Function estimating number of pairs of primes (p, q) for all $z \in \mathbb{N}$ of form $z = p + q$

Dmitri Abramov
ddmitri.abramov@gmail.com

October 14, 2013

Abstract

This paper derives a function that estimates number of unique ways you can write z as $z = p + q$, where p and q are prime numbers, for every $z \in \mathbb{N}$ that can be written in that form.

1 Introduction

During my previous work on possible ways of solving the Goldbach's conjecture via distribution and probabilities of prime numbers I discovered an interesting function that might describe number of (p, q) pairs for all z that can be written as sum of two primes. Although the method of deriving it is dubious it shows some interesting properties that might merit further insight.

2 Derivation

The Prime Number Theorem defines $\pi(x)$ as number of primes less than or equal to x. One approximation for $\pi(x)$ is

$$\pi(x) \sim \frac{x}{ln(x)}. \tag{1}$$

Let us define two sets, **M** and **N** so that

$$\mathbf{M} = \{1, 2, \ldots, z-2, z-1\}$$

$$\mathbf{N} = \{z-1, z-2, \ldots, 2, 1\}$$

so that $\mathbf{M}_a + \mathbf{N}_a = z$. Since the length of $\mathbf{M} = \mathbf{N} = z - 1$ there are approximately $\dfrac{z-1}{ln(z-1)}$ primes in each set. Thus, the probability of picking a prime at random is number of primes divided by the amount of numbers in $\mathbf{M}$ and $\mathbf{N}$:

$$P(\mathbf{M}_{prime}) = P(\mathbf{N}_{prime}) \approx \frac{1}{z-1} \cdot \frac{z-1}{ln(z-1)}$$

$$P(\mathbf{M}_{prime}) = P(\mathbf{N}_{prime}) \approx \frac{1}{ln(z-1)} \tag{2}$$

Let us pick a at random. For the set $\mathbf{M}$ the probability that $\mathbf{M}_a$ is prime is $P(\mathbf{M}_p) = 1/(ln(z-1))$. Analogously, $P(\mathbf{N}_p) = 1/(ln(z-1))$. If those two probabilities are independent, then probability that both numbers are prime is

$$P(\mathbf{M}_p, \mathbf{N}_p) = P(\mathbf{M}_p) \cdot P(\mathbf{N}_p)$$

$$P(\mathbf{M}_p, \mathbf{N}_p) \approx \frac{1}{ln^2(z-1)} \tag{3}$$

Whether or no are these two probabilities independent is open to debate, as they are connected via $\mathbf{M}_a + \mathbf{N}_a = z$, although the a was picked randomly and independantly from the interval $[0, z-1]$. For the moment, let's assume that the probabilities is independent and that (3) is valid.

3 Comparison with real-world data

Interestingly, there are numbers that can't be written as sum of two primes. Since odd z can only be written as $p + q = z$ if $p = 2$, all odd numbers for which $z - 2$ isn't a prime can't be written as a sum of two primes and thus has zero (p, q) pairs. Since (3) is only valid for z that can be written as $p + q = z$ all z that can't be written in that way are here disregarded.

Number of (p,q) pairs is (3) times number of members of items in sets $\mathbf{M}$ and $\mathbf{N}$ times two:

$$\#pairs \approx \frac{2(z-1)}{ln^2(z-1)}$$

We are multiplying by 2 because using two sets gives us symmetrical pairs $(p, q) = (q, p)$. In order to get number of *unique* pairs we divide the equation by 2 to get

$$\#pairs \approx \frac{z-1}{ln^2(z-1)} \tag{4}$$

The plot shown in Figure 1 shows scatter plot of number of (p, q) pairs (y axis) for a number z (x axis). Function (4) is in red, while fitted quadratic curve was obtained by Mathematica's `Fit[data, {1, x, x^2}, x]` is in green. Similar results were obtained using `Fit[data, {1, x, Log[x]}, x]` and `Fit[data, {1, x, x/Log[x]},x]` — Figures 2 and 3 respectively.

