Experimental Mathematics, 20(4):400–411, 2011

ISSN: 1058-6458 print / 1944-950X online
DOI: 10.1080/10586458.2011.565256

Taylor & Francis
Taylor & Francis Group

Frequencies of Successive Pairs of Prime Residues

Avner Ash, Laura Beltis, Robert Gross, and Warren Sinnott

CONTENTS

We consider statistical properties of the sequence of ordered pairs obtained by taking the sequence of prime numbers and reducing modulo *m*. Using an inclusion/exclusion argument and a cutoff of an infinite product suggested by Pólya, we obtain a heuristic formula for the "probability" that a pair of consecutive prime numbers of size approximately *x* will be congruent to $(a, a + d)$ modulo *m*. We demonstrate some symmetries of our formula. We test our formula and some of its consequences against data for *x* in various ranges.

1. INTRODUCTION

In a beautiful paper from 1959, Pólya presented a heuristic formula to approximate the number of primes $p < x$ such that $q = p + d$ is also prime, where d is a positive even integer [Pólya 59]. His formula is a special case of the Bateman–Horn conjecture [Bateman and Horn 62], and had already been proposed in [Hardy and Littlewood 23]. However, his method of derivation is much more elementary.

Pólya uses the standard probabilistic model of the primes. His innovation is in deciding where to cut off a certain product, which he determines by invoking Mertens's theorem. In our treatment, this cutoff step, involving Euler's constant, occurs in (3–1).

Pólya does not care whether there are primes between p and q. In contrast, we ask about *consecutive* prime pairs, and we are interested in gaps not only equal to d but congruent to d modulo a given positive integer m. This interest stems from our earlier work [Ash et al. 09], in which we studied the statistics of the sequence of prime numbers modulo m. For this reason, we actually look at the finer count of consecutive prime pairs congruent to a given $(a, a + d) \pmod m$.

We can thus state our problem as follows:

Problem 1.1. Given positive integers a, d, and m and a positive number x, define $N(a, d, m, x)$ to be the number of consecutive prime pairs $p < q$ such that $p < x$, $p \equiv$

2000 AMS Subject Classification: 11N05, 11K45, 11N69
Keywords: prime pairs, Bateman–Horn, Hardy–Littlewood, Pólya

a, and $q \equiv a + d \pmod{m}$. What are the asymptotics of $N(a, d, m, x)$ as x tends to infinity?

Dirichlet's theorem tells us that if we look only at single primes, the number of primes $p < x$ with $p \equiv a \pmod{m}$ is approximately $\pi(x)/\varphi(m)$, independent of a (provided that a is relatively prime to m). Here φ is the Euler φ-function. One might suppose that the residues modulo m of consecutive prime pairs would be "independent" in a pseudoprobabilistic sense, and therefore that $N(a, d, m, x)$ would be approximately independent of a and d (provided that a and $a + d$ are relatively prime to m). However, this appears not to be the case in the numerical evidence compiled in [Ash et al. 09] and in this paper.

For example, suppose $m = 4$ and consider consecutive pairs of primes congruent to $(1, 1)$ and $(1, 3)$ respectively. Our data show that there is a definite tendency for the $(1, 1)$-pairs to occur considerably less frequently than the $(1, 3)$-pairs. We do not know whether this tendency persists as x increases.

The results in [Hardy and Littlewood 23] suggest an explanation of the imbalance. If d is a positive even integer, then Hardy and Littlewood propose [Hardy and Littlewood 23, p. 42, Conjecture B] that the probability that x and $x + d$ are prime is asymptotic to

$$\prod_{p|d, p>2} \frac{p-1}{p-2} \cdot \frac{2C_2}{(\log x)^2}.$$

Here

$$C_2 = \prod_{p>2} \left(1 - \frac{1}{(p-1)^2}\right).$$

In particular, the probability that x and $x + 2$ are prime is asymptotic to

$$\frac{2C_2}{(\log x)^2}.$$

We observe that the dependence on d occurs only through the factor $\prod_{p|d, p>2} \frac{p-1}{p-2}$, and therefore this factor should tell us the *relative* frequency of primes p such that $p + d$ is also prime when compared to the frequency of twin primes. Here is a table of values for small d:

d	2	4	6	8	10	12	14	16
$\prod_{p\|d, p>2} \frac{p-1}{p-2}$	1	1	2	1	$\frac{4}{3}$	2	$\frac{6}{5}$	1

Thus, for example, primes p such that $p + 6$ is also prime should occur twice as often as twin primes.

If p is a prime congruent to 1 mod 4, and the next prime q is congruent to 3 mod 4, then $q = p + d$ for $d \in \{2, 6, 10, 14, \dots\}$, while if the next prime q is congruent to 1 mod 4, then $q = p + d$ for $d \in \{4, 8, 12, 16, \dots\}$. The table above suggests that $(p, q) \equiv (1, 3) \pmod{4}$ would be more likely. However, the suggestion raises further questions, since even if $p + d$ is prime, it does not mean that it is the next prime after p. For example, it is perfectly possible for each of p, $p + 4$, and $p + 6$ to be prime. Thus there are interactions among the various events "p and $p + d$ are prime" that need to be taken into account if we want to predict whether $(p, q) \equiv (1, 3) \pmod{4}$ is more likely than $(p, q) \equiv (1, 1) \pmod{4}$.

To the best of our knowledge, Problem 1.1 is wide open, and cannot be treated using L-functions, unlike the case of Dirichlet's theorem. In this paper, we attempt a heuristic solution to Problem 1.1 by applying an inclusion/exclusion argument, together with Pólya's heuristic argument, in Section 3. This leads to a formula for a function we call $P(a, d, m, x)$, which heuristically should be the "probability" that x is a prime congruent to $a \pmod{m}$ and the next largest prime is congruent to $a + d \pmod{m}$. Thus, we should have the approximation $N(a, d, m, x) \approx P(a, d, m, x)x$, in the sense that

$$\lim_{x\to\infty} \frac{P(a, d, m, x)x}{N(a, d, m, x)} = 1. \tag{1–1}$$

However, (1–1) is not the appropriate conjecture to verify. For the same reasons that the approximation $\pi(x) \approx x/\log x$ is less accurate than $\pi(x) \approx \int_2^x dy/\log y$, we instead compare $N(a, d, m, X) - N(a, d, m, x)$ with $\sum_{y=x+1}^{X} P(a, d, m, y)$. We do not attempt to provide a measure of how good this approximation should be, since there is no actual underlying probability distribution we are sampling. Nor did Pólya do so in [Pólya 59].

