \sigma_{\mathcal{P}} = m \log\log x + O(1)$ where m is the number of distinct irreducible factors of f. Again this example was first considered by Halberstam (Halberstam, 1956).

Alladi (Alladi, 1987) proved an Erdős–Kac theorem for integers without large prime factors. Proposition 3 reduces this problem to obtaining information about multiples of d in this set of "smooth numbers." We invite the reader to fill in this information.

In place of $\omega(a)$ we may study more generally the distribution of values of $g(a)$ where g is an "additive function." Recall that an additive function satisfies $g(1) = 0$, and $g(mn) = g(m) + g(n)$ whenever m and n are coprime. Its values are determined by the prime-power values $g(p^k)$. If in addition $g(p^k) = g(p)$ for all $k \ge 1$ we say that the function g is "strongly additive." The strongly additive functions form a particularly nice subclass of additive functions and for convenience we restrict ourselves to this subclass.

PROPOSITION 4. *Let $\mathcal{A}$ be a (multi)-set of x integers, and let $h(d)$ and r_d be as above. Let $\mathcal{P}$ be a set of primes, and let g be a real-valued, strongly additive function with $|g(p)| \le M$ for all $p \in \mathcal{P}$. Let*

$$\mu_{\mathcal{P}}(g) = \sum_{p\in\mathcal{P}} g(p)\frac{h(p)}{p}, \quad \text{and} \quad \sigma_{\mathcal{P}}(g)^2 = \sum_{p\in\mathcal{P}} g(p)^2\frac{h(p)}{p}\Big(1-\frac{h(p)}{p}\Big).$$

Then, uniformly for all even natural numbers $k \le (\sigma_{\mathcal{P}}(g)/M)^{2/3}$,

$$\sum_{a\in\mathcal{A}}\Big(\sum_{\substack{p\mid a\\ p\in\mathcal{P}}} g(p) - \mu_{\mathcal{P}}(g)\Big)^k = C_k x\sigma_{\mathcal{P}}(g)^k\Big(1+O\Big(\frac{k^3M^2}{\sigma_{\mathcal{P}}(g)^2}\Big)\Big) + O\Big(M^k\Big(\sum_{p\in\mathcal{P}}\frac{h(p)}{p}\Big)^k \sum_{d\in D_k(\mathcal{P})}|r_d|\Big),$$

while for all odd natural numbers $k \le (\sigma_{\mathcal{P}}(g)/M)^{2/3}$,

$$\sum_{a\in\mathcal{A}}\Big(\sum_{\substack{p\mid a\\ p\in\mathcal{P}}} g(p) - \mu_{\mathcal{P}}(g)\Big)^k \ll C_k x\sigma_{\mathcal{P}}(g)^k\frac{k^{3/2}M}{\sigma_{\mathcal{P}}(g)} + M^k\Big(\sum_{p\in\mathcal{P}}\frac{h(p)}{p}\Big)^k\sum_{d\in D_k(\mathcal{P})}|r_d|.$$

Proof. We follow closely the proofs of Propositions 2 and 3, making appropriate modifications. Let $f_r(n)$ be as in the proof of Proposition 3. Then we wish to evaluate

$$\sum_{a\in\mathcal{A}}\Big(\sum_{p\in\mathcal{P}} g(p)f_p(a)\Big)^k = \sum_{p_1,\dots,p_k\in\mathcal{P}} g(p_1)\cdots g(p_k)\sum_{a\in\mathcal{A}} f_{p_1\cdots p_k}(a).$$

We may now input the results (13, 14, 15) here. Consider first the error terms that arise. Since $|g(p)| \le M$ for all $p \in \mathcal{P}$ this contribution is at most M^k times the corresponding error in Proposition 3. To wit, the error terms are

$$\ll M^k\Big(\sum_{p\in\mathcal{P}}\frac{h(p)}{p}\Big)^k\sum_{d\in D_k(\mathcal{P})}|r_d|.$$

As for the main term, note that $G(r) = 0$ unless r is square-full and so if $q_1 < q_2 < \dots < q_s$ are the distinct primes among the $p_1, \dots, p_k$ our main term is

$$x\sum_{s\le k/2}\sum_{\substack{q_1<\dots<q_s\\ q_i\in\mathcal{P}}}\sum_{\substack{\alpha_1,\dots,\alpha_s\ge 2\\ \sum_i\alpha_i=k}}\frac{k!}{\alpha_1!\cdots\alpha_s!}\prod_{i=1}^{s} g(q_i)^{\alpha_i}G(q_1^{\alpha_1}\cdots q_s^{\alpha_s}). \tag{16}$$

