



Minimal Elements for the Prime Numbers

Curtis Bright^a, Raymond Devillers^b, and Jeffrey Shallit^a

^aSchool of Computer Science, University of Waterloo, Waterloo, ON, Canada; ^bDépartement d'Informatique, Université Libre de Bruxelles, Bruxelles, Belgium

ABSTRACT

We say a string of symbols s is *minimal* for a language L if s is a member of L , and it is not possible to obtain another member of L by striking out one or more symbols from s . Although the set $M(L)$ of minimal strings is necessarily finite, determining it explicitly for a given L can be a difficult computational problem. We use some number-theoretic heuristics to compute $M(L)$, where L is the language of base- b representations of the prime numbers, for $2 \leq b \leq 30$.

KEYWORDS

prime number; minimal element; heuristic algorithm

1. Introduction

Problems about the digits of prime numbers have a long history, and many of them are still unsolved. For example, are there infinitely many primes, all of whose base-10 digits are 1? Currently, there are only five such “repunits” known [Williams and Dubner 86], corresponding to $(10^p - 1)/9$ for $p \in \{2, 19, 23, 317, 1031\}$. It seems likely that four more are given by $p \in \{49081, 86453, 109297, 270343\}$, but this has not yet been rigorously proven.

Another problem on the digits of primes was introduced by the third author [Shallit 01]. To describe it, we need some definitions. We say that a string x is a *subword* of a string y , and we write $x \triangleleft y$, if one can strike out zero or more symbols of y to get x . For example, string is a subword of Meistersinger. (In the literature, this concept is sometimes called a “scattered subword” or “substring” or “subsequence”.) A *language* is a set of strings. A string s is *minimal* for L if (a) $s \in L$ and (b) if $x \in L$ and $x \triangleleft s$, then $x = s$. The set of all minimal strings of L is denoted $M(L)$.

In this article, we describe a heuristic technique for determining $M(L_b)$ in the case where L_b consists of the representations, in base b , of the prime numbers $\{2, 3, 5, \dots\}$. We obtain a complete characterization of $M(L_b)$ for bases $2 \leq b \leq 16$ and $b = 18, 20, 22, 23, 24$, and 30. For the remaining bases $b = 17, 19, 21$, and $25 \leq b \leq 29$, we obtain results that allow us to “almost” completely characterize this set.

The same technique can also be applied to find minimal sets for subsets of prime numbers. For example, we were able to determine the minimal set for primes of the

form $4n + 1$ represented in base 10 (and similarly for those of the form $4n + 3$). This successfully completes the sequences A111055 and A111056 in the *Encyclopedia of Integer Sequences* [OEIS Foundation Inc. 96], which had been incomplete since their introduction to the Encyclopedia in 2005.

Finally, we also show that determining the minimal set for the composite numbers represented in base b is a significantly easier problem and explicitly compute the minimal set for all $2 \leq b \leq 30$.

1.1. Notation

In what follows, if x is a string of symbols over the alphabet $\Sigma_b := \{0, 1, \dots, b-1\}$, we let $[x]_b$ denote the evaluation of x in base b (starting with the most significant digit), and $[\epsilon]_b := 0$ where ϵ is the empty string. This is extended to languages as follows: $[L]_b := \{[x]_b : x \in L\}$. We use the convention that $A := 10$, $B := 11$, and so forth, to conveniently represent strings of symbols in base $b > 10$. We let $(x)_b$ be the canonical representation of x in base- b , that is, the representation without leading zeroes.

Finally, as usual, for a language L , we let $L^n := \overbrace{LL \cdots L}^n$ and $L^* := \bigcup_{i \geq 0} L^i$.

2. Why minimal sets are interesting

One reason why the minimal set $M(L)$ of a language L is interesting is because it allows us to compute two natural and related languages, defined as follows:

$\text{sub}(L) := \{x \in \Sigma^* : \text{there exists } y \in L \text{ such that } x \triangleleft y\};$
 $\text{sup}(L) := \{x \in \Sigma^* : \text{there exists } y \in L \text{ such that } y \triangleleft x\}.$

An amazing fact is that $\text{sub}(L)$ and $\text{sup}(L)$ are always regular. This follows from the following classical theorem due to Higman [Higman 52] and Haines [Haines 69].

Theorem 1. *For every language L , there are only finitely many minimal strings.*

Indeed, we have $\text{sup}(L) = \text{sup}(M(L))$ and $\Sigma^* - \text{sub}(L) = \text{sup}(M(\Sigma^* - \text{sub}(L)))$, and the superword language of a finite language is regular, since

$$\text{sup}(\{w_1, \dots, w_n\}) = \bigcup_{i=1}^n \Sigma^* w_{i,1} \Sigma^* \cdots \Sigma^* w_{i,|w_i|} \Sigma^*$$

where $w_i = w_{i,1} \cdots w_{i,|w_i|}$ with $w_{i,j} \in \Sigma$.

3. Why the problem is hard

Determining $M(L)$ for arbitrary L is in general unsolvable and can be difficult even when L is relatively simple [Gruber et al. 07, Gruber et al. 09].

The following is a “semi-algorithm” that is guaranteed to produce $M(L)$, but it is not so easy to implement:

- (1) $M := \emptyset$
- (2) while $(L \neq \emptyset)$ do
 - (3) choose x , a shortest string in L
 - (4) $M := M \cup \{x\}$
 - (5) $L := L - \text{sup}(\{x\})$

In practice, for arbitrary L , we cannot feasibly carry out step (5). Instead, we work with L' , some regular overapproximation to L , until we can show $L' = \emptyset$ (which implies $L = \emptyset$). In practice, L' is usually chosen to be a finite union of sets of the form $L_1 L_2^* L_3$, where each of L_1, L_2, L_3 is finite. In the case we consider in this article, we then have to determine whether such a language contains a prime or not.

However, it is not even known if the following simpler decision problem is recursively solvable:

Problem 2. Given strings x, y, z , and a base b , does there exist a prime number whose base- b expansion is of the form $x \overbrace{yy \cdots y}^n z$ for some $n \geq 0$?

An algorithm to solve this problem, for example, would allow us to decide if there are any additional Fermat primes (of the form $2^{2^n} + 1$) other than the known ones (corresponding to $n = 0, 1, 2, 3, 4$). To see this, take $b := 2$, $x := 1$, $y := 0$, and $z := 0^{16}1$. Since if $2^n + 1$ is prime, then n must be a power of two, a prime of the form $[xy^*z]_b$ must be a new Fermat prime.

Therefore, in practice, we are forced to try to rule out prime representations based on heuristics such as modular techniques and factorizations. This is discussed in the next section.

4. Some useful lemmas

It will be necessary for our algorithm to determine if families of the form $[xL^*z]_b$ contain a prime or not. We use two different heuristic strategies to show that such families contain no primes.

In the first strategy, we mimic the well-known technique of “covering congruences” [Choi 71], by finding some finite set S of integers $N > 1$ such that every number in a given family is divisible by some element of S . In the second strategy, we attempt to find a difference-of-squares or difference-of-cubes factorization.

