PATTERN AVOIDANCE IN WORDS

ANSHUL MANTRI

Abstract

This paper is about combinatorics on words, which is the study of finite and infinite sequences of symbols. We will use free monoids and morphisms to generate infinite words. The main focus of this paper is patterns, which are words made of variables. We will explore whether certain patterns are necessarily present in all infinite words. We will disucss the infinite Thue-Morse word and the patterns it avoids. Finally, we will discuss how to check if a pattern is unavoidable and characterize all unavoidable patterns using sesquipowers and Zimin's reducing algorithm.

1. Introduction

Combinatorics on words is a relatively new field of mathematics that studies finite and infinite sequences of symbols, which are called words. This topic seems very fundamental, but surprisingly it was not extensively studied until quite recently. The study of combinatorics on words started in the early 20th century when Axel Thue studied squarefree words. Since then, there have been many contributions and developments by mathematicians including van der Waerden, de Brujin, Zimin, and many others. More recently, a group of authors by the name M. Lothaire wrote a comprehensive series of books about combinatorics on words, one of which is found at [Lot02]. Combinatorics on words has useful applications in computer science, information theory, cryptography, and even bioinformatics. The main topic of this paper, patterns, is related to Ramsey theory, which is about finding the minimal size of a structure that guarantees the existence of some substructure.

In this paper, we will begin with basic definitions about words. We will also consider alternative ways to view words using algebraic structures like monoids and semigroups.

Next, we will define morphisms, which we will find to be helpful later on in the paper. We will also consider how to generate infinite fixed points from certain morphisms.

After that, we will formally define what patterns are, and what it means to encounter or avoid them. We will also define some other terms related to patterns

Later, we will talk about powers, which are the simplest patterns, and power-free words. We will show that the Thue-Morse sequence on two letters is cubefree and overlap-free, and we will use this fact to generate an infinite squarefree word on three letters.

Lastly, we will define the sesquipowers, and consider their pattern analogy using variables, which are called the Zimin patterns. We will prove that the Zimin patterns are unavoidable on all infinite words, and we will show that reducibility of a pattern via the Zimin algorithm implies unavoidability.

Date: July 2024.

2. Preliminaries

In this section we will define some important notations that will be used later on in the paper. Before we can formally define what words are, we must first discuss what they are made of.

Definition 2.1. An *alphabet* is a finite set of symbols, and is typically denoted by Σ. The elements of Σ are called *letters*.

We are now ready to introduce *words*.

Definition 2.2. A *word* is a sequence of letters from an alphabet A.

Words are generally finite in length, but this paper will also discuss words that are infinite in one direction. There also exist words that are infinite in both directions. For a finite word w, we denote its length as $|w|$.

Definition 2.3. A *factor* or *subword* of a word w is a contiguous subsequence of w of any length.

If a factor u of a word w contains the first letter of w, then u is a *prefix* of w, and if it contains the final letter of w, then u is a *suffix* of w. The term "factor" alludes that a word is equal to a product of some of its subwords. This is true under the operation of *concatenation*, i.e., $(u)(v) = (uv)$ for words u, v. It is important to note that concatenation is not commutative, since $(u)(v)$ does not necessarily equal $(v)(u)$. However, the operation of concatenation has some other useful properties.

Definition 2.4. An operation, say $\cdot$, between two objects of a set Σ is *associative* if for any $a, b, c \in \Sigma$, we have $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.

Definition 2.5. An element of a set Σ with a binary operation is called the *identity element*, denoted as ε, if for all $a \in \Sigma$, we have $\varepsilon a = a\varepsilon = a$.

There can only be one identity element, because if there were two identity elements x and y, then their product would have to equal both x and y, so $x = y$ after all. However, it is not necessary for an identity element to exist.

Definition 2.6. A *monoid* is a set with an associative binary operation that it is closed under, and an identity element.

Example. The whole numbers with the operation of addition and identity element of 0 is a monoid. It is denoted as $(\mathbb{N}, +, 0)$.

