Combinatorics on words:

Avoidability of patterns

Kirby A. Baker

UCLA

GSO Seminar
March 4, 2002

I. Axel Thue and square-free words.

Work with a finite alphabet Γ.

A *word* is a nonempty string of symbols from Γ.

Γ^+ is the set of all finite words from Γ.

A word w is *square-free* if w has no repeated block. For example, `banana` is not square-free.

Theorem (Thue). If $|\Gamma| = 3$ there are infinitely many square-free words.

Ex. $\texttt{02012021020121012021020}\ldots$

Equivalently: There exists an infinite square-free word from Γ.

If $|\Gamma| = 2$? No, $\texttt{010}$ can't be continued.

On $\Gamma_4 = \{0, 1, 2, 3\}$, an easy construction of an infinite square-free word:

Let φ be the substitution $\begin{cases} 0 \mapsto 01 \\ 1 \mapsto 21 \\ 2 \mapsto 03 \\ 3 \mapsto 23 \end{cases}$

Start from 0 and iterate:

0

01

0121

01210321

0121032101230321

...

The "union" of these images is an infinite word Ω.

Proof of square-freeness of
$\Omega = 0121032101230321012103230123032 1\ldots$

Suppose $\Omega = \ldots XX \ldots$, where X is some block.

Choose X to have minimum possible length.

Observe that the symbols in Ω alternate even-odd.

Case 1: X starts with an even symbol.

Then we can pull back via φ by taking a pre-image, to get a shorter squared block.

Observe that the even symbols $0, 2$ alternate.

Case 2: X starts with an odd symbol.
We can shift left one position to obtain Case 1. $\square$

II. Avoiding or encountering a pattern

Let $\Sigma = \{a, b, \ldots, z\}$ be a second alphabet, used to express patterns.

For example, we say that Ω (as above) *avoids* the pattern word xx , while $02\underbrace{012021}_{X}\underbrace{012021}_{X}02$ *encounters* xx

For a word α to encounter xyx means that α has blocks $\ldots XYX \ldots$.

Similarly, for α to encounter $xyxzxyx$ means that α has blocks $\ldots XYXZXYX \ldots$.

III. Avoidable and unavoidable words

Given a pattern such as xx, xyx, $xyxzxyx$, we can ask:

Does there exist a finite alphabet Γ and an infinite word α on Γ avoiding the pattern?

Ex. 1: xx? Yes, $|\Gamma| = 3$ (Thue)

We say xx is an *avoidable* pattern.

Ex. 2: xyx? No, some symbol repeats, say 0, and then α would have $\dots 0 \dots 0 \dots$, giving

$$\cdots \underbrace{0}_{X} \underbrace{\cdots}_{Y} \underbrace{0}_{X} \cdots$$

We say xyx is an *unavoidable* pattern.

Ex. 3: $xyxzxyx$?

Unavoidable!

Ex. 4: $xyxzyxy$? Fact: Avoidable.

IV. Conditions for avoidability?

Sufficient conditions: A pattern word w is avoidable if

(1) Each letter of w appears at least two times:

$abcacb$

(2) $w \in \Sigma^+$, $|\Sigma| = k$, $|w| \geq 2^k$.

$babcabab$

(3) The adjacency graph of w is connected:

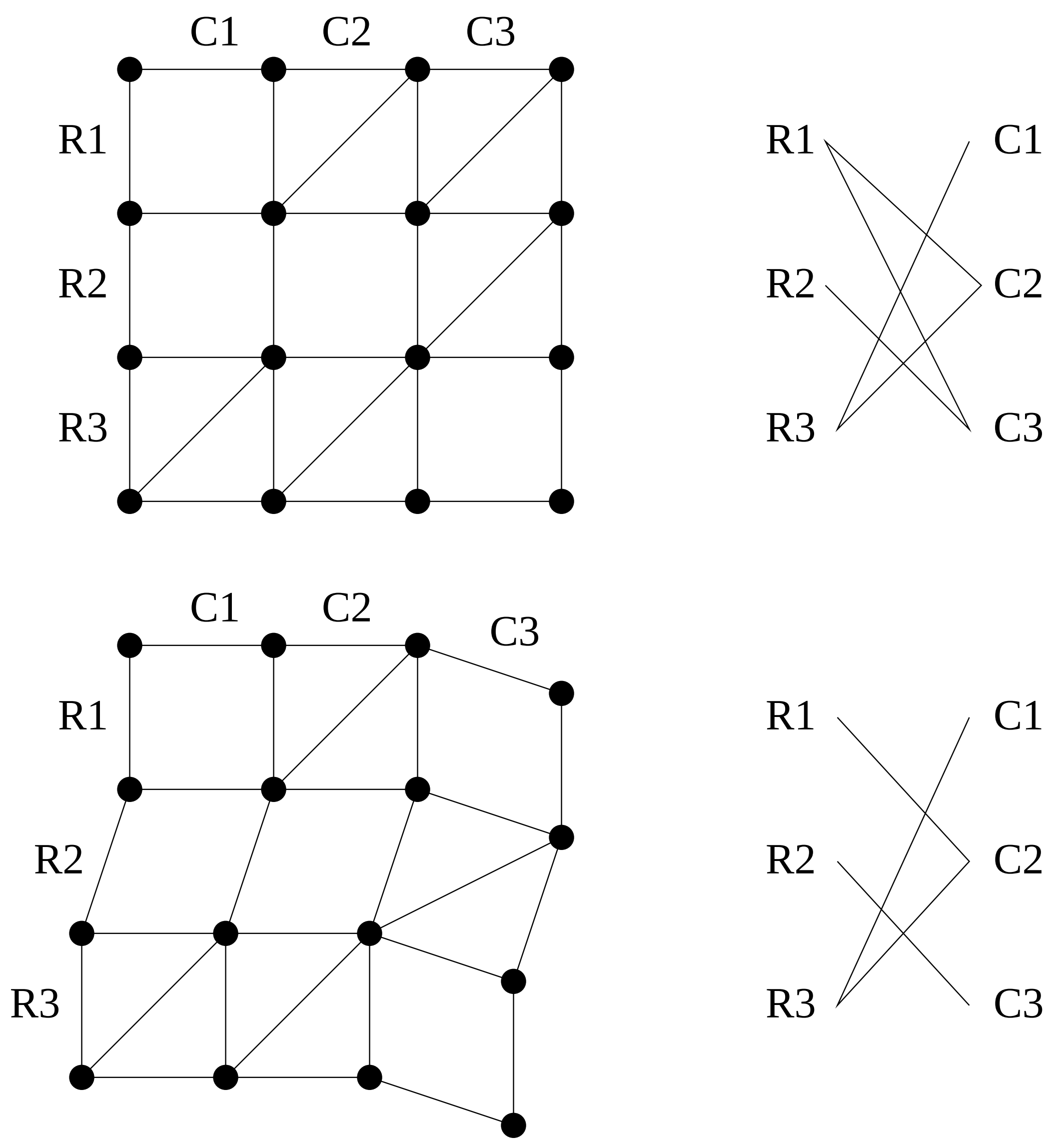
$abcacb$

a	a
b	b
c	c

Say "w is locked".

In this case Ω on four symbols avoids w (KB-McNulty-Taylor).

?? Connection with rigid/flexible planar frameworks?



Theorem: The framework is rigid if and only if the row/column graph is connected.

A necessary and sufficient condition for avoidability

(Zimin; Bean-Ehrenfeucht-McNulty):

(4) w is avoidable if and only if you can't reach the empty word by a sequence of steps from among these two options:

(i) merging two letters

(ii) erasing a letter whose two occurrences in the adjacency graph are in separate components.

Another necessary and sufficient condition:

(5) w is avoidable if and only if w does not encounter one of the Zimin words x, xyx, $xyx\ z\ xyx$, $\ldots$,

or better, $w_1 = x_1$ and $w_{n+1} = w_n x_{n+1} w_n$.

V. k-avoidability

Means: Avoidability by an infinite word from Γ with $|\Gamma| = k$.

Do examples of the following kind exist?

1. w is 2-avoidable but 1-unavoidable?

Yes, xxx (Thue)

2. w is 3-avoidable but 2-unavoidable?

Yes, xx (Thue)

3. w is 4-avoidable but 3-unavoidable?

Yes, KB-McNulty-Taylor:

$$w_{\triangle} = ab\ x\ bc\ y\ ca\ z\ ac\ u\ cb.$$

Also $abcadacb$.

4. w is 5-avoidable but 4-unavoidable?

Previously not known.

Newly discovered by R. Clark:

$$w_C = ab\ x\ ba\ y\ ac\ z\ bc\ u\ cda\ v\ dcd.$$

5. w is 6-avoidable but 5-unavoidable?

Not known!

For convenience, say the *index* of an avoidable pattern word is the least alphabet size on which it is avoidable. Thus xx has index 3, w_Δ has index 4.

(Say that an *un*avoidable pattern word has index ∞.)

So have seen examples of pattern words of indices 2, 3, 4, 5, while the existence of higher examples is not known.

Why is $w_{\triangle}$ 4-avoidable?

Locked:

a	a
b	b
c	c
x	x
y	y
z	z
w	w

Why is $w_{\triangle}$ 3-unavoidable?

Later.

VI. Formulas

The idea:

To say "α encounters the formula $aba \wedge bab$" means that there are blocks A, B such that α has $\ldots ABA \ldots$ somewhere and also $\ldots BAB \ldots$ somewhere.

In general:

A (conjunctive) formula has the form
$f = w_1 \wedge \cdots \wedge w_k$, $w_i \in \Sigma^+$.

Say α encounters f if there is a homomorphism $\mu : \Sigma^+ \to \Gamma^+$ such that each $\mu(w_i)$ is a factor of α.

Otherwise, say α avoids f.

Definition. For a pattern word $w = a_1 \dots a_n$, the *dissociation* of w, denoted $D(w)$, is the formula obtained by replacing each uniquely occurring letter by $\wedge$.

Ex. $w = abacbab \Rightarrow D(w) = aba \wedge bab$.

Proposition. w and $D(w)$ have the same index.