4.1. The first strategy

We start with the simplest version of the idea: to find an $N > 1$ that divides each element of the family $[xL^*z]_b$. At first glance, this would require checking that N divides $xL^n z$ for $n = 0, 1, 2, \dots$. However, the following lemma shows that it is only necessary to check the two cases $n = 0$ and 1 . Although divisibility based on digital considerations has a long history (e.g., [Dickson 52, Chap. XII]), we could not find these kinds of results in the literature.

Lemma 3. *Let $x, z \in \Sigma_b^*$, and let $L \subseteq \Sigma_b^*$. Then N divides all numbers of the form $[xL^*z]_b$ if and only if N divides $[xz]_b$ and all numbers of the form $[xLz]_b$.*

Proof. Let $y = y_1 \cdots y_n \in L^*$, where $y_1, \dots, y_n \in L$. By telescoping we have

$$[xyz]_b - [xz]_b = \sum_{i=1}^n ([xy_i y_{i+1} \cdots y_n z]_b - [xy_{i+1} \cdots y_n z]_b).$$

Cancelling the final $|y_{i+1} \cdots y_n z|$ base- b digits in the summand difference—which are identical—this becomes

$$[xyz]_b = [xz]_b + \sum_{i=1}^n b^{|y_{i+1} \cdots y_n z|} ([xy_i]_b - [x]_b).$$

But $b^{|z|}([xy_i]_b - [x]_b) = [xy_i z]_b - [xz]_b$ by adding and subtracting $[z]_b$, so we have

$$[xyz]_b = [xz]_b + \sum_{i=1}^n b^{|y_{i+1} \cdots y_n|} ([xy_i z]_b - [xz]_b).$$

Since $N \mid [xz]_b$ and $N \mid [xy_i z]_b$ for each $1 \leq i \leq n$, it follows that $N \mid [xyz]_b$.

The other direction is clear, since $[xz]_b$ and numbers of the form $[xLz]_b$ are both of the form $[xL^*z]_b$. \square

In practice, our algorithm employs this lemma with $L := \{y_1, \dots, y_n\} \subseteq \Sigma_b$, and all numbers of the form $[xL^*z]_b$ are shown to be composite with the following corollary.

Corollary 4. *If $1 < \gcd([xz]_b, [xy_1z]_b, \dots, [xy_nz]_b) < [xz]_b$ then all numbers of the form $[x\{y_1, \dots, y_n\}^*z]_b$ are composite.*

Proof. By Lemma 3, we know that the number $N := \gcd([xz]_b, [xy_1z]_b, \dots, [xy_nz]_b) > 1$ divides all numbers of the form $[x\{y_1, \dots, y_n\}^*z]_b$. By the size condition N is strictly less than each such number, and so is a nontrivial divisor. \square

Example 5. Since $\gcd(49, 469) = 7$, every number with base-10 representation of the form 46^*9 is divisible by 7. Since $1 < 7 < 49$, each such number is composite.

We also generalize this to the following corollary in the case where a single divisor does not divide each number in the family.

Corollary 6. *Let $L := \{y_1, y_2, \dots, y_n\}$. If*

$$N_0 := \gcd(\{[xz]_b\} \cup [xL^2z]_b)$$

and

$$N_1 := \gcd([xLz]_b \cup [xL^3z]_b)$$

*lie strictly between 1 and $[xz]_b$, then all numbers of the form $[xL^*z]_b$ are composite.*

Proof. By Lemma 3 applied to $[x(L^2)^*z]_b$, we know that N_0 divides all numbers of the form $[xL^*z]_b$ in which an even number of y_i appear. By Lemma 3 on $[xy_i(L^2)^*z]_b$ for each $1 \leq i \leq n$, we know that N_1 divides all numbers of the form $[xL^*z]_b$ for which an odd number of y_i appear. By the size conditions, N_0 and N_1 are nontrivial divisors. \square

Example 7. Since $\gcd([6]_9, [611]_9) = 2$, every number with base-9 representation of the form 61^* of odd length is divisible by 2. Since $\gcd([61]_9, [6111]_9) = 5$, every number with base-9 representation of the form 61^* of even length is divisible by 5. Since these numbers lie strictly between 1 and 6, every number with base-9 representation of the form 61^* is composite.

We also note that it is simple to generalize Corollary 6 to apply to check if there are divisors N_0, N_1, \dots, N_{k-1} such that N_i divides all numbers of the form $[x\{y_1, \dots, y_n\}^*z]_b$ in which the number of y_i appearing is congruent to $i \pmod k$.

Example 8. Let $b := 16$. Then 7 divides $[8A01]_b$ and $[8A0AAA1]_b$. Furthermore, 13 divides $[8A0A1]_b$ and $[8A0AAAA1]_b$, and 3 divides $[8A0AA1]_b$ and

$[8A0AAAAA1]_b$. Thus, all numbers with base-16 representation of the form $8A0A^*1$ are divisible by either 7, 13, or 3, depending on their length mod 3.

A version of Lemma 3 which applies to the most general kind of family we need to consider ($x_1L_1^* \cdots x_mL_m^*$, where we allow the case $L_m^* = \emptyset$) is formulated in Lemma 9.

Lemma 9. *Let x_1, \dots, x_m be strings in Σ_b^* , and $L_1, \dots, L_m \subseteq \Sigma_b^*$. Then N divides all numbers of the form $[x_1L_1^*x_2L_2^* \cdots x_mL_m^*]_b$ if and only if N divides $[x_1 \cdots x_m]_b$ and all numbers of the form $[x_1L_1x_2x_3 \cdots x_m]_b, \dots, [x_1 \cdots x_{m-1}x_mL_m]_b$.*

Proof. Say $w \in x_1L_1^*x_2L_2^* \cdots x_mL_m^*$; then there exist $y_{i,1}, \dots, y_{i,n_i} \in L_i$ such that

$$w = x_1y_{1,1} \cdots y_{1,n_1}x_2y_{2,1} \cdots y_{2,n_2} \cdots x_my_{m,1} \cdots y_{m,n_m}$$

for $1 \leq i \leq m$. As in the proof of Lemma 3, we have that

$$\begin{aligned} [w]_b &= [x_1 \cdots x_m]_b + \sum_{i=1}^m \sum_{j=1}^{n_i} b^{|y_{i,j+1} \cdots y_{i,n_m}|} \\ &\quad \times ([x_1 \cdots x_i y_{i,j} x_{i+1} \cdots x_m]_b - [x_1 \cdots x_m]_b) \end{aligned}$$

from which the claim follows. \square

As in Lemma 3, we typically apply this lemma in the case where each $L_i \subseteq \Sigma_b$ and show that all numbers of the form $[x_1L_1^*x_2L_2^* \cdots x_mL_m^*]_b$ have a divisor.

Example 10. Take $(L_1, L_2, L_3) := (\{0\}, \{0\}, \emptyset)$ and $(x_1, x_2, x_3) := (9, 8, 1)$. Since 9 divides 981, 9081, and 9801, it follows that 9 divides every number with base-10 representation of the form 90^*80^*1 .

More generally, if a single divisor doesn't work for every number, Lemma 9 can also be applied in the case where all numbers of the form $[x_1L_1^* \cdots x_i(L_i^2)^* \cdots x_mL_m^*]_b$ have one divisor, and all numbers of the form $[x_1L_1^* \cdots x_iL_i(L_i^2)^* \cdots x_mL_m^*]_b$ have another divisor.

