LINES IN THE PRIME NUMBER GRAPH

Scott Duke Kominers
Harvard University, Cambridge, Massachusetts, USA
kominers@fas.harvard.edu

Rudi Mrazović
University of Zagreb, Zagreb, Croatia
Rudi.Mrazovic@math.hr

Carl Pomerance
Dartmouth College, Hanover, New Hampshire, USA
carlp@math.dartmouth.edu

Patrick Solé
I2M, (CNRS, Aix-Marseille University), Marseille, France
patrick.sole@telecom-paris.fr

Abstract

The prime number graph is the set of points (n, p_n) where p_n denotes the n^{th} prime. Let $L(n)$ be the minimum number of straight lines needed to cover the first n points in this set. Let $B(n)$ be the largest number of points (k, p_k) with $k \leq n$ covered by a single line. Recently Sloane conjectured that $L(n) = O(n/\log n)$. We prove a much stronger bound, as well as upper and lower estimates for $B(n)$. Our proofs use the Prime Number Theorem with remainder and are considerably improved with the assumption of the Riemann Hypothesis.

AMS Subject Classification: 52C10, 11A41, 11N05.
Keywords: Prime Number Theorem, prime number graph, awkward prime.

1. Introduction

Let $p_1, p_2, \ldots$ denote the sequence of primes. A **prime point** is a point of the plane of the form (k, p_k) for some k. This graphical representation of the primes was considered in [8]. It is interesting to look at sets of these prime points that are collinear, such as

$$(6, 13),\ (7, 17),\ (10, 29),\ (12, 37),\ (13, 41),\ (16, 53),\ (18, 61),\ (21, 73)$$

which are all on the line $y = 4x - 11$. Let $L(n)$ be the minimum number of lines needed to cover the first n prime points. For example, $L(2) = 1$ and $L(3) = 2$.

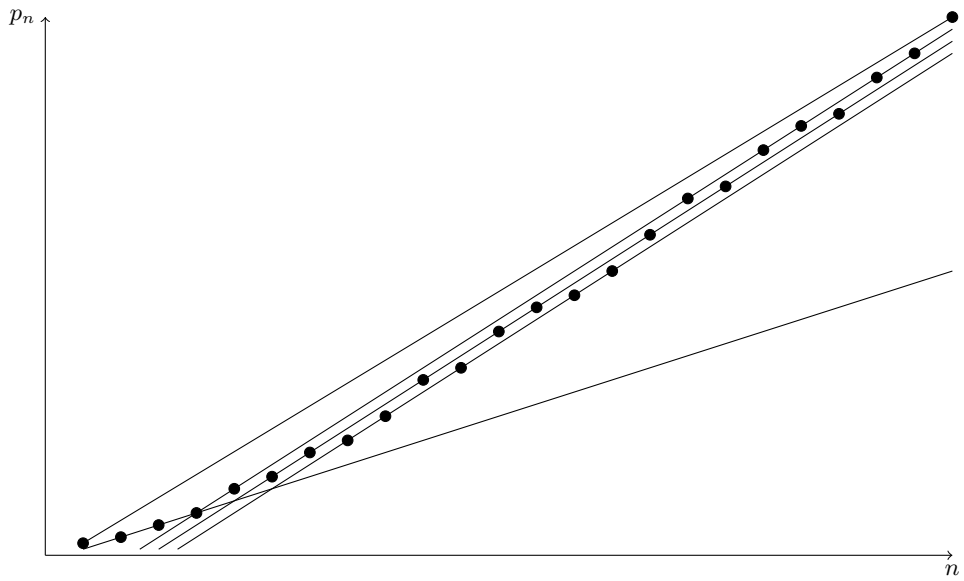

Figure 1: First 24 points of the prime number graph covered by $L(24) = 5$ lines.

A prime p_n is **awkward** if $L(n) > L(n-1)$. These concepts were introduced in recent Numberphile videos [1], [2]. See [5] for numerics of $L(n)$ (called there $a(n)$) for small n, and see [6] for the list of the first awkward primes.

