STRINGS OF CONGRUENT PRIMES

D. K. L. SHIU

In 1920 Chowla made the following conjecture. Let p_n denote the nth prime; if $q \geqslant 3$, $(q, a) = 1$ then there are infinitely many pairs of consecutive primes p_n and p_{n+1} such that

$$p_n \equiv p_{n+1} \equiv a \bmod q.$$

By considering the sum

$$\sum_p \chi(p),$$

where χ is the non-principal character modulo 4 or 6, it is possible to prove the conjecture for $q = 4$ and $q = 6$ ($a = \pm 1$). In this paper we prove Chowla's conjecture for all q and a with $(q, a) = 1$. Moreover, we shall show that for any k there exist 'strings' of congruent primes such that

$$p_{n+1} \equiv p_{n+2} \equiv \ldots \equiv p_{n+k} \equiv a \bmod q.$$

For each modulus q the method used applies best to the following two sets of residue classes:

$$A_+ := \{a : \forall\, p | q, a \equiv 1 \bmod p\}$$
$$A_- := \{a : \forall\, p | q, a \equiv -1 \bmod p\}.$$

Larger values of k in terms of p_{n+i} can be found for residue classes belonging to these sets.

THEOREM 1. (i) *For each q and $a \in A_\pm$ and large x, there exists a string of primes*

$$p_{n+1} \equiv p_{n+2} \equiv \ldots \equiv p_{n+k} \equiv a \bmod q,$$

where $p_{n+k} < x$ and

$$k \gg \left(\frac{\log\log x}{\log\log\log x}\right)^{1/\phi(q)}.$$

(ii) *For each q, a with $(q, a) = 1$ and large x, there exists a string of primes*

$$p_{n+1} \equiv p_{n+2} \equiv \ldots \equiv p_{n+k} \equiv a \bmod q,$$

where $p_{n+k} < x$ and

$$k \gg \left(\frac{\log\log x \log\log\log\log x}{(\log\log\log x)^2}\right)^{1/\phi(q)}.$$

Received 28 April 1997; revised 10 January 1998.

2000 *Mathematics Subject Classification* 11N05 (primary), 11N13 (secondary).

It is natural to ask how frequently such strings occur. If for given q, k we define

$$\varepsilon_1(x) := C(q)\,k\left(\frac{\log\log\log x}{\log\log x}\right)^{1/\phi(q)}$$

and

$$\varepsilon_2(x) := C'(q)\,k\left(\frac{(\log\log\log x)^2}{\log\log x\log\log\log\log x}\right)^{1/\phi(q)},$$

where $C(q)$, $C'(q)$ are constants depending on at most q, then we have the following theorem.

THEOREM 2. (i) *For given q, k and $a \in A_{\pm}$ and large x the number B of strings of the form*

$$p_{n+1} \equiv p_{n+2} \equiv \ldots \equiv p_{n+k} \equiv a \bmod q,$$

where $p_{n+k} < x$ satisfies

$$B \gg x^{1-\varepsilon_1(x)}.$$

(ii) *For given q, k, a with $(q, a) = 1$ and large x the number B of strings of the form*

$$p_{n+1} \equiv p_{n+2} \equiv \ldots \equiv p_{n+k} \equiv a \bmod q,$$

where $p_{n+k} < x$ satisfies

$$B \gg x^{1-\varepsilon_2(x)}.$$

In particular the number B_2 of pairs of consecutive congruent primes congruent to $a \bmod q$ satisfies

$$B_2 \gg x\exp\left(\left(-\frac{\theta(q)\log\log\log x}{\log\log x}\right)^{1/\phi(q)}\right)$$

if $a \in A_{\pm}$ and

$$B_2 \gg x\exp\left(\left(-\frac{\theta(q)\,(\log\log\log x)^2}{\log\log x\log\log\log\log x}\right)^{1/\phi(q)}\right)$$

otherwise ($\theta(q)$ is a constant depending on at most q).

1. *Outline of the proof*

The proof is adapted from Maier's proof of the existence of chains of large gaps between consecutive primes [**5**]. We shall define $Q(y)$ as the product of a subset of the primes $p < y$. These primes shall be chosen in such a way that the L-functions modulo Q have no Siegel zeros. It then follows from a theorem of Gallagher (Lemma 2) that the distribution of primes in arithmetic programmes mod $Q(y)$ is regular.

The variable x in Theorem 1 will be of the order of $Q(y)^D$, where D is a fixed constant. Since $y \approx \log Q(y)$, we have $y \approx \log x$. We introduce a further variable z; our method will be to examine a set of intervals of length yz. The set of intervals will be dense in primes congruent to a mod q and thin in other primes. One of the intervals can be shown to have a string required for the proof of Theorem 1.

We shall choose $Q(y)$ and an interval $(m, m+yz]$ such that the elements of $(m, m+yz]$ which are relatively prime to $Q(y)$ fall mainly in the residue class $a \bmod q$. We now use the matrix construction of Maier. Our matrix is a set of integers arranged in $Q(y)^{D-1}$ rows and yz columns. Each row is an interval of yz consecutive integers. Each column is an arithmetic progression with common difference $Q(y)$. The upper left-hand element of the matrix is $m+1+Q(y)$.

We consider the set of primes contained in the matrix. Clearly any prime in the matrix must lie in a column whose elements are relatively prime to $Q(y)$. Our choices of interval and $Q(y)$ will mean that most of these columns are made up of numbers congruent to $a \bmod q$. By Lemma 2, each column will contain about the expected number of primes. It follows that the majority of primes in the matrix are congruent to $a \bmod q$.