In fact, our expression for P is an infinite series, and we do not know whether this series converges. Thus, we choose an integer $J \geq 0$ and truncate P at the Jth term to obtain a finite expression $P_J(a, d, m, x)$. Our heuristic proposes that $P_J(a, d, m, x)$ should be the "probability" that x is a prime congruent to $a \pmod{m}$ and that the next prime is $x + d + jm$ for some $j = 0, \dots, J$. Thus, if J is sufficiently large (given x), our heuristic suggests that $P_J(a, d, m, x)$ should be a "density function" for $N(a, d, m, x)$ in the sense given above. Our heuristic does not tell us exactly how to choose J. We suggest below that $d + Jm \approx 4 \log x$ is a reasonable choice.

We first investigate properties of the heuristic formula for P_J for a fixed J and test them against some actual data. Below we present data for only a few values of m, which suffice to convey the general picture.

The function $P_J(a,d,m,x)$ possesses two exact symmetries. First, Proposition 4.2 states that

$$P_J(a,d,m,x) = P_J(-a-d,d,m,x).$$

We call this "antidiagonal symmetry" because of its appearance when we arrange the values of $P_J(a,d,m,x)$ in a matrix indexed by a and $a+d$. Second, Proposition 4.1 says that when m is a power of 2, $P_J(a,d,2^k,x)$ is independent of a.

We test these two symmetry predictions against actual data in Sections 5 and 6 respectively. The idea is that $\sum_{y=x+1}^{X} P_J(a,d,m,y)$ should closely approximate $N(a,d,m,X) - N(a,d,m,x)$. Therefore the matrix of $N(a,d,m,X)$, indexed by $(a,a+d)$, should show approximately these two symmetries.

In addition to these symmetries, the observations at the start of Section 4 imply that if we combine the terms in $1/(\log x)^2$ from $P_J(a,d,m,x)$, we obtain a sum that is independent of a. Heuristically, we might expect this term to be dominant as $x \to \infty$, and therefore independence of a should appear for large x. We test this prediction numerically in Section 8.

It would be interesting to have analytical proofs of any of these symmetries for the ratio $N(a,d,m,x)/\pi(x)$ in the limit as $x \to \infty$. As stated above, we do not know whether $N(a,d,m,x)/\pi(x)$ might be independent of both a and d in this limit.

We compare the actual values of

$$\sum_{y=x+1}^{X} P_J(a,d,m,y) \quad \text{and} \quad N(a,d,m,X) - N(a,d,m,x)$$

for $x = 10^3$, $X = 10^6$, and various small values of a, d, and m in Section 7. We begin our sum at 10^3, because our heuristic relies on x being large relative to a, d, and m. The alternating sums (1–9) become too large for feasible computation when X gets much beyond 10^6.

For any x, $P_J(1,2,2,x)$ should approximate $(\log x)^{-1}$. We test this for $J = 28$ and various x in Section 4.

We also discuss what happens when J varies. We prove an internal consistency, called "vertical compatibility," in Proposition 4.3 and Corollary 4.4. This asserts that if $m \mid n$ and J' is given, there exists J such that $P_J(a,d,m,x)$ equals a certain sum of $P_{J'}(a',d',n,x)$'s.

The function P_J appears to stabilize surprisingly quickly as a function of J. Experimentally, for small values of x, P_J still appears to remain approximately constant, even when J is much larger than $4\log x$. See the remarks at the end of Section 4. This stabilization as J varies is somewhat surprising; our heuristic is based on a probabilistic picture, whereas, for example, we know that the next prime after p is certainly less than $2p$. It would be interesting to have a theoretical grasp of how P_J varies with J.

We could try to eliminate the dependence on J by looking at

$$P(a,d,m,x) := \lim_{J\to\infty} P_J(a,d,m,x).$$

However, we see no way of proving that the limit exists. The sum defined by the limit is certainly not absolutely convergent (see Section 4 below). Our heuristic makes less and less sense if x is fixed and J increases.

2. NOTATION

To simplify the job of the reader, we record all of our considerable notation at the outset. Fix an integer $m > 1$ and a positive integer a. Let S be a finite set of nonnegative numbers containing 0. Let p be a prime. Let $x > 2$ be an integer. We present a list of some of our notation in Table 1.

The definition of n_p in (1–3) means the number of distinct residue classes modulo p in the set S. Once p is larger than the largest element in S, we have $n_p(S) = n(S)$. The product in (1–5) is infinite but converges, because for $p > n$, both the numerator and denominator have leading terms $1 - \frac{n}{p}$. The product in (1–6) contains only finitely many terms that do not equal 1, because once $p > n$ and $p > \max(S)$, the numerator and denominator are equal.

Below, we will typically think of a as a positive integer. However, a appears only as the first argument in $\beta(a,m,S)$. In turn, the definition of $\beta(a,m,S)$ depends only on a in testing the condition $(a+s,m) = 1$. Thus, $\beta(a,m,S)$, $Q(a,d,m,x,j)$, and $P_J(a,d,m,x)$ depend only on the value of $a \pmod m$.

3. THE HEURISTIC

We apply Pólya's heuristic to the problem of determining the following probabilities:

1. Given a positive integer d, the probability that x and $x+d$ are successive primes.
2. Given integers $a, d \geq 1$ and $m \geq 2$, the probability that x and $x+d$ are successive primes and $x \equiv a \pmod m$.

$$n = n(S) = \operatorname{card}(S) \tag{1–2}$$

$$n_p = n_p(S) = \operatorname{card}(S \bmod p) \tag{1–3}$$

$$\left(1 - \frac{n}{p}\right)^{\natural} = \begin{cases} 1 & \text{if } p \le n \\ 1 - \dfrac{n}{p} & \text{otherwise} \end{cases} \tag{1–4}$$

$$A_n = \prod_p \frac{\left(1 - \dfrac{n}{p}\right)^{\natural}}{\left(1 - \dfrac{1}{p}\right)^{n}} \tag{1–5}$$

$$\alpha(S) = \prod_p \frac{\left(1 - \dfrac{n_p(S)}{p}\right)}{\left(1 - \dfrac{n(S)}{p}\right)^{\natural}} \tag{1–6}$$

$$\beta(a,m,S) = \begin{cases} \frac{1}{m} \prod_{p \mid m} \frac{p}{p - n_p(S)} & \text{if } (a+s,m) = 1 \text{ for all } s \in S \\ 0 & \text{otherwise} \end{cases} \tag{1–7}$$

$$S_k = \{0, 1, \ldots, k\} \tag{1–8}$$

$$Q(a,d,m,x,j) = \sum_{\{0,d+jm\} \subseteq S \subseteq S_{d+jm}} (-1)^{n(S)} \alpha(S) \beta(a,m,S) \frac{A_{n(S)}}{(\log x)^{n(S)}} \tag{1–9}$$

$$P_J(a,d,m,x) = \sum_{j=0}^{J} Q(a,d,m,x,j) \tag{1–10}$$

TABLE 1. Summary of notation.