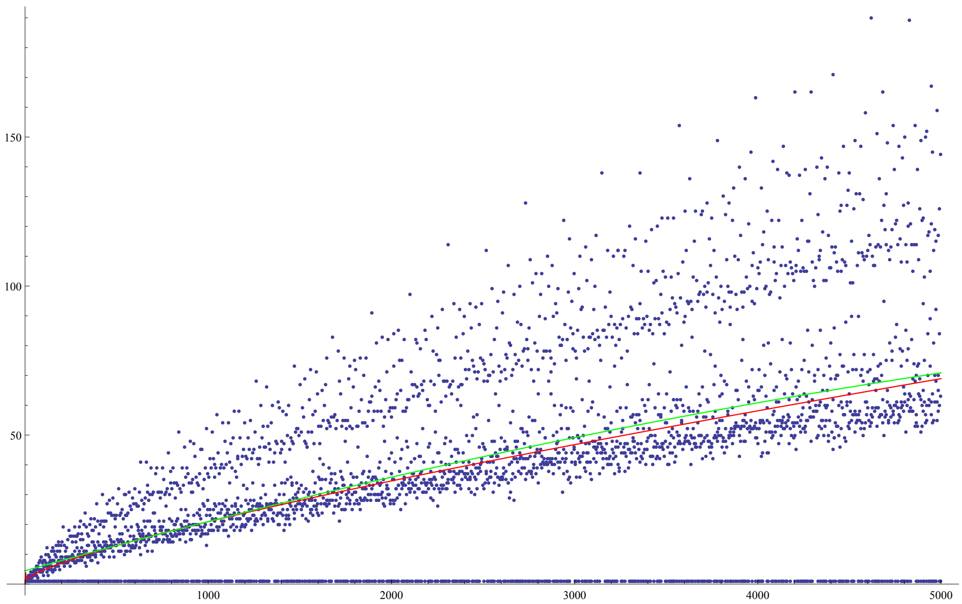

Figure 1: Real data and fitted function `Fit[data, {1, x, x^2}, x]` (green) compared to $x/ln^2(x)$ (red)

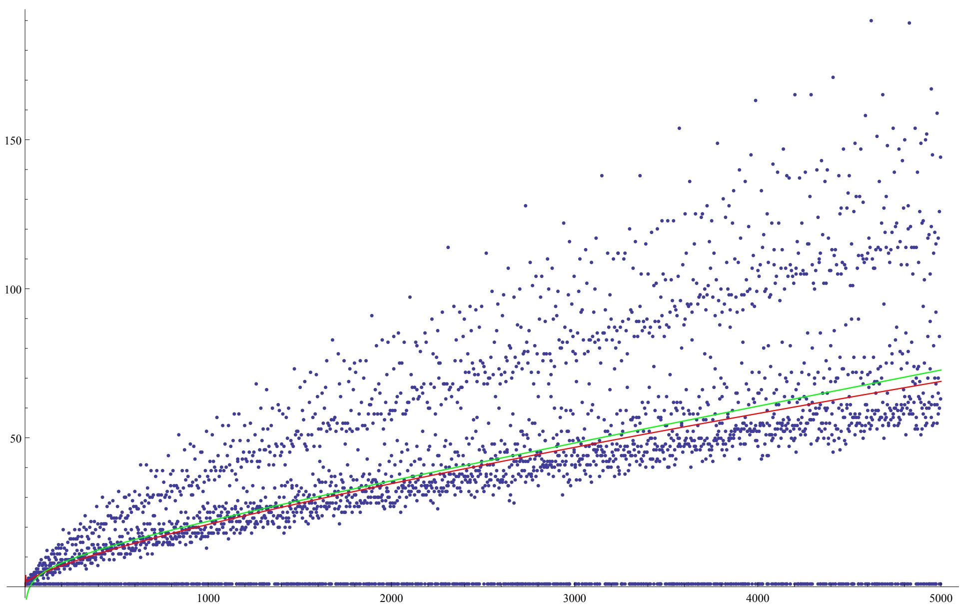

Figure 2: Real data and fitted function `Fit[data, {1, x, Log[x]}, x]` (green) compared to $x/ln^2(x)$ (red)