We now must find one of our $Q(y)^{D-1}$ intervals which contains a string. The number of primes congruent to $a \bmod q$ exceeds other primes by a factor $\alpha(y,z)$. The expected number of primes in each interval is $yz/\log x \gg z$. It follows that there is either an interval where the number of primes congruent to $a \bmod q$ exceeds other primes by a factor $\alpha(y,z)/2$, or an interval with $k \gg z$ primes congruent to $a \bmod q$ and no other primes. In the first case there would be a string of length $k \gg \alpha(y,z)$ and in the second a string of length $k \gg z$. Theorem 1 follows.

To prove Theorem 2 we use a similar construction but allow D to vary with x and k. The number of strings B will satisfy $B \gg x/Q(y) = Q(y)^{D-1}$. We shall show that a positive proportion of our $Q(y)^{D-1}$ intervals contain such strings. For a given $0 < \theta_2 < 1$ we can choose z so that the number of primes not congruent to $a \bmod q$ in the matrix is at most $\theta_2 Q(y)^{D-1}$. This means the proportion of intervals containing such a prime is at most θ_2. We shall then show that there exist D and $\theta_1 > \theta_2$ such that $\theta_1 Q(y)^{D-1}$ intervals of our matrix contain a string of length k. If such a θ_1 did not exist, then our primes congruent to $a \bmod q$ would lie mostly in a small number of intervals. These intervals would have an unusually large number of pairs of primes p_i, p_j such that $q|(p_i - p_j)$. Using sieve methods to estimate the number of such pairs in our matrix we find that this is not the case. It follows that a proportion $\theta_1 - \theta_2$ of our intervals contain strings of length k and Theorem 2 follows provided that $1/D < \varepsilon_1, 1/D < \varepsilon_2$ respectively.

2. *Basic lemmas*

We require estimates for the number of small primes in arithmetic progressions modulo $Q(y)$. These are given in Lemma 2, provided that the L-functions modulo $Q(y)$ have no zero in a region

$$1 \geqslant \Re s > 1 - \frac{C}{\log[Q(y)(|\Im s|+1)]}.$$

Lemma 1 will be used to show that we can choose $Q(y)$ to satisfy this criterion. We shall also require estimates for the number of elements of the interval $(m, m+yz]$ which are coprime to $Q(y)$. Our choice of m and $Q(y)$ shall be such that to obtain such estimates we must estimate the number of elements of $(0, yz]$ whose prime factors are all congruent to $1 \bmod q$. We shall estimate these in Lemma 3 which is a generalisation of Landau's work counting sums of two squares. Additionally, for the cases when $a \notin A_{\pm}$ we shall require an estimate for elements of $(0, yz]$ all of whose prime factors are less than a parameter t. This estimate is given in Lemma 4 and is due to de Bruijn.

We define

$$P(y,p_0) := q\prod_{\substack{p\leqslant y\\ p\neq p_0}} p.$$

LEMMA 1. *There exists a fixed constant C such that for all natural q and large X there exists y and a prime $p_0 \gg \log y$ such that none of the L-functions modulo $P(y,p_0)$ has a zero in the region*

$$1 \geqslant \Re s > 1 - \frac{C}{\log[P(y,p_0)(|\Im s|+1)]},$$

and

$$X < P(y,p_0) \ll X(\log X)^2.$$

Proof. Page's theorem states that there exists a constant C_1 such that for any modulus q' the L-functions mod q' have no zeros in the region

$$1 \geqslant \Re s > 1 - \frac{C_1}{\log[q'(|\Im s|+1)]},$$

with the exception of at most one real zero of an L-function generated by a real character. Consider the product

$$P'(y) := q\prod_{p\leqslant y} p.$$

Suppose that there exists a character χ_1 mod $P'(y)$ whose L-function has a real zero β in the range

$$1 \geqslant \beta \geqslant 1 - \frac{C_1}{\log P'}.$$

χ_1 is induced by a character χ_1' mod P'' where $P''|P'(y)$. The L-function generated by χ_1' will also have a zero at β. By the class number formula there exists a constant C_2 such that

$$\beta < 1 - \frac{C_2}{P''^{1/2}}.$$

We deduce that $P'' \gg (\log\ P'(y))^2$. Since $P''|P'(y)$ we know that P'' is squarefree in all prime factors greater than q and hence has a prime divisor p_0 satisfying $p_0 \gg \log P'' \gg \log\log P'(y) \gg \log y$ (provided that $P'(y)$ is sufficiently large relative to q). We use this p_0 in our definition of $P(y,p_0)$ whenever β exists.

Suppose now that there is a real character χ_2 mod $P(y,p_0)$ whose L-function has a real zero β' in the range

$$1 > \beta' > 1 - \frac{C_1}{2\log P(y,p_0)}.$$

Then χ_2 induces a character χ_2' mod $P'(y)$ and the L-function generated by χ_2' will have a zero at β'. We observe that $\log P(y,p_0) = \log P'(y) - \log p_0 > \log P'(y)/2$ and so β' lies in the range

$$1 > \beta' > 1 - \frac{C_1}{\log P'(y)}.$$

We further observe that $p_0|P'', p_0 \nmid P(y,p_0) \Rightarrow P'' \nmid P(y,p_0)$ and deduce that $\chi_2' \neq \chi_1$. β and β' are therefore zeros to two different L-functions in the interval $(1-C_1/\log P'(y), 1]$. This contradicts Page's theorem. We conclude that no such χ_2 exists. If χ_1 does not exist, we simply take p_0 to be any prime greater than $\log y$ and because $P(y,p_0)|P'(y)$ it has no zero in the interval $(1-C_1/2\log P(y,p_0), 1]$. We take $C = C_1/2$ and it only remains to show that such a $P(y,p_0)$ exists in the range $X < P(y,p_0) \ll X(\log X)^2$. We consider the sequence of products $P'(p_n)$ where p_n denotes the nth prime. We see that $P'(p_n) = P'(p_{n+1})/p_{n+1} > P'(p_n)/2\log P'(p_n)$. It follows that there exists a y with

