Large Gaps Between Primes

Everyone's heard about small gaps, but what about big ones?

Shihan Kanungo

Euler Circle, Palo Alto, CA 94306

Abstract

Let p_n be the n-th prime, and define the maximal prime gap $G(x)$ as

$$G(x) = \max_{p_n \leq x} (p_{n+1} - p_n).$$

We give a summary of the upper bounds that have been obtained for $G(x)$ over the last century, particularly Rankin's lower bound, and the improvement to it discovered independently by Ford-Green-Konyagin-Tao and Maynard in 2014. We go over a sketch of Rankin's, Ford-Green-Konyagin-Tao's, and Maynard's proofs of their bounds, omitting technical details but still presenting the main ideas.

1 Introduction

Prime numbers have long been known to be fundamental in the study of number theory. Around 300BC, the Greek Mathematician Euclid was one of the first to provide a treatise on prime numbers and proved several key facts about primes, including that there are infinitely many such prime numbers, as well as proving the well-known fundamental theorem of arithmetic. This was all written up in his textbook, *The Elements*, in what is considered today to be one of the most influential textbooks of all time. Since then, mathematicians have long held a keen interest in prime numbers, with the study of the distribution of primes being central to analytic number theory.

Notation and Conventions

As is standard convention, we shall use the notation $\log_n x$ to mean the n-th iterated logarithm. That is $\log_1 x = \log x$ and

$$\log_{n+1} x = \log(\log_n x) \quad \text{for all } n \geq 1.$$

We also make frequent use of Vinogradov's notation $f \ll g$ to mean $f = \mathcal{O}(g)$ and $f \gg g$ to mean $g = \mathcal{O}(f)$.

Let p_n denote the n-th prime number. For all $X \geq 2$, we define $G(X)$ as the maximal prime gap

$$G(X) = \max_{p_n \leq X} (p_{n+1} - p_n).$$

Since the early 20th century, many mathematicians have studied the growth rate of $G(X)$ and several results have been proven regarding the function. We shall first give a short history of the lower bounds that have been proven for $G(X)$ to date. We also consider some conjectural results about the growth rate of $G(X)$, as well as look at some recent computational results.

We start by giving an overview of the history of lower bounds for $G(X)$, in Table 1.

Bound for $G(X)$	Proved by
$G(X) \geq (1+o(1))\log X$	PNT and the pigeonhole principle
$G(X) \geq (2+o(1))\log X$	Backlund in 1929
$G(X) \geq (4+o(1))\log X$	Brauer-Zeitz in 1930
$G(X) \gg c\log X \frac{\log_3 X}{\log_4 X}$	Westzynthius in 1931
$G(X) \gg c\log X \frac{\log_2 X}{(\log_3 X)^2}$	Erdős in 1935
$G(X) \gg c\log X \frac{\log_2 \log_4 X}{(\log_3 X)^2}$	Rankin in 1938

Table 1. Early Bounds for $G(X)$

The c in Rankin's bound was originally $\frac{1}{3}$, but it was improved by several other authors.

Constant c	Proved by
$c = \frac{1}{2}e^{\gamma}$	Schönhage (1963)
$c = e^{\gamma}$	Ricci (1952), Rankin (1963)
$c = 1.31256e^{\gamma}$	Maier–Pomerance (1990)
$c = 2e^{\gamma}$	Pintz (1997)

Table 2. Early Bounds for c

Then, in 2014, Ford-Green-Konyagin-Tao and Maynard independently proved that the c in Rankin's bound could be taken to be arbitrarily large. In fact, they proved

$$G(X) \gg \frac{\log X \log_2 X \log_4 X}{\log_3 X}, \tag{1}$$

which is even stronger. In addition to outlining a proof of Rankin's bound, we shall also outline the key new contribution which is a generalisation of a hypergraph covering theorem from Pippenger, Spencer. This result is based on the *Rödl nibble* method.

In 1920, Cramér used a probabilistic model of the primes to conjecture that

$$G(X) \gg \log^2 X.$$

The model was tweaked by Granville in 1995, and the new model predicted the lower bound

$$G(X) \geq (2e^{-\gamma} - o(1))\log^2 X.$$

The weaker version $G(X) = \mathcal{O}(\log^2 X)$ is called **Cramér's conjecture**.