Example 11. Let $b := 11$. Since 3 divides each of $[44A1]_b, [44A111]_b, [440A1]_b$, it follows that every number of the form $[440^*(11)^*1]_b$ is composite. Since 2 divides each of $[44A11]_b, [44A1111]_b, [440A11]_b$, we know every number of the form $[440^*(11)^*11]_b$ is composite. It follows that all numbers of the form $[440^*A1^*1]_b$ are composite.

Lemma 9 can also be applied to the case when all even-length strings under consideration have one divisor, and all the odd-length strings have another divisor. One such case is, for example, if numbers of the form $[x_1(L_1^2)^*x_2(L_2^2)^*x_3]_b$ and $[x_1L_1(L_1^2)^*x_2L_2(L_2^2)^*x_3]_b$ have one divisor, and

numbers of the form $[x_1 L_1(L_1^2)^* x_2 (L_2^2)^* x_3]_b$ and $[x_1 (L_1^2)^* x_2 L_2 (L_2^2)^* x_3]_b$ have another divisor.

Example 12. Let $b := 9$. Since 2 divides each of $[6]_b$, $[116]_b$, $[611]_b$, $[161]_b$, $[11161]_b$, $[16111]_b$, every odd-length string of the form $1^* 6 1^*$ is composite. Since 5 divides each of $[16]_b$, $[1116]_b$, $[1611]_b$, $[61]_b$, $[1161]_b$, $[6111]_b$, every even-length string of the form $1^* 6 1^*$ is composite.

4.2. The second strategy

A second way of proving that families of the form xL^*z do not contain a prime is via algebraic factorizations, such as a difference-of-squares factorization.

Lemma 13. Let $x, z \in \Sigma_b^*$, $y \in \Sigma_b$, and let $g := \gcd([y]_b, b-1)$, $X := ([y]_b + (b-1)[x]_b)/g$, and $Y := (b^{|z|}[y]_b - (b-1)[z]_b)/g$. If b, X , and Y are all squares and $\sqrt{b^{|z|}X} - \sqrt{Y} > (b-1)/g$, then all numbers of the form $[xy^*z]_b$ are composite.

Proof. Evaluating the base- b expansion of $xy^n z$, we get

$$\begin{aligned} [xy^n z]_b &= b^{|z|+n}[x]_b + b^{|z|} \frac{b^n - 1}{b-1} [y]_b + [z]_b \\ &= \frac{b^{|z|+n}X - Y}{(b-1)/g}. \end{aligned}$$

Since b, X , and Y are all squares the numerator factors as a difference of squares. By the size condition, both factors are strictly larger than the denominator, and so the factorization is nontrivial. \square

Example 14. Let $b := 16$, $x := 4$, $y := 4$, and $z := 1$. Then, $g = 1$, $X = 8^2$, $Y = 7^2$, and

$$[44^n 1]_b = \frac{(4^{n+1} \cdot 8 + 7)(4^{n+1} \cdot 8 - 7)}{15}.$$

Since $4 \cdot 8 - 7 > 15$, this factorization is nontrivial and no number of the form $[44^n 1]_b$ is prime.

It is also possible to combine Lemma 9 with Corollary 6 to construct a test which also applies to bases which are not squares.

Corollary 15. Using the same setup as in Lemma 9, if $b^{|z|}X$ and Y are squares, $\sqrt{b^{|z|}X} - \sqrt{Y} > (b-1)/g$, and $1 < \gcd([xyz]_b, [xy^3z]_b) < [xz]_b$, then all numbers of the form $[xy^*z]_b$ are composite.

Proof. Say $n = 2m$ is even. Then from the factorization in Lemma 9,

$$[xy^n z]_b = \frac{(b^m \sqrt{b^{|z|}X} + \sqrt{Y})(b^m \sqrt{b^{|z|}X} - \sqrt{Y})}{(b-1)/g}$$

which is nontrivial by the size condition.

Alternatively, if n is odd then as in Corollary 6 we have that $\gcd([xyz]_b, [xy^3z]_b)$ divides $[xy^n z]_b$, and by the size condition this divisor is nontrivial. \square

Example 16. Let $b := 17$, $x := 19$, $y := 9$, and $z := 9$. Then $g = 1$, $b^{|z|}X = 85^2$, $Y = 3^2$, and

$$[xy^{2n} z]_b = \frac{(17^n \cdot 85 + 3)(17^n \cdot 85 - 3)}{16}.$$

Since $85 - 3 > 16$, this factorization is nontrivial. Furthermore, all numbers of the form $[xy^{2n+1} z]_b$ are even, so all numbers of the form $[199^*9]_b$ are composite.

Finally, we present a variant of Lemma 13 which applies to a difference-of-cubes factorization.

Lemma 17. Let $x, z \in \Sigma_b^*$, $y \in \Sigma_b$, and let $g := \gcd([y]_b, b-1)$, $X := ([y]_b + (b-1)[x]_b)/g$, and $Y := (b^{|z|}[y]_b - (b-1)[z]_b)/g$. If b, X , and Y are all cubes and $\sqrt[3]{b^{|z|}X} - \sqrt[3]{Y} > (b-1)/g$, then all numbers of the form $[xy^*z]_b$ are composite.

Proof. As in Lemma 13, we have

$$[xy^n z]_b = \frac{((b^{|z|+n}X)^{1/3} - Y^{1/3})((b^{|z|+n}X)^{2/3} + (b^{|z|+n}XY)^{1/3} + Y^{2/3})}{(b-1)/g}.$$

The second factor is at least as large as the first (except in the single case $b^{|z|+n}X = 1$ and $Y = -1$, which is not possible by construction of X and Y), so by the size condition both factors are strictly larger than the denominator, and the factorization is nontrivial. \square

Example 18. Let $b := 8$, $x := 1$, $y := 0$, and $z := 1$. Then $g = 7$, $X = 1$, $Y = -1$, and

$$[10^n 1]_b = (2^{n+1} + 1)(4^{n+1} - 2^{n+1} + 1).$$

Since $2 - (-1) > 1$, this factorization is nontrivial and no number of the form $[10^n 1]_b$ is prime.

5. Our heuristic algorithm

As previously mentioned, in practice to compute $M(L_b)$, one works with an underapproximation M of $M(L_b)$ and an overapproximation L of $L_b - \sup(M)$. One then refines such approximations until $L = \emptyset$ from which it follows that $M = M(L_b)$.

For the initial approximation, note that every minimal prime in base b with at least 4 digits is of the form xy^*z , where $x \in \Sigma_b - \{0\}$, $z \in \Sigma_b$, and

$$Y := \Sigma_b - \{y : (p)_b \triangleleft xyz \text{ for some prime } p\}.$$

Making use of this, our algorithm sets M to be the set of base- b representations of the minimal primes with at most 3 digits (which can be found simply by brute force) and L to be $\bigcup_{x,z} xy^*z$, as described above.

All remaining minimal primes are members of L , so to find them we explore the families in L . During this process, each family will be decomposed into possibly multiple other families. For example, a simple way of exploring the family xY^*z where $Y := \{y_1, \dots, y_n\}$ is to decompose it into the families $xY^*y_1z, \dots, xY^*y_nz$. If the smallest member (say xy_iz) of any such family happens to be prime, it can be added to M and the family xY^*y_iz removed from consideration. Furthermore, once M has been updated, it may be possible to simplify some families in L . In this case, xY^*y_jz (for $j \neq i$) can be simplified to $x(Y - \{y_i\})^*y_jz$ since no minimal prime contains xy_iz as a proper subword.