In this note we study how the function $L(n)$ behaves for large n. Since the primes have asymptotic density 0, it is clear that no line can contain infinitely many prime points (k, p_k). We introduce the function $B(n)$ which is the largest number of prime points among the first n of them covered by a single line, and derive upper and lower bounds for it. Our arguments for $L(n)$ and $B(n)$ are based on the Prime Number Theorem with remainder; we therefore also obtain a direct strengthening if we assume the Riemann Hypothesis.

The material is arranged as follows. The next section recalls some known results on the topic of the prime number graph and the Prime Number Theorem with remainder. Sections 3 and 4 study the functions $L(n)$ and $B(n)$, respectively. Section 5 presents concluding remarks and underlines some challenging open problems.

2. Background results

We begin by recording an observation that is originally due to Erdős and quoted without proof in [8].

Theorem 1. *For any positive integer k, almost all prime points (n, p_n) lie on a line with k other prime points. That is, the set of primes for which this is so has relative density* 1 *in the set of primes.*

From there, we deduce an asymptotic upper bound on $L(n)$.

Corollary 1. *For $n \to \infty$ we have $L(n) = o(n)$.*

In the sequel, we present a detailed proof of a more quantitative version of Theorem 1 and Corollary 1. We also derive a lower bound for $B(n)$, using a proof strategy strongly based on the proof of [8, Theorem 4.1].

We note that based on the experimental data in [5], Sloane conjectured in [1] that $L(n) = O(\frac{n}{\log n})$. This in particular motivates Theorem 3 of the next section, where we in fact obtain a bound stronger than Sloane's conjecture.

As mentioned in the Introduction, our proofs strongly use the Prime Number Theorem with remainder. In particular, let

$$\mathrm{li}(x) = \int_0^x \frac{dt}{\log t}$$

denote the logarithmic integral function (where the principal value is taken for the singularity at $t = 1$). Then $\mathrm{li}(x) \sim x/\log x \sim \pi(x)$ as $x \to \infty$, but the $\mathrm{li}(x)$ approximation to $\pi(x)$ is much more accurate. In particular, we have

$$|\pi(x) - \mathrm{li}(x)| \leq x/\exp(c(\log x)^{3/5}(\log\log x)^{-1/5}) \tag{2.1}$$

for a positive constant c and all large x. If the Riemann Hypothesis is assumed, then (2.1) improves to

$$|\pi(x) - \mathrm{li}(x)| \leq x^{1/2}\log x$$

for all $x \geq 2$. (See [3] and [4]; and see also [7] for an asymptotically weaker, but numerically explicit version of (2.1).) In general, let $R(x)$ be any smooth function with $R' > 0$ and $R'' < 0$ and

$$|\pi(x) - \mathrm{li}(x)| \leq R(x) \tag{2.2}$$

for all sufficiently large values of x. Our subsequent results are all stated in terms of $R(x)$.

3. The quantitative Erdős observation and awkward primes

In this section we give a quantitative proof of Theorem 1. In particular, we will prove the following.

Theorem 2. *Let $N(x)$ denote the number of $n \le x$ such that (n, p_n) lies on a line with at least $(n/R(n))^{1/4}$ prime points in total. Then $N(x) \sim x$ as $x \to \infty$.*

We prove Theorem 2 as a corollary of the following result.

Theorem 3. *We have $L(n) = O(n^{3/4} R(n)^{1/4}/(\log n)^{1/2})$.*

Proof. Let Q be a large integer and consider the Farey sequence of level Q. Let $a/b, a'/b'$ be two consecutive terms. Then

$$a'/b' - a/b = 1/bb', \quad b, b' \le Q, \quad b + b' > Q, \quad \gcd(b, b') = 1.$$

Let k be a large integer, and let $u = e^{k+a/b}$, $u' = e^{k+a'/b'}$, with I the interval $(u, u']$. We consider **inverse prime points** (p_n, n) with $p_n \in I$. The length $|I|$ of I has

$$|I| = u' - u = e^{k+a'/b'} - e^{k+a/b} = e^{k+a/b}(e^{1/bb'} - 1) = (1 + o(1))u/bb', \quad Q \to \infty.$$