2 Some Elementary Bounds

To get a lower bound for $G(X)$, we can start by using the Pigeonhole principle and the Prime Number Theorem. The Prime Number Theorem tells us that there are $\sim \frac{X}{\log X}$ primes in $[1, X]$. If the gap between them were to be less than $(1 + o(1)) \log X$, then the spacing between the first and last primes in $[1, X]$ would be strictly less than

$$\frac{X}{\log X} \cdot (1 + o(1)) \log X = X + o(X).$$

This is false because we can choose X to be a prime number, and the spacing would be equal to $X - 2 = X + o(X)$. Thus, $G(X) \geq 1 + o(1) \log X$.

We now look at another way to show that the are arbitrarily long gaps in the primes. Consider the sequence

$$n! + 2, n! + 3, \ldots, n! + n.$$

Note that every number in this sequence must be composite, since 2 divides the first number, 3 divides the second number, and so on. Thus the biggest prime less than $n! + 2$ and the smallest prime more than $n! + n$ are consecutive, and separated by a gap of at least n. Thus $G(X) = n$, where $p < X$, and p is the smallest prime more than $n! + n$. Using Stirling's formula, we see that

$$X \approx n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

Then, we see that

$$n = \frac{n \log n}{\log n} \approx \frac{n \log \left(\frac{n}{e}\right)}{\log \left(n \log \left(\frac{n}{e}\right)\right)} \approx \frac{\log \left(\sqrt{2\pi n} \left(\frac{n}{e}\right)^n\right)}{\log \log \left(\sqrt{2\pi n} \left(\frac{n}{e}\right)^n\right)} \approx \frac{\log X}{\log \log X}.$$

Thus we get the bound

$$G(X) \geq (1 + o(1)) \frac{\log X}{\log \log X}.$$

This is in fact much worse than the bound given by the Prime Number Theorem, but our implementation of the idea is nowhere as efficient as it can be. The first step is to replace the factorial $n! = \prod_{j \leq n} j$ with the **primorial** $n\# := \prod_{p \leq n} p$. We have

$$\log n\# = \sum_{p \leq n} p = \theta(n) \sim n, \qquad \text{(by PNT)}$$

so $n\# \sim e^n$. Then, replacing the factorial with the primorial in our previous argument, we get $G(X) \geq n$, with $X \approx n\# = e^n$. Thus, we have the bound

$$G(X) \geq (1 + o(1)) \log X.$$

Notice that we have recovered our original bound for $G(x)$!

3 Improving the Bound

Now let us pause for a moment and consider why the primorial strategy works. The key point is that the numbers

$$2 + n\#,\ 3 + n\#,\ \ldots,\ n + n\#,$$

are all composite since each number from 2 to n is divisible by a prime $p \leq n$. Thus the residue classes $0 \pmod p$ for $p \leq n$ cover the set $\{2, \ldots, n\}$.

To generalize this idea, suppose that the residue classes $a_p \pmod p$ for $p \le x$ cover the set $\{1, \ldots, y\}$. By the Chinese Remainder Theorem, there exists a number $k \le \prod_{p \le x} p = x\#$ with $k \equiv -a_p \pmod p$ for $p \le x$. Then, for each $i \in \{1, \ldots, y\}$ there is $p \le x$ with $i \equiv a_p \pmod p$, so $k + i \equiv 0 \pmod p$; hence $k + i$ is composite. Thus the set $\{k+1, \ldots, k+y\}$ consists of entirely composite numbers. Repeating our previous argument, this gives us that $G(k) \ge y$. Since $k \le x\#$, we have $G(x\#) \ge y$. To generalize our old argument, suppose that $Y(x)$ is the maximum possible y. Then, since $x\# \sim e^x$, we have $G(e^x) \ge Y((1-\varepsilon)x)$ for any $\varepsilon > 0$ and sufficiently large x. Taking logs, we get

$$G(x) \ge Y((1-\varepsilon)\log x)$$

for sufficiently large x and $\varepsilon > 0$. This lets us get a bound for $G(x)$ if we have a bound for $Y(x)$. Note that if $Y(x) \sim x$, we get our old bound. But we can do better!

