Recent progress in multiplicative number theory

Kaisa Matomäki
Maksym Radziwiłł

August 4, 2018

Single averages

Analytic number theory has a well developed theory for understanding single averages such as,

$$1)\sum_{p\leq x} 1$$

$$2)\sum_{p\leq x} (a(p)-(p+1))^2$$

$$3)\sum_{n\leq x} \#\{(a,b)\in\mathbb{Z}^2 : a^2+b^2=n\}$$

where

$$a(p)=\#\{(x,y)\in\mathbb{F}_p^2 : y^2\equiv x^3+69x-5290 \pmod p)\}$$

Single averages

1. In each case the problem is reduced to understanding analytic properties of the underlying L-function (i.e a generating series of the form $\sum a(n)n^{-s}$ that has both an Euler product and a functional equation) : location of zeros, functional equations, growth rate on vertical lines, ...
2. Each of these properties is sufficiently well understood that we can almost always give a satisfying qualitative answer as to the behavior of
$$\sum_{n\leq x} a(n)$$
where $a(n)$ is some sequence of "arithmetical interest"
3. When we cannot the difficulty is often algebraic in nature (for instance, no modularity is known)

Single averages

The many challenging question that remain are of quantitative nature:

1. What is the best bound towards,

$$\sum_{p \leq x} 1 - \int_2^x \frac{dt}{\log t} \quad ?$$

2. Is there a $\delta > 0$ such that,

$$\sum_{n \leq x} \lambda_f(n) \chi_{-d}(n) \ll x^{1-\delta}$$

as soon as $x > d^{1-\delta}$ where f is a fixed holomorphic modular form and χ_{-d} is the Kronecker symbol with conductor d?

These questions should not be underestimated: An optimal bound in the first question is equivalent to the Riemann Hypothesis. An affirmative answer to the second question is (morally) equivalent to proving the equidistribution of lattice points lying on the surface of a 3-dimensional sphere of radius d (and now a theorem thanks to Duke, Fomenko-Golubeva and Iwaniec)

Correlations

1. However we lack a similar structural understanding for questions related to <u>correlations</u>,

$$\sum_{n \leq x} a(n)a(n+h)$$

   with $a(n)$ a sequence of “arithmetical interest”.

2. The most notorious problem that falls under this category is the twin prime conjecture, which in its qualitative form asserts that,

$$\sum_{\substack{p \leq x \\ p+2 \text{ is prime}}} 1 \sim \frac{Cx}{\log^2 x}$$

   as $x \to \infty$ with $C > 0$ an absolute constant.

3. The best result that we have in this direction results from the work of Zhang, Maynard, Tao and the polymath group. It asserts that there are at least $x(\log x)^{-2018}$ primes $p \leq x$ for which there exists another prime q with $|p - q| \leq 300$.

Correlations

1. There appears to be at first no connection to L-functions, for instance the Dirichlet series

$$\sum_{p+2 \text{ is prime}} \frac{1}{p^s}$$

has no functional equation and no Euler product. It's hardly an L-function and thus the framework used to understand single averages is of little use.

Correlations

1. Instead to tackle a problem such as
$$\sum_{\substack{p \leq x \\ p+2 \text{ is prime}}} 1$$
a natural instinct is to appeal to combinatorial formulas.
2. A "combinatorial formula" decomposes primes into linear combinations of hopefully simpler objects. These come roughly in two flavors.
$$(\mathit{Linnik}) \quad \mathbf{1}_{n \text{ is prime}} = -\sum_{k} \frac{(-1)^k}{k} d_k^{\star}(n)$$
$$(\mathit{Trivial}) \quad \frac{\log n}{k} \cdot \mathbf{1}_{n = p^k} = \sum_{d \mid n} \log d \cdot \mu(d)$$
3. Here $d_k(n) = \sum_{n = n_1 \ldots n_k} 1$ is the number of ways of writting n as a product of k integers, and $d_k^{\star}(n)$ is the variant that excludes from the count products involving 1. The Möebius function $\mu(n)$ is defined as $(-1)^{\omega(n)}$ when n is square-free ($\omega(n)$ is the number of prime factors of n) and 0 otherwise.

Correlations

1. Roughly speaking these combinatorial formulas show that if we could estimate,
$$\sum_{n\leq x} d_k(n)d_k(n+h) \tag{1}$$
for all integer $k\geq 1$ and shifts $h\in\mathbb{N}$ then we would be able to establish twin primes.
2. Alternatively (still morally speaking) if we could show that,
$$\sum_{n\leq x}\mu(n)\mu(n+h)=o(x) \tag{2}$$
then we also should be able to obtain twin primes.
3. The functions $d_k(n)$ and $\mu(n)$ are multiplicative, that is they satisfy $f(ab)=f(a)f(b)$ whenever a,b are co-prime. We have thus gained a little bit of additional structure in this reduction.
4. Handling (2) appears harder than (1) since μ is connected to the zeros of the Riemann zeta-function while d_k is more connected to the growth of the Riemann zeta function on vertical lines.