Another way of decomposing the family xY^*z is possible if one knows that a digit of Y can only occur a certain number of times. For example, if xy_iy_jz has a proper prime subword then the digit y_i can occur at most once in any minimal prime of the form xY^*z , and we can split xY^*z into the two families

$$x(Y - \{y_i\})^*z \quad \text{and} \quad x(Y - \{y_i\})^*y_i(Y - \{y_i\})^*z.$$

Last, the family xY^*z can be decomposed by considering digits of Y which are mutually incompatible, i.e., they cannot occur simultaneously in a minimal prime. For example, if xy_iy_jz and xy_jy_iz ($i \neq j$) both have proper prime subwords then the digits y_i and y_j cannot occur simultaneously in any minimal prime of the form xY^*z , and we can split xY^*z into the two families

$$x(Y - \{y_i\})^*z \quad \text{and} \quad x(Y - \{y_j\})^*z.$$

Sometimes, it is not possible to show two digits are mutually incompatible, but it is possible to know that one digit must appear before the other. For example, if xy_iy_jz has a proper prime subword then the digit y_j must appear before y_i in any minimal prime of the form xY^*z , and we can replace xY^*z with the family

$$x(Y - \{y_i\})^*(Y - \{y_j\})^*z.$$

Similarly, if xy_jy_iz has a proper prime subword, then we can split xY^*yY^*z into the two families

$$x(Y - \{y_i\})^*yY^*z \quad \text{and} \quad xY^*y(Y - \{y_j\})^*z.$$

We now formulate these arguments for the most general kind of family we need to consider, namely $x_1L_1^* \cdots x_mL_m^*$. For simplicity, we only specify the decompositions as applying to $L_1 := \{y_1, \dots, y_n\}$, but it is straightforward to generalize these decompositions to also apply to L_i for any $1 \leq i \leq m$.

Lemma 19. *Every minimal prime of the form $x_1L_1^* \cdots x_mL_m^*$ must also be of the form $x_1x_2L_2^* \cdots x_mL_m^*$ or $x_1y_iL_1^*x_2L_2^* \cdots x_mL_m^*$ for some $1 \leq i \leq n$.*

Proof. Follows from the fact that $L_1^* = \{\epsilon\} \cup \bigcup_{i=1}^n y_iL_1^*$. \square

Similarly, one can generalize Lemma 19 to apply to adding characters to the right of L_1 rather than to the left.

Example 20. The family $10\{0, 1\}^*61^*1$ splits into the three families 1061^*1 , $10\{0, 1\}^*061^*1$, and $10\{0, 1\}^*161^*1$ by exploring $\{0, 1\}^*$ on the right.

Lemma 21. *If $x_1y_i^kx_2 \cdots x_m$ contains a prime proper subword (for some $k \geq 1$) then every minimal prime of the form $x_1L_1^* \cdots x_mL_m^*$ is of the form*

$$x_1(L_1 - \{y_i\})^*(y_i(L_1 - \{y_i\})^*)^jx_2L_2^* \cdots x_mL_m^*$$

for some $0 \leq j < k$.

Proof. If $w \in x_1L_1^* \cdots x_mL_m^*$ then $w \in x_1yx_2L_2^* \cdots x_mL_m^*$ for some $y \in L_1^*$. If y contains k or more instances of y_i , then by assumption it follows that w contains a proper prime subword, and therefore is not a minimal prime. So if w is a minimal prime, then y must contain less than k instances of y_i , i.e., y must be of the form $(L_1 - \{y_i\})^*(y_i(L_1 - \{y_i\})^*)^j$ for some $0 \leq j < k$, from which the claim follows. \square

Example 22. The string 661 represents a prime in base 9 and is a proper subword of 10661. It follows that the family $10\{0, 1, 6\}^*1$ splits into the families $10\{0, 1\}^*1$ and $10\{0, 1\}^*6\{0, 1\}^*1$ in base 9.

Lemma 21 is especially useful when it can be applied with $k = 1$, since in that case the family $x_1L_1^* \cdots x_mL_m^*$ is replaced by a single strictly simpler family, in contrast to the other lemmas we will describe.

Lemma 23. *If $x_1y_iy_jx_2 \cdots x_m$ and $x_1y_jy_ix_2 \cdots x_m$ contain prime proper subwords (where $i \neq j$) then every minimal prime of the form $x_1L_1^* \cdots x_mL_m^*$ is of the form*

$$x_1(L_1 - \{y_i\})^*x_2L_2^* \cdots x_mL_m^* \quad \text{or} \\ x_1(L_1 - \{y_j\})^*x_2L_2^* \cdots x_mL_m^*.$$

Proof. If $w \in x_1L_1^* \cdots x_mL_m^*$ then $w \in x_1yx_2L_2^* \cdots x_mL_m^*$ for some $y \in L_1^*$. If y contains both y_i and y_j , then by assumption it follows that w contains a proper prime subword, and therefore is not a minimal prime. So if w is a minimal prime, then y cannot contain y_i and y_j simultaneously, i.e., y must be of the form $(L - \{y_i\})^*$ or $(L - \{y_j\})^*$, from which the claim follows. \square

Example 24. The string 4611 represents a prime in base 8 and is a proper subword of 446411 and 444611. It follows that the family $44\{4, 6\}^*11$ splits into the families 444^*11 and 446^*11 in base 8.

Lemma 25. *If $x_1y_iy_jx_2 \cdots x_m$ contains a prime proper subword (where $i \neq j$) then every minimal prime of the form $x_1L_1^* \cdots x_mL_m^*$ is of the form*

$$x_1(L_1 - \{y_i\})^*(L_1 - \{y_j\})^*x_2L_2^* \cdots x_mL_m^*.$$

Proof. If $w \in x_1 L_1^* \cdots x_m L_m^*$, then $w \in x_1 y x_2 L_2^* \cdots x_m L_m^*$ for some $y \in L_1^*$. If y contains y_i before y_j then by assumption it follows that w contains a proper prime subword and therefore is not a minimal prime. So if w is a minimal prime, then y cannot contain y_i before y_j , i.e., y must be of the form $(L - \{y_i\})^* (L - \{y_j\})^*$, from which the claim follows. \square

Example 26. The string 10 represents a prime in base 11 and is a proper subword of 90101. It follows that the family $90\{0, 1, 9\}^*1$ splits into the family $90\{0, 9\}^*\{1, 9\}^*1$ in base 11.