To visualize the problem, consider the following "shooting" metaphor. Suppose we have y ducks numbered $\{1, \ldots, y\}$. Our goal is to shoot all the ducks. To do this, we are given a special rifle for each prime $p \le x$. When we fire the rifle for the prime p, we can shoot all the ducks corresponding to one residue class modulo p. We can shoot each rifle only once. Then, to find the maximum possible value of y, we want to find the best (most efficient) "shooting strategies." To do this, we will have to use many tricks and observations.

Observation 1 (Most of the Ducks)

The first observation is to note that we only need to shoot *most* of the ducks, rather than *all* of them, provided we use slightly fewer primes.

Suppose we have used the primes $p \le \frac{x}{2}$ to shoot most of the ducks, and suppose there are N ducks remaining. What is the largest value of N we can have that still allows us to use the remaining primes from $\frac{x}{2}$ to x to shoot the N ducks?

Well, we don't know anything about these N ducks, so it seems hard to show that we can can shoot two birds with one stone, i.e use a single prime to shoot two or more ducks. But, it is obvious that we can always shoot a single duck in each shot. Thus if there are M primes from $\frac{x}{2}$ to x, we can shoot down at least M ducks. To find out what M is, we just need to remember the Prime Number Theorem, which tells us that

$$N = M \approx \frac{x}{2\log x}.$$

Thus, if we use just the primes $\le \frac{x}{2}$, we only need to shoot $N - \frac{x}{2\log x}$ of the ducks.

Observation 2 (Sieve of Eratosthenes)

One way to think about our setup of shooting ducks is as a sort of sieve: we try to sieve out *all* the numbers from $\{1, \ldots, y\}$ using residue classes of the primes. This reminds us of the Sieve of Eratosthenes.

The Sieve of Eratosthenes is a process to sieve out all the composite numbers from 2 to x, leaving us with the primes from 2 to x. What we do is start from the first number, which is 2, add it to our list of primes, and delete all the other multiples of 2. Then we take the next number, 3, add it to our list of primes, and delete all other multiples of 3. Continue this process until we reach the end of the list, and we have sieved out all the primes. This is illustrated in Figure 1.

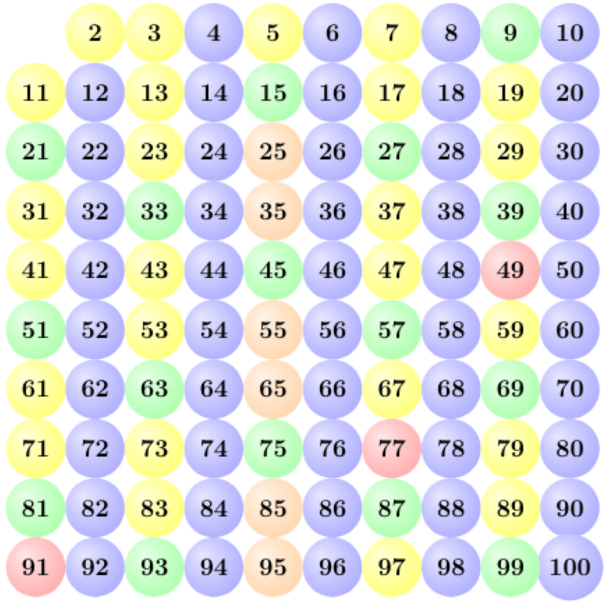

FIGURE 1. The Sieve of Eratosthenes

To modify this for the duck-shooting problem, we do the same process, but we also delete the prime from the list. This is equivalent to picking the residue classes $0 \pmod p$ for all the primes p. Concretely, if we remove $0 \pmod p$ for all $p \leq \frac{x}{4}$ (say), then the only remaining ducks are 1 and the primes between $\frac{x}{4}$ and y, since all the composite numbers get sieved out[1].