$$X\log X < P'(y) \ll X(\log X)^2.$$

Removing a prime $p_0 \leqslant y$ from our product as above we have

$$X < P(y,p_0) \ll X(\log X)^2. \qquad \square$$

LEMMA 2. *Let C be a constant and let q' be a natural number such that the L-functions induced by characters* mod q' *have no zeros in the region*

$$1 \geqslant \Re s > 1 - \frac{C}{\log[q'(|\Im s|+1)]}.$$

Then there exists a constant D depending on at most C such that the estimates

$$\frac{x}{\phi(q')\log x} \ll \pi(x;q',a') \ll \frac{x}{\phi(q')\log x}$$

hold uniformly for $(q',a') = 1$ and $x \geqslant q'^D$.

Proof. The upper bound is a weak form of the Brun–Titchmarsh inequality. An unconditional proof of this for $D = 2+\varepsilon$ can be found in [**2**]. The lower bound is a theorem of Gallagher, a proof of which can be found in [**4**]. $\square$

LEMMA 3. *Let q be a natural number and let $\mathscr{S}(x)$ denote the set of positive integers $n \leqslant x$ which are composed only of primes congruent to* 1 mod q. *Then as $x \to \infty$ we have*

$$|\mathscr{S}(x)| = \left(c_0 + O\left(\frac{1}{\log x}\right)\right)\frac{x}{\log x}(\log x)^{1/\phi(q)},$$

where

$$c_0 := \frac{1}{\Gamma(1/\phi(q))}\lim_{s\longrightarrow 1}(s-1)^{1/\phi(q)}\prod_{p\equiv 1 \bmod q}\left(1-\frac{1}{p^s}\right)^{-1}$$

is a constant depending on at most q.

Proof. The proof is a generalisation of Landau's work on sums of two squares (see [3]). We define

$$\mathscr{S}' := \{n : p|n \Rightarrow p \equiv 1 \bmod q\},$$

the characteristic function

$$a(n) := \begin{cases} 1 & \text{for } n \in \mathscr{S}' \\ 0 & \text{otherwise,} \end{cases}$$

and the Dirichlet series generated by $a(n)$

$$f(s) := \sum_{n=1}^{\infty} \frac{a(n)}{n^s} = \prod_{p \equiv 1 \bmod q} \left(1 - \frac{1}{p^s}\right)^{-1}.$$

We write

$$\Theta(s) := \frac{\prod_{\chi \bmod q} L(s,\chi)}{(f(s))^{\phi(q)}}.$$

It follows that

$$\begin{aligned}
\log \Theta(s) &= \log\left(\prod_{\chi \bmod q} L(s,\chi)\right) - \phi(q) \log f(s) \\
&= \sum_{\chi} \sum_{p^m} \frac{\chi(p^m)}{mp^{ms}} - \phi(q) \sum_{p \equiv 1 \bmod q} \sum_{m} \frac{1}{mp^{ms}} \\
&= \sum_{p^m \equiv 1 \bmod q} \frac{\phi(q)}{mp^{ms}} - \sum_{\substack{p^m \\ p \equiv 1 \bmod q}} \frac{\phi(q)}{mp^{ms}} \\
&= \sum_{\substack{p^m \equiv 1 \bmod q \\ p \not\equiv 1 \bmod q}} \frac{\phi(q)}{mp^{ms}}.
\end{aligned}$$

We note that in our expression for $\log \Theta(s)$ there is no term with $m = 1$. We deduce that

$$\begin{aligned}
\Theta(s) &= \exp(\log \Theta(s)) \\
&= \prod_{p \not\equiv 1 \bmod q} \left(\exp\left(\sum_{\substack{m \geqslant 2 \\ p^m \equiv 1 \bmod q}} m^{-1} p^{-ms} \right)\right)^{\phi(q)} \\
&= \prod_{p \not\equiv 1 \bmod q} \left(1 + \left(\sum_{\substack{m \geqslant 2 \\ p^m \equiv 1 \bmod q}} m^{-1} p^{-ms} \right) + \frac{(\sum_{\substack{m \geqslant 2 \\ p^m \equiv 1 \bmod q}} m^{-1} p^{-ms})^2}{2!} + \dots \right)^{\phi(q)} \\
&= \sum_{n=1}^{\infty} \frac{b(n)}{n^s},
\end{aligned}$$

where $b(n) = 0$ if n is divisible by a prime congruent to 1 mod q or if there exists a prime $p : p | n, p^2 \nmid n$. This means that $\Theta(s)$ is regular, non-zero and absolutely convergent for $\Re s > 1/2$. The product $\Pi_{\chi} L(s,\chi)$ has a simple pole at $s = 1$ and is non-zero in the region

$$D : \left\{ s : 2 \geqslant \Re s \geqslant 1 - \frac{c}{\log T} \right\} \cap \{ s : |\Im s| < T \},$$

for all large T. We can therefore write

$$f(s) = (s-1)^{-1/\phi(q)} g(s),$$

where $g(s)$ is a function regular in D.

By the usual argument [**5**, Lemma 3.19] we have

$$|\mathscr{S}(x)| = \frac{1}{2\pi i} \int_{1+(1/\log x)-iT}^{1+(1/\log x)+iT} f(s) \frac{x^s}{s} ds + O(1) + O(x \log x / T).$$

Consider the contour shown in Figure 1.