Lemma 27. If $x_1 y_i x_2 y_{2,j} x_3 \cdots x_m$ contains a prime proper subword (where $y_{2,j} \in L_2$) then every minimal prime of the form $x_1 L_1^* \cdots x_m L_m^*$ is of the form

$$x_1 (L_1 - \{y_i\})^* x_2 L_2^* \cdots x_m L_m^* \quad \text{or} \\ x_1 L_1 x_2 (L_2 - \{y_{2,j}\})^* x_3 L_3^* \cdots x_m L_m^*.$$

Proof. If w is a member of the language $x_1 L_1^* \cdots x_m L_m^*$ then $w \in x_1 y x_2 y' x_3 L_3^* \cdots x_m L_m^*$ for some $y \in L_1^*$ and $y' \in L_2^*$. If y contains y_i and y' contains $y_{2,j}$, then by assumption it follows that w contains a proper prime subword, and therefore is not a minimal prime. So if w is a minimal prime then either y cannot contain y_i or y' cannot contain $y_{2,j}$, i.e., y is of the form $(L_1 - \{y_i\})^*$ and y' is of the form L_2^* , or y is of the form L_1^* and y' is of the form $(L_2 - \{y_{2,j}\})^*$, from which the claim follows. \square

Example 28. The string 60411 represents a prime in base 8, and is a proper subword of 604101. It follows that the family $60\{0, 4\}^*10^*1$ splits into the families 600^*10^*1 and $60\{0, 4\}^*11$ in base 8.

We call families of the form xy^*z (where $x, z \in \Sigma_b^*$ and $y \in \Sigma_b$) *simple families*. Our algorithm then proceeds as follows:

- (1) $M := \{\text{minimal primes in base } b \text{ of length } \leq 3\}$
 $L := \bigcup_{x,z \in \Sigma_b} xY^*z$, where $x \neq 0$ and Y is the set of digits y such that xyz has no subword in M
- (2) While L contains non-simple families:
 - (a) Explore each family of L by applying Lemma 19, and update L .
 - (b) Examine each family of L :
 - (i) Let w be the shortest string in the family. If w has a subword in M , then remove the family from L . If w represents a prime, then add w to M and remove the family from L .
 - (ii) If possible, simplify the family by applying Lemma 21 with $k = 1$.
 - (iii) Using the techniques of Section 4, check if the family can be proven to only contain composites, and if so then remove the family from L .

- (c) Apply Lemmas 21, 23, 25, and 27 to the families of L as much as possible and update L ; after each split, examine the new families as in (b).

The process of exploring/examining/splitting a family can be concisely expressed in a tree of decompositions. A sample tree of decompositions, for the family $1\{0, 1, 6\}^*1$ in base 9, is displayed in Figure 1. When applying Lemma 19 on a family with only one nonempty L_i , we don't show the first decomposition (simply removing L_i^*) since that results in the shortest word in the family, which is always implicitly checked for primality/minimal subwords at each step anyway.

5.1. Implementation

An implementation of the above algorithm was written in C using the GMP library [Granlund et al.] by the first author [Bright 14]. Independently, the second author [Devillers 15] wrote an implementation in C++ using the MIRACL library [Scott 03] using the same ideas and derived the same results (with a less extensive search for primes in the simple families, but with results for bases $b \leq 50$).

Note that when exploring the families in line (a), one should not always apply Lemma 19 directly as stated by adding characters to the left of L_1 ; in our implementation, we alternate between adding characters to the left and right of $L_{1+k \bmod m}$, where k is the number of times we have previously passed through line (a). Also, the application of Lemma 25 tended to initially produce an excess number of cases, so it was useful to only apply it following a certain number of iterations (6 in our implementation). It also led to duplicate families after simplifying, for example families of the form $xy^*y^n y^*z$ for varying n , but these could be detected and removed.

At the conclusion of the algorithm described, L will consist of simple families (of the form xy^*z) which have not yet yielded a prime, but for which there is no obvious reason why there can't be a prime of such a form. In such a case, the only way to proceed is to test the primality of larger and larger numbers of such form and hope a prime is eventually discovered.

The numbers in simple families are of the form $(ab^n + c)/d$ for some fixed a, b, c, d where $d \mid ab^n + c$ for all n . Except in the special case $c = \pm 1$ and $d = 1$, when n is large the known primality tests for such a number are too inefficient to run. In this case, one must resort to a probable primality test such as a Miller–Rabin test, unless a divisor of the number can be found. Since we are testing many numbers in an exponential sequence, it is possible to use a sieving process to find divisors rather than using trial division.

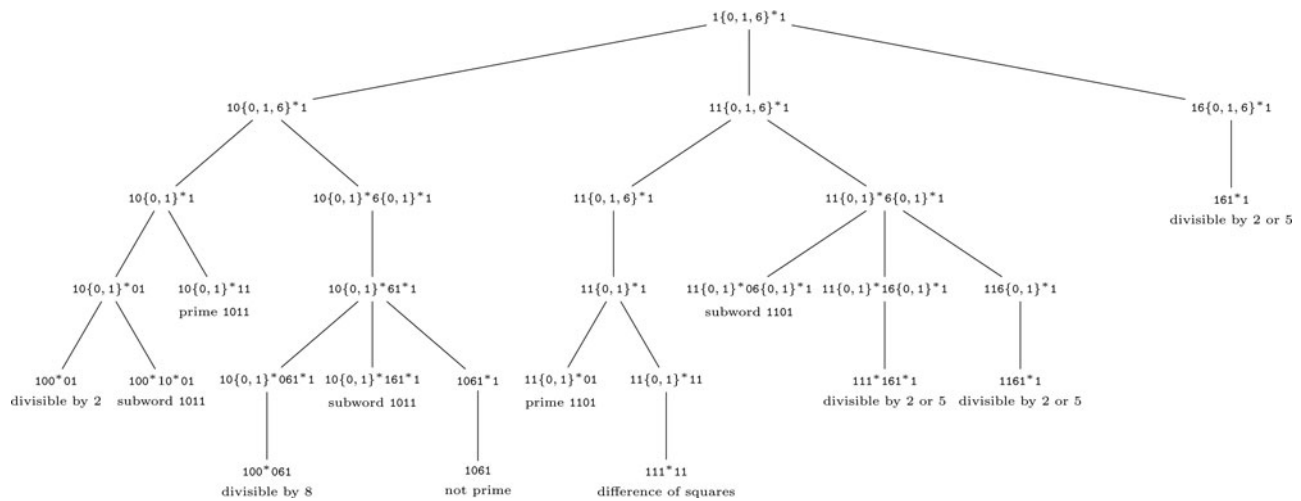


Figure 1. Tree of decompositions for $1\{0, 1, 6\}^*1$ in base 9.

To do this, we made use of Geoffrey Reynolds' `srsieve` software [Reynolds 10]. This program uses the baby-step giant-step algorithm to find all primes p which divide $ab^n + c$ where p and n lie in a specified range. Since this program cannot handle the general case $(ab^n + c)/d$ when $d > 1$, we only used it to sieve the sequence $ab^n + c$ for primes $p \nmid d$, and initialized the list of candidates to not include n for which there is some prime $p \mid d$ for which $p \mid (ab^n + c)/d$. The program had to be modified slightly to remove a check which would prevent it from running in the case when a , b , and c were all odd (since then $2 \mid ab^n + c$, but 2 may not divide $(ab^n + c)/d$).

Once the numbers with small divisors had been removed, it remained to test the remaining numbers using a probable primality test. For this, we used the software LLR by Jean Penné [Penné 15]. Although undocumented, it is possible to run this program on numbers of the form $(ab^n + c)/d$ when $d \neq 1$, so this program required no modifications. A script was also written which allowed one to run `srsieve` while LLR was testing the remaining candidates, so that when a divisor was found by `srsieve` on a number which had not yet been tested by LLR it would be removed from the list of candidates.

