The integrated explicit analytic number theory network

Terence Tao

May 2026

Mathematics at scale

- The traditional unit of mathematical research output is the paper.
- Mathematical literature as a whole consists of many papers (and monographs, etc.) linked together.
- But aside from some bibliographic databases (MathSciNet, ZbMath, etc.) or mathematical databases (OEIS, LMFDB, etc.) there are few systematic efforts to organize the literature at larger scales than an individual paper.

- The Integrated Explicit Analytic Number Theory Network is an ongoing experiment to see if advances in autoformalization can create a useful new type of mathematical artefact at the scale of multiple papers, rather than a single paper.
- The vision is to combine a large number of papers in this area into a single living "spreadsheet", formalized in Lean.

What is explicit analytic number theory?

- A core part of analytic number theory is concerned with asymptotic bounds involving quantities relating to the prime numbers.
- For instance, one form of the prime number theorem with classical error term that

$$\theta(x) = x + O(x \exp(-c\sqrt{\log x}))$$

  for some $c > 0$, where $\theta(x) := \sum_{p \leq x} \log p$ is the Chebyshev theta function.
- The implied constants in the asymptotic notation are usually left unspecified, as they are quite tedious to compute.

Generally speaking, these sorts of asymptotic bounds can be satisfactory when analyzing sufficiently large numbers, so long as one does not care to make the threshold for "sufficiently large" explicit. A typical application is

Vinogradov's theorem

Every sufficiently large odd number can be expressed as the sum of three primes.

But when one wants to use analytic number theory to prove a result for *all* numbers, not just the sufficiently large ones, then one needs to make the implied constants in the asymptotic bounds *explicit*, combined with numerical verification to handle small or medium cases. For instance, this was used t prove the non-asymptotic form of Vinogradov's theorem:

Odd Goldbach conjecture (Helfgott 2013)

Every odd number greater than 5 can be expressed as the sum of three primes.

Here is a typical result in explicit analytic number theory:

Explicit prime number theorem (Fiori-Kadiri-Swidinsky 2022)

One has

$$|\theta(x) - x| \leq 121.0961 \left(\frac{\log x}{R}\right)^{3/2} \exp(-2\sqrt{\log x/R})$$

for all $x \geq 2$, where $R = 5.5666305$.

In practice, such results only are useful for extremely large x, e.g., $x \geq e^{100}$, which is too large a threshold to cover remaining ranges by purely numerical methods. So one also seeks "medium-range" estimates, such as the following:

Explicit prime number theorem (Fiori-Kadiri-Swidinsky 2022)

One has

$$|\theta(x) - x| \leq 5.9651 \times 10^{-7} x$$

when $x \geq e^{30}$.

The paper of Fiori-Kadiri-Swidinsky 2022 contains large tables of results of this type.

The explicit analytic number theory literature is highly interdependent. For instance, there is a broad pipeline from results in each of the categories below to the next:

- Zeta function estimates: zero-free regions, zero density estimates, etc. for the Riemann zeta function $\zeta(s)$.
- Primary explicit estimates: explicit bounds on the Chebyshev function $\psi(x) = \sum_{n \leq x} \Lambda(n)$ or the Mertens function $M(x) = \sum_{n \leq x} \mu(n)$, both of which are tied to the zeta function by various "explicit formulae".
- Secondary explicit estimates: explicit bounds on other expresions relating to primes, such as the second Chebyshev function $\theta(x) = \sum_{p \leq x} \log p$, the prime counting function $\pi(x)$, or bounds on prime gaps.
- Applications: explicit results in number theory that rely on the above estimates

A typical zeta function estimate:

Explicit zero-free region (Mossinghoff–Trudgian 2015)

There are no zeroes $\zeta(s)$ of the zeta function with

$$\mathrm{Re}(s) \geq 1 - \frac{R}{\mathrm{Im}(s)}$$

and $R = 5.5666305$.

A typical primary explicit estimate:

Explicit prime number theorem (Fiori-Kadiri-Swidinsky 2022)

One has

$$|\psi(x) - x| \leq 121.096 \left(\frac{\log x}{R}\right)^{3/2} \exp(-2\sqrt{\log x/R})$$

for all $x \geq 2$, where $R = 5.5666305$.

A typical secondary explicit estimate:

Explicit prime number theorem (Fiori-Kadiri-Swidinsky 2022)

One has

$$|\theta(x) - x| \leq 121.0961 \left(\frac{\log x}{R}\right)^{3/2} \exp(-2\sqrt{\log x/R})$$

for all $x \geq 2$, where $R = 5.5666305$.