Observation 3 (Smooth Numbers)

We can actually do better than the sieve of Eratosthenes. The thing to note is that the big primes in the sieve contribute negligibly. For example, in the sieve pictured in Figure 1, the prime 7 only deletes one number, 49.

Thus, instead of using all the primes from 1 to $\frac{x}{4}$, let's only use the primes p with $p \leq \frac{y}{x}$ (these are the small primes that contribute a lot) and the primes $z < p \leq \frac{x}{4}$, where z is a number that will be chosen later. Then, along with the primes from $\frac{x}{4}$ to y, it is evident that the additional survivors must be ***z*-smooth**, i.e. their prime factors are all less than or equal to z.

It turns out that if we choose z correctly, the number of additional ducks that survive is negligible. The optimal choice is

$$z = x^{\frac{c \log_3 x}{\log_2 x}}$$

where c is a small positive number. The reason for this choice arises from the asymptotics of smooth numbers, and it leads to all the logarithms in our final bounds.

---

[1]One thing to note is that this only works if $y < \left(\frac{x}{4}\right)^2$, but it turns out that our best bound for y is much smaller than that.

4 Finishing the Bound and Improvements

We now have two sets of primes left: the medium-sized primes from $\frac{y}{x}$ and z, and the large primes from $\frac{x}{4}$ and $\frac{x}{2}$.

It turns out that for the medium sized primes, the best thing to do is to just randomly select residue classes. This actually is pretty good at reducing the number of survivors. For the large primes, as we discussed earlier, we will just use each prime to shoot one duck.

The steps we have discussed are more or less Rankin's argument for his bound, and they result in the bound

$$G(X) \geq (c - o(1)) \log X \frac{\log_2 X \log_4 X}{(\log_3 X)^2},$$

where c is a small number.

Now, how do we improve this bound? Well, remember how we were only able to shoot one duck with each of the large primes? It turns out that we actually *can* shoot two (or more) birds with one stone. Thus, if we want to make the constant c larger, a natural thing to do is to look for residue classes $a_p \pmod{p}$ that hit many primes in $\{1, \ldots, y\}$.

There are in fact two ways to do this, discovered independently by Ford-Green-Konyagin-Tao on one hand, and Maynard on the other. We give an overview of both approaches.

Ford-Green-Konyagin-Tao's Approach

In this way to prove that c can get arbitrarily large, we start with **Green-Tao's Theorem**, which locates long arithmetic progressions $a, a+r, \ldots, a+(k-1)r$ of prime numbers. What Ford, Green, Konyagin, and Tao did was to modify this result to the case where r itself is a small multiple of a large prime p. Then, the residue class $a \pmod{p}$ contains many primes. It turns out that this is enough to make the c in Rankin's bound arbirarily large.

Maynard's Approach

Maynard's approach was to modify an argument that he and others had used to reduce the bound for the twin prime conjecture to 246. The small prime gaps arguments produce many numbers n such that $n + h_1, \ldots, n + h_k$ are prime. It turns out that a variation of this argument can show that for any large prime p, one can find many n for which many of $n + h_1 p, \ldots, n + h_k p$ are prime. Then the residue class $n \pmod{p}$ hits many primes in $\{1, \ldots, y\}$, and this gives an alternate approach in making the c in Rankin's bound arbitrarily large.

However, neither of these strategies reach the current best lower bound of

$$G(x) \geq (c - o(1)) \log X \frac{\log_2 X \log_4 X}{\log_3 X}.$$

This is due to a "lack of coordination" between the residue classes $a_p \pmod{p}$ for large p. What this means is that each of these residue classes contain many primes, but some of them intersect each other, leading to a loss in efficiency.

This problem is a special case of a **hypergraph coloring problem**. In general, a hypergrpah coloring problem can be stated as follows: given a collection of subsets of a large set V, what is the most efficient way to cover most of V using as few subsets as possible?

The Rödl Nibble

The Rödle Nibble is an efficient covering algorithm developed by Pippenger and Spencer in 1989. We give a brief outline of the method.