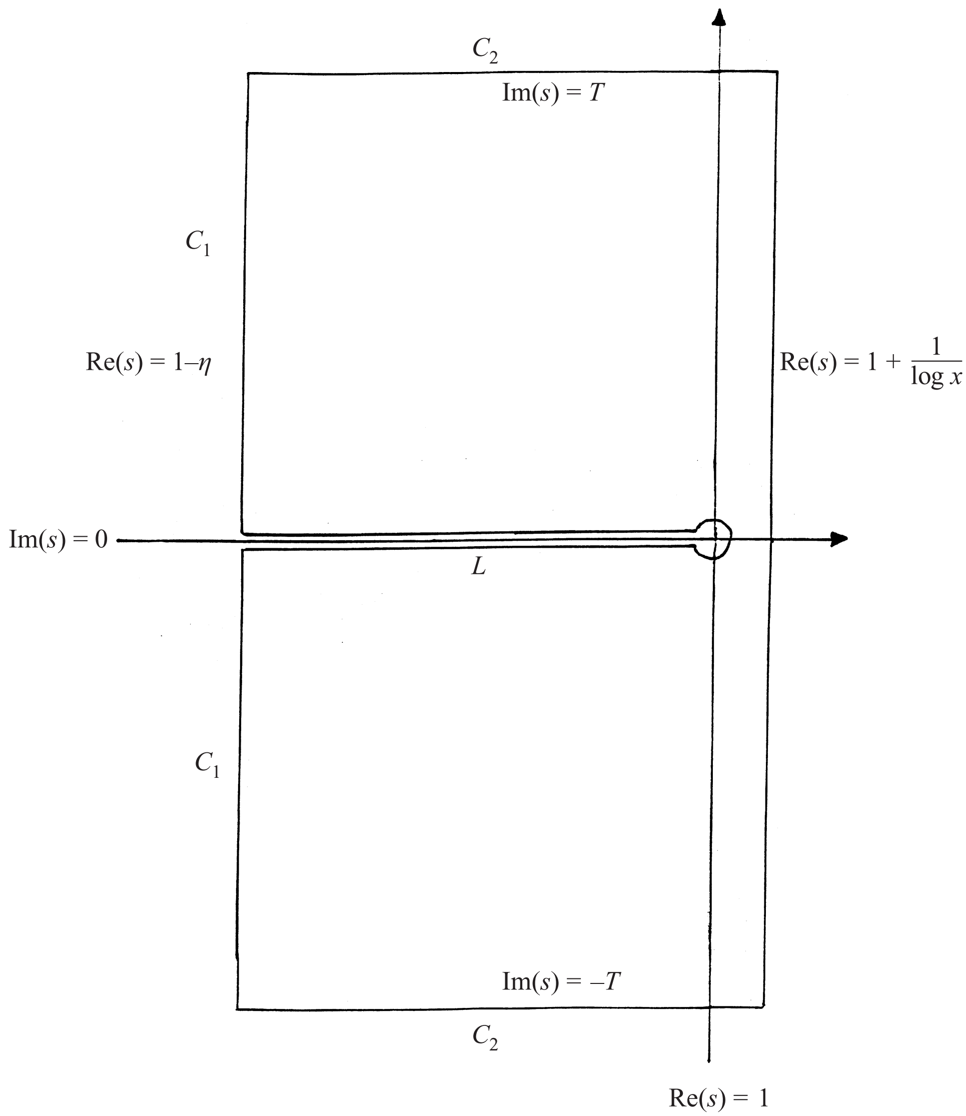

FIGURE 1.

We have

$$|\mathscr{S}(x)| = \frac{1}{2\pi i}\int_{L+C_1+C_2} f(s)\frac{x^s}{s}\,ds.$$

The main term of our theorem is due to the integral around L. Writing

$$h(s) := \frac{g(s)}{g(1)},$$

we have $h(s) = 1 + O(|s-1|)$ along L and upon shrinking the loop of the contour to zero we have

$$\begin{aligned}\frac{1}{2\pi i}\int_L f(s)\frac{x^s}{s} = ds\frac{g(1)}{2\pi i}\int_L \frac{x^s}{(s-1)^{1/\phi(q)}}h(s)\,ds \\ = \frac{g(1)\sin(\pi/\phi(q))}{\pi}\int_{1-\eta}^{1}\frac{x^s}{(1-s)^{1/\phi(q)}}ds + O\left(\frac{x(\log x)^{1/\phi(q)}}{(\log x)^2}\right).\end{aligned} \tag{1}$$

Removing a factor $x(\log x)^{1/\phi(q)}$ from the last integrand gives

$$\begin{aligned}\int_{1-\eta}^{1}\frac{x^s}{(1-s)^{1/\phi(q)}}ds &= x(\log x)^{1/\phi(q)}\int_{1-\eta}^{1}\frac{x^{s-1}}{((1-s)\log x)^{1/\phi(q)}}ds \\ &= \frac{x(\log x)^{1/\phi(q)}}{\log x}\int_0^{\eta\log x}\exp(-u)\,u^{-1/\phi(q)}du.\end{aligned} \tag{2}$$

The last integral is derived from the substitution $u = (1-s)\log x$ and we now approximate it by $\Gamma(1-1/\phi(q))$:

$$\begin{aligned}\int_0^{\eta\log x}\exp(-u)\,u^{-1/\phi(q)}du &= \int_0^{\infty}\exp(-u)\,u^{-1/\phi(q)}du - \int_{\eta\log x}^{\infty}\exp(-u)\,u^{-1/\phi(q)}du \\ &= \Gamma\left(1-\frac{1}{\phi(q)}\right) + O(x^{-\eta}).\end{aligned} \tag{3}$$

The identity $\Gamma(\theta)\Gamma(1-\theta) = \pi\,\mathrm{cosec}(\pi\theta)$ combined with (1), (2) and (3) give us

$$\frac{1}{2\pi i}\int_L f(s)\frac{x^s}{s}ds = (c_0 + O(x^{-\eta} + 1/\log x))\frac{x(\log x)^{1/\phi(q)}}{\log x}.$$

