Subwords, Regular Languages, and Prime Numbers

Jeffrey Shallit
School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada
shallit@cs.uwaterloo.ca
https://www.cs.uwaterloo.ca/~shallit

Joint work with Curtis Bright and Raymond Devillers.

Partial orders

Recall: a *partial order* "$\leq$" on a set S is a subset $T \subseteq S \times S$ satisfying three properties (where we write $x \leq y$ if $(x, y) \in T$):

1. Reflexive: $\forall x \; x \leq x$
2. Transitive: $\forall x, y, z \; x \leq y$ and $y \leq z$ implies $x \leq z$
3. Anti-symmetric: $\forall x, y \; x \leq y$ and $y \leq x$ implies $x = y$

So partial orders mimic the behavior of "$\leq$" on the real numbers.

Comparable and incomparable elements

We say $x, y \in S$ are *comparable* according to the partial order if either $x \leq y$ or $y \leq x$.

Otherwise they are *incomparable*.

An *antichain* is a list of pairwise incomparable elements.

Some partial orders have infinite antichains and some do not...

Antichains in $\mathbb{N}^k$

Consider the following partial order on k-tuples of natural numbers ($\mathbb{N}^k$):

a point $(a_1, a_2, \ldots a_k)$ is $\leq_p (b_1, b_2, \ldots, b_k)$

if $a_1 \leq b_1$, $a_2 \leq b_2$, ..., $a_k \leq b_k$.

Are there infinite antichains in this partial order?

Antichains in $\mathbb{N}^k$

No! We prove this by induction on k.

For $k = 1$ this is clear: any two elements of $\mathbb{N}$ are comparable.

Otherwise assume true for $k - 1$ and we prove for k.

Let $p_1, p_2, \ldots$ be an infinite antichain in $\mathbb{N}^k$.

Since each of $p_2, p_3, \ldots$ are incomparable to p_1, each p_i has some coordinate where it is less than the corresponding coordinate of p_1.

Since there are only k coordinates, some coordinate has the property that infinitely many of the p_i are less than p_1 in that coordinate.

Without loss of generality, let it be the first coordinate.

Antichains in $\mathbb{N}^k$

Call these infinitely many p_i

$$q_1, q_2, q_3, \ldots.$$

Now there are only finitely many non-negative integers less than the first coordinate of p_1, so there is some non-negative integer such that infinitely many of the q_i have their first coordinate equal (say equal to d for some $d <$ first coordinate of p_1).

Call these $r_1, r_2, \ldots$.

Now delete the first coordinate of each of the r_i to get infinitely many pairwise incomparable elements in $\mathbb{N}^{k-1}$, a contradiction.

That completes the proof.

Partial orders on words

There are a number of obvious partial orders on words:

$x \leq y$ if $|x| \leq |y|$

$x \leq y$ if x is a *factor* of y (a contiguous block sitting inside y, the way `ore` is a factor of `theorem`)

$x \leq y$ if x precedes y in alphabetic order

$x \leq y$ if x is a *subword* of y (alternatively, x is obtained from y by striking out 0 or more letters of y, the way `them` is a subword of `theorem`)

Note: "subword" is also called "scattered subword" or "substring" or "subsequence".

The factor partial order has infinite antichains

For example, the set

$$\{ab^n a \ : \ n \geq 1\} = \{aba, abba, abbba, \ldots\}$$

is an infinite set in which no two words are factors of each other.

Higman-Haines theorem: the subword partial order has no infinite antichains

Write $x \triangleleft y$ for the partial order "x is a subword of y" and $x \not\triangleleft y$ for "x is not a subword of y".

Proof strategy: assume there is an infinite antichain.

This implies the weaker result that there is an infinite *division-free sequence* of words $(f_i)_{i \geq 1}$, i.e., a sequence of strings $f_1, f_2, \ldots$ such that $i < j \implies f_i \not\triangleleft f_j$.