A typical tertiary explicit estimate. A highly abundant number N is a number for which the sum of divisors $\sigma(N)$ is larger than $\sigma(M)$ for all $M < N$.

Is $\mathrm{LCM}(1, \ldots, n)$ highly abundant? (MathOverflow, 2025)

The least common multiple of the first n positive integers is highly abundant if and only if $n \leq 70$, $81 \leq n \leq 96$, $125 \leq n \leq 148$, and $169 \leq n \leq 172$.

Secondary explicit estimates were needed to cover the range $n \geq 89693^2$; the remaining cases require computer computation. The entire theorem is verified in Lean (contingent on a tertiary explicit estimate of Dusart (2018)).

- Because of this interdependence, numerical errors in one explicit paper could propagate to other papers, leading to potential trust issues with the literature.
- On the other hand, these results are quite useful for various applications (e.g., Erdős problems).
- This suggests a good use case for formalization.
- However, the task of carefully going through the arguments is quite tedious, particularly with regards to the extensive mechanical computations involved.
- This suggests a good use case for auto-formalization.

- The literature updates over time due to various numerical or technological improvements, and many of the intermediate results are of no interest to highly curated formal repositories such as Mathlib.
- Furthermore, these formalization tasks are unlikely to compete with existing human-led projects.
- This makes for a opportunity to deploy larger amounts of AI assistance than would otherwise be advisable for a mathematical project.

For this discussion it may be useful to distinguish several levels of quality of proofs:

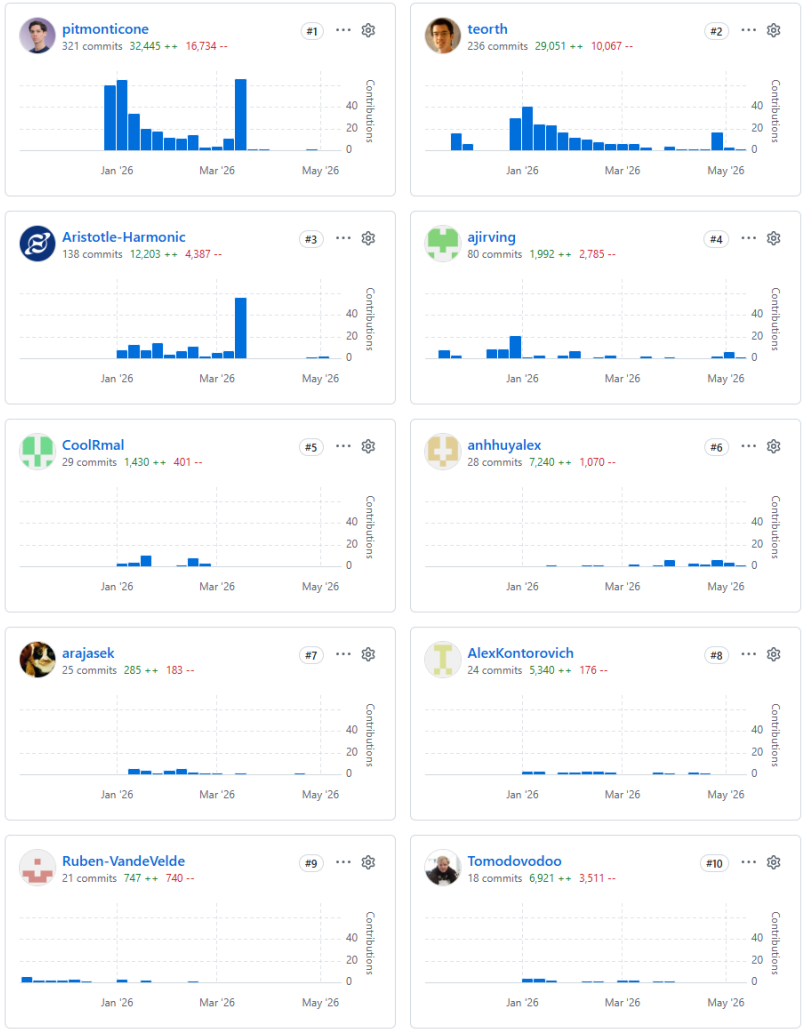
- Canonical-level: Carefully curated proofs that prioritize reusability, maintainability, broad applicability, and coherent style. (Examples: Mathlib, Bourbaki)
- Publication-level: Still carefully curated, but prioritizing specific applications. Many components of at this level could eventually be upstreamed to a canonical-level repository. (Examples: PNT+, published papers)
- Prototype-level: Loosely organized formalization that prioritizes rapid development and iteration, with some components eventually upstreamed to higher-level repositories. (Examples: IEANTN, lecture notes)
- Bare proof: Completely unregulated in structure; establishes that a theorem is true or that formalization is possible, but proofs are difficult to directly upstream to higher-level repositories.