The usual error estimates give

$$\int_{C_1} f(s)\frac{x^s}{s}ds = O(x^{1-\eta}T(\log T)^A)$$
$$\int_{C_2} f(s)\frac{x^s}{s}ds = O(x(\log x)^{1/\phi(q)}T^{-1}),$$

for a fixed positive constant A. We take

$$T = (\log x)^3, \quad \eta = \frac{c}{\log T}.$$

Our contour lies in D as required and our error terms are $O(x(\log x)^{-2})$. Lemma 3 follows. □

LEMMA 4. *Let $\Psi(x, y)$ denote the number of positive integers $n \leqslant x$ which are composed only of primes $p \leqslant y$. For $y \leqslant x$ and y approaching infinity with x, we have*

$$\Psi(x, y) \leqslant x(\log y)^2\exp(-u\log u - u\log\log u + O(u)),$$

where $u := \log x/\log y$.

Proof. This is a result of de Bruijn and a proof can be found in [**1**]. □

3. *Proof of Theorem* 1

For a given q, a, x and sufficiently large D we use Lemma 1 to choose y and p_0 such that

$$x^{1/D} \leqslant P(y, p_0) \ll x^{1/D}(\log x)^2$$

and such that there is no L-function modulo $P(y, p_0)$ with an exceptional zero. We introduce variables $z < y$ and $t \leqslant (yz)^{1/2}$ and define

$$\mathscr{P}_a = \begin{cases} \{p \leqslant y : p \neq p_0, p \not\equiv 1 \bmod q\} & \text{for } a \text{ in } A_{\pm}, \\ \{p \leqslant y : p \neq p_0, p \not\equiv 1, a \bmod q\} & \\ \cup\, \{t \leqslant p \leqslant y : p \neq p_0,\ p \equiv 1 \bmod q\} & \\ \cup\, \{p \leqslant yz/t : p \neq p_0,\ p \equiv a \bmod q\} & \text{otherwise.} \end{cases}$$

We now define $Q(y)$ mentioned in our outline by

$$Q(y) := q \prod_{p \in \mathscr{P}_a} p.$$

We note that $Q(y) | P(y, p_0)$ and that $\log P < 3 \log Q$. We conclude that the L-functions mod $Q(y)$ have no zeros in the region

$$1 \geqslant \Re s > 1 - \frac{C}{3 \log[Q(y)(|\Im s| + 1)]},$$

because such a zero would induce a zero in an L-function mod $P(y, p_0)$ at the same point, contradicting our choice of y, p_0. Also, for this choice of $Q(y)$ we have $x^{1/2D} < Q(y) < x^{1/D}$ which enables us to prove Theorem 1 for any large x. We define the interval I by

$$I := \begin{cases} (m, m + yz] & \text{for } a \in A_+ \\ [n - yz, n) & \text{for } a \in A_- \\ (0, yz] & \text{otherwise,} \end{cases}$$

where

$$m \equiv \begin{cases} 0 & \bmod p \text{ for } pq | Q \\ a - 1 & \bmod q, \end{cases}$$

$$n \equiv \begin{cases} 0 & \bmod p \text{ for } pq | Q \\ a - 1 & \bmod q. \end{cases}$$

We note that by our definitions of A_+ and A_- we have $m, n \equiv 0 \bmod p$ for all $p | Q$. We now construct the matrix described in Section 1. We label this matrix M; it is defined as the following set of integers:

$$M := \bigcup_{k=1}^{Q(y)^{D-1}} \bigcup_{i \in I} (i + kQ(y)).$$

We also define the sets

$$S := \{i \in I : (i, Q) = 1, i \equiv a \bmod q\},$$
$$T := \{i \in I : (i, Q) = 1, i \not\equiv a \bmod q\},$$
$$P_1 := \{p \in M : p \equiv a \bmod q, p \text{ prime}\},$$
$$P_2 := \{p \in M : p \not\equiv a \bmod q, p \text{ prime}\}.$$

We shall estimate $|P|$ and $|P_2|$ by combining Lemma 2 with estimates for $|S|$ and $|T|$ respectively. We obtain estimates for $|S|$ and $|T|$ from Lemma 3.

In the case of Theorem 1(i), we have $a \in A_+$, $i \in S \Leftrightarrow i-m \equiv 1 \bmod q$ and $(i, Q) = 1 \Leftrightarrow (i-m, Q) = 1$. Similarly, we have $a \in A_-$, $i \in S \Leftrightarrow n-i \equiv 1 \bmod q$ and $(i, Q) = 1 \Leftrightarrow (n-i, Q) = 1$. It follows that

$$\begin{aligned} a \in A_{\pm} \Rightarrow |S| &= |\{j \in (0, yx] : (j, Q) = 1, j \equiv 1 \bmod q\}|, \\ |T| &= |\{j \in (0, yz] : (j, Q) = 1, j \not\equiv 1 \bmod q\}|. \end{aligned}$$

Since $p = 1 \bmod q \Rightarrow p \nmid Q$ and the product of primes congruent to 1 $\bmod q$ is congruent to 1 mod q we have

$$|S| \geqslant |\mathscr{S}(yz)| \gg \frac{yz(\log y)^{1/\phi(q)}}{\log y},$$

where $\mathscr{S}(x)$ is defined as in Lemma 3. If $j \in (0, yz]$, $j \not\equiv 1 \bmod q$ then there exists a prime $p|j$ such that $p \not\equiv 1 \bmod q$. We estimate $|T|$ by estimating the number of elements $pn : p > y, n \in \mathscr{S}(z)$ and multiples of p_0 in the interval $(0, yz]$. There are $O(yz/\log y)$ multiples of p_0 in $(0, yz]$ and so we concentrate on the elements of the first type. We split the interval $(y, yz]$ into $O(\log z)$ intervals of length $2^l y$ and deduce that

$$\begin{aligned} |T| &\ll \sum_{l \ll \log z} \sum_{2^{l-1} y < p \leqslant 2^l y} \sum_{\substack{n \leqslant z/2^l \\ n \in S(z)}} \\ &\ll \sum_{l \ll \log z} \frac{2^{l-1} y}{\log y} \frac{z(\log z)^{1/\phi(q)}}{2^l \log z} \\ &\ll \frac{yz(\log z)^{1/\phi(q)}}{\log y}. \end{aligned}$$