Proof. We know that the set is closed under addition because the sum of two natural numbers is always a natural number. Next, addition is associative on this set because for all $a, b, c \in \mathbb{N}$, we have $(a+b)+c = a+(b+c)$. Finally, 0 is the identity element since $a+0 = 0+a = a$. ∎

Going back to concatenation of words, we can easily verify that concatenation is associative. Moreover, there exists an identity word, namely, the empty word, which we denote as ε. Therefore, the set of words given an alphabet is a monoid, which has a special name.

Definition 2.7. The *free monoid* of an alphabet Σ is the monoid defined with the binary operation of concatenation, and with the empty word ε as the identity element. The free monoid of Σ is denoted as Σ^*.

In other words, the free monoid of Σ is the set of all words of any length whose letters are in Σ. Sometimes we do not want to include the empty word in this set, so we use a different set that excludes it.

Definition 2.8. A *semigroup* is a set closed under an associative binary operation.

Semigroups are the same as monoids, except they need not contain an identity element.

Definition 2.9. The *free semigroup* over an alphabet Σ is the semigroup defined with the operation of concatenation. It is denoted as Σ^+.

The sets Σ^* and Σ^+ only differ by the empty word, so it is true that $\Sigma^* = \Sigma^+ \cup \{\varepsilon\}$.

Definition 2.10. Given an alphabet Σ, a language L is a subset of the set Σ^*.

We may define languages for various reasons, but they are useful for categorizing words with certain properties.

We will use *morphisms* very often in this paper to define words. The following is the broad definition of a morphism.

3. Morphisms

Morphisms are special functions that can be used between words. We will use many morphisms in definitions and proofs later on, so this section is devoted to showing how morphisms work, and how they can generate infinite words with special properties.

Definition 3.1. A *morphism* g between two free monoids Σ^* and Γ^* is a function that maps each element of Σ^* to an element of Γ^* such that for any $a, b \in \Sigma^*$, we have $g(ab) = g(a)g(b)$. It is denoted as $g : \Sigma^* \to \Gamma^*$.

Morphisms can be defined between various algebraic strucutres, but for the purpose of combinatorics on words, we will only use morphisms between free monoids. We call Σ^* the *domain*, and Γ^* the *codomain*, of the morphism. If $\Sigma = \Gamma$, we call the morphism an *endomorphism*. Additionally, if $g(a) = b$, then b is the *image* of a and a is a *preimage* of b with respect to g.

Since $\varepsilon \in \Sigma^*$, by letting $a = b = \varepsilon$, we get

$$g(\varepsilon\varepsilon) = g(\varepsilon) = g(\varepsilon)g(\varepsilon).$$

Hence, $g(\varepsilon)$ must be the identity of Γ^*, so the identity of the domain maps to the identity of the codomain in all morphisms. However, since morphisms are not necessarily injective, it is still possible for a nonempty element of Σ^* to map to ε. If we want to forbid this, we can define a *nonerasing morphism*, which prevents any element of the domain other that ε to map to ε.

We can apply a morphism to a word to generate a new word.

Definition 3.2. If for some endomorphism $h : \Sigma^* \to \Sigma^*$, we have that $w = h(w)$, then we call w a *fixed point* of h.

For example, ε is a fixed point of every morphism. We are interested in morphisms that have infinite words as fixed points.

Definition 3.3. If we can apply a nonerasing morphism indefinitely to a starting letter to generate an infinite word, we say the morphism is *prolongable*, and the infinite word generated by the morphism is *morphic*.

Formally, a morphic word w starting with the letter a generated by a prolongable morphism is defined as

$$w = \lim_{n\to\infty} h^n(a),$$

where $h^n(a)$ represents the composition on h, n times. Clearly, w is a fixed point of h, and it is unique. However, this limit does not exist for non-prolongable morphisms, and it might not exist for some letters of Σ even with a prolongable morphism.

Proposition 3.4. *A nonerasing endomorphism $h : \Sigma^* \to \Sigma^*$ is prolongable if and only if there exists a letter $a \in \Sigma$ such that $h(a) = au$ where $u \in \Sigma^+$. Furthermore, there exists a unique morphic word w that begins with a for each satisfactory letter a.*

Proof. Assume for the sake of contradiction that $w = \lim_{n\to\infty} h^n(a)$ exists and begins with a. If $h(a)$ does not begin with a, then $h(w)$ will not begin with a either, which contradicts the fact that w is a fixed point of h. Next, if $h(a) = a$, then $w = a$, which contradicts the fact that w is infinite.