3. Given integers $a, d \ge 1$ and $m \ge 2$, the probability that x is prime, $x \equiv a \pmod m$, and the next prime is congruent to $a + d \pmod m$.

3.1. The Probability That x and $x + d$ Are Successive Primes

Let S be a nonempty finite set of nonnegative integers. Let p be a prime number. Then for integers x,

$$\operatorname{Prob}(p \nmid x + k \text{ for } k \in S) \sim \frac{p - n_p}{p},$$

where $n_p = n_p(S)$ is defined in (1–3). Note that $1 \le n_p \le p$. Furthermore, if $n_p = p$, then each residue class modulo p is represented by some $k \in S$; hence $p \mid x + k$ for some $k \in S$, so that $\operatorname{Prob}(p \nmid x + k \text{ for } k \in S)$ is indeed 0. Also $n_p = n := \operatorname{card} S$ as soon as p is sufficiently large (p larger than the largest element in S will suffice).

In [Pólya 59] a way is proposed to pass from the probability that "$p \nmid x + k$ for $k \in S$" to the probability that "$x + k$ is prime for $k \in S$." Pólya argues that the events "$p \nmid x + k$ for $k \in S$" for various primes p should be considered heuristically independent as long as p is small compared with x, so that we may conclude that

$$\operatorname{Prob}(p \nmid x + k \text{ for } k \in S \text{ for all } p < y) \sim \prod_{p<y} \frac{p - n_p}{p}$$

as long as x is sufficiently larger than y. Now if $x > y > x^{1/2}$, then

$$\begin{aligned} &p \nmid x + k \text{ for } k \in S \text{ for all } p < y \\ &\quad \iff x + k \text{ is prime for } k \in S, \end{aligned}$$

at least if x is large enough that the size of the elements of S can be ignored. Pólya proposed that the correct value of y should be x^{μ}, where $\mu = e^{-\gamma}$, and $\gamma = 0.5772\ldots$ is Euler's constant. His justification for this "trick of the magic μ" is simply and solely that it works when $S = \{0\}$: indeed, by Mertens's theorem,

$$\prod_{p<x^{\mu}} \frac{p-1}{p} \sim \frac{1}{\log x},$$

while the probability that x is prime is also $\frac{1}{\log x}$, by the prime number theorem. Accordingly we propose

$$\operatorname{Prob}(x + k \text{ is prime for each } k \in S) \sim \prod_{p<x^{\mu}} \frac{p - n_p}{p}. \tag{3–1}$$

We rewrite (3–1) as follows:

$$\begin{aligned}
&\prod_{p<x^\mu} \frac{p-n_p}{p} \\
&\quad\sim \prod_{p\le n} \frac{p-n_p}{p} \prod_{n<p<x^\mu} \left(\frac{p-n_p}{p-n}\left(1-\frac{n}{p}\right)\right) \\
&\quad\sim \prod_{p\le n} \frac{p-n_p}{p} \prod_{n<p<x^\mu} \frac{p-n_p}{p-n} \prod_{n<p<x^\mu} \frac{\left(1-\frac{n}{p}\right)}{\left(1-\frac{1}{p}\right)^n} \\
&\qquad\times \prod_{n<p<x^\mu} \left(1-\frac{1}{p}\right)^n \\
&\quad\sim \prod_{p\le n} \frac{p-n_p}{p} \prod_{n<p<\infty} \frac{p-n_p}{p-n} \prod_{p<x^\mu} \frac{\left(1-\frac{n}{p}\right)^\natural}{\left(1-\frac{1}{p}\right)^n} \\
&\qquad\times \prod_{p<x^\mu} \left(1-\frac{1}{p}\right)^n. \qquad (3\text{–}2)
\end{aligned}$$

The second product may be viewed as a finite product, because $n_p = n$ for p sufficiently large.

The first two products are $\alpha(S)$, as defined in (1–6). The third product approaches A_n, as defined in (1–5). The fourth product satisfies

$$\prod_{p<x^\mu} \left(1-\frac{1}{p}\right)^n \sim \frac{1}{(\log x)^n}$$

by Mertens's theorem. Putting all this together gives

$$\text{Prob}(x+k \text{ is prime for } k\in S) \sim \alpha(S)\frac{A_n}{(\log x)^n}. \quad (3\text{–}3)$$

As a check, let d be a positive integer and take $S = \{0, d\}$. We should retrieve the Hardy–Littlewood formula. In fact,

$$\begin{aligned}
A_2 &= \prod_p \frac{\left(1-\frac{2}{p}\right)^\natural}{\left(1-\frac{1}{p}\right)^2} = 4\prod_{p>2} \frac{\left(1-\frac{2}{p}\right)}{\left(1-\frac{1}{p}\right)^2} \\
&= 4\prod_{p>2}\left(1-\frac{1}{(p-1)^2}\right),
\end{aligned}$$

and $n_p(S) = 1$ if $p \mid d$, with $n_p(S) = 2$ otherwise. Hence,

$$\alpha(\{0,d\}) = \begin{cases} 0 & \text{if } d \text{ is odd,} \\ \frac{1}{2}\prod_{p\mid d, p>2} \frac{p-1}{p-2} & \text{if } d \text{ is even,} \end{cases}$$

which agrees with the formula from [Hardy and Littlewood 23] quoted in the introduction (note that our A_2 is four times Hardy and Littlewood's C_2).