In the case of Theorem 1(ii), elements of S are now of the form $pn : p > yz/t$, $p \equiv a \bmod q$, $n \in \mathscr{S}(t)$ in the interval $(0, yz]$. By another splitting argument we have

$$|S| \gg \frac{yz(\log t)^{1/\phi(q)}}{\log y}.$$

Elements of T are multiples of p_0 or multiples of a prime greater than y or composed solely of primes less than t and congruent to 1 mod q. We estimate the first two types as before. The third type we estimate by Lemma 4. We take

$$t = \exp\left(\theta \frac{\log y \log\log\log y}{\log\log y}\right).$$

Our third type of elements is therefore of order

$$\Psi(yz, t) < yz(\log t)^2 \exp(-\theta^{-1} \log\log y + o(\log\log y)) \ll \frac{yz}{\log y},$$

provided that θ is sufficiently small (in fact we can take $\theta = 1/4$). We therefore have the estimate

$$|T| \ll \frac{yz(\log z)^{1/\phi(q)}}{\log y},$$

in both Theorem 1(i) and (ii).

Each element of S and T is the first term in an arithmetic progression mod $Q(y)$. Each of these arithmetic progressions will contain primes. It is now possible to estimate $|P_1|$ and $|P_2|$ by Lemma 2. We recall that $Q(y)$ is a 'good' modulus and so we can choose D such that

$$|P_1| \gg |S| \frac{x}{\phi(Q)\log x}, \quad |P_2| \ll |T| \frac{x}{\phi(Q)\log x},$$

for $x \geqslant Q^D$. We now argue two possible cases. We write M' for the subset of intervals of M which contain a prime belonging to P_2. Then either there exists an interval in M' where the primes belonging to P_1 exceed those belonging to P_2 by a factor $|P_1|/2|P_2|$ or the number of primes in $M \backslash M'$ is at least $|P_1|/2$, that is, either there exists

$$I_0 \in M' : |I_0 \cap P_1| \geqslant \frac{1}{2}\frac{|P_1|}{|P_2|}|I_0 \cap P_2|,$$

or

$$|(M \backslash M') \cap P_1| \geqslant \frac{1}{2}|P_1|.$$

We know that one of these cases must arise, otherwise

$$\begin{aligned} |P_1| &= |P_1 \cap M'| + |P_1 \cap (M \backslash M')| \\ &= \sum_{I \in M'} |P_1 \cap I| + |P_1 \cap (M \backslash M')| \\ &< \frac{1}{2}\frac{|P_1|}{|P_2|} \sum_{I \in M'} |I \cap P_2| + \frac{1}{2}|P_1| \\ &= \frac{1}{2}\frac{|P_1|}{|P_2|}|P_2| + \frac{1}{2}|P_1| = |P_1|. \end{aligned}$$

This is clearly a contradiction and so one of our two cases arises. In the first case it follows that our interval I_0 contains a string of length k where $k \gg |P_1|/|P_2|$. In the second case we note that there are at most x/Q intervals in $M \backslash M'$ and so one of them must contain a string of length k where $k \gg Q|P_1|/x$. Now

$$\frac{|P_1|Q}{x} = |S| \frac{Q}{\phi(Q)\log x}.$$

Returning to our definition of Q we have

$$\frac{Q}{\phi(Q)} = \frac{q}{\phi(q)} \prod_{p \in \mathscr{P}} \frac{p}{p-1} = \frac{q}{\phi(q)} \prod_{p \in \mathscr{P}} \left(1 - \frac{1}{p}\right)^{-1}.$$

By a generalisation of Merten's theorem we have $Q/\phi(Q) \gg (\log y)^{1/\phi(q)}/\log y$ if $a \in A_{\pm}$ and $Q/\phi(Q) \gg (\log t)^{1/\phi(q)}/\log y$ otherwise. It follows that

$$\frac{|P_1|Q}{x} \gg \frac{yz}{\log x} \gg z,$$

because $\log x \ll Q \ll y$.

It follows that there exists in M a string of length k where

$$k \gg \min\left(\frac{|P_1|}{|P_2|}, z\right).$$

In case (i)

$$k \geqslant \min\left(\left(\frac{\log y}{\log z}\right)^{1/\phi(q)}, z\right),$$

and in case (ii)

$$k \geqslant \min\left(\left(\frac{\log t}{\log z}\right)^{1/\phi(q)}, z\right).$$