Now iteratively choose a minimal such sequence, as follows:

- Let f_1 be a shortest word beginning an infinite division-free sequence;
- Let f_2 be a shortest word such that f_1, f_2 begins an infinite division-free sequence;
- Let f_3 be a shortest word such that f_1, f_2, f_3 begins an infinite division-free sequence; etc.

Higman-Haines theorem: the subword partial order has no infinite antichains

By the pigeonhole principle, there exists an infinite subsequence of the f_i, say $f_{i_1}, f_{i_2}, f_{i_3}, \ldots$ such that each of the strings in this subsequence starts with the same letter, say a.

Define x_j for $j \geq 1$ by $f_{i_j} = ax_j$. Then

$$f_1, f_2, f_3, \ldots, f_{i_1-1}, x_1, x_2, x_3, \ldots$$

is an infinite division-free sequence which precedes $(f_i)_{i \geq 1}$, contradicting the supposed minimality of $(f_i)_{i \geq 1}$.

To see this, note that $f_i \not\lhd f_j$ for $1 \leq i < j < i_1$ by assumption.

Next, if $f_i \lhd x_j$ for some i with $1 \leq i < i_1$ and $j \geq 1$, then $f_i \lhd ax_j = f_{i_j}$, a contradiction.

Finally, if $x_j \lhd x_k$, then $ax_j \lhd ax_k$, and hence $f_{i_j} \lhd f_{i_k}$, a contradiction. That completes the proof.

The difference between infinite and very large

Notice that although we have proved there are no infinite pairwise incomparable sets for the subword ordering, there are arbitrarily large such sets.

For example, the language $\{0,1\}^n$ consists of 2^n strings that are pairwise incomparable.

Two operations on languages

We now introduce two operations on languages, the *subword* and *superword* operations.

Let $L \subseteq \Sigma^*$.

We define

$$\begin{aligned} \text{sup}(L) &= \{x \in \Sigma^* \ : \ \text{there exists } y \in L \text{ such that } y \lhd x\} \\ \text{sub}(L) &= \{x \in \Sigma^* \ : \ \text{there exists } y \in L \text{ such that } x \lhd y\} \end{aligned}$$

Our goal is to prove that if L is a language, then $\text{sub}(L)$ and $\text{sup}(L)$ is regular.

Lemma

Let $L \subseteq \Sigma^*$. Then

(a) $L \subseteq \mathrm{sup}(L)$;

(b) $L \subseteq \mathrm{sub}(L)$;

(c) $\mathrm{sub}(L) = \mathrm{sub}(\mathrm{sub}(L))$.

Minimal elements

Let R be a partial order on a set S.

Then we say $x \in S$ is *minimal* if

$$yRx \implies y = x$$

for $y \in S$.

Basic properties of minimal elements

Let $D(y)$ be the set $\{x \in S \ : \ xRy\}$.

Lemma

Let R be a partial order on a set S.

(a) *If x, y are distinct minimal elements, then x, y are incomparable.*

(b) *Suppose the set $D(y)$ is finite. Then there exists a minimal y' such that $y'Ry$.*

The result for sup

Lemma

Let $L \subseteq \Sigma^*$. Then there exists a finite subset $M \subseteq L$ such that $\sup(L) = \sup(M)$.

Proof.

Let M be the set of minimal elements of L.

We proved that the elements of M are pairwise incomparable. Hence M is finite.

It remains to see that $\sup(L) = \sup(M)$.

Clearly $\sup(M) \subseteq \sup(L)$. Now suppose $x \in \sup(L)$.

Then there exists $y \in L$ such that $y \lhd x$. By lemma above, there exists $y' \in M$ such that $y' \lhd y$.

Then $y' \lhd y \lhd x$, and so $x \in \sup(M)$.

The second lemma

Lemma

Let $L \subseteq \Sigma^*$. Then there exists a finite subset $G \subseteq \Sigma^*$ such that $\mathrm{sub}(L) = \Sigma^* - \mathrm{sup}(G)$.

Proof.

Let $T = \Sigma^* - \mathrm{sub}(L)$. I claim that $T = \mathrm{sup}(T)$.