In the cases where the elements of $M(L_b)$ could be proven prime rigorously, we employed PRIMO by Marcel Martin [Martin 11], an elliptic curve primality proving implementation.

6. Results

A summary of the results of our algorithm is presented in Figure 2; it was able to completely solve all bases up to 30 except for 17, 19, 21, and those between 25 and 29. The results in base 29 required some additional strategies, as described in Section 7.

b	$ M(L_b) $	$\max_{x \in M(L_b)} x $	Number of unsolved families	Searched height
2	2	2	0	—
3	3	3	0	—
4	3	2	0	—
5	8	5	0	—
6	7	5	0	—
7	9	5	0	—
8	15	9	0	—
9	12	4	0	—
10	26	8	0	—
11	152	45	0	—
12	17	8	0	—
13*	228	32,021	0	—
14	240	86	0	—
15	100	107	0	—
16	483	3545	0	—
17*	≥ 1279	$\geq 111,334$	1	967,000
18	50	33	0	—
19*	≥ 3462	$\geq 110,986$	1	686,000
20	651	449	0	—
21*	≥ 2600	$\geq 479,150$	1	475,000
22	1242	764	0	—
23*	6021	800,874	0	—
24	306	100	0	—
25*	$\geq 17,597$	$\geq 136,967$	12	292,000
26	≥ 5662	≥ 8773	2	461,000
27*	$\geq 17,210$	$\geq 109,006$	5	355,000
28*	≥ 5783	$\geq 94,538$	1	524,000
29*	$\geq 57,283$	$\geq 174,240$	14	233,000
30	220	1024	0	—

*Data based on results of probable primality tests.

Figure 2. Summary of results for the prime numbers for each base b .

For every solved base, we give the size $|M(L_b)|$ and the “width” $\max_{x \in M(L_b)} |x|$ of the corresponding family. For the unsolved bases, we give a lower bound on the size and width of $M(L_b)$, along with the number of families of the form xy^*z for which no prime member could be found, nor could the family be ruled out as only containing composites. Since such simple families can contain at most

one minimal prime, an upper bound on $|M(L_b)|$ is given by the sum of the lower bound for $|M(L_b)|$ and the number of unsolved families. Additionally, we give the height to which the simple families were searched for primes; if there are any more primes in $M(L_b)$, they must have at least this many digits in base b .

The family 80^*1 in base 23, corresponding to the generalized Proth numbers $8 \cdot 23^n + 1$, was already known to be prime for minimal $n = 119215$ in the process of solving the generalized Sierpiński conjecture in base 23 [Barnes 07].

The largest probable prime we found was the number $9E^{800873}$ (expressed as a base 23 string) or $(106 \cdot 23^{800873} - 7)/11$ in standard notation. It contains 1,090,573 decimal digits and at the time of discovery was the tenth largest known probable prime according to Henri and Renaud Lifchitz's ranking [Lifchitz and Lifchitz 00].

6.1. Unsolved families

There were 37 families for which we were unable to determine if they contain a prime or not. They are explicitly described in Figure 3 in both their “base b ” string representation xy^*z and “algebraic” form $(ab^n + c)/d = [xy^*z]_b$. The numbers a and c are given in their factorized form so as to help spot algebraic factorizations.

6.2. Primes of the form $4n + 1$ and $4n + 3$

The heuristic algorithm described in Section 5 may also be easily modified to apply to recursive subsets of prime numbers, e.g., those congruent to 1 mod 4 (alternatively, 3 mod 4). The modifications necessary are to the initialization of M in line 1, and to update the check “ w represents a minimal prime” to also check that w is in the subset under consideration. When searching for minimal primes of the form $4n + 1$, it was necessary to remove families only containing numbers of the form $4n + 3$, and vice versa. A modified Lemma 3 can be used to check this, e.g., $[x(L^0 \cup L)z]_b$ are all congruent to $c \bmod d$ implies that $[xL^*z]_b$ are all congruent to $c \bmod d$.

The lemmas of Section 4 can be used without modification, since a family which contains only composites of course does not contain any primes of a special form, either. For simplicity, the lemmas of Section 5 are stated to apply when the minimal set consists of primes, but actually apply to general minimal sets and also need no modification.

When run on the primes of the form $4n + 1$ represented in base 10, our implementation successfully computes the minimal set consisting of 146 elements, the largest of which contains 79 digits. The *Encyclopedia of*

Base	Family	Algebraic form
17	F19*	$(5 \cdot 821 \cdot 17^n - 3^2)/16$
19	EE16*	$(2^2 \cdot 13 \cdot 307 \cdot 19^n - 1)/3$
21	G0*FK	$2^4 \cdot 21^{n+2} + 5 \cdot 67$
25	6MF*9	$(1381 \cdot 25^{n+1} - 53)/8$
	CM1*	$(59 \cdot 131 \cdot 25^n - 1)/24$
	EE1*	$(8737 \cdot 25^n - 1)/24$
	E1*E	$(337 \cdot 25^{n+1} + 311)/24$
	EFO*	$2 \cdot 3 \cdot 61 \cdot 25^n - 1$
	F1*F1	$(19^2 \cdot 25^{n+2} + 37 \cdot 227)/24$
	FO*KO	$3 \cdot 5 \cdot 25^{n+2} + 2^2 \cdot 131$
	FOK*O	$(5 \cdot 11 \cdot 41 \cdot 25^{n+1} + 19)/6$
	LOL*8	$(53 \cdot 83 \cdot 25^{n+1} - 3 \cdot 37)/8$
	M1*F1	$(23^2 \cdot 25^{n+2} + 37 \cdot 227)/24$
	M1O*8	$19 \cdot 29 \cdot 25^{n+1} + 2^3$
	OL*8	$(199 \cdot 25^{n+1} - 3 \cdot 37)/8$
26	A*6F	$(2 \cdot 26^{n+2} - 7 \cdot 71)/5$
	I*GL	$(2 \cdot 3^2 \cdot 26^{n+2} - 11 \cdot 113)/25$
27	80*9A	$2^3 \cdot 27^{n+2} + 11 \cdot 23$
	999G*	$(101 \cdot 877 \cdot 27^n - 2^3)/13$
	CL*E	$(3^2 \cdot 37 \cdot 27^{n+1} - 7 \cdot 29)/26$
	EI*F8	$(191 \cdot 27^{n+2} - 2^3 \cdot 149)/13$
	F*9FM	$(3 \cdot 5 \cdot 27^{n+3} - 113557)/26$
28	OA*F	$(2 \cdot 7 \cdot 47 \cdot 28^{n+1} + 5^3)/27$
29	1A*	$(19 \cdot 29^n - 5)/14$
	68LO*6	$7 \cdot 757 \cdot 29^{n+1} + 2 \cdot 3$
	AMP*	$(8761 \cdot 29^n - 5^2)/28$
	C*FK	$(3 \cdot 29^{n+2} + 2 \cdot 331)/7$
	F*OPF	$(3 \cdot 5 \cdot 29^{n+3} + 139 \cdot 1583)/28$
	FKI*	$(6379 \cdot 29^n - 3^2)/14$
	F*OP	$(3 \cdot 5 \cdot 29^{n+2} + 7573)/28$
	LP09*	$(31 \cdot 16607 \cdot 29^n - 3^2)/28$
	OOPS*A	$2 \cdot 10453 \cdot 29^{n+1} - 19$
	PC*	$(2 \cdot 89 \cdot 29^n - 3)/7$
	PPPL*O	$(87103 \cdot 29^{n+1} + 3^2)/4$
	Q*GL	$(13 \cdot 29^{n+2} - 3 \cdot 1381)/14$
	Q*L0	$(13 \cdot 29^{n+2} - 19 \cdot 109)/14$
	RM*G	$(389 \cdot 29^{n+1} - 5 \cdot 19)/14$

Figure 3. The unsolved families listed in their “base b ” and “algebraic” representations.