Theorem 1 follows upon taking $z = \log\log x$. □

4. *Proof of Theorem* 2

The previous section took D to be constant. If we allow D to vary, our estimates remain the same apart from that for y. Instead of writing $y \gg \log x$ we now have to write $y \gg (\log x)/D$. For a given x and k we shall take ε such that there exists a Q with $\log Q/\log x = \varepsilon$ which produces strings of length at least k in our construction. If a proportion $\theta > 0$ of our x/Q intervals contain such a string, then we have $B \geqslant \theta x/Q = \theta x^{1-\varepsilon}$. Theorem 2(i) will follow provided that $\varepsilon < \varepsilon_1$ and similarly Theorem 2(ii) will follow provided that $\varepsilon < \varepsilon_2$. In our adapted construction we shall choose $z = \alpha(\log t/\log\log t)^{1/\phi(q)}$ for some $\alpha > 0$ (throughout this section, if $a \in A_{\pm}$ read $t = y$). Given $\Theta > 0$ we can choose $\alpha > 0$ such that for any given ε our construction with the new choice of Q and z yields

$$P_2 \leqslant \Theta \frac{xy}{Q \log x}.$$

Hence θ_2, the proportion of intervals in M', satisfies $\theta_2 < \Theta\varepsilon$. We can find a constant C such that for any sufficiently large D our altered construction yields

$$P_1 \geqslant Cz \frac{x \log Q}{Q \log x}$$

for $Q < x^D$. We choose a Q such that $2k < Cz \log Q/\log x < 3k$ and set $\varepsilon = \log Q/\log x$. Note that this means that, on average, intervals of M contain at least $2k$ elements of P_1. This choice of ε and z means we can choose $C(q), C'(q)$ such that $\varepsilon < \varepsilon_1$ for Theorem 2(i) and $\varepsilon < \varepsilon_2$ for Theorem 2(ii). We define M^* to be the subset of intervals of M which contain at least k primes. Our intention is to show that the number of intervals in M^* is at least $\theta_1 x/Q$ for some $\theta_1 > \theta_2$. We note that the number of primes in $M \backslash M^*$ is less than kx/Q. We deduce that M^* contains at least kx/Q primes. For any interval $J \in M^*$ we write $\alpha_J k$ for the number of primes in J. Note that $\alpha_J > 1$ and that

$$\sum_{J \in M^*} \alpha_J \geqslant \frac{x}{Q}.$$

We write $U(J)$ for the set of ordered pairs of primes in J:

$$U(J) := \{p_1, p_2 \in J : p_1 < p_2\}.$$

For $J \in M^*$ we have $|U(J)| \gg (\alpha_J k)^2$. We use this to obtain a lower bound for the number of ordered pairs of primes in intervals of M:

$$V := \{p_1, p_2 \in M : 0 < p_2 - p_1 < yz\},$$

and we have

$$|V| \geqslant \bigcup_{J \in M^*} U(J)| \gg k^2 \sum_{J \in M^*} \alpha_J^2.$$

If we take N to be a natural number such that the number of intervals in M^* is at least x/QN we have by Cauchy's inequality

$$\left(\frac{x}{Q}\right)^2 \leqslant \left(\sum_{J \in M^*} \alpha_J\right)^2 \leqslant \left(\sum_{J \in M^*} \alpha_J^2\right)\left(\sum_{J \in M^*} 1\right) \leqslant \frac{x}{QN} \sum_{J \in M^*} \alpha_J^2.$$

This gives us an estimate $\Sigma_J \alpha_J^2 \geqslant Nx/Q$ and a lower bound for $|V|$:

$$|V| \gg \frac{Nx}{Q} k^2.$$

We now bound $|V|$ from above using the following lemma.

Lemma 5. *Let i,j be integers coprime to Q with $0 < (j-i) < y^2$. Let Z approach infinity. Then $|W(Z,i,j)|$, the number of pairs of primes of the form $(kQ+i, kQ+j)$ with $1 \leqslant k \leqslant Z$, satisfies*

$$|W(Z,i,j)| \ll \frac{Z}{(\log Z)^2}\left(\frac{Q}{\phi(Q)}\right)^2 \left(\sum_{\substack{d|(j-i)\\(d,Q)=1}} \frac{1}{d}\right).$$

Proof. The result is proved by applying [**2**], Theorem 7.2. We write $\mathscr{B}$ for the set of odd primes and

$$\mathscr{A} := \bigcup_{1 \leqslant k \leqslant Z} \{n_1 n_2 : n_1 = kQ+i, n_2 = kQ+j\},$$

$$\omega(p) := \begin{cases} 0 & \text{for } p|Q, \\ 2 & \text{for } p \nmid Q, p \nmid (j-i), \\ 1 & \text{for } p \nmid Q, p|(j-i). \end{cases}$$

In the terminology of [**2**] we satisfy the conditions Ω_1 and $\Omega_2(2)$ and we have

$$|W(Z,i,j)| \leqslant \mathscr{S}(\mathscr{A}; \mathscr{B}, Z^{1/3}) \ll Z \prod_{p \leqslant Z^{1/3}} \left(1 - \frac{\omega(p)}{p}\right).$$

We rewrite the product as

$$\prod_{p \leqslant Z^{1/3}} \left(1 - \frac{2}{p}\right) \prod_{p|Q} \left(1 + \frac{2}{p-2}\right) \prod_{p|(j-i),\,(p,Q)=1} \left(1 + \frac{1}{p-2}\right).$$

We conclude that

$$\prod_{p \leqslant Z^{1/3}} \left(1 - \frac{\omega(p)}{p}\right) \ll \frac{1}{(\log Z)^2}\left(\frac{Q}{\phi(Q)}\right)^2 \sum_{\substack{d|(j-i)\\(d,Q)=1}} \frac{1}{d},$$

and the lemma follows. □

V, the number of ordered prime pairs in intervals of M, can be written as

$$V = \bigcup_{\substack{0<i<j<yz\\(ij,Q)=1}} W(x/Q, i, j).$$

Using Lemma 5, we see that

$$|V| \ll \frac{x}{Q(\log x)^2}\left(\frac{Q}{\phi(Q)}\right)^2 \sum_{i,j} \sum_{\substack{d \mid (j-i) \\ (d,Q)=1}} \frac{1}{d}.$$

We rewrite the double sum as

$$\sum_{d:(d,Q)=1} \frac{1}{d} \sum_{\substack{0<i<j<yz,\,(ij,Q)=1 \\ i \equiv j \bmod d}} 1.$$