Clearly $T \subseteq \mathrm{sup}(T)$.

Suppose $\mathrm{sup}(T) \not\subseteq T$.

Then there exists an $x \in \mathrm{sup}(T)$ with $x \notin T$.

Since $T = \Sigma^* - \mathrm{sub}(L)$, this means $x \in \mathrm{sub}(L)$.

Since $x \in \mathrm{sup}(T)$, there exists $y \in T$ such that $y \lhd x$.

Hence, by a lemma, we have $y \in \mathrm{sub}(L)$.

The second lemma

But then $y \notin T$, a contradiction.

Finally, by part (a) there exists a finite subset G such that $\sup(G) = \sup(T)$.

Then $\sup(G) = \sup(T) = T = \Sigma^* - \text{sub}(L)$, and so $\text{sub}(L) = \Sigma^* - \sup(G)$.

The main result

Theorem

Let L be a language (not necessarily regular). Then both $\mathrm{sub}(L)$ *and* $\mathrm{sup}(L)$ *are regular.*

Proof.

Clearly $\mathrm{sup}(L)$ is regular if $L = \{w\}$ for some single word w.

This is because if $w = a_1 a_2 \cdots a_k$, then

$$\mathrm{sup}(\{w\}) = \Sigma^* a_1 \Sigma^* a_2 \Sigma^* \cdots \Sigma^* a_k \Sigma^*.$$

Similarly, for any finite language $F \subseteq \Sigma^*$, $\mathrm{sup}(F)$ is regular because

$$\mathrm{sup}(F) = \bigcup_{w \in F} \mathrm{sup}(\{w\}).$$

The main result

Now let $L \subseteq \Sigma^*$, and let M and G be defined as in the proof before.

Then $\sup(L) = \sup(M)$, and so $\sup(L)$ is regular, since M is finite.

Also, $\text{sub}(L) = \Sigma^* - \sup(G)$, and so $\text{sub}(L)$ is regular since G is finite. That completes the proof.

Representations of integers

We'll represent integers in base k using the digits $0, 1, \ldots, k-1$.

We'll write $(n)_k$ for the word giving the canonical representation of the integer n in base k (with no leading zeroes).

We'll write $[w]_k$ for the integer represented by the word w in base k (where w can have leading zeroes).

Minimal elements for the prime numbers

Consider the language

$$P_3 = \{2, 10, 12, 21, 102, 111, 122, 201, 212, 1002, \ldots\},$$

which represents the primes in base 3.

I claim that the minimal elements of P_3 are $\{2, 10, 111\}$.

Clearly each of these are in P_3 and no proper subword is in P_3.

Now let $x \in P_3$.

If $2 \not\lhd x$, then $x \in \{0, 1\}^*$.

If further $10 \not\lhd x$, then $x \in 0^*1^*$.

An example involving prime numbers

Since x represents a number, x cannot have leading zeroes.

It follows that $x \in 1^*$.

But the numbers represented by the strings 1 and 11 are not primes.

However, 111 represents the number 13, which is prime.

It now follows that

$$\sup(P_3) = \Sigma^* 2 \Sigma^* \cup \Sigma^* 1 \Sigma^* 0 \Sigma^* \cup \Sigma^* 1 \Sigma^* 1 \Sigma^* 1 \Sigma^*$$

where $\Sigma = \{0, 1, 2\}$.

On the other hand, $\mathrm{sub}(P_3) = \Sigma^*$. This follows from Dirichlet's theorem on primes in arithmetic progressions, which states that every arithmetic progression of the form $(a + nb)_{n \geq 0}$, $\gcd(a, b) = 1$, contains infinitely many primes.

The base-10 case

THE PRIME GAME

Ask a friend to write down a prime number.
Bet them that you can always strike out 0 or
more digits to get a prime on this card.

2, 3, 5, 7, 11, 19, 41, 61, 89, 409, 449, 499, 881, 991,
6469, 6949, 9001, 9049, 9649, 9949, 60649,
666649, 946669, 60000049, 66000049, 66600049

Minimal elements for the primes in other bases

A computationally difficult problem! No algorithm is known that is guaranteed to halt.