Let d be a positive integer. We want to apply the formula (3–3) to find the probability that x and $x+d$ are successive primes. Let $S_d = \{0, 1, \ldots, d\}$. Then by inclusion/exclusion,

$$\begin{aligned}
&\text{Prob}(x \text{ is prime and } x+d \text{ is the } \textit{next} \text{ prime}) \\
&\quad= \sum_{\{0,d\}\subseteq S\subseteq S_d} (-1)^{n(S)}\alpha(S)\frac{A_{n(S)}}{(\log x)^{n(S)}}.
\end{aligned}$$

3.2. The Probability That x and $x+d$ Are Successive Primes and $x \equiv a \pmod m$

We start over with a congruence condition. Let $m \ge 2$ and $a \ge 1$ be integers. Then

$$\text{Prob}(x \equiv a \ (\text{mod} m)) \sim \frac{1}{m}.$$

Suppose that p is a prime divisor of m. Then the congruence $x \equiv a \bmod m$ determines whether $p \mid x+k$ for any $k \in S$. Indeed, if $a+k$ is not relatively prime to m for some $k \in S$, then

$$\text{Prob}(x \equiv a \ (\text{mod } m) \text{ and } x+S \text{ are primes}) = 0.$$

Here $x+S$ denotes $\{x+s : s \in S\}$. On the other hand, if $p \nmid m$, the conditions "$p \nmid x+k$ for $k \in S$" should be independent of congruences modulo m (and independent of each other), heuristically speaking. So if $a+k$ *is* relatively prime to m for every $k \in S$, then

$$\begin{aligned}
&\text{Prob}(x \equiv a \ (\text{mod } m) \text{ and } x+S \text{ are primes}) \\
&\quad\sim \frac{1}{m}\prod_{\substack{p<x^\mu \\ p\nmid m}} \frac{p-n_p}{p}.
\end{aligned}$$

So

$$\begin{aligned}
&\text{Prob}(x \equiv a \ (\text{mod } m) \text{ and } x+S \text{ are primes}) \\
&\quad\sim \begin{cases} \dfrac{1}{m}\displaystyle\prod_{p<x^\mu\, p\nmid m} \frac{p-n_p}{p} & \text{if } (a+k,m)=1 \ \forall k\in S, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

In the first case, when the elements of $a+S$ are all relatively prime to m, we have $n_p < p$ for all $p \mid m$, and therefore

$$\begin{aligned}
\frac{1}{m}\prod_{\substack{p<x^\mu \\ p\nmid m}} \frac{p-n_p}{p} &= \frac{1}{m}\prod_{p\mid m}\frac{p}{p-n_p}\prod_{p<x^\mu}\frac{p-n_p}{p} \\
&\sim \frac{1}{m}\prod_{p\mid m}\frac{p}{p-n_p}\alpha(S)\frac{A_n}{(\log x)^n}.
\end{aligned}$$

Thus

$$\begin{aligned}
&\text{Prob}(x \equiv a \ (\text{mod } m) \text{ and } x+S \text{ are primes}) \\
&\quad\sim \beta(a,m,S)\alpha(S)\frac{A_n}{(\log x)^n},
\end{aligned}$$

where $\beta(a, m, S)$ is defined in (1–7).

Let d be a positive integer. Then by inclusion/exclusion,

$$\begin{aligned}
&\text{Prob}(x \text{ is prime, } x \equiv a \pmod m, \\
&\qquad \text{and } x + d \text{ is the } \textit{next} \text{ prime}) \\
&\sim \sum_{\{0,d\} \subseteq S \subseteq S_d} (-1)^{n(S)-2} \text{Prob}(x \equiv a \pmod m \\
&\qquad \text{and } x + S \text{ are primes}) \\
&\sim \sum_{\{0,d\} \subseteq S \subseteq S_d} (-1)^{n(S)} \alpha(S) \beta(a, m, S) \frac{A_{n(S)}}{(\log x)^{n(S)}}. \\
&= Q(a, d, m, x, 0).
\end{aligned}$$

3.3. The Probability That x Is Prime, $x \equiv a \pmod m$, and the Next Prime Is $\equiv a + d \pmod m$

Finally, suppose that $1 \le d \le m$ and that $(a, m) = 1$ and $(a + d, m) = 1$. Then

$$\begin{aligned}
&\text{Prob}(x \equiv a \pmod m, \ x \text{ prime}, \\
&\qquad \text{and the next prime is} \equiv a + d \bmod m) \\
&\sim P_J(a, d, m, x) \\
&= \sum_{j=0}^{J} \sum_{S} (-1)^{n(S)} \alpha(S) \beta(a, m, S) \frac{A_{n(S)}}{(\log x)^{n(S)}},
\end{aligned}$$

where the second sum is over $\{0, d + jm\} \subseteq S \subseteq S_{d+jm}$. One possibility is to let j range from 0 to ∞. However, there are various finite upper bounds that can be used. We need to use a large enough value of J to ensure that the set $x + S_{d+jm}$ contains the next prime after x (assuming that x is prime). Bertrand's postulate guarantees that choosing J such that $d + Jm \ge x$ is sufficient. We can safely use a much smaller value of J, however. The average gap between primes of size x is $\log x$, and the standard deviation is also $\log x$. We can be *almost* certain that $x + S_{d+jm}$ will contain the next prime if $d + Jm \ge 4 \log x$.

It is worth recalling Pólya's remark in [Pólya 59, p. 384, footnote]:

> When we consider a fixed number of primes, the "probabilities" introduced can be regarded as "independent," but they cannot be so regarded when the number of primes increases in an arbitrary manner.

Unfortunately, as a practical matter, we are only able to choose J such that $d + Jm$ is smaller than 55 so that $J = \lfloor (55 - d)/m \rfloor$; otherwise, the number of subsets of S_{d+jm} is too large to be computationally feasible. Thus in computing our heuristic, we need to limit ourselves to $4 \log x \le 55$, or x less than roughly 10^6. We will discuss this further below.

4. SOME OBSERVATIONS

We observe the following:

- If S contains any odd integers, then $\alpha(S) = 0$, because $n_2(S) = 2$.
- If $(a, m) \ne 1$, then $\beta(a, m, S) = 0$, because $0 \in S$. Therefore, we will henceforth assume that $(a, m) = 1$.
- If $\{0, d + jm\} \subseteq S$ and $(a + d, m) \ne 1$, then $\beta(a, m, S) = 0$.
- If $a \equiv a' \pmod m$, then $\beta(a, m, S) = \beta(a', m, S)$.
- If $\beta(a, m, S) \ne 0$, then $\beta(a, m, S)$ is independent of a.