First, we select at random a small number of the residue classes $a_p \pmod p$. This is a "nibble." Then, we delete all the possible residue classes that intersect these ones, so all the remaining residue classes we choose won't intersect these ones. Then, we take another "nibble," and repeat the same process. Continuing this operation gives us a covering.

The Rödl nibble method eliminates almost all of the losses coming from overlapping residue classes. Implementing this method leads to the final bound

$$G(X) \geq (c - o(1)) \log X \frac{\log_2 X \log_4 X}{\log_3 X},$$

for some $c > 0$. This is the currently the best known lower bound for $G(X)$. Also, note that this improves Rankin's original bound by a factor of $\log_3 x$.

5 Computational Results

With the advent of computers in the mid-1900s, the computation of increasingly larger primes, hence prime gaps, has become much more attainable. Table 3 summarizes the values of $G(x)$ calculated for $x \leq X$.

Upper Bound X	**Attained by**	**Year**
$3 \cdot 10^6$	Glaisher	1878
$1 \cdot 10^7$	Western	1934
$3.7 \cdot 10^7$	Lehmer	1957
$1.044 \cdot 10^8$	Gruenberger, Armerding, Baker	1959
$1.096 \cdot 10^{10}$	Lander, Parkin	1967
$4.444 \cdot 10^{12}$	Brent	1980
$7.263 \cdot 10^{13}$	Yound, Potler	1989
$1 \cdot 10^{15}$	Nicely	1999
$5 \cdot 10^{16}$	Nyman	2003
$4 \cdot 10^{18}$	Oliveira e Silva, Herzog, Pardi	2012
2^{64}	PGS (Mersenne Forum)	2018

TABLE 3. Summary of values of $G(x)$ calculated for $x \leq X$.

Referring back to Cramér's conjecture, we can use these results to calculate the **Cramér-Shanks-Granville ratio**

$$\max_{7<x\leq 2^{64}} \frac{G(x)}{\log^2 x} = 0.9206.$$

This is still below the conjectured value of 1 by Cramér, and $2e^{-\gamma} \approx 1.12$ by Granville.

Others have made more conjectured growth rates that are more fine-tuned than the simple $\log^2 x$ bounds predicted by Cramér. These are displayed in Figure 2 below.