Let P denote the parallelogram bounded by the vertical lines $x = u$, $x = u'$ and the lines with slope $1/\log u = 1/(k + a/b)$ through $(u, \mathrm{li}(u) - w)$ and $(u, \mathrm{li}(u) + w)$, where

$$w = \frac{|I|^2}{u \log^2 u} = (1 + o(1)) \frac{u}{(bb')^2 \log^2 u}.$$

These lines gain $|I|/\log u$ on the interval $I = (u, u']$. This is about the same gain as $\mathrm{li}(x)$ on the interval. Note that by Taylor's theorem,

$$\mathrm{li}(u') - \mathrm{li}(u) = \frac{|I|}{\log u} - \left(\frac{1}{2} + o(1)\right) \frac{|I|^2}{u \log^2 u} = \frac{|I|}{\log u} - \left(\frac{1}{2} + o(1)\right) w,$$

and hence the region

$$|y - \mathrm{li}(x)| \le w/4, \quad x \in I$$

lies wholly in P. For a given large number k we want to choose Q as large as possible so that the graph of $y = \pi(x)$ for $x \in I$ also lies in P. Since $bb' < Q^2$, we have $w/4 \ge e^k/(8Q^4k^2)$ for k large, and so by (2.2) this will be accomplished if we take

$$Q = \left\lfloor \left(\frac{e^k}{8R(e^{k+1})k^2} \right)^{1/4} \right\rfloor.$$

We can also do a similar construction by using u' instead of u to determine the slope. Namely, let P' be the parallelogram bounded by $x = u$, $x = u'$ and the lines with slope $1/\log u'$ through $(u', \mathrm{li}(u') - w)$ and $(u', \mathrm{li}(u') + w)$. With the same choice of Q, the prime count $y = \pi(x)$ for $x \in I$ also lies in P' for k large.

Hence all of the inverse prime points (p_n, n) with $p_n \in I$ lie in both P and P'.

If $b < b'$, we consider those lines with slope $1/\log u = 1/(k + a/b) = b/(bk + a)$ that pass through a lattice point in P. Equations of these lines are $(bk+a)y-bx = C$ for different integers C, and so there are at most

$$(2 + o(1))w \cdot (bk + a) \sim 2wb \log u \sim \frac{2u}{bb'^2 \log u} \ll \frac{e^k}{bb'^2 k}$$

of them. If $b > b'$, we take those lines with slope $1/\log u' = b'/(b'k + a')$ which pass through a lattice point in P'; there are $\ll e^k/(b^2 b'k)$ of them. So if we consider all of the lines appearing in this argument for primes in $(e^k, e^{k+1}]$, the number of them is bounded by a constant times

$$\frac{e^k}{k} \sum \frac{\min\{b, b'\}}{(bb')^2}, \tag{3.1}$$

where the sum is over the full Farey dissection of level Q. If $b, b' > Q/2$, then the summand in (3.1) is of magnitude $1/Q^3$, and there are fewer than Q^2 such pairs, so the contribution is bounded by $1/Q$. So, assume $\min\{b, b'\} \le Q/2$. The contribution to the sum in (3.1) is at most

$$2 \sum_{b \le Q/2} \frac{1}{b} \sum_{Q - b < b' \le Q} \frac{1}{b'^2} < 2 \sum_{b \le Q/2} \frac{1}{b} \left(\frac{1}{Q - b} - \frac{1}{Q} \right) = 2 \sum_{b \le Q/2} \frac{1}{(Q - b)Q} \ll \frac{1}{Q}.$$

Thus, the sum in (3.1) is $O(1/Q)$ and so all of the inverse prime points (p_n, n) with $p_n \in (e^k, e^{k+1}]$ are covered by $O(e^k/kQ)$ lines. With our choice for Q and noting that $R(e^k) \sim R(e^{k+1})$, we have these inverse prime points covered by $O(e^{3k/4} R(e^k)^{1/4}/k^{1/2})$ lines. Summing this for $k \le K - 1$, we have that the total number of lines that contain some (p_n, n) for $n \le e^K$ is $O(e^{3K/4} R(e^K)^{1/4}/K^{1/2})$. Thus, if $n \in (e^{K-1}, e^K]$, then $L(n) = O(n^{3/4} R(n)^{1/4}/(\log n)^{1/2})$. This completes the proof. □

As a corollary we obtain Theorem 2.