Correlations

1. The fundamental difficulty when investigating correlations like these is that we are asking for information about the multiplicative structure of consecutive integers.
2. When $a(n)$ is defined in a multiplicative way we hope to show that $a(n)$ and $a(n+h)$ are on average independent of each other.
3. As a rough heuristic, in those cases, we expect that

$$\frac{1}{x}\sum_{n\le x} a(n)a(n+h) \approx \Big(\frac{1}{x}\sum_{n\le x} a(n)\Big)^2$$

That is, on average, $a(n)$ and $a(n+h)$ are independent of each other. Specifically in the previous cases we expect,

$$\frac{1}{x}\sum_{n\le x} d_k(n)d_k(n+h) \approx \Big(\frac{1}{x}\sum_{n\le x} d_k(n)\Big)^2 \approx (\log x)^{2k-2}$$

and that,

$$\frac{1}{x}\sum_{n\le x} \mu(n)\mu(n+h) \approx \Big(\frac{1}{x}\sum_{n\le x} \mu(n)\Big)^2 = o(1).$$

Correlations of the divisor function

1. The intuition that estimating $\sum_{n \leq x} d_k(n) d_k(n+h)$ should be easier than $\sum_{n \leq x} \mu(n)\mu(n+h)$ is correct at least when $k = 2$.
2. It turns out that the Dirichlet series,

   $$\sum_{n} \frac{d(n)d(n+h)}{(n(n+h))^s}$$

   admits an explicit and beautiful analytic continuation to the region $0 < \Re s < 1$ (work on this and related issues by Selberg, Kuznetsov, Vinogradov-Taktadzyan, Sarnak, Goldfeld, Jutila, ...).
3. The analytic continuation is described explicitely in terms of eigenvalues of the hyperbolic Laplacian on a certain arithmetic hyperbolic manifold (i.e $\mathrm{SL}_2(\mathbb{Z}) \backslash \mathbb{H}$).

Correlations of the divisor function

1. Using the analytic continuation of

$$\sum_{n} \frac{d(n)d(n+h)}{(n(n+h))^s}$$

and bounds for its growth inside the critical strip $0 < \Re s < 1$ one is able to obtain optimal results of the form,

$$\sum_{n \le X} d(n)d(n+h)\Big(1 - \frac{n}{X}\Big) = XP_h(\log X) + O(X^{1/2})$$

where P_h is a polynomial of degree 2. Moreover one can show that the error term cannot be improved.

Correlations of higher divisor functions

1. One might reasonably hope that a similar story will hold for
$$\sum_{n \leq x} d_k(n) d_k(n+h)$$
once we understand the spectral theory of automorphic forms on $GL(k)$ with $k \geq 3$ (and already just doing so for $k = 3$ would be important). If we could do so for every k we would most likely be able to obtain twin primes through the relation provided by the combinatorial formulas.
2. This is subject of intense on-going current (and past) research.
3. Unfortunately so far we are still not even able to obtain an asymptotic for
$$\sum_{n \leq x} d_3(n) d_3(n+h).$$
4. It appears that there are significant differences between $GL(2)$ and $GL(k)$ with $k \geq 3$.

Correlation - non-automorphically

1. In recent years it appears that we might have found a way to contourn the automorphic approach all-together (at least in important special cases).
2. Three years ago a somewhat widespread belief was that showing that
$$\Big| \sum_{n \leq x} \mu(n)\mu(n+h) \Big| \leq (1 - \delta_h) \sum_{n \leq x} |\mu(n)\mu(n+h)|$$
for all sufficiently large x and $\delta_h > 0$ is (morally) equivalent to twin primes
3. ... and that showing
$$\sum_{n \leq x} \mu(n)\mu(n+h) = o(x)$$
should be equivalent to quantitative twin primes.
4. What justifies this "moral belief" are the combinatorial formulas relating primes to $\mu(n)$. The implications are rigorous under a somewhat more stringent quantification
5. Moreover $\mu(n)$ is tightly related to zeros of $\zeta(s)$ just like the primes. Indeed $\sum \mu(n)n^{-s}$ and $\sum_p \log p \cdot p^{-s}$ share poles at exactly the same locations (except for $s = 1$), that is at the zeros of $\zeta(s)$.