Proposition 4.1. *If m is a power of 2, then $\alpha(S)\beta(a, m, S)$ is independent of a. Hence, if $m = 2^k$ and a is odd, then $P_J(a, d, m, x)$ is independent of a.*

Proof. Because $(a, m) = 1$, we know that a is odd. If S contains any odd elements, then $\alpha(S) = 0$. If S contains only even numbers, then $a + S$ will contain only odd numbers, and hence will be relatively prime to m, regardless of the value of a. □

It is tempting to rearrange the terms in (1–10) in order of increasing $n(S)$ and let $J \to \infty$. However, the sum of the terms with $n(S) = 2$ is divergent. This is true in general, but it is simplest to see when $m = 2^k$. In that case, we must have a odd and d even in order for $\alpha(S)\beta(a, m, S) \ne 0$. The terms with $n(S) = 2$ all have the form $S = \{0, d + jm\}$. In this situation, we can easily compute that

$$\beta(a, m, S) = \left(\frac{1}{2^k}\right)\left(\frac{2}{1}\right) = \frac{1}{2^{k-1}}.$$

On the other hand, $\alpha(S)$ will always be larger than $\frac{1}{2}$, so all of the (infinitely many) terms with $n(S) = 2$ will be larger than $\frac{A_2}{2^k (\log x)^2}$, and hence will diverge.

Proposition 4.2. ("Antidiagonal symmetry.")

$$P_J(a, d, m, x) = P_J(-a - d, d, m, x).$$

Proof. We will show that for each set $\{0, d+jm\} \subseteq S \subset S_{d+jm}$ in the sum in (1–10) for $P_J(a,d,m,x)$, there is a set $\{0, d+jm\} \subseteq S' \subseteq S_{d+jm}$ with $n(S) = n(S')$, $\alpha(S) = \alpha(S')$, and $\beta(a,m,S) = \beta(-a-d,m,S')$.

In particular, set $S' = (d+jm) - S$. We automatically have $\{0, d+jm\} \subseteq S' \subseteq S_{d+jm}$. Obviously, $n(S) = n(S')$, and for any prime p, $n_p(S) = n_p(S')$. This shows that $\alpha(S) = \alpha(S')$.

Finally, suppose that for every element $s \in S$, we have $(a+s,m) = 1$, so $\beta(a,m,S) \neq 0$. Then $(-a-s,m) = 1$, and so $(jm-a-s,m) = 1$. Therefore, $((d+jm-s)+(-a-d),m) = 1$, showing that $\beta(-a-d,m,S') \neq 0$. We know that $n_p(S) = n_p(S')$, and therefore if $\beta(a,m,S) \neq 0$, we see that $\beta(a,m,S) = \beta(-a-d,m,S')$. Conversely, if $\beta(a,m,S) = 0$, the same argument shows that $\beta(-a-d,m,S') = 0$. We may conclude that $P_J(a,d,m,x) = P_J(-a-d,d,m,x)$. □

Proposition 4.3. ("Vertical compatibility.") *Fix J'. Suppose that n is a positive integer and $m \mid n$. Then*

$$P_J(a,d,m,x) = \sum_{\substack{a' \equiv a \pmod m \\ d' \equiv d \pmod m \\ 1 \le a', d' \le n}} P_{J'}(a',d',n,x),$$

where $(J+1)m = (J'+1)n$.

Proof. We will first show that each set S_{d+jm} used in the sum on the left-hand side of the equation appears in the sum on the right-hand side of the equation for a unique d'. To see this, choose $d' \equiv d + jm \pmod n$ with $1 \le d' \le n$. Write $(d+jm) - d' = j'n$, and then $d + jm = d' + j'n$. On the other hand, if we start with $d' \equiv d \pmod m$ and $j' \ge 0$, it is clear that we can find a unique j such that $d + jm = d' + j'n$. We know that $j' \le J'$ and $d' \le n + d - m$. Therefore, the upper bound for J is $\frac{(J'+1)n-m}{m}$.

This leaves us needing to show that

$$\beta(a,m,S) = \sum_{\substack{a' \equiv a \pmod m \\ 1 \le a' \le n}} \beta(a',n,S).$$

Notice that if $\beta(a,m,S) = 0$, then automatically $\beta(a',n,S) = 0$ for all possible values of $a' \equiv a \pmod m$. So we may assume that $\beta(a,m,S) \neq 0$.

It suffices to consider the case $n = qm$, with q prime. Notice that the set $\{a, a+m, a+2m, \ldots, a+(q-1)m\}$ contains the complete set of integers a' with $a' \equiv a \pmod m$ and $1 \le a' \le qm$. In particular, there are q such integers a'.

There are two cases to finish the proof:

Case 1: $q \mid m$. If $a + S$ contains only elements relatively prime to m, and $a' \equiv a \pmod m$, then $a' + S$ contains only elements relatively prime to m. Hence, $a' + S$ contains only elements relatively prime to qm. Therefore, $\beta(a,m,S) = 0 \iff \beta(a',qm,S) = 0$. Note also that if p is prime, then $p \mid qm \iff p \mid m$.

If $\beta(a',qm,S) \neq 0$, then we know that $\beta(a',qm,S)$ is independent of a'. There are q possible values of a', and we have

$$\begin{aligned}
\sum_{a'} \beta(a',qm,S) &= \sum \Big(\frac{1}{qm} \prod_{p \mid qm} \frac{p}{p - n_p(S)} \Big) \\
&= \sum \Big(\frac{1}{qm} \prod_{p \mid m} \frac{p}{p - n_p(S)} \Big) \\
&= q \Big(\frac{1}{qm} \prod_{p \mid m} \frac{p}{p - n_p(S)} \Big) \\
&= \frac{1}{m} \prod_{p \mid m} \Big(\frac{p}{p - n_p(S)} \Big) = \beta(a,m,S).
\end{aligned}$$

Case 2: $q \nmid m$. In this case, the elements of the set $\{a, a+m, \ldots, a+(q-1)m\}$ are distributed over all q of the residue classes modulo q. Suppose that $n_q(S) = c$. Then there are $q - c$ possible residue classes r modulo q such that $r + S$ will be relatively prime to q, and there are c residue classes such that $r + S$ will have a factor in common with q (namely $r \equiv -s \pmod q$ for some $s \in S$). Let r' be one of the residue classes with $r' + S$ relatively prime to q. Recall as usual that if $\beta(a,qm,S) \neq 0$, then it is independent of a.