Proof of Theorem 2. Consider the $L(n)$ lattice lines that cover all of the points (p_j, j) for $j \le n$. Those lines that cover fewer than $(n/R(n))^{1/4}$ inverse prime points together cover $O(n/(\log n)^{1/2})$ points. This leaves still asymptotically all n inverse prime points, where each such point is contained in a line with at least $(n/R(n))^{1/4}$ other inverse prime points. □

Theorem 4. *The number of awkward primes among the first n primes is $L(n)$. Thus, the reciprocal sum of the awkward primes is finite.*

Proof. Let $L(0) = 0$. For each positive integer j, we have $L(j) - L(j-1)$ equal to 0 or 1, where the value 1 occurs if and only if p_j is awkward (because adding the j^{th}

point can always be handled by adding one line). So, we have

$$L(n) = \sum_{j \le n} (L(j) - L(j-1)).$$

Thus, it is clear then that $L(n)$ is the number of awkward primes p_j with $j \le n$. Theorem 2, together with a partial summation argument, then shows that their reciprocal sum is finite. □

4. The function $B(n)$

We know by [8, Theorem 4.1] and Theorem 2 that $B(n)$ is not bounded above. In fact, we have the following estimate.

Theorem 5. *There is a positive constant c_1 such that for all large n we have $B(n) \ge c_1 \sqrt{n/R(n)}/\log n$.*

Proof. First, we note that there is a simple combinatorial relation connecting the functions L and B, namely

$$L(n)B(n) \ge n. \tag{4.1}$$

This is immediate by considering a covering of the first n prime points by $L(n)$ line segments. Each line contains at most $B(n)$ prime points, which then gives (4.1). Thus, from Theorem 3 we have $B(n) = \Omega((n/R(n))^{1/4}(\log n)^{1/2})$. But if we use the proof of Theorem 3, then we can do better. In that proof, we used the Farey dissection of level Q to obtain a dissection of the interval $(e^k, e^{k+1}]$. Now we use only the first (and longest) piece of the dissection—namely, $(e^k, e^{k+1/Q}]$. The length of this interval is $\sim e^k/Q$ and $w \sim e^k/(Qk)^2$. Now the constraint on Q being large is somewhat relaxed and we can take Q as an integer near $\sqrt{e^k/R(e^k)}/k$. The number of primes in the interval is $\sim e^k/kQ$ and the number of lines that cover them is $O(wk) = O(e^k/Q^2k)$. Thus, the average number of inverse prime points per line is $\Omega(Q)$. So there is a positive constant c_1 such that $B(n) \ge c_1\sqrt{n/R(n)}/\log n$. □

We remark that using the interval $(e^k, e^{k+1/k}]$ to show that some lines have many prime points was used in the proof of [8, Theorem 4.1].

A trivial upper bound for $B(n)$ is n, of course. We can do considerably better.

Theorem 6. *For n sufficiently large, we have*

$$B(n) = O(\sqrt{nR(n)}).$$

Proof. From the discussion above, we may assume that $R(x) = o(\mathrm{li}(x))$ as $x \to \infty$ and that both functions $y = \mathrm{li}(x) + R(x)$ and $y = \mathrm{li}(x) - R(x)$ are smooth, strictly increasing, and strictly concave down. Thus, a line may intersect these two curves

in at most two points each. In fact, a line can intersect the region $|y - \mathrm{li}(x)| \le R(x)$ at most twice, i.e., either for one bounded interval I on the positive x-axis or for two disjoint bounded intervals. A calculation shows that the length of such an interval is $O((xR(x))^{1/2} \log x)$. Since the number of primes in such an interval is $O((xR(x))^{1/2})$, the theorem follows. □

Assuming the Riemann Hypothesis (RH) we can give a more explicit upper bound on $B(n)$, and provide a similar refinement for $L(n)$.