Correlations - non-automorphically

The situation starts to change in 2015 and the state of the art as of now is that

1. For each $h \neq 0$ there exists a $\delta_h > 0$ such that for all x sufficiently large
$$\Big| \sum_{n \le x} \mu(n)\mu(n+h) \Big| \le (1 - \delta_h) \sum_{n \le x} |\mu(n)\mu(n+h)|$$
2. We have,
$$\sum_{n \le x} \frac{\mu(n)\mu(n+h)}{n} = o(\log x).$$
Consequently there exists an infinite subsequence $x \to \infty$ along which,
$$\sum_{n \le x} \mu(n)\mu(n+h) = o(x)$$
3. The first result follows from our work (i.e Matomäki-Radziwill), where-as the second is a result of Tao which builds on our result and a novel use of entropy in analytic number theory. I will now discuss both.

Correlations - non-automorphically

The main difference compared to the automorphic approach is that while the later is very heavy on automorphy, our approach almost exclusively relies on multiplicativity. This is however also its limitation. One would hope there is a way to combine both ...

Correlations - non-automorphically

1. Let us suppose now that we know nothing in the direction of showing that,
$$\sum_{n \leq X} \mu(n)\mu(n+h) = o(X) \tag{3}$$
and we would like to find a simpler problem that would allow us to "benchmark" our progress.
2. A problem of this type would be to attempt to bound,
$$\sum_{|h| \leq H} (H - |h|) \sum_{n \leq X} \mu(n)\mu(n+h) \tag{4}$$
for H as small as possible in terms of X.
3. The trivial bound is $H^2 X$. The bound we are aiming at is $o(H^2 X)$ because this is what we would get if we knew (3). The goal is to obtain the bound $o(H^2 X)$ for (4) for H as small as possible in terms of X. The smallest that we could hope for is H going to infinity with X arbitrarily slowly. Note also that the case $H = X$ is trivial from the prime number theorem.

Correlations - non-automorphically

Let us then consider this "easier problem" in more detail. Note that, for $H \geq 1$,

$$\sum_{|h| \leq H} (H - |h|) \sum_{n \leq X} \mu(n)\mu(n+h) = \sum_{x \leq X} \Big| \sum_{x \leq n \leq x+H} \mu(n) \Big|^2 \qquad (5)$$

1. This identity gives an equivalent way of viewing the problem of bounding the left-hand side of (5), namely we can re-phrase the problem as that of showing that for a density 1 subset of x we have,

$$\sum_{x \leq n \leq x+H} \mu(n) = o(H)$$

2. This is a consequence of our work when H is a function of X that grows to infinity arbitrarily slowly (and thus the smallest rate for which we could hope).

Main theorem

Theorem (Matomäki-Radziwill)

Let $f : \mathbb{N} \to \mathbb{N}$ be a multiplicative function with $|f| \leq 1$. Then, for a sequence of x of density 1 and for any monotonic $h = h(x)$ going to infinity arbitrarily slowly with x,

$$\frac{1}{h} \sum_{x \leq n \leq x+h} f(n) - \frac{1}{x} \sum_{x \leq n \leq 2x} f(n) = o(1)$$

1. Previously for $f = \mu$ this was known only for $h > x^{1/3}$ unconditionally and for $h > (\log x)^A$ under the Riemann Hypothesis (for some large fixed $A > 0$).
2. In view of the identity,

   $$\sum_{|h| \leq H} (H - |h|) \sum_{n \leq X} f(n) f(n+h) = \sum_{x \leq X} \Big| \sum_{x < n < x+H} f(n) \Big|^2$$

   the Theorem says that we are able to understand correlations, at least on average, from the knowledge of the behavior of the single average $\sum_{n \leq x} f(n)$. This is consistent with our previous heuristic.

Corollaries

Our theorem has several immediate consequences.

1. A multiplicative function f has a positive proportion of sign changes if and only if f is non-zero for a positive proportion of integers and there exists an integer $n \geq 1$ at which $f(n) < 0$.
2. For every $\varepsilon > 0$ there exists an $C(\varepsilon) > 0$ such that there is an integer $n \in [x, x + C(\varepsilon)\sqrt{x}]$ all of whose prime factors are $\leq n^{\varepsilon}$.
3. The last corollary is related to Lenstra's elliptic curve factoring algorithm which requires such a result with $C(\varepsilon) < 2$ (and unfortunately with an ε tending to zero with x rather rapidly...)