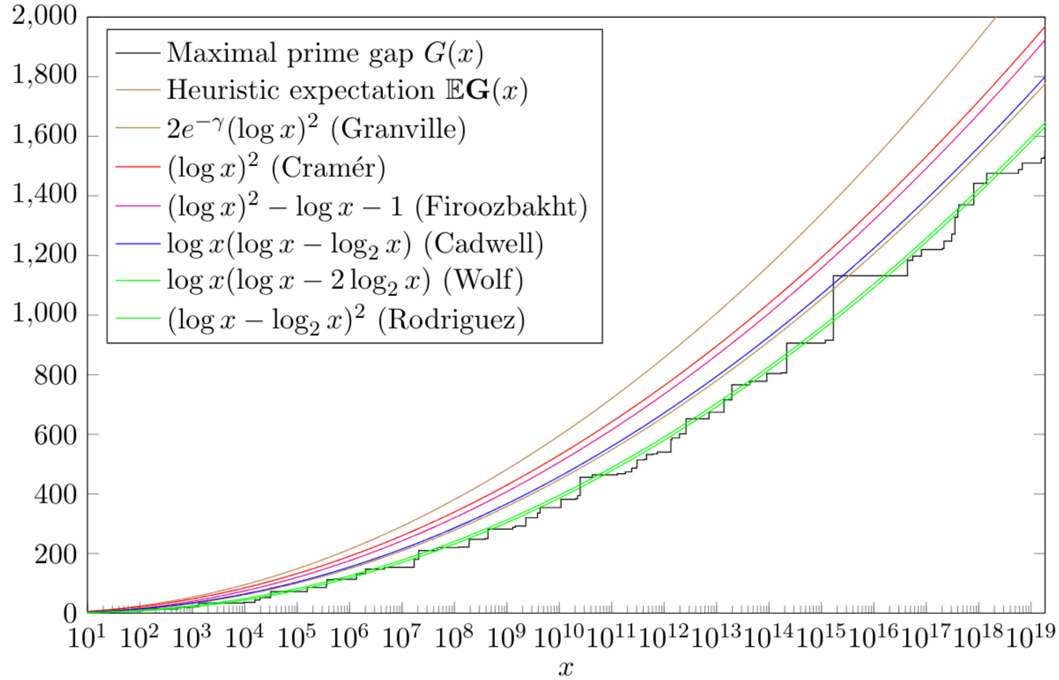

FIGURE 2. Comparison of the true value of $G(x)$ with the heuristic expectation $\mathbb{E}\mathbf{G}(x)$ calculated by Cramér's model, along with other conjectured growth rates.

Now we look at computational results for $Y(x)$, which by comparison to $G(x)$, has been going much slower. This is because even the best algorithms do not run substantially faster than brute force, which is $\mathcal{O}(x\#) = \mathcal{O}(e^x)$. $Y(x)$ has been computed up till $p_{57} = 269$, the 57-th prime number.

Since it is hard to compute $Y(x)$ exactly, we will calculate some sub-optimal lower bounds based on a greedy approach. We will also modify this approach based on Rankin's proof of his bound for $G(X)$.

To calculate this lower bound, we will calculate a lower bound inverse for $Y(x)$. What this means is that for an integer y, we let $q(y)$ denote the minimal prime number p such that $\{1, \ldots, y\}$ can be sieved out using only primes from 2 to p. We employ a greedy strategy starting from the smallest prime.

Algorithm 1 Greedy strategy

$\mathcal{N} \leftarrow \{1, \ldots, y\}$

$p \leftarrow 2$ (start with the smallest prime)

while $\mathcal{N} \neq \emptyset$ **do**

  Choose residue class $a_p \pmod p$ which maximizes

$$|\{n \in \mathcal{N} : n \equiv a_p \pmod p\}|$$

  Sieve out these elements from $\mathcal{N}$: $\mathcal{N} \mapsto \{n \in \mathcal{N} : n \not\equiv a_p \pmod p\}$

  if $\mathcal{N} \neq \emptyset$ **then**

    increment the prime p to the next prime

  else if $\mathcal{N} = \emptyset$ **then**

    output p as $q(y)$

  end if

end while

Then we can define the inverse function

$$Y'_{\mathrm{gr}}(x) = \max\{y : q(y) = x\}.$$

However, $q(y)$ is not non-decreasing (for example $q(18) = 17$ and $q(21) = 13$), so neither is Y'_{gr}. We can rectify this problem by setting

$$Y_{\mathrm{gr}}(x) = \max\{y : q(y) \leq x\}.$$

This function, by definition, is non-decreasing.

To further improve $Y_{\mathrm{gr}}(x)$, we modify the greedy algorithm by choosing residue classes $a_p = 0$ for small primes $p \leq L$, where L is a suitably chosen small value. Thus, this strategy mimics Rankin's bound.

To do this, let $r(y, L)$ be the minimal prime p such that $\{1, \ldots, y\}$ can be sieved out using primes from 2 to p by taking $a_p = 0$ for $p \leq L$ and a_p chosen greedily for $p > L$. Then we define

$$Y_{\mathrm{Ran}}(x) = \max\{y : \text{there exists } L \leq y \text{ such that } r(y, L) \leq x\}.$$

This is a marginal improvement over the greedy algorithm. See Figure 3 for details.

One might wonder why there is no graph of $\mathbb{E}\mathbf{Y}(x)$, obtained from Cramér's model, in Figure 3. The reason is very simple: it's infinite!

The reason for this is that it happens with nonzero probability that $2, 3, 4, 6, 12$ are chosen to be "prime" in Cramér's model. Then the residue classes $a_2 = 0, a_3 = 2, a_4 = 1, a_6 = 3$, and $a_{12} = 11$ cover all the integers, which means $\mathbf{Y}(x)$ is infinite in this case. Thus $\mathbb{E}\mathbf{Y}(x)$ is infinite for $x \geq 12$.