Corollary 2. *Under RH we have for n large that*

$$n^{1/4}/(\log n)^{3/2} \ll B(n) \ll n^{3/4}(\log n)^{1/2},$$

and

$$n^{1/4}/(\log n)^{1/2} \ll L(n) \ll n^{7/8}/(\log n)^{1/4}.$$

Proof. As we noted in Section 2, we can take $R(x) = \sqrt{x} \log x$ under RH. The results then follow from Theorems 2, 5, and 6, and (4.1). □

5. Conclusion and open problems

In this note we have studied the covering properties of line segments in the prime number graph. We have derived an asymptotic upper bound for the minimum size of a cover, as well as estimates for the largest number of prime points on a single segment. Our estimates seem far from optimal, as is also suggested from the numerical work in [5], [6]. In particular, regarding the functions $L(n)$ and $B(n)$ it would be nice to reduce the huge gaps between our upper and lower bounds.

Numerical experiments indicate that $L(n)$ is achieved by lines many of which are parellel to each other. Our proof of Theorem 3 also utilizes such sets of lines. This motivates considering the quantity $L_{\mathrm{np}}(n)$—the minimal number of lines that cover the first n points of the prime number graph and have pairwise different slopes. It seems that even proving that $L_{\mathrm{np}}(n) = o(n)$ is nontrivial.

It seems straightforward to generalize our results to primes in a fixed residue class, where the Extended Riemann Hypothesis (namely, the RH for Dirichlet L-functions) plays a role. Likewise one can also look at primes of a particular splitting type in an algebraic number field, using the Chebotarev density theorem. More interestingly, one can ask about general increasing integer sequences. For example, it follows from [9] that if $a_1 < a_2 < \dots$ is a sequence of positive integers with $\liminf a_n/n < \infty$, then for every k there are k collinear points (n, a_n). Is this true under the weaker hypothesis that $\sum 1/a_n = \infty$? Given x, what is the largest number n for which there is an integer sequence $0 < a_1 < a_2 < \cdots < a_n \le x$ such that no three points (j, a_j) are collinear? This holds for the $\lfloor \sqrt{x} \rfloor$ squares in $[1, x]$, can one do better?

Acknowledgments

S.D.K. is a Research Partner at a16z crypto. This work was conducted while he was visiting the Technological Innovation, Entrepreneurship, and Strategic Management (TIES) Group at the MIT Sloan School of Management; he greatly appreciates their hospitality.

R.M. was supported by the Croatian Science Foundation under the project no. HRZZ-IP-2022-10-5116 (FANAP) and by the European Union – NextGenerationEU through the National Recovery and Resilience Plan 2021-2026 Institutional grant of University of Zagreb Faculty of Science (IK IA 1.1.3. Impact4Math).

References

[1] B. Haran and N. Sloane, Awkward primes, Numberphile video, 2026. https://www.youtube.com/watch?v=VFoIPlUalRY&t=525s

[2] B. Haran and N. Sloane, Primes at the end of the line, Numberphile video, 2026. https://www.youtube.com/watch?v=u-_8wX4cECo

[3] H. von Koch, Sur la distribution des nombres premiers, Acta Math. **24** (1901), 159–182.

[4] H. L. Montgomery and R. C. Vaughan, *Multiplicative number theory I. Classical Theory*, Cambridge University Press, 2007.

[5] Online Encyclopedia of Integer Sequences, A373813.

[6] Online Encyclopedia of Integer Sequences, A393445.

[7] D. J. Platt and T. S. Trudgian, The error term in the prime number theorem, Math. Comp. **90** (2021), 871–881.

[8] C. Pomerance, The prime number graph, Math. Comp. **33** (1979), 399–408.

[9] C. Pomerance, Collinear subsets of lattice point sequences—an analog of Szemerédi's theorem, J. Combinatorial Theory (A) **28** (1980), 140–149.