Now, if $h(a) = au$, then since h is nonerasing, the length of a word containing a increases by at least $|u|$ if h is applied, so the word w is indeed infinite. ■

We have shown when a morphic word can exist, that it exists if so, and that it is unique. However, we don't actually know how to find what the word is. This is because with our current definition, the infinite word can look very different depending on the order of the letters to which we apply the morphism. This can be remediated easily, since we just need to find a single construction for the word.

Proposition 3.5. *Given a nonerasing endomorphism $h : \Sigma^* \to \Sigma^*$ with $a \in \Sigma$ such that $h(a) = au$ where $u \in \Sigma^+$, we have*

$$\lim_{n\to\infty} h^n(a) = auh(u)h^2(u)h^3(u)\cdots .$$

Proof. Note that

$$\begin{aligned} h(auh(u)h^2(u)h^3(u)\cdots) &= h(a)h(u)h(h(u))h(h^2(u))\cdots \\ &= auh(u)h^2(u)h^3(u)\cdots , \end{aligned}$$

so the word $auh(u)h^2(u)h^3(u)\cdots$ is a fixed point of h. By 3.4, there is a unique infinite fixed point starting with a that is equal to $\lim_{n\to\infty} h^n(a)$, so we have found it. ■

Now, we are able to calculate any arbitrary letter in an infinite morphic word of a morphism. To make this process easier to understand, we will go through a concrete example.

Example. Let the alphabet $\Sigma = \{a, b\}$. Consider the endomorphism $\varphi : \Sigma^* \to \Sigma^*$, such that

$$\begin{aligned} \varphi(a) &= ab \\ \varphi(b) &= a \end{aligned}$$

The infinite word f is generated by starting with the initial symbol, a, which we will call f_0. On each iteration, we apply the morphism φ to the previous word to get the new word, so

$f_{n+1} = \varphi(f_n)$ for all $n > 0$. These are the first few terms:

$$\begin{aligned} f_0 &= a \\ f_1 &= ab \\ f_2 &= aba \\ f_3 &= abaab \\ f_4 &= abaababa \\ &\vdots \\ f &= abaababaabaab\ldots \end{aligned}$$

The infinite word f is defined as $ab\varphi(b)\varphi^2(b)\cdots$, but a finite prefix of it is equal to the word generated by applying φ some finite number of times, and the limit of both words as they go to infinity are equal. For the specific morphism φ, this word is called the *Fibonacci word*.

Another example of a morphic word is the Thue-Morse word, which we will discuss extensively later on in the paper.

4. Patterns

One of the most important and natural types of regularities in words, and the focus of this paper, is a pattern. In this section, we will explore patterns in words, and what it means for a pattern to be avoidable or unavoidable.

Recall that for a word to contain another word as a factor, it must contain the exact word. The difference with patterns is that the letters of patterns can correspond to more than one letter in the word.

To formally define patterns, we will define a new alphabet Δ. The elements of Δ are called *variables*, and the words in Δ^* are called *patterns*.

Definition 4.1. Given a pattern p and an alphabet Σ, the *pattern language* P is the language on Σ consisting of all words $w \in \Sigma^*$ such that there exists a nonerasing morphism $h : \Delta^* \to \Sigma^*$ where $h(p) = w$.

In other words, a word w is in P if and only if it is the result of substituting each distinct variable of p with a word in Σ^+. Note in particular that since the morphism is nonerasing, a variable cannot map to the empty word ε.

Definition 4.2. A word is said to *encounter* a pattern p if it contains a factor that is in the pattern language of p. On the other hand, if a word does not encounter a pattern, it *avoids* the pattern.

This is illustrated in the following example.

Example. Consider the pattern $p = xxy$. The word $u = abcacabb$ encounters p since the image of p under the morphism $h : \{x, y\}^* \to \{a, b, c\}^*$ with $h(x) = ca$ and $h(y) = b$ gives $h(p) = cacab$, which is a factor of u. On the other hand, the word $v = bacbcaba$ avoids p.

An important fact is that the pattern morphism is not necessarily injective, so multiple distinct variables can map to the same word. Thus, if a factor of a word w is repeated consecutively at least $|p|$ times, then w necessarily encounters p, no matter what the pattern p is.