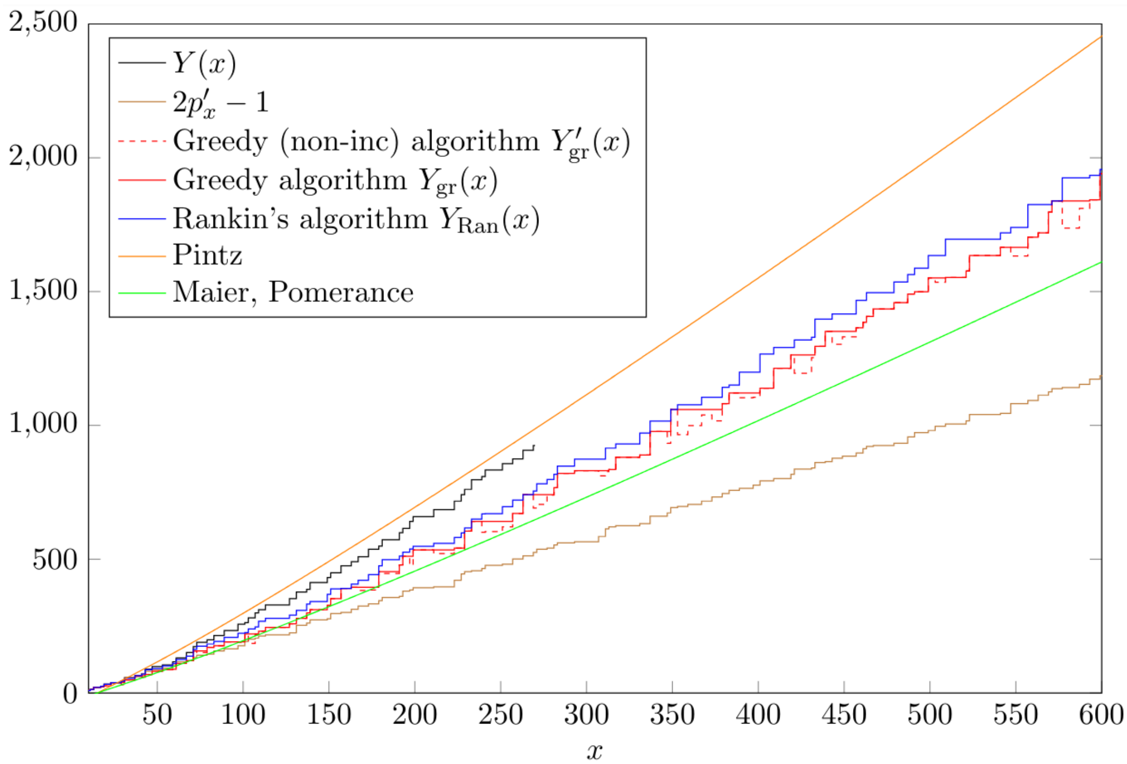

FIGURE 3. Graph of known values of $Y(x)$, along with several lower bounds.

Even if we fix the small primes, as is done in Granville's improvement to Cramér's model, it turns out that $\mathbb{E}\mathbf{Y}(x)$ is infinite for sufficiently large x. However, there is still hope to be able to analyze $Y(x)$ using heuristic models, especially given some results based on a new probabilistic model published by Banks, Ford, and Tao.

References

[1] Robin Visser (2020). *Large Gaps Between Primes.* Link

[2] Terence Tao (2015). *Small and Large Gaps between Primes.* Slides

[3] W. Banks, K. Ford, T. Tao (2019). *Large prime gaps and probabilistic models.* arXiv:1908.08613

[4] K. Ford, B. Green, S. Konyagin ,J. Maynard, T. Tao (2015). *Long gaps between primes.* arXiv:1412.5029

[5] James Maynard (2015). *Small gaps between primes.* arXiv:1311.4600

[6] James Maynard (2016). *Dense clusters of primes in subsets.* arXiv:1405.2593

[7] James Maynard (2016). *Large gaps between primes.* arXiv:1408.5110

[8] Simon Rubinstein-Salzedo. 2024. *Analytic Number Theory.* Lecture notes for Euler Circle, Spring 2024.