Integer Sequences [OEIS Foundation Inc. 96] contains the elements for this minimal set in entry A111055. Prior to our work, the longest 11 elements were missing from this listing.

When run on the primes of the form $4n + 3$ represented in base 10, our implementation successfully computes the minimal set consisting of 113 elements, the largest of which contains 19,153 digits. Because of its size, this number is not easily proven prime, but François Morain successfully produced a primality certificate for it [Morain 94]. The second-largest number has 50 digits, so the remaining elements are easily proven prime. The *Encyclopedia of Integer Sequences* [OEIS Foundation Inc. 96] contains the elements for this minimal set in entry A111056. Prior to our work, the longest 10 elements were missing from this listing.

7. Some additional strategies

The strategies discussed so far suffice to restrict the possible forms of minimal primes to a finite number of simple families in all bases $2 \leq b \leq 28$. However, as b increased, in addition to the calculations becoming more costly, it was found to be necessary to use increasingly complicated strategies. We now describe some additional strategies which we found sufficient to solve all non-simple families in base 29.

Lemma 29. *If every number of the form $x_1(L_1 - \{y_i\})^*x_2L_2^* \cdots x_mL_m^*$ is composite, then every minimal prime of the form $x_1L_1^* \cdots x_mL_m^*$ must also be of the form $x_1L_1^*y_iL_1^*x_2L_2^* \cdots x_mL_m^*$.*

Proof. If $w \in x_1L_1^* \cdots x_mL_m^*$ then $w \in x_1y_iL_1^* \cdots x_mL_m^*$ for some $y \in L_1^*$. If y does not contain a y_i then w is composite by assumption. Therefore if w is a prime then y contains a y_i , i.e., $y \in L_1^*y_iL_1^*$, from which the result follows. (Note that one could improve this result via $y \in (L_1 - \{y_i\})^*y_iL_1^* \cup L_1^*y_i(L_1 - \{y_i\})^*$, but this was found to be unnecessary for our purposes.) \square

Example 30. The numbers represented by the family $F\{0, 9, F\}^*F$ in base 29 are divisible by 3, so every minimal prime of the form $F\{0, 9, F, P\}^*F$ must also be of the form $F\{0, 9, F, P\}^*P\{0, 9, F, P\}^*F$.

Lemma 31. *If $x_1y_iy_jy_ix_2 \cdots x_m$ contains a prime proper subword (where $i \neq j$) then every minimal prime of the form $x_1L_1^* \cdots x_mL_m^*$ is of the form*

$$x_1(L_1 - \{y_i\})^*(L_1 - \{y_j\})^*(L_1 - \{y_i\})^*x_2L_2^* \cdots x_mL_m^*.$$

Proof. If $w \in x_1L_1^* \cdots x_mL_m^*$ then $w \in x_1y_iy_jy_ix_2L_2^* \cdots x_mL_m^*$ for some $y \in L_1^*$. If y contains $y_iy_jy_i$ then by assumption it follows that w contains a proper prime subword, and therefore is not a minimal prime. So if w is a minimal prime, then either y does not contain a y_j or y contains a y_j and all y_i s in y either come before or after the y_j . In each case, y is of the form $(L - \{y_i\})^*(L - \{y_j\})^*(L - \{y_i\})^*$, from which the claim follows. \square

Example 32. The string QLQ represents a prime in base 29, and is a proper subword of $LQLQL$. It follows that the family $L\{L, Q\}^*L$ splits into the family $LL^*Q^*L^*L$ in base 29.

This rule may also be generalized to apply to the case when $x_1y_iy_jy_ix_2 \cdots x_m$ contains a prime proper subword (where $i \neq j$).

Example 33. The string $LL9L9LQL$ represents a prime in base 29. It follows that the family $LL\{9, L\}^*Q^*QL$ splits into the family $LLL^*9^*L^*9^*Q^*QL$ in base 29.

In Section 4.1, we described a number of strategies for determining if every member of a family has a divisor, but for some families divisors exist which will not be found using those tests. The following lemma can help one discover when this is the case—in particular when every member of a family is divisible by some small prime. For a language L , we use the notation $[L]_b \bmod N$ to denote the finite set of residues $\{[x]_b \bmod N : x \in L\}$.

Lemma 34. *If $1 < \gcd(k, N)$ for every $k \in [L]_b \bmod N$ and $p \notin [L]_b$ for every prime p which divides N , then all numbers of the form $[L]_b$ are composite.*

Proof. Let x be an arbitrary member of L . Then $\gcd([x]_b, N) = \gcd([x]_b \bmod N, N) > 1$ by assumption, so $[x]_b$ cannot be prime unless $[x]_b = \gcd([x]_b, N)$. But in that case $[x]_b$ would be a prime which divides N , and so $[x]_b \notin [L]_b$ by the second assumption, in contradiction to $x \in L$. \square

To make use of this lemma, we need to be able to compute $[x_1L_1^* \cdots x_mL_m^*]_b \bmod N$. To do this, we can make use of the following relations:

$$[Lx]_b \bmod N = (b \cdot ([L]_b \bmod N) + [x]_b) \bmod N$$

$$[L\{y_1, \dots, y_n\}]_b \bmod N = \bigcup_{i=1}^n [Ly_i]_b \bmod N$$

$$[L\{y_1, \dots, y_n\}^*]_b \bmod N = \bigcup_{i=0}^{\infty} [L\{y_1, \dots, y_n\}^i]_b \bmod N$$

In the final case, the union may be taken to be finite over $i < l$ where l is chosen such that $\bigcup_{i=0}^{l-1} [L\{y_1, \dots, y_n\}^i]_b \bmod N = \bigcup_{i=0}^{l-1} [L\{y_1, \dots, y_n\}^i]_b \bmod N$. Using these relations, we can compute $[x_1L_1^* \cdots x_mL_m^*]_b \bmod N$ progressively, working left to right and starting from $[\emptyset]_b \bmod N = \{0\}$. To solve base 29, it was sufficient to use $N = 2 \cdot 3 \cdot 5$.

Example 35. Let $b := 29$ and $N := 30$. Then $[L1^*61^*LK^*K]_b \bmod N = \{4, 5, 6, 14, 15, 16, 25\}$, so if k is in this set then $\gcd(k, N) \in \{2, 5, 6, 15\}$ and $\gcd(k, N) > 1$. Since $2, 3, 5 \notin [L1^*61^*LK^*K]_b$, by Lemma 34 all numbers of the form $L1^*61^*LK^*K$ are composite.