Proof. Let's show w is k-unavoidable $\Leftrightarrow$ $D(w)$ is k-unavoidable.

For $\Rightarrow$: Trivial, since $D(w)$ encounters w.

For $\Leftarrow$: by example—Consider the case

$$w = w_{\Delta} = ab\ x\ bc\ y\ ca\ z\ ac\ u\ cb$$

and its dissociation

$$D(w) = ab \wedge bc \wedge ca \wedge ac \wedge cb.$$

On Γ_k, $D(w)$ avoids at most a finite number of words, so there is some N such that $D(w)$ encounters all Γ-words of length N. In fact, $N = 7$ in this example.

Given an infinite word α on k symbols, chop it into blocks of lengths alternating between N and 1:

$$\alpha = \underbrace{\ldots}_{N}\underbrace{\cdot}_{1}\underbrace{\ldots}_{N}\underbrace{\cdot}_{1} \text{ etc.}$$

$D(w)$ encounters each N-block in one of finitely many ways.

Thus there are infinitely many separated N-blocks having identical encounters with $D(w)$.

Patch parts of these together into an encounter of w with α.

In the example, take AB from one, BC from the next along, and so on, obtaining an encounter of $w = w_{\Delta}$. $\square$

Applications:

(1) $xyxzxyx$?

Same as $xyx \wedge xyx$, or simply xyx, which is the same as $x \wedge x$, which is the same as x: unavoidable.

(2) $w_{\triangle} = ab\ x\ bc\ y\ ca\ z\ ac\ u\ cb$? Just done.

To check 3-unavoidability of $w_{\triangle}$ directly by computer is difficult or impossible, since can get words up to length 100 or more avoiding $w_{\triangle}$.

To check 3-unavoidability of $D(w)$ is easy; square-free and maximum length avoiding is 7, as mentioned.

(3) $w_C = ab\ x\ ba\ y\ ac\ z\ bc\ u\ cda\ v\ dcd$,

$D(W_C) = ab \wedge ba \wedge ac \wedge bc \wedge cda \wedge dcd$.

Still difficult, but makes an enormous problem tractable.

Words avoiding a given pattern formula

Contrasts:

(1) xx is easy to avoid on 3 symbols.

The number of words of length n avoiding xx grows exponentially with n.

(2) $w_\triangle$ is just barely avoidable on 4 symbols.

The number of words of length n avoiding $w_\triangle$ grows polynomially with n.

They look pretty much like Ω but with garbage on the front.

If take bi-infinite words, no garbage.

$\ldots 012103210123032101210323012303 21 \ldots$ or

$\ldots 023201320231013202320131023101 32 \ldots$ or

$\ldots$

Key endomorphisms:

For $|\Gamma| = 2$, have Thue-Morse: $\begin{cases} 0 \mapsto 01 \\ 1 \mapsto 10 \end{cases}$

Generates infinite word avoiding xxx.

For $|\Gamma| = 3$, have $\begin{cases} 0 \mapsto 012 \\ 1 \mapsto 02 \\ 2 \mapsto 1 \end{cases}$

Generates infinite word avoiding xx.

For $|\Gamma| = 4$, have φ: $$\begin{cases} 0 \mapsto 01 \\ 1 \mapsto 21 \\ 2 \mapsto 03 \\ 3 \mapsto 23 \end{cases}$$

Generates infinite word avoiding all locked words and formulas.

For $|\Gamma| = 5$, have φ: $$\begin{cases} 0 \mapsto 01 \\ 1 \mapsto 02 \\ 2 \mapsto 3204 \\ 3 \mapsto 31 \\ 4 \mapsto 3234 \end{cases}$$

Generates infinite word avoiding w_C.

These are minimal in numbers of image elements, but beyond that they seem to play a key role in some sense not yet understood.

Some names:

A. I. Zimin

D. Bean, A. Ehrenfeucht, G. McNulty

M. Sapir, applications to varieties of semigroups and Burnside problem.

J. Cassaigne thesis—classified all pattern words on 3 letters

J. Currie—cash problems

R. Clark – formulas on 3, 4 letters; formulas of index 5.

Introductory References:

M. V. Sapir, *Combinatorics on Words*, Birkhauser (to appear)

J. Currie, *Open problems in pattern avoidance*, Amer. Math. Monthly 100 (1993), 790-793.

J. Currie, web page
`http://www.uwinnipeg.ca/~currie/wordtext.html`

Example of Bean-Ehrenfeucht-McNulty. Said:

Theorem. w is avoidable if and only if you can't reach the empty word by a sequence of steps from among these two options:

(i) merging two letters

(ii) erasing a letter whose two occurrences in the adjacency graph are in separate components.

Ex. 1. $xyxzzxyx$? Graph is

x	x
y	y
z	z

Can delete x, leaving yzy, graph

y	y
z	z

Now delete y, then z, so unavoidable.

Ex. 2. $xyxzyxy$? Graph is

x	x
y	y
z	z

Try deleting x, get $yzyy$, graph

y	y
z	z

The yy will prevent reduction to empty. Other routes also fail.

So $yxyzyxy$ is avoidable.