We have

$$\begin{aligned}
&\sum_{\substack{a' \equiv a \pmod m \\ 1 \le a' \le qm}} \beta(a',qm,S) \\
&\quad = (q - n_q(S)) \beta(r',qm,S) \\
&\quad = (q - n_q(S)) \frac{1}{qm} \prod_{p \mid qm} \frac{p}{p - n_p(S)} \\
&\quad = (q - n_q(S)) \frac{1}{qm} \left(\frac{q}{q - n_q(S)} \right) \prod_{p \mid m} \frac{p}{p - n_p(S)} \\
&\quad = \frac{1}{m} \prod_{p \mid m} \frac{p}{p - n_p(S)} = \beta(a,m,S).
\end{aligned}$$

□

Corollary 4.4. *Choose an even integer $m > 2$, fix J', and let $J = \frac{(J'+1)m-2}{2}$. Then*

$$\sum_{\substack{1 \le a \le m \\ 1 \le d \le m}} P_{J'}(a,d,m,x) = P_J(1,2,2,x). \tag{4–1}$$

Proof. Proposition 4.3 reduces the left-hand side of (4–1) to

$$P(1,2,2,x) + P(1,1,2,x) + P(2,1,2,x) + P(2,2,2,x).$$

The last three terms in this sum are all 0.

Note that if for some reason it is desirable to begin with odd m, we may replace m with $2m$ without changing the sum on the left-hand side of (4–1), again because of Proposition 4.3. □

Theoretically, $P_J(1,2,2,x)$ should give the probability that x is prime, which according to our heuristic is $(\log x)^{-1}$. We can sum only to $J = 28$, because the computational time involved becomes inordinate, and we tabulate our results here:

x	$P_{28}(1,2,2,x)$	$(\log x)^{-1}$	Ratio
10	0.425586	0.434294	0.9799
10^2	0.217147	0.217147	1.0000
10^3	0.144765	0.144765	1.0000
10^4	0.108567	0.108574	0.9999
10^5	0.0867455	0.0868589	0.9987
10^6	0.0719549	0.0723824	0.9941

We conjecture that the relatively poor agreement with prediction when $x = 10$ is caused by replacing a finite product over the range $p < x^\mu$ with the infinite product defining A_n in the penultimate term of formula (3–2). The product defining A_n converges relatively rapidly, but there is still an appreciable error introduced by including so many more terms when $x = 10$. The fact that the ratio is slowly tending away from 1 might show that for $x \geq 10^4$, J should be larger than 28 for best results.

More interestingly, the computations for $x = 10^2$ and $x = 10^3$ converged relatively rapidly and *did not change significantly* when more terms were included in the sum. For $x = 10^2$, the series attained the value $0.217\ldots$ for $J = 13$. Extending the sum over larger values of J did not change the first three significant digits. The sum took the value $0.21715\ldots$ at $J = 20$, and that did not change up to our maximum value of $J = 28$. For $x = 10^3$, the summation takes values varying between $0.144765\ldots$ and $0.144767\ldots$ when J runs from 16 to 28. It is hard to understand this apparent stabilization of P_J for varying J; our heuristic makes less and less sense as J increases for fixed x.

5. NUMERICAL CHECKS OF ANTIDIAGONAL SYMMETRY

Take, for example, $m = 5$. We count residues modulo 5 of consecutive prime pairs $p < q$ with $p < 982451653$. (982451653 is the 50 millionth prime number). We then record our results in a matrix with $\varphi(m)$ rows and columns. The results for $m = 5$ are

2289170 3778890 3732547 2698886
3189954 2190360 3386288 3734018
2995506 3535854 2191584 3777311
4024863 2995514 3189838 2289414

where a_{ij} records the count of ($p \equiv i \pmod 5$, $q \equiv j \pmod 5$). (The pairs $(3,5)$ and $(5,7)$ are thus omitted from the counts in this table.)

To check antidiagonal symmetry, we take the ratio of each entry to its reflection across the antidiagonal. We obtain (rounded to six places)

0.999893 1.000418 0.999606 1.000000
1.000036 0.999441 1.000000 1.000394
0.999997 1.000000 1.000559 0.999582
1.000000 1.000003 0.999964 1.000107

The entries are all within 0.0006 of 1, while the ratio of the largest to the smallest entry in the original array is about 1.84.

If $m = 12$, we have the following table of counts of prime pair residues ($p \bmod 12$, $q \bmod 12$), where $p < q$ are consecutive primes with $p < 982451653$; we omit the pairs $(2,3)$ and $(3,5)$ that are not relatively prime to 12:

2265842 3746000 3296656 3190380
2944446 2266005 3994598 3295820
3294707 3190968 2268555 3745878
3993884 3297895 2940299 2268064

For example, the entry 3994598 records the number of pairs of consecutive primes congruent to $(5,7) \pmod{12}$.

Again, our heuristic predicts that this matrix should be symmetric across the antidiagonal. Taking the ratio of each entry to its reflection across the antidiagonal, we obtain (rounded to six places)

0.999020 1.000033 1.000254 1.000000
1.001410 0.998876 1.000000 0.999746
0.999033 1.000000 1.001125 0.999967
1.000000 1.000968 0.998592 1.000981

409015	997313	**843077**	1082276	749027	721892	804113	642323
643202	408017	995715	**843672**	1083407	748537	722996	805076
803581	643323	408783	996936	**842431**	1082780	749737	723085
721382	805281	642633	408177	997195	**843473**	1081932	749402
750223	721241	804604	642735	407612	997681	**843417**	1082432
1082176	750074	723110	803572	643435	407018	996935	**843628**
842474	1082071	751349	721474	804368	643327	408482	996565
996983	**843301**	1081386	750633	722470	805240	642498	407695

TABLE 2. Residues modulo 16 of consecutive prime pairs $p < q$ with $p < 982451653$.

The entries are all within 0.002 of 1, while the ratio of the largest to the smallest entry in the original array is about 1.76.

6. NUMERICAL CHECKS OF POWER-OF-2 PREDICTION

As noted above, when $m = 2^k$, our heuristic predicts that $N(a, d, 2^k, x)$ should be independent of a. Taking $m = 16$, we count residues modulo 16 of consecutive prime pairs $p < q$ with $p < 982451653$, with the result shown in Table 2.