8. Composite numbers

One can also consider the companion problem of determining the minimal elements for the composite numbers $\{4, 6, 8, 9, 10, 12, \dots\}$. Here, in contrast with the primes, we have

Theorem 36. *The following decision problem is recursively solvable: given a base b and a DFA M accepting a language*

L of strings in a base- b canonical representation, does M accept the base- b representation of a composite number?

This follows immediately from

Theorem 37. *Suppose L is a regular language, accepted by a deterministic finite automaton M of n states. Then if M accepts a composite number expressed in base b , it must accept one whose base- b representation has at most $n(b^{2n} + 1)$ digits.*

Proof. If L is finite, then the longest string accepted by M has at most $n - 1$ digits, and $n - 1 < n(b^{2n} + 1)$.

Otherwise L is infinite. Then M accepts a string of length ℓ with $n < \ell \leq 2n$ that can be pumped (as in the pumping lemma). That is, there exists $x \in L$ such that $x = uvw$ with $|uv| \leq n$. Then a classical proof of the non-regularity of the prime numbers (e.g., [Hopcroft and Ullman 79, Example 3.2, p. 57]) shows that either x is the representation of a composite number (in which case $|x| \leq 2n \leq n(b^{2n} + 1)$) or $uv^p w$ is composite, where $p = \lceil x \rceil_b$. Our bound now follows. \square

Write $S_b := \{ (n)_b : n \geq 4 \text{ is composite} \}$. For our purposes, the following theorem gives a better bound.

Theorem 38. *Every minimal element of S_b is of length at most $b + 2$.*

Proof. Consider any word w of S_b of length $\geq b + 3$. Since there are only b distinct digits, some digit d is repeated at least twice, so that $dd \triangleleft w$. If $d > 1$, the number $[dd]_b$ is composite, as it is divisible by $[11]_b$ but not equal to it. If $d = 0$, then some nonzero digit c precedes it in w , so $c00 \triangleleft w$ and $[c00]_b$ is divisible by b^2 , which is composite. Finally, if no digit other than 1 is repeated, then $1111 \triangleleft w$, and $[1111]_b = [11]_b \cdot [101]_b$, and hence is composite. \square

Theorem 38 turns the computation of the minimal elements for S_b into a finite search. Our results are given in Figure 4.

9. Conclusion and perspectives

We have (sometimes partially) solved the problem of finding the minimal elements for the prime numbers represented with the first bases. However, the problem rapidly becomes very complex: the number of members increases rapidly, some bases need sophisticated strategies, and some members are very long. Hence, to complete the cases where some (simple) families resisted our analysis, and to go further in the bases, we need better exploration strategies, and better primality tests for numbers expressed by a simple family in some base.

b	$ M(S_b) $	$\max_{x \in M(S_b)} x $
2	3	4
3	4	3
4	9	3
5	10	3
6	19	4
7	18	3
8	26	3
9	28	2
10	32	3
11	32	3
12	46	4
13	43	3
14	52	3
15	54	2
16	60	3
17	60	3
18	95	4
19	77	3
20	87	3
21	90	2
22	94	3
23	97	3
24	137	4
25	117	2
26	111	3
27	115	2
28	131	3
29	123	3
30	207	4

Figure 4. Summary of results for the composite numbers for each base b .

Another problem left open is to delineate the asymptotic behaviour of the size and width of those sets $M(L_b)$ when b increases. A rough estimation is that the size increases exponentially, but this increase is lower when the base has many different factors.

To illustrate once more the difficulty of computing $M(L)$, we recall an open problem from [Shallit 01]:

Problem 39. Let $L := \{1, 2, 4, 8, 16, 32, 64, \dots\}$, the base-10 representation of the powers of 2. Is it the case

that

$$M(L) = \{1, 2, 4, 8, 65536\}?$$

This would follow, for example, if we could prove that every power of 16 greater than 65,536 contained at least one of the digits $\{1, 2, 4, 8\}$. But this seems beyond current capabilities.

Acknowledgments

We are very grateful to François Morain for proving the primality of $2 \cdot (10^{19153} - 1)/9 + 77$, which enabled us to complete the classification of the minimal elements of primes of the form $4n + 3$ in base 10.

References

- [Barnes 07] G. Barnes. “Sierpinski Conjectures and Proofs.” www.noprimeleftbehind.net/crus/Sierp-conjectures.htm, 2007.
- [Bright 14] C. Bright. “MEPN Implementation.” github.com/curtisbright/mepn, 2014.
- [Choi 71] S. L. G. Choi. “Covering the Set of Integers by Congruence Classes of Distinct Moduli.” *Math. Comp.* 25 (1971), 885–895.
- [Devillers 15] R. Devillers. “Results in Searching for Minimal Primes Wrt Subword Order.” github.com/RaymondDevillers/primes, 2015.
- [Dickson 52] L. E. Dickson. *History of the Theory of Numbers. Volume I: Divisibility and Primality*. New York: Chelsea Publishing Company, 1952.
- [Granlund et al.] T. Granlund, et al. “The GNU Multiple Precision Arithmetic Library.” gmplib.org.
- [Gruber et al. 07] H. Gruber, M. Holzer, and M. Kutrib. “The Size of Higman-Haines Sets.” *Theoret. Comput. Sci.* 387 (2007), 167–176.
- [Gruber et al. 09] H. Gruber, M. Holzer, and M. Kutrib. “More on the Size of Higman-Haines Sets: Effective Constructions.” *Fundam. Inform.* 91 (2009), 105–121.
- [Haines 69] L. H. Haines. “On Free Monoids Partially Ordered by Embedding.” *J. Comb. Theory* 6 (1969), 94–98.
- [Higman 52] G. Higman. “Ordering by Divisibility in Abstract Algebras.” *Proc. London Math. Soc.* 2: 3 (1952), 326–336.
- [Hopcroft and Ullman 79] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Reading, MA: Addison-Wesley, 1979.
- [Lifchitz and Lifchitz 00] H. Lifchitz and R. Lifchitz. “PRP Records.” www.primenumbers.net/prptop/prptop.php, 2000.
- [Martin 11] M. Martin. “Primo for Linux.” www.ellipsa.eu/public/primo/primo.html, 2011.
- [Morain 94] F. Morain. “Some Primes Proven by My Programs.” www.lix.polytechnique.fr/Labo/Francois.Morain/Primes/myprimes.html, 1994.
- [OEIS Foundation Inc. 96] OEIS Foundation Inc. “The On-Line Encyclopedia of Integer Sequences.” oeis.org, 1996.
- [Penné 15] J. Penné. “LLR Version 3.8.15.” jpenn.free.fr/index2.html, 2015.
- [Reynolds 10] G. Reynolds. “Sierpinski/Riesel sieve version 0.6.17.” sites.google.com/site/geoffreywalterreynolds/programs/srsieve, 2010.
- [Scott 03] M. Scott. “Multiprecision Integer and Rational Arithmetic Cryptographic Library.” www.certivox.com/miracl, 2003.
- [Shallit 01] J. Shallit. “Minimal Primes.” *J. Recreational Math.* 30: 2 (2001), 113–117, 2003.
- [Williams and Dubner 86] H. C. Williams and H. Dubner. “The Primality of R_{1031} .” *Math. Comput.* 47 (1986), 703–711.