Patterns can be labeled as either avoidable or unavoidable on alphabets of specific sizes.

Definition 4.3. If there exists an infinite word on an alphabet with k letters that avoids a pattern p, then p is *k-avoidable.* Otherwise, p is *k-unavoidable.*

More generally, if a pattern is unavoidable on all alphabets, then we simply say it is *unavoidable.*

Definition 4.4. The *avoidability index* of a pattern p is the smallest integer k such that p is k-avoidable. If p is unavoidable, then the avoidability index of p is valued as ∞.

5. Powers

The simplest patterns in words are when a single subword is repeated some number of times. Patterns of this form have a special name.

Definition 5.1. A pattern is a *power* if it has exactly one distinct variable.

In fact, we have even more specific names for basic powers.

Definition 5.2. A word is a *square* if it is in the pattern language of the pattern xx. Similarly, a word is a *cube* if it is in the pattern language of the pattern xxx.

Definition 5.3. A word that avoids squares is *squarefree*, and a word that avoids cubes is *cubefree.*

In this section, we will compute the avoidability indices of these two patterns. First, we will consider an alphabet with two letters. It is easy to see that the pattern xx is 2-unavoidable.

Lemma 5.4. *There are no infinite squarefree words on the alphabet $\{a, b\}$.*

Proof. Assume for the sake of contradiction that there exists an infinite binary word w that is squarefree. Without loss of generality, the first letter is a. To avoid a square, any two adjacent letters must be distinct. Hence, the first four letters are $abab$. However, this subword contains the square $(ab)^2$, which is a contradiction. ■

Corollary 5.5. *All binary words with length at least 4 contain a square.*

While there are no infinite squarefree words on an alphabet with 2 symbols, there are numerous cubefree words. The most notable of them, which is also one of the most important infinite words in all of combinatorics on words, is called the Thue-Morse sequence, also known as the Prouhet-Thue-Morse sequence. It was independently discovered by many mathematicians including Prouhet, Thue, Morse, and others, but Thue was the first to use in the context of combinatorics on words.

The Thue-Morse word is morphic, since it is generated by the endomorphism $\mu : \Sigma* \to \Sigma*$ where $\Sigma = \{a, b\}$, with starting term a and the following equations:

$$\mu(a) = ab$$
$$\mu(b) = ba$$

Definition 5.6. The *Thue-Morse word*, which we refer to as t, is the infinite word on the alphabet $\Sigma = \{a, b\}$ that is the fixed point starting with a of the morphism μ.

The first few iterations of the morphism, which we will denote as t_i, are listed below.

$$\begin{aligned}
t_0 &= a\\
t_1 &= ab\\
t_2 &= abba\\
t_3 &= abbabaab\\
t_4 &= abbabaabbaababba\\
t_5 &= abbabaabbaababbabaababbaabbabaab\\
&\vdots\\
t &= abbabaabbaababbabaababbaabbabaab\ldots
\end{aligned}$$

This sequence can also be generated in a variety of other ways.

For example, for $n \geq 1$, the word t_n appears to be the concatenation of t_{n-1} and the image of t_{n-1} under the morphism $h : \Sigma^* \to \Sigma^*$ where $h(a) = b$ and $h(b) = a$. For simplicity, we call $h(w) = \overline{w}$ for all $w \in \Sigma^*$. Indeed, this is true, and can be proved with the morphism.

Proposition 5.7. *For $n \geq 1$, we have $t_n = t_{n-1}\overline{t_{n-1}}$.*

Proof. We will proceed with induction. For the base case, it is true that $t_1 = t_0\overline{t_0} = ab$. The inductive hypothesis is that $t_n = t_{n-1}\overline{t_{n-1}}$.

By the symmetry of the morphism μ, the word $\overline{t_n}$ is equal to the word $\mu^n(b)$, which has starting letter b instead of a. Hence,

$$\begin{aligned}
t_{n+1} &= \mu(t_n)\\
&= \mu(t_{n-1}\overline{t_{n-1}})\\
&= \mu(t_{n-1})\mu(\overline{t_{n-1}})\\
&= \mu(\mu^{n-1}(a))\mu(\mu^{n-1}(b))\\
&= \mu^n(a)\mu^n(b)\\
&= t_n\overline{t_n}.
\end{aligned}$$

By induction, the proposition is true. ■

To prove that t is indeed cubefree, we need the following lemma.