There is a "sort-of" algorithm:

(1) $M := \emptyset$
(2) while $(L \neq \emptyset)$ do

(3) choose x, a shortest string in L

(4) $M := M \cup \{x\}$

(5) $L := L - \sup(\{x\})$

It's hard to carry out step (5)!

In practice we work with L', a regular "over-approximation" of L, and we assume L' is the union of sets of the form $L_1 L_2^* L_3$, and use heuristics.

Heuristics

We have to rule out prime numbers in various regular languages.

One method is to find an N such that N divides each of the numbers $[xL^*z]_b$.

You might think you have to check $[xL^iz]_b$ for all i.

But in fact

Lemma

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

Heuristics

Corollary

If $1 < \gcd([xz]_b, [xy_1z]_b, \ldots, [xy_nz]_b) < [xz]_b$, *then all numbers of the form* $[x\{y_1, y_2, \ldots, y_n\}^*z]_b$ *are composite.*

Example: since $\gcd(49, 469) = 7$, every number with base-10 representation of the form 46^*9 is divisible by 7 and hence composite.

Other heuristics

Difference-of-squares factorization:

An example: since

$$[44^n1]_{16} = \frac{(4^{n+1} \cdot 8 + 7)(4^{n+1} \cdot 8 - 7)}{15},$$

it follows that all numbers of the form $[44^n1]_{16}$ are composite.

Our results

We were able to find the minimal elements for the primes in all bases up to 16, and some additional bases up to 30.

Sometimes we had to do primality tests on very large numbers (with thousands of digits).

For primes of the form $4n + 3$ in base 10, the set of minimal elements consists of 13 elements, with the largest having 19153 decimal digits! This was proved prime by François Morain.

Minimal elements for the composite numbers

By contrast, for computing the minimal elements for the composite numbers, there is an algorithm (Devillers).

Write $S_b := \{ (n)_b : n \geq 4 \text{ is composite} \}$.

Theorem

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 $dd \lhd w$.

If $d > 1$, the number $[dd]_b$ is composite, as it is divisible by $[11]_b$ but not equal to it.

Minimal elements for the composite numbers

If $d = 0$, then some nonzero digit c precedes it in w, so $c00 \lhd w$ and $[c00]_b$ is divisible by b^2, which is composite.

Finally, if no digit other than 1 is repeated, then $1111 \lhd w$, and $[1111]_b = [11]_b \cdot [101]_b$, and hence is composite.

Minimal elements for the composite numbers, base 10

They are:

$$\{4, 6, 8, 9, 10, 12, 15, 20, 21, 22, 25, 27, 30, 32, 33, 35, 50,$$
$$51, 52, 55, 57, 70, 72, 75, 77, 111, 117, 171, 371, 711, 713, 731\}$$

Some open problems

1. What are the minimal elements for the powers of 2, expressed in base 10? Probably

$$\{1, 2, 4, 8, 65536\}$$

   but nobody knows how to prove this!
2. Are there infinitely many primes whose base-10 representation consists of all 1's? The only known "repunit" primes are of the form $(10^p - 1)/9$ for $p = 2, 19, 23, 317, 1031$. It seems likely that those for $p = 49081, 86453, 109297, 270343$ are also prime, but this has not been rigorously proven.

Some open problems

3. Is the following problem decidable? Given a finite automaton A accepting (say) numbers expressed in base 2, does A accept the base-2 representation of at least one prime number? By contrast, the same problem with "prime" replaced with "composite" is decidable.

4. Is the following even weaker variant decidable? Given a regular expression of the form xy^*z, does it represent the base-2 expansion of at least one prime number? If this were decidable, in principle we could determine if there exists another Fermat prime in addition to $2^{2^i}+1$ for $i = 0, 1, 2, 3, 4$. (Choose $x = 1$, $y = 0$, and $z = 0^{16}1$.)