When k is even there is a term with $s = k/2$ and all $\alpha_i = 2$ which is the leading contribution to (16). This term contributes

$$x\frac{k!}{2^{k/2}(k/2)!}\sum_{\substack{q_1,\dots,q_{k/2}\in\mathcal{P}\\ q_i \text{ distinct}}}\prod_{i=1}^{k/2} g(q_i)^2\frac{h(q_i)}{q_i}\Big(1-\frac{h(q_i)}{q_i}\Big).$$

If we fix $q_1, \dots, q_{k/2-1}$, then the sum over $q_{k/2}$ is $\sigma_{\mathcal{P}}(g)^2 + O(M^2k)$, since $|g(p)| \le M$ for all $p \in \mathcal{P}$, and $0 \le h(p) \le p$. Therefore the contribution of the term $s = k/2$ to (16) is

$$C_k x(\sigma_{\mathcal{P}}(g)^2 + O(M^2k))^{k/2} = C_k x\sigma_{\mathcal{P}}(g)^k\Big(1 + O\Big(\frac{M^2k^2}{\sigma_{\mathcal{P}}(g)^2}\Big)\Big),$$

since $kM \le \sigma_{\mathcal{P}}(g)$.

Now consider the terms $s < k/2$ in (16). Since $|G(q_1^{\alpha_1}\cdots q_s^{\alpha_s})| \le \prod_{i=1}^{s}(h(q_i)/q_i)(1-h(q_i)/q_i)$, and $\prod_{i=1}^{s}|g(q_i)|^{\alpha_i} \le M^{k-2s}\prod_{i=1}^{s}|g(q_i)|^2$, we see that

these terms contribute an amount whose magnitude is

$$\le x \sum_{s<k/2} \frac{k!}{s!} M^{k-2s} \Big(\sum_{q\in\mathcal{P}} |g(q)|^2 \frac{h(q)}{q}\Big(1-\frac{h(q)}{q}\Big)\Big)^s \sum_{\substack{\alpha_1,\dots,\alpha_s\ge 2\\ \sum \alpha_i = k}} \frac{1}{\alpha_1!\cdots\alpha_s!}$$

$$\le x \sum_{s<k/2} \frac{k!}{s!2^s} \binom{k-s}{s} M^{k-2s} \sigma_{\mathcal{P}}(g)^{2s},$$

using that $\binom{k-s}{s}$ equals the number of ways of writing $k = \sum \alpha_i$ with each $\alpha_i \ge 2$. The proposition follows.

One way to apply Proposition 4 is to take $\mathcal{P}$ to be the set of all primes below z with $|g(p)|$ small. If there are not too many values of p with $|g(p)|$ large, then we would expect that $g(a)$ is roughly the same as $g_{\mathcal{P}}(a)$ for most a. In such situations, Proposition 4 which furnishes the distribution of $g_{\mathcal{P}}(a)$ would also furnish the distribution of $g(a)$. In this manner one can deduce the result of Kubilius and Shapiro (Shapiro, 1956) which is a powerful generalization of the Erdős–Kac theorem for additive functions. Indeed we can derive such a Kubilius–Shapiro result in the more general sieve theoretic framework given above, and for all additive functions rather than only for the subclass of strongly additive functions.

There are many other interesting number theory questions in which an Erdős–Kac type theorem has been proved. We have collected some of these references below[1] and invite the reader to determine which of these Erdős–Kac type theorems can be deduced from the results given herein. The reader may also be interested in the textbooks (Elliott, 1979; Kubilius, 1964; Tenenbaum, 1995) for a more classical discussion of some of these issues, and to the elegant essays (Billingsley, 1973; Kac, 1959).

Acknowledgements

Le premier auteur est partiellement soutenu par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada. The second author is partially supported by the National Science Foundation.

References

Alladi, K. (1987) An Erdős–Kac theorem for integers without large prime factors, *Acta Arith.* **49**, 81–105.

Billingsley, P. (1973) Prime numbers and Brownian motion, *Amer. Math. Monthly* **80**, 1099–1115.

[1] Thanks are due to Yu-Ru Liu for her help with this.

David, C. and Pappalardi, F. (1999) Average Frobenius distributions of elliptic curves, *Internat. Math. Res. Notices* **1999**, 165–183.

Elliott, P. D. T. A. (1979) *Probabilistic number theory.* Vol. I. and II, Vol. 239 and 240 of *Grundlehren Math. Wiss.*, New York–Berlin, Springer.