Lemma 5.8. *Every two letters of t, starting from the first two and not skipping or repeating any letters, contains exactly one a and one b.*

Proof. Since t is morphic, applying the morphism to t results in itself. After applying the morphism to t, each letter of t is replaced with either ab or ba. Thus, every pair of two letters in t is either ab or ba. ■

Corollary 5.9. *The word formed by taking the left letter in each pair is a copy of t.*

Proof. The unique preimage of ab is a, and the unique preimage of ba is b. In both cases, the preimage is the left letter of the pair. ■

Proposition 5.10. *The Thue-Morse word is cubefree.*

Proof. We will use a proof by contradiction, so we assume that t encounters the pattern xxx. If this is true, then there is a minimal cube factor in t, say uuu where $u \in A^+$, such that no other cube in t has a smaller length.

The word uuu must have a factor of either aa or bb, because if it were alternating between a and b, that would imply the existence of either aaa or bbb in a preimage of t with respect to the morphism μ. Without loss of generality, aa is a factor of uuu, which implies that it is a factor of uu. By 5.8, the word aa can only appear in every other factor of t of length 2, but we know that it appears in two identical factors uu of uuu that are u letters apart. Thus, $|u|$ is even.

Let u' be the word formed by taking every other letter of u in such a way that every chosen letter is the left letter of the pair as defined in 5.8. Since $|u|$ is even, this is still true if we take the same u' of each u in the word uuu. By 5.9, the word $u'u'u'$ is a factor of t. This is a cube of length $\frac{|uuu|}{2}$. However, this contradicts the minimality of the cube uuu, so it is impossible for there to be a cube in t. ■

We can actually use a similar proof to get a stronger result about the powers in the Thue-Morse word.

Definition 5.11. An *overlap* is an occurence of the pattern $xyxyx$.

We call this pattern an overlap because encountering this pattern in a word w is equivalent to containing the same factor in two distinct places in w such that the two factors overlap.

Proposition 5.12. *The Thue-Morse word is overlap-free.*

Proof. This proof is very similar to the one in 5. We start in the same way, by assuming the existence of an overlap $uvuvu$ of minimal length.

We know that neither $ababa$ nor $babab$ are factors of t since they would imply the existence of either aaa or bbb in a preimage of t with respect to μ. Hence, aa or bb must be a factor of $uvuvu$ and thus a factor of uvu. This time, the two factors uvu are vu letters apart, so $|vu|$ is even.

The rest of the proof is similar to 5, but we have to take a few extra precautions because we aren't guaranteed that $|u|$ is even. If $|u| \geq 2$, we can use the last two letters of u along with $vuvu$ to force the existence of a shorter overlap in their preimage. If $|u| = 1$, then if the leftmost u is the left letter in its pair, then this still works. If it is the right letter, we need to use the letter directly to the left. By 5.8, this letter is guaranteed to be equal to the letter to the left of the rightmost u, so we are still guaranteed to have a smaller overlap. ■

Since the Thue-Morse sequence is overlap-free, it is as close to being squarefree as possible, without being squarefree. This is because any square factor uu of t that starts with the letter a can be written as $avav$ for a word v. Then, to prevent an overlap of the form $avava$, the next letter cannot be a. In other words, no matter the size of a square in t, the power cannot continue for even one letter more. For this reason, we say that that overlap-free words are 2^+-power free.

There are, however, many infinite squarefree words on an alphabet with 3 symbols. We will look at an example of such a word. This example is due to [BK03].

Example. Define a morphism $g : \{a, ab, abb\}^* \to \{a, b, c\}$, where

$$g(a) = a$$
$$g(ab) = b$$
$$g(abb) = c$$

This is not a regular morphism because the letters of the domain are entire words. This is okay, since the morphism is properly defined as long as the input has a unique representation in $\{a, ab, abb\}^*$.