Main idea

The main inputs in our Theorem are the following:

1. Use of "anatomy of integers" : most integers n have a small prime factors p of size about n^{ε}.
2. Use of harmonic analysis to convert the problem to that of bounding,

$$\int_{\log X}^{X/H} \Big| \sum_{X \leq n \leq 2X} \frac{f(n)}{n^{1+it}} \Big|^2 dt.$$

3. Use of multiplicativity and the typical existence of a small prime divisor to introduce a bilinear structure in the Dirichlet polynomial $\sum_{X \leq n \leq 2X} f(n) n^{-1-it}$.
4. Handling of the bilinear structure through an 1) iterative scheme with (somewhat opaque) origins from sieve methods 2) distributional results for Dirichlet polynomials. Here 1) is important only for very small H which are the (crucial) cases that go beyond the Riemann Hypothesis.

Correlations on average

1. Soon after we came in contact with Terence Tao. He thaught us how to use our result to obtain stronger results on correlations of the form,
$$\sum_{|h|\le H}\Big|\sum_{n\le X}\mu(n)\mu(n+h)\Big| = o(HX)$$
for any H growing arbitrarily slowly with X.
2. Notice the absolute values!
3. The proof relies on being able to show that,
$$\sum_{x\le X}\Big|\sum_{x\le n\le x+H}\mu(n)e(\alpha n)\Big| = o(HX)$$
uniformly in $\alpha\in\mathbb{R}$ and for H growing to infinity arbitrarily slowly with X. For fixed $\alpha\in\mathbb{Q}$ this is an extension of my result with Matomäki. For fixed $\alpha\in\mathbb{Q}$ this is a consequence of ideas of Daboussi-Delange, inspired by ideas of Vinogradov.

Correlations on average

1. These result establish

$$\sum_{n\leq x}\mu(n)\mu(n+h)=o(x)$$

   for a density one subset of $|h|\leq H$ with H going to infinity with x arbitrarily slowly.

2. Can we obtain similar results for

$$\sum_{n\leq x}d_k(n)d_k(n+h)$$

   and

$$\sum_{p\leq x}\log p\cdot d_k(p+h)\ ?$$

3. Both are again naturally related to twin prime type conjectures. An additional difficulty is the presence of a main term and the unboundedness (in fact "sparsity of support") of $d_k(n)$.

Correlations on average

1. It is conjectured that for each k there exists a $\delta_k > 0$ and polynomials P_k and Q_k respectively of degree $2k - 2$ and $k - 1$ such that,
$$\sum_{n \leq x} d_k(n) d_k(n + h) = x P_k(\log x) + O(x^{1 - \delta_k})$$
and
$$\sum_{p \leq x} \log p \cdot d_k(p + h) = x Q_k(\log x) + O(x^{1 - \delta_k})$$
2. Morally speaking both conjecture imply a quantitative form of twin primes.
3. On average thus far they are known only when averaging over at least $h > x^{0.24242424\ldots}$ shifts h with strong error terms.

Theorem (Matomäki-Radziwill-Tao)

For any $k \geq 2$ integer, we have,

$$\sum_{|h| \leq H} \Big| \sum_{n \leq X} d_k(n) d_k(n+h) - XP_k(\log X) \Big| = o(HX(\log X)^{2k-2})$$

$$\sum_{|h| \leq H} \Big| \sum_{p \leq X} \log p \cdot d_k(p+h) - XQ_k(\log X) \Big| = o(HX(\log X)^{k-2})$$

for $H > (\log X)^{10000k \log k}$ where P_k is a polynomial of degree $2k-2$ and Q_k a polynomial of degree $k-1$.

1. The correlations
$$\sum_{n \leq x} d_k(n) d_k(n+h)$$
arise in averaged form in the problem of estimating moments of the Riemann-zeta function $\int_T^{2T} |\zeta(\frac{1}{2} + it)|^{2k} dt$.
2. In fact our results are sufficient to imply corollaries for moments of Dirichlet L-functions (specifically an asymptotic for the eight moment averaged over characters and moduli).

Removing the averaging

1. Despite the very short averaging, without new ideas one cannot pass from the results on correlations “on average” to results for individual correlations.
2. A key insight was introduced in the work of Tao, and allowed him to prove the following theorem.