Elliott, P. D. T. A. and Sárkőzy, A. (1997) The distribution of the number of prime divisors of numbers of form $ab + 1$, In *New trends in probability and statistics.* Vol. 4, Palanga, 1996, pp. 313–321, VSP, Utrecht.

Erdős, P. (1935) On the normal order of prime factors of $p-1$ and some related problems concerning Euler's φ-functions, *Quart. J. Math.(Oxford)* **6**, 205–213.

Erdős, P. and Kac, M. (1940) The Gaussian law of errors in the theory of additive number theoretic functions, *Amer. J. Math* **62**, 738–742.

Erdős, P., Maier, H., and Sárkőzy, A. (1987) On the distribution of the number of prime factors of sums $a + b$, *Trans. Amer. Math. Soc* **302**, 269–280.

Erdős, P. and Pomerance, C. (1985) On the normal number of prime factors of $\varphi(n)$, *Rocky Mountain J. Math* **15**, 343–352.

Erdős, P. and Wintner, A. (1939) Additive arithmetical functions and statistical independence, *Amer. J. Math.* **61**, 713–721.

Halberstam, H. (1955) On the distribution of additive number theoretic functions. I, *J. London Math. Soc.* **30**, 43–53.

Halberstam, H. (1956) On the distribution of additive number theoretic functions. III, *J. London Math. Soc.* **31**, 15–27.

Hardy, G. H. and Ramanujan, S. (1917) The normal number of prime factors of a number n, *Quar. J. Pure. Appl. Math* **48**, 76–97.

Hensley, D. (1994) The number of steps in the Euclidean algorithm, *J. Number Theory* **49**, 142–182.

Hildebrand, A. (1987) On the number of prime factors of integers without large prime divisors, *J. Number Theory* **25**, 81–106.

Kac, M. (1959) *Statistical independence in probability, analysis and number theory*, Vol. 12 of *Carus Math. Monogr.*, New York, Math. Assoc. America.

Khan, R. (2006) On the distribution of normal numbers, preprint.

Kubilius, J. (1964) *Probabilistic methods in the theory of numbers*, Vol. 11 of *Transl. Math. Monogr.*, Providence, RI, Amer. Math. Soc.

Kuo, W. and Liu, Y.-R. (2006) Erdős–Pomerance's conjecture on the Carlitz module, to appear.

Li, S. and Pomerance, C. (2003) On generalizing Artin's conjecture on primitive roots to composite moduli, *J. Reine Angew. Math.* **556**, 205–224.

Liu, Y.-R. (2004) A generalization of the Erdős–Kac theorem and its applications, *Canad. Math. Bull.* **47**, 589–606.

Liu, Y.-R. (2005)a A prime analogue of Erdős–Pomerance's conjecture for elliptic curves, *Comment. Math. Helv.* **80**, 755–769.

Liu, Y.-R. (2005)b Prime divisors of the number of rational points on elliptic curves with complex multiplication, *Bull. London Math. Soc* **37**, 658–664.

Mauduit, C. and Sárkőzy, A. (1996) On the arithmetic structure of sets characterized by sum of digits properties, *J. Number Theory* **61**, 25–38.

Montgomery, H. and Soundararajan, K. (2004) Primes in short intervals, *Comm. Math. Phys.* **252**, 589–617.

Murty, M. R. and Saidak, F. (2004) Non-abelian generalizations of the Erdős–Kac theorem, *Canad. J. Math* **56**, 356–372.

Murty, V. K. and Murty, M. R. (1984)a An analogue of the Erdős–Kac theorem for Fourier coefficients of modular forms, *Indian J. Pure Appl. Math.* **15**, 1090–1101.

Murty, V. K. and Murty, M. R. (1984)b Prime divisors of Fourier coefficients of modular forms, *Duke Math. J.* **51**, 57–76.

Sathe, L. G. (1953) On a problem of Hardy on the distribution of integers having a given number of prime factors. II., *J. Indian Math. Soc. (N.S.)* **17**, 83–141.

Selberg, A. (1954) Note on a paper by L. G. Sathe, *J. Indian Math. Soc. (N.S.)* **18**, 83–87.

Shapiro, H. (1956) Distribution functions of additive arithmetic functions, *Proc. Nat. Acad. Sci. USA* **42**, 426–430.

Tenenbaum, G. (1995) *Introduction to analytic and probabilistic number theory*, Vol. 46 of *Cambridge Stud. Adv. Math.*, Cambridge, Cambridge University Press.

Turán, P. (1934) On a theorem of Hardy and Ramanujan, *J. London Math. Soc.* **9**, 274–276.

INDEX