Define the word w as $d(t)$, where t is the Thue-Morse word. We know w is defined because all words on $\{a, b, \}^*$ without a factor of bbb are in the domain of g, and by 5, t does not contain the factor bbb.

Claim 5.13. *The word w is an example of an ternary squarefree word.*

Proof. For the sake of contradiction, we assume that w contains a factor that is a square. This square can be written as uu, where u is word of length $x \geq 1$. No matter what the first letter of u is, its preimage with respect to d begins with the letter a. The same is true for the letter directly after uu in w. This implies the existence of an overlap in t, which is a contradiction by 5.12. ■

Powers can be generalized to accounts for overlaps. This introduces the notion of fractional powers, which is illustrated in the following example.

Example. The word *abcabcaba* contains an $\frac{8}{3}$ power. This is because the word *abc* is repeated 2 times, and the first $\frac{2}{3}$ of *abc* is directly after.

Definition 5.14. The *repetition threshold* of an integer $k > 1$ is the largest exponent e such that there is no e-power free word on an alphabet with k letters.

By 5.4, the repetition threshold of 2 is at least 2. By 5.12, it is at most 2. Therefore, we have shown that the repetition threshold of 2 is equal to 2. We also showed in 5.13 that the repetition threshold of 3 is less than 2. In fact, this can be improved a fair amount, and Dejean showed that the repetition threshold of 3 is equal to $\frac{7}{3}$. A proof of this is given in [Ram07]. Dejean also conjectured a general formula for the repetition threshold, parts of which were proven by various mathematicians until the general case was proved in 2009 in [CR09].

Theorem 5.15 (Dejean's Theorem)**.** *The value of the repetition threshold as a function of the number of letters $k \geq 2$ is defined as follows.*

$$RT(k) = \begin{cases} \frac{7}{4} & \textit{if } k = 3 \\ \frac{7}{5} & \textit{if } k = 4, \\ \frac{k}{k-1} & \textit{if } k = 2, k \geq 5. \end{cases}$$

6. Zimin Patterns

It turns out that there are some patterns that are unavoidable in an infinite word with any finite number of distinct letters.

Proposition 6.1. *The pattern xyx is unavoidable.*

Proof. In an infinite word w on a finite alphabet, there must be some letter that appears at least 3 times. If this letter is a, then w can be rewritten as $w_0aw_1aw_2aw_3$. If $x = a$ and $y = w_1aw_2$, then xyx is a factor of w, so it is unavoidable. ■

Corollary 6.2. *On an alphabet with k symbols, all words of length at least $2k+1$ encounters the pattern xyx. This bound is sharp.*

Proof. By the Pigeonhole Principle, there must be exist a letter that appears at least three times in any $2k + 1$ letters. The rest of the proof is the same as 6.1.

This is not true for a word of length $2k$, due to the following counterexample: if the alphabet is $\{a_1, a_2, \dots a_k\}$, then the word $a_1a_1a_2a_2 \dots a_ka_k$ avoids the pattern xyx. ■

We can generalize this idea to work for entire patterns.

Lemma 6.3. *Let p be a pattern that is unavoidable on an alphabet Σ. If x is a variable that does not appear in p, then the pattern pxp is also unavoidable on Σ.*

Proof. Even though the pattern p is encountered multiple times in all infinite words, the pattern does not necessarily represent the same word, so we can't use the same proof as 6.1.

Let k be the number of symbols in Σ. Since p is unavoidable, it is encountered in every infinite word on Σ. Hence, there is some finite length l such that all k^l words in Σ^l encounter p.

Consider a word w with $(k^l+1)l+k^l$ letters. This word can be viewed as the concatenation of k^l+1 words of length l, with one letter separating any adjacent words. By the Pigeonhole Principle, at least two of the words of length l are the same. For any such word, since it has length l, it encounters p. Thus, we can find identical occurrences of p in two distinct locations in w. There is at least one letter between these two occurrences of p, so the subword between them is a valid value for x. Therefore, the pattern pxp is unavoidable. ■

Using this proposition, we can recursively create an infinite class of patterns.

Definition 6.4. The *sesquipowers*, also called the *Zimin words*, are defined as follows. Let $Z_0 = \varepsilon$, and for every positive integer n, let a_n be a new symbol in an alphabet Σ. Then, $Z_{n+1} = Z_na_nZ_n$.