We define

$$G(d, i_0) := |\{i \leqslant yz : (i, Q) = 1, i \equiv i_0 \bmod d\}|,$$

and use this to estimate the inner sum. Writing

$$\mathscr{A} := \{n \leqslant yz : n \equiv i_0 \bmod d\}, \quad \mathscr{B} := \{p : p | Q\},$$

we have, in the notation of [**2**], $\mathscr{S}(\mathscr{A}; \mathscr{B}, v) \geqslant |G(d, i_0)|$ for any $v \leqslant (yz)^{1/2}$. For $p \in \mathscr{B}$ we take $\omega(p) = 1$, $v = (yz)^{1/5}$ and the sieving conditions Ω_1 and $\Omega_2(1)$ are satisfied. We apply [**2**, Theorem 4.1] and we have

$$|G(d, i_0)| \ll \frac{yz\phi(Q)}{dQ},$$

for $d < (yz)^{1/2}$. For $d \geqslant (yz)^{1/2}$ we use the trivial estimate $|G(d, i_0)| \geqslant yz/d$. Writing

$$\sum_{\substack{0<i<j<yz,\,(ij,Q)=1 \\ i \equiv j \bmod d}} 1 = \sum_{i_0 \bmod d} \frac{1}{2}(|G(d,i_0)|^2 - |G(d,i_0)|) \ll d\left(\frac{yz\phi(Q)}{dQ}\right)^2,$$

we have

$$\begin{aligned}\sum_{d:(d,Q)=1} \frac{1}{d} \sum_{\substack{0<i<j<yz,\,(ij,Q)=1 \\ i \equiv j \bmod d}} 1 &\ll y^2z^2\left(\frac{\phi(Q)}{Q}\right)^2 \sum_{d<(yz)^{1/2}} \frac{1}{d^2} + \sum_{d \geqslant (yz)^{1/2}} \frac{|G(d,i_0)|^2}{d} \\ &\ll y^2z^2\left(\frac{\phi(Q)}{Q}\right)^2 + yz \sum_{d \geqslant (yz)^{1/2}} \frac{1}{d^3} \\ &\ll y^2z^2\left(\frac{\phi(Q)}{Q}\right)^2.\end{aligned}$$

We conclude that

$$|V| \ll \frac{xy^2z^2}{Q(\log x)^2} \ll \frac{x}{Q}\left(\frac{z \log Q}{\log x}\right)^2 \ll \frac{x}{Q}k^2.$$

Comparing this with our lower bound $|V| \geqslant xk^2/QN$ we deduce that there exists an effective constant N_0 such that $N \leqslant N_0$. We now set α in our definition of z so that

$$\theta_2(\alpha) < \frac{1}{N_0}.$$

We now have $\theta_1 x/Q$ intervals containing at least k primes and $\theta_2 x/Q$ intervals containing a prime not congruent to a mod q, where

$$\theta_1 > \frac{1}{N} \geqslant \frac{1}{N_0} > \theta_2.$$

Theorem 2 follows. □

5. *Observations and conjectures*

The theorems also hold if we examine strings of primes $p_j, n < j \leqslant n+k$ for which $\chi(p_j) = \chi(a)$ (χ being a character mod q) instead of congruence classes. The results are analogous with the exponent $1/\phi(q)$ being replaced by $1/d$ throughout (here $d{:}\chi^d(n) = \chi_0(n)$). In particular it is possible to find $(\log\log x/\log\log\log x)^{1/2}$ consecutive primes all of which are quadratic residues (respectively non-residues) mod q.

On a heuristic level, we expect the probability of a given prime being congruent to a mod q to be $1/\phi(q)$. Assuming congruence classes of consecutive primes to be independent, we would then expect the probability that the prime p_j is the start of a string of k primes to be $1/\phi(q)^k$. It would seem unlikely that we would find such a prime in the first $\phi(q)^k$ primes, but we might expect to find one shortly thereafter. We might therefore conjecture that our upper bound x for the least such prime would satisfy

$$\frac{x}{\log x} \ll \phi(q)^k.$$

This would give us an estimate for our string length of $k \gg \log x/\log\phi(q)$. In our construction this would correspond to using the primes up to y to construct $Q(y)$ and to produce an interval such that the number of elements relatively prime to $Q(y)$ and congruent to a mod q is of size $y^2/\log y$ (and at most $y/\log y$ other relatively prime elements). We would expect such an interval to be of length y^2. This would be analogous to constructing a prime gap of size $(\log x)^2$. Our conjecture therefore seems reasonable, but beyond current methods.

We might also hope to remove the dependence on a from our theorems, and that the lengths and densities demonstrated for $a \in A_\pm$ could be demonstrated for all a. There appears to be little hope of this using the de Bruijn estimates. Some variation on the reflection/translation of the interval $(0, yz]$ might be workable, though no such variation occurs to the author.

Acknowledgements. I would like to thank D. R. Heath-Brown for his invaluable supervision of this work.

References

1. N. G. De Bruijn, 'On the number of positive integers $\leqslant x$ and free of prime factors $\geqslant y$' *Indag. Math.* 13 (1951) 50–60.
2. H. Halberstam and H. E. Richert, *Sieve methods* (Academic Press, London, 1974).
3. G. H. Hardy, *Ramanujan* (Cambridge University Press, 1940).
4. H. Maier, 'Chains of large gaps between consecutive primes', *Adv. Math.* 39 (1981) 257–269.
5. E. C. Titchmarsh, *The theory of the Riemann zeta-function* (Oxford University Press, 1986) (revised by D. R. Heath-Brown).

Department of Mathematics
University of Georgia
Athens
GA 30602
USA