Theorem (Tao)

Let $h \neq 0$ be given. Then, as $x \to \infty$,

$$\sum_{n \leq x} \frac{1}{n} \cdot \mu(n)\mu(n+h) = o(\log x).$$

Consequently for each fixed $h \neq 0$ there exists an subsequence $x_n \to \infty$ along which we have

$$\sum_{m \leq x_n} \mu(m)\mu(m+h) = o(x_n).$$

(On-going work of Tao-Teräväinen to say something about the density of such x and uniformity in h)

Tao's ideas

1. Tao's idea is to use the logarithmic weights and multiplicativity to introduce a third variable,
$$\sum_{n\leq x}\frac{\mu(n)\mu(n+h)}{n}\approx\frac{1}{\log P}\sum_{\substack{n\leq x\\ p\approx P}}\frac{1}{n}\cdot\mu(n)\sum_{p|n}\mu(n+p)$$
2. The second idea is the introduction of an "entropy decrement argument" which allows to show the existence of a very small P at which for most integers $n\leq x$,
$$\sum_{p|n}\mu(n+p)\approx\sum_{p\approx P}\frac{1}{p}\cdot\mu(n+p)$$
i.e there is a scale P at which the event $p|n$ with $p\approx P$ and $n\leq x$ is independent from the sign patterns of $(\mu(n),\mu(n+1),\ldots,\mu(n+P))$.
3. The third idea is the use of my work with Matomäki to handle the resulting ternary problem,
$$\frac{1}{\log P}\sum_{p\approx P}\sum_{n\leq x}\frac{\mu(n)\mu(n+p)}{n}=o(\log x)$$

Erdös discrepancy problem

1. This result was used to resolve the Erdös discrepancy problem in combinatorics. The conjecture states that for any sequence x_n of ± 1 and any $\Delta > 0$ there exists a $k \in \mathbb{N}$ and a $N \in \mathbb{N}$ such that,
$$\Big| \sum_{n \leq N} x_{kn} \Big| \geq \Delta.$$
2. It turns out that this problem can be reduced to the case of $x_n = f(n)$ with f a completely multiplicative function (i.e $f(ab) = f(a)f(b)$). Thus it's enough to show that it is impossible to have for all N and h,
$$\Big| \sum_{N \leq n \leq N+h} f(n) \Big| = O(1) \tag{6}$$
3. But squaring and averaging (6) with logarithmic weights implies that
$$\sum_{n \leq N} \frac{1}{n} f(n) f(n+h) \gg \log N.$$
4. Inspecting Tao's proof one can see that this implies that f needs to "pretend" to be a character. These cases can be ruled out by direct computation. Further work on this by Klurman and Mangerel.

Higher correlations

1. There are still many things aspects of these methods that we do not understand.
2. The case of higher correlations, such as

$$\sum_{n \leq x} \mu(n+h_1) \ldots \mu(n+h_k)$$

   remains poorly understood. There has been recent progress by Tao-Teräväinen when k is odd. To address this problem for all k one needs to establish "local Fourier uniformity",

$$\sum_{x \leq X} \sup_{\alpha} \Big| \sum_{x \leq n \leq x+h} \mu(n) e(\alpha n^{\ell}) \Big| = o(XH)$$

   for all $\ell \geq 1$ and already $\ell = 1$ remains a very significant challenge.
3. The case of higher correlations is essentially equivalent to Sarnak's conjecture asserting that

$$\sum_{n \leq N} \mu(n) x_n = o(N)$$

   for any sequence of x_n of topological entropy 0

Primes

1. A remaining and important limitation is that despite all the progress on correlations of the form
$$\sum_{n\leq x}\mu(n)\mu(n+h) \text{ or } \sum_{n\leq x}d_k(n)d_k(n+h)$$
these method do not (yet?) apply to directly attack primes.
2. One can use this method to obtain new results for products of exactly k primes (as soon as $k\geq 2$, by works of Goudout and Teräväinen) thus going beyond what the sieve delivers. Nonetheless the method fails short at the primes.
3. The basic reason for this failure is that we rely on the existence of a small prime factor p of size n^{ε} for most integers in our sets of interest. This is patently not true for primes.
4. For instance an important problem that remains untouched is that of showing that there exists a $\delta > 0$ such that,
$$\sum_{x\leq p\leq x+x^{1/6-\delta}} 1 \sim \frac{x^{1/6-\delta}}{\log x}$$
for almost all x (that is a subsequence of density 1).

Problems

In my opinion the central problems that remain (and unfortunately we are still at a stage where there are more problems than solutions) are to

1. Establish "local Fourier uniformity",

$$\sum_{x \le X} \sup_{\alpha} \Big| \sum_{x \le n \le x+H} \mu(n) e(\alpha n^k) \Big| = o(HX)$$

   that is address the case of higher correlations (either with logarithmic weights or not).

2. Show that,

$$\sum_{n \le x} \mu(n)\mu(n+h)$$

   that is remove the logarithmic weights.

3. Show that there exists a sequence of $x \to \infty$ on which

$$\sum_{n \le x} d_k(n) d_k(n+h) \sim Cx(\log x)^{2k-2}$$

4. ... and to find a way to address the case of prime numbers (and anything new and non-trivial is welcome).