We are interested in the Zimin patterns, which are sesquipowers in the pattern alphabet Δ.

Theorem 6.5. *The Zimin patterns are unavoidable.*

Proof. We will proceed with induction. The base case is true because every word has the empty word as a factor. For the inductive hypothesis, assume that the pattern Z_n is unavoidable. By 6.3, the new pattern Z_{n+1} is also unavoidable. By induction, all Zimin patterns are unavoidable. ■

Corollary 6.6. *All patterns that are encountered in the sesquipowers are unavoidable patterns.*

Surprisingly, the converse of this statement is also true. This was first proved by Zimin in 1984 in [Zim84].

Theorem 6.7 (Zimin). *A pattern is unavoidable if and only if it is encountered in a Zimin pattern.*

To do this, Zimin used an algorithm which became known as the Zimin algorithm to determine whether a word is avoidable. A similar algorithm was found by Bean et al. a few years earlier, which is shown in [BEM79], but Zimin's algorithm is preferred because it is simpler. In this paper, we will not prove the other direction of 6.7, but Zimin's proof is given in [Zim84]. We will, however, discuss Zimin's algorithm. To do this, we must first define some new terminology.

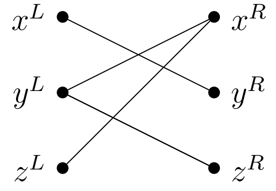
Definition 6.8. Given a pattern $p \in \Delta^*$, the *adjacency graph* of p is a bipartite graph between the sets of vertices ^{L} and Δ^R. Each variable $x_i \in \Delta$ corresponds to a vertex $x_i^L \in \Delta^L$ and a vertex $x_i^R \in \Delta^R$. There is an edge between vertices x_m^L and x_n^R if and only if the subpattern $x_m x_n$ is a factor of p.

Definition 6.9. A subset F of Δ is called a *free set* if for any $x, y \in \Delta$, the vertices x^L and y^R are disconnected in the adjacency graph.

By "disconnected," we mean that there is no path between the vertices, so it is possible for vertices to be connected even if there is not an edge between them.

Definition 6.10. A pattern p *reduces* to a pattern q if q is the result of repeating any number of times the operation of deleting all variables in a free set. If p reduces to ε, then p is *reducible.*

Example. Consider the pattern $p = xyzxyx$. The following is the adjacency graph of p.



In this pattern, the set $\{x\}$ is a free set because the vertices x_L and x_R are disconnected. Hence, we may delete x from the pattern, and it reduces to yzy. We may continue deleting free sets from the new graph to determine if the pattern is reducible, but we already know that the pattern yzy is unavoidable since it is a Zimin pattern.

Lemma 6.11. *If a pattern p reduces in one step to an unavoidable pattern q, then p is also unavoidable*

Proof. We will prove that p is unavoidable on all alphabets by induction on the size of the alphabet. The base case of an alphabet of size 1 is trivial because all patterns are unavoidable on this alphabet. We may assume that p is unavoidable on the alphabet Σ' with k symbols.

We will now prove that p is unavoidable on the alphabet $\Sigma = \Sigma' \cup \{a\}$, where a is a symbol not in Σ'. Let L be the set of words on Σ' that avoid p, which is finite since by our assumption p is unavoidable on all infinite words in Σ'^*. Let M be the set of words on Σ that start with a and avoid p.

Next, we will define an alphabet where the letters are words, but we can treat them as single letters. Let N be the set of words in the form $a^i w a^j$, where $i, j < |p|$ and $w \in L$. Since L is finite, the set N is also finite. We can write any word in M that is not a power of a as the product of at least one word in N because a word that avoids p cannot have $|p|$ consecutive a's, and w always avoids p.

Let Δ be the alphabet of p, and define a new alphabet $\Delta' = \Delta \cup \{x\}$ where x is a variable not in p. Since the pattern q is unavoidable, and x does not occur in q either, the pattern

qx is also unavoidable. Thus, every sufficiently long word w in N^* encounters the pattern qx. This implies that there is a morphism $g : \Delta'^* \to N^*$ such that $g(qx)$ is a factor of w. Also, since for any variable y we have $g(y) \in N^+$, we also know that $g(y) \in aA^+$. Now, we will define a morphism $h : \delta^* \to \Sigma^*$ based on a free set F of the adjacency graph, with the following rules. Here, a^{-1} represents the deletion of a from a word.