The Integrated Explicit Analytic Number Theory Network

I launched the Integrated Explicit Analytic Number Theory Network (IEANTN) in 2025, hosted within the existing Prime Number Theorem and more (PNT+) formalization project.

- Phase 1 of the project seeks to collect a large number of inter-related results in explicit analytic number theory and formalize them all in a single Lean repository.
- Phase 2 of the project seeks to use AI tools to auto-update the constants in one formalized paper given changes in the constants in the input papers, e.g., to feed in a new zero free region and see how downstream constants change.
- The dream is to obtain a "living spreadsheet" of explicit constants in the subject that can be updated on the scale of hours or days, rather than years, as well as rolled back in time.

Dozens of contributors (including three students under my co-supervision, Xinjie He, Rushil Raghavan, and Hyunsik Chae)



What have we formalized so far?

- Chebyshev's theorems $\psi(x), \theta(x) \asymp x$, $\pi(x) \asymp \frac{x}{\log x}$ (upstreamed to Mathlib!)
- Mertens' theorems (in progress, planned for Mathlib)
- Costa-Pereira's inequalities
- Applications of prime gap bounds to Goldbach-type problems
- Ramanujan's inequality $\pi(x)^2 < \frac{ex}{\log x}\pi\left(\frac{x}{e}\right)$ for sufficiently large x
- Erdős' 392 conjecture on well-balanced factorizations of $n!$

What have we formalized so far?

- First-order Euler-Maclaurin formula (planned for Mathlib)
- Statements (without proofs) of many literature results in the subject, including many from the TME-EMT wiki, and numerical work such as tables of prime gaps
- Classification of the highly abundant least common multiples of $1, \dots, n$
- (Conditional) "primary to secondary" pipelines of Fiori–Kadiri–Swidinsky and Buthe (in progress)
- General "zeta to primary" pipeline of Chirre and Helfgott (in progress)

The ongoing work on formalizing Riemann-Stieltjes integration will be quite helpful.

Some observations while formalizing:

- We needed to perform extensive numerical computations, for instance calculating numerical inequalities involving logarithms and other special functions.
- We found that resurrecting the ancient practice of logarithm tables to be useful: standalone Lean files consisting of nothing more than upper and lower bounds for numerical logarithms and similar quantities. Similarly we have a file of prime tables.
- The third-party packages LeanCert and PrimeCert were used to help generate these tables.
- More generally, moving computation-heavy tasks to files with minimal imports helped reduce build time for contributors, especially when working on unrelated parts of the project.

Some observations while formalizing:

- We uncovered a few errors in the literature. Most were minor. A few required consultation with the original authors, and some resulted in slight numerical weakening of the results.
- One issue was that many numerical tables in the literature were computed with floating-point arithmetic and were found to be formally incorrect in the final digit.
- To address this, the Lean formalization of each table now comes with a safety margin (e.g., 1.01, 1.02, etc.). Each bound is worsened by its safety margin, which is set to slowly increase when moving downstream in the pipeline. Thus, roundoff errors can be quickly resolved with acceptable losses by tweaking safety margins appropriately.

- Autoformalizers were very useful, but often exploited misformalized sorried statements from the literature that were false and could in fact be used to prove any statement.
- Human review was usually able to catch these issues immediately (the proof was suspiciously easy); in a few cases they only got detected when some later difficult results also were mysteriously easy to autoformalize.
- Autoformalizers were also strong at detecting misformalizations that made a statement false due to easy counterexamples.
- Some experimental uses of autoformalizer tools to try to auto-detect such misformalizations in the entire repository had reasonable success, but such sweeps were far from comprehensive.

- My three students Rushil Raghavan, Xinjie He, and Hyunsik Chae are developing highly AI-assisted pipelines that can take, say, a large number of existing bounds on $\psi(x)$ and produce good bounds on (say) $\theta(x)$ for various ranges on x, by modifying existing arguments in the repository.
- Currently, the pipelines are slow, with several hours of AI computation needed to locate working modified arguments. And they are not guaranteed to be optimal.
- But when the parameters are only perturbed by a small amount, they work well; and my students are finding ways to efficiently preprocess the repository to make the pipelines more efficient.

Future goals

- Explicit zero-free regions
- Explicit zero-density estimates
- Tighter integration with the TME-EMT wiki
- More upstreaming of the cleaner results to Mathlib

Thanks for listening!