Our conjecture predicts that the numbers on the "broken diagonals" parallel to the main diagonal should be approximately equal. For example, the numbers in boldface in the table form a "broken diagonal." We can test our conjecture by noting that the counts in the matrix range from 407018 to 1083407, with a ratio approximately 2.66, while the counts in the broken diagonal in boldface range from 842431 to 843672, with a ratio of 1.00147.... The next broken diagonal (beginning with 1082276) varies from 1081386 to 1083407, with a ratio of 1.00186.... The remaining broken diagonals exhibit similar behavior.

7. NUMERICAL CHECKS OF THE HEURISTIC

In Sections 5 and 6, we tested the heuristic indirectly, by observing symmetries in the heuristic and seeing whether the same symmetries appeared in the counts of residues of prime pairs. In this section we compare our heuristic formula directly with such counts. This requires computing the heuristic numerically, which, as noted above, becomes rapidly unwieldy as J increases, so that we must restrict J so that $d + Jm < 55$ in our calculations, and correspondingly restrict x to be at most 10^6.

We will use $m = 4$, 5, 11, and 12 as typical examples. We count prime pairs reduced modulo m between 10^3 and 10^6:

$m = 5$

3207 6536 6284 3550
5053 2868 5430 6224
4383 5800 2810 6630
6934 4371 5100 3150

$m = 12$

2942 6315 5253 5018
4358 2959 7034 5216
5257 5010 2918 6438
6970 5283 4419 2940

We compute $\sum_{x=10^3}^{10^6} P_J(a, d, m, x)$ for all possible values of a and d for each value of m above. Because of limited computer time, we chose J maximal with $d + Jm < 55$. The results are as follows:

$m = 5$

3136.4 6562.3 6237.6 3570.7
5081.4 2738.6 5464.4 6237.6
4360.4 5838.2 2738.6 6562.3
6946.5 4360.4 5081.4 3136.4

$m = 12$

2960.0 6418.9 5258.5 4877.2
4279.9 2960.0 7013.5 5258.5
5258.5 4877.2 2960.0 6418.9
7013.5 5258.5 4279.9 2960.0

To gauge the success of the heuristic in predicting the counts, we take the ratios of the corresponding entries in these tables, that is, we calculate, for each congruence class $(i, j) \pmod m$, the ratio (actual count)/(predicted

$m = 11$

263	1071	1242	522	1019	426	1625	617	771	285
819	200	836	1181	485	1106	329	1603	501	765
346	814	217	987	1227	675	1063	324	1555	588
834	328	817	272	834	1221	499	1060	318	1640
682	1015	362	1028	271	1095	1223	663	1049	448
1618	534	775	256	794	275	866	1219	485	1036
382	1826	594	784	253	1004	262	1004	1230	518
1055	366	1599	596	779	387	803	198	867	1208
604	1072	316	1795	577	1004	355	836	192	1062
1238	599	1038	403	1597	665	832	333	845	273

$\sum_{x=10^3}^{10^6} P_J(a, d, 11, x)$

263.7	1048.5	1242.9	497.9	1029.7	434.8	1642.4	606.3	778.9	257.8
802.7	191.4	851.5	1219.9	486.6	1077.4	313.4	1595.3	488.1	778.9
335.3	823.1	193.7	983.5	1205.5	655.8	1069.6	331.0	1595.3	606.3
801.8	332.3	800.4	237.2	865.3	1250.3	493.8	1069.6	313.4	1642.4
656.4	1000.8	359.0	1018.3	230.0	1117.3	1250.3	655.8	1077.4	434.8
1627.2	560.9	777.0	252.1	780.6	230.0	865.3	1205.5	486.6	1029.7
389.7	1806.0	592.5	805.6	252.1	1018.3	237.2	983.5	1219.9	497.9
1072.6	322.0	1596.1	592.5	777.0	359.0	800.4	193.7	851.5	1242.9
615.4	1103.7	322.0	1806.0	560.9	1000.8	332.3	823.1	191.4	1048.5
1245.0	615.4	1072.6	389.7	1627.2	656.4	801.8	335.3	802.7	263.7

Ratios

1.00	1.02	1.00	1.05	0.99	0.98	0.99	1.02	0.99	1.11
1.02	1.04	0.98	0.97	1.00	1.03	1.05	1.00	1.03	0.98
1.03	0.99	1.12	1.00	1.02	1.03	0.99	0.98	0.97	0.97
1.04	0.99	1.02	1.15	0.96	0.98	1.01	0.99	1.01	1.00
1.04	1.01	1.01	1.01	1.18	0.98	0.98	1.01	0.97	1.03
0.99	0.95	1.00	1.02	1.02	1.20	1.00	1.01	1.00	1.01
0.98	1.01	1.00	0.97	1.00	0.99	1.10	1.02	1.01	1.04
0.98	1.14	1.00	1.01	1.00	1.08	1.00	1.02	1.02	0.97
0.98	0.97	0.98	0.99	1.03	1.00	1.07	1.02	1.00	1.01
0.99	0.97	0.97	1.03	0.98	1.01	1.04	0.99	1.05	1.04

TABLE 3. Actual counts and predicted values for $m = 11$.

count), with the following results, rounded to two places:

$m = 5$

1.02 1.00 1.01 0.99
0.99 1.05 0.99 1.00
1.01 0.99 1.03 1.01
1.00 1.00 1.00 1.00

$m = 12$

0.99 0.98 1.00 1.03
1.02 1.00 1.00 0.99
1.00 1.03 0.99 1.00
0.99 1.00 1.03 0.99

By comparison, the ratio of the largest to the smallest count is $6934/2810 \approx 2.47$ (for $m = 5$), and $7034/2918 \approx 2.41$ (for $m = 12$).

Here are the results for $m = 4$, again for the range 10^3 to 10^6:

$m = 4$: 16574 22521
22520 16715

$\sum_{x=10^3}^{10^6} P_J(a, d, 4, x)$: 16618.8 22407.7
22407.7 16618.8

ratios: 1.00 1.01
1.01 1.01

Finally, for $m = 11$, the calculations appear in Table 3. The counts for $m = 11$ vary from 192 to 1826, with ratio $1826/192 \approx 9.51$. Our predicted values are for the most part quite close to the observed counts, though the predictions for a few entries (notably, and curiously, $(3,3), (4,4), (5,5), (6,6), (7,7)$) are a bit low, only 85%–90% of their observed counts. This is partly explained by the size of the entries themselves—the main diagonal entries are the smallest, in both the table of counts and the table of predictions, so a discrepancy there between prediction and observation shows up as a larger percentage error than the same discrepancy elsewhere. For example, the observed count for $(5,5)$ is 41 above the predicted count, or 18%; the observed count for $(10,9)$ is 42 above predicted, or 5%; the observed count for $(2,4)$ is 39 below predicted, or 3%.