(1) If $y \in F$, then $h(y) = a$.
(2) If y^L and y^R are both disconnected to all left vertices of variables in F, then $h(y) = g(y)$.
(3) If y^R is connected to a left vertex in F, but y^L is not, then $h(y) = a^{-1}g(y)$.
(4) If y^L is connected to a left vertex in F and y^R is not, but $y \notin F$, then $h(y) = g(y)a$.
(5) If y^L and y^R are both connected to a (possibly distinct) left vertex of F, then $h(y) = a^{-1}g(y)a$.

These cases are exclusive of each other and they cover every possible scenario, so this morphism is properly defined for all inputs. Now we must prove that $h(p)$ is a factor of $g(qx)$. This can be proved by induction on k, the length of prefixes p_k and q_k of p and q respectively, where we assume that p_k reduces to q_k. We want to prove that $rh(p_k) = g(q_k)s_k$, where r and s_k are either ε or a depending on the first and last letters of p_k. The base case $k = 1$ is true by the definition of h. For the inductive step, let the last letter of p_k be α and the last letter of p_{k+1} be β, so there is an edge from α^L to β^R. We must show that $rh(p_{k+1}) = g(q_{k+1})s_{k+1}$, where s_{k+1} is ε or a depending on whether β^L is connected to a left vertex of F. By the inductive hypothesis,

$$rh(p_{k+1}) = rh(p_k)h(\beta) = g(q_k)s_k h(\beta).$$

If $\beta \notin F$, then this reduces to $g(q_k)g(\beta)s_{k+1} = g(q_{k+1})s_{k+1}$, and if $\beta \in F$, it reduces to $g(q_k)s_{k+1}$. In both cases, by the definition of h, the value of s_{k+1} will be a if α^L is connected to a left vertex of F, and ε otherwise.

Now, since $h(p)$ is a factor of $g(qx)$, all sufficiently long words $w \in N^*$ encounter p. Therefore, the set M is finite, so p is unavoidable on A. ■

Proposition 6.12. *All Zimin patterns are reducible.*

Proof. Begin by drawing the adjacency graph of the Zimin pattern. Note that every other letter of a sesquipower is the same. In an arbitrary Zimin pattern p, we will call this variable x. Every other variable in the pattern is preceded and followed by x, so there x^L has an edge to all right vertices, and x_R has an edge to all left vertices, except there is no edge between x_L and x_R. In fact, x_L is disconnected from x_R, so $\{x\}$ is a free set. Thus, we may remove all occurences of x from the pattern. The resulting pattern is another sesquipower with the remaining variables. Inducting downwards, the pattern eventually becomes $Z_0 = \varepsilon$. ■

Corollary 6.13. *All factors of Zimin patterns are reducible.*

Proof. The edges of the adjacency graph of a factor of a Zimin pattern p are some subset of the edges of the adjacency graph of p. If two vertices are disconnected, then they will still be disconnected after removing some of the edges. Hence, we can delete the same free sets in the same order as the Zimin pattern, and it will necessarily reduce any factor of the Zimin pattern. ■

Acknowledgements

The author would like to thank Sawyer Dobson and Simon Rubenstein-Salzedo for helping him create this paper.

References

[BEM79] Dwight Bean, Andrzej Ehrenfeucht, and George McNulty. Avoidable patterns in strings of symbols. *Pacific Journal of Mathematics*, 85(2):261–294, 1979.

[BK03] Jean Berstel and Juhani Karhumäki. Combinatorics on words: a tutorial. *Bulletin of the EATCS*, 79(178):9, 2003.

[CR09] James Currie and Narad Rampersad. A proof of dejean's conjecture, 2009.

[Lot02] M. Lothaire. *Algebraic Combinatorics on Words*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2002.

[Ram07] Narad Rampersad. *Overlap-free words and generalizations*. PhD thesis, University of Winnipeg, 2007.

[Zim84] AI Zimin. Blocking sets of terms. *Mathematics of the USSR-Sbornik*, 47(2):353, 1984.