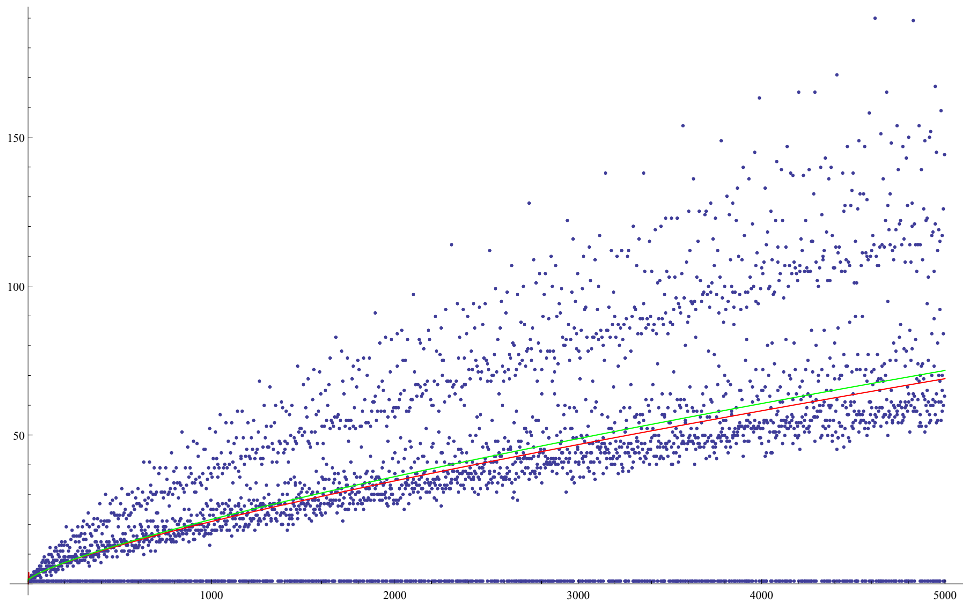

Figure 3: Real data and fitted function
`Fit[data, {1, x, x/Log[x]}, x]` (green) compared to $x/ln^2(x)$ (red)

4 Analysis and discussion

Function $f(z) = (z-1)/ln^2(z-1)$ has moderate correlation of $r = 0.556$ with real world data, which should be taken with grain of salt because of the non-linear growth of both $f(z)$ and largest number of pairs. If we define $g(z)$ as difference between actual number of prime pairs for z and $f(z) = \frac{z-1}{ln^2(z-1)}$ (Equation (4)) rounded to the closest integer then the average of $g(z)$ values, or average of mistakes, between 0 and 100 is -0.2877. Between 0 and 1000 the average mistake is 0.2973 and continues to rise up to 1.5633 in the interval from 0 to 5000. Correlation between z and average mistake is $r = 0.9782$, a very strong positive correlation which implies that (4) becomes weaker and weaker as $z \to \infty$. This is an interesting behavior, as one would expect (4) to become more accurate as z grows since (1) increases in accuracy as $x \to \infty$. It suggests that there might be an additional constant c multiplying (4), which Mathematica's `Fit[data, {(x-1)/(Log[x-1])^2}, x]` returns as $c = 1.041$ for $z \leq 5000$. As logarithmic integral $Li(x)$ converges to $\pi(x)$ faster than (1) more insight might be gained by repeating the derivation of (4) using $Li(x)$ although there are obvious issues with method of derivation itself, which can not be ignored. Results obtained using $Li(x)$ will be reported in a follow-up paper.

The shape of scatter plot merits it's own lengthy discussion, but I'll only note some observations. First of all, the plot shows five distinct features where points behave in an interesting way. The most obvious one is at

$y = 1$. Although the average y grows with x there seems to be a significant portion of numbers that can be written as only one unique pair of primes. Except for 4 and 6, only odd numbers have $y = 1$ — only way to write odd z as sum of two primes is when $p = 2$ and $q = z - 2$. If $z - 1$ isn't a prime, a number can't be written as sum of two primes. Such numbers become less and less common since we encounter less and less primes as we go to infinity. In context of Goldbach's conjecture it might be best to explore graph with only even z as most of them (or, if the conjecture is true, all of them) can be written as sum of two primes whereas growing portion of odd numbers can not.

Second prominent feature of the plot is dense band of points in the middle, around and below function (4) as can be seen on Figure 3. Function describing that portion of data seems to grow slightly slower than (4) but it also seems to be bound by two unknown logarithmic functions.

Third feature is slightly less dense band above the first one. There seems to be a lot of variance though, mostly on the upper side. It also seems to correlate with a logarithmic function growing faster than (4).

Fourth one is thin but distinct top band that shows numbers with highest number of pairs. They seem to be thinly separated from third feature with small amount of whitespace in between. Function that describes that section of the graph is also the upper bound for number of possible pairs.

Fifth and maybe not so obvious one is the whitespace itself. Distribution of data points isn't uniform but, if we ignore $y = 1$ points, has a both lower and upper bound. If only odd numbers have $y = 1$ then for all even numbers z there are two bounding functions that goven the maximum and minimum number of pairs for z. It is obvious that the study of these functions and graphs may give more insight to the ideas and problems behind Goldbach's conjecture and providing a function that describes the lower boundary would prove Goldbach's conjecture.