8. ANOTHER PREDICTION

If we let $x \to \infty$ and fix J, we can view the dominant terms in $P_J(a,d,m,x)$ as those with $n(S) = 2$. (In reality, the estimation is more complex, because the value of J should increase with x.) If $n(S) = 2$, then $S = \{0, d + jm\}$. As long as $(a,m) = (a+d,m) = 1$, these "dominant" terms in $P_J(a,d,m,x)$ will be independent of a.

To see whether in fact this prediction appears to be true, we compute prime pair counts modulo 5 and 12 for the first 10^7 (probable) primes larger than 10^{100}. The results are

$$m = 5$$

$$\begin{array}{cccc} 300655 & 318723 & 320000 & 310947 \\ 313970 & 300412 & 315948 & 319266 \\ 312837 & 317325 & 300976 & 318355 \\ 322830 & 313721 & 314012 & 300023 \end{array}$$

$$m = 12$$

$$\begin{array}{cccc} 301604 & 318324 & 315239 & 314667 \\ 313320 & 300633 & 319515 & 316479 \\ 315294 & 315144 & 300873 & 318316 \\ 320319 & 315883 & 314055 & 300335 \end{array}$$

These data are consistent with prediction: for example, for $m = d = 5$ (respectively $m = d = 12$), part of the prediction is that the entries on the main diagonal should tend to equality; an ad hoc way to judge this is to take the ratio of the largest to the smallest terms on the diagonal for both our first set of residues, involving primes less than 10^6, and this set. For $m = 5$ (respectively $m = 12$), the ratio for the "small-prime" data set is $\frac{1638}{1390} \approx 1.18$ (respectively $\frac{1501}{1430} \approx 1.05$), while for the "large-prime" data set, the ratio is $\frac{300976}{300023} \approx 1.003$ (respectively $\frac{301604}{300335} \approx 1.004$).

9. FURTHER OPEN QUESTIONS

One obvious question is whether the different residue classes of prime pairs occur asymptotically equally often: in other words, given any a, d, a', d' with $a, a+d, a', a'+d'$ relatively prime to m, do we have

$$\frac{N(a,d,m,x)}{N(a',d',m,x)} \to 1$$

as $x \to \infty$?

This corresponds to asking whether all of the entries in the matrices in the previous section are tending toward equality as x tends to infinity. One way to gauge the variation in our tables is to take the ratio of the largest to the smallest counts. In Section 7, the ratio for $m = 5$ is approximately 2.47 (respectively 2.39 for $m = 12$), while for the "large prime" data set in Section 8, the ratio for $m = 5$ is approximately 1.08 (respectively 1.07). Obviously, the terms appear to be getting closer together, but of course we cannot tell whether they are tending toward a limiting ratio of 1.

If the different residue classes of prime pairs do occur asymptotically equally often, we are in a "prime race" situation, as studied in [Rubinstein and Sarnak 94] and [Granville and Martin 06]. For example, take $m = 4$ and count prime pairs congruent to $(1,1) \pmod 4$ versus prime pairs congruent to $(1,3) \pmod 4$. Even if the counts are asymptotically equal, i.e., if $\lim_{x\to\infty} N(1,4,4,x)/N(1,2,4,x) = 1$, we observe for finite chunks of primes a decided tendency for the total of $(1,3)$'s to exceed the total of $(1,1)$'s. We can keep score, as we search through the sequence of odd primes $3, 5, 7, \ldots$, and see when the $(1,1)$'s are ahead of the $(1,3)$'s and when the opposite holds. It seems quite possible that $N(1,4,4,x) > N(1,2,4,x)$ for almost all values of x.

Another question concerns the appropriate value of J to use in our heuristic. The computation mentioned at the end of Section 4 shows that in at least one case, $P_J(a,d,m,x)$ is nearly independent of J for a range of values of J. A precise understanding of the dependence of $P_J(a,d,m,x)$ on J would go a long way toward establishing the solidity of our heuristic.

ACKNOWLEDGMENTS

We thank Jenny Baglivo and K. Soundarajan for their help, and the referee for helpful comments. The first and second authors thank the National Science Foundation for partial funding of this research under grant DMS-0455240.

REFERENCES

[Ash et al. 09] A. Ash, B. Bate, and R. Gross. "Frequencies of Successful Tuples of Frobenius Classes." *Experiment. Math.* 18 (2009), 55–63.

[Bateman and Horn 62] Paul T. Bateman and Roger A. Horn. "A Heuristic Asymptotic Formula Concerning the Distribution of Prime Numbers." *Math. Comp.* 16 (1962), 363–367.

[Granville and Martin 06] Andrew Granville and Greg Martin. "Prime Number Races." *Amer. Math. Monthly* 113 (2006), 1–33.

[Hardy and Littlewood 23] G. H. Hardy and J. E. Littlewood. "Some Problems of Partitio Numerorum III: On the Expression of a Number as a Sum of Primes." *Acta Math.* 44 (1923), 1–70.

[Pólya 59] G, Pólya. "Heuristic Reasoning in the Theory of Numbers." *Amer. Math. Monthly* 66 (1959), 375–384.

[Rubinstein and Sarnak 94] Michael Rubinstein and Peter Sarnak. "Chebyshev's Bias." *Experiment. Math.* 3 (1994), 173–197.

Avner Ash, Department of Mathematics, Boston College, Chestnut Hill, MA 02467-3806, USA (ashav@bc.edu)

Laura Beltis, Actuarial Analysis Department, Tufts Health Plan, Watertown, MA 02472, USA (lbeltis@gmail.com)

Robert Gross, Department of Mathematics, Boston College, Chestnut Hill, MA 02467-3806, USA (gross@bc.edu)

Warren Sinnott, Department of Mathematics, The Ohio State University, Columbus, OH 43210-1174, USA (sinnott@math.ohio-state.edu)

Received February 2, 2010; accepted May 11, 2010.