On the Existence of an Infinite Cube-Free Word on Two Letters

Jack Conger

June 1, 2014

1 Introduction

Axel Thue's first paper on infinite sequences of letters, or words [3], establishes some of the first substantial results in the field of combinatorics on words. Often read with another paper published in 1912 [4], this work establishes possible properties of infinitely long words of alphabets of certain amounts of letters. Particularly, his ultimate theorem in [3] is the central theorem of this paper: that there exists an infinitely long cube-free word on an alphabet of just two letters.

A main focus of the field of combinatorics on words deals with the nature of long words—what characteristics are unavoidable for words of sufficient length, and which are avoidable. This paper's central theorem is an example of such an investigation: must an infinite word have internal repetition, that is, a square, or a cube, or even greater repetition? This intrigues me personally because it is an elementary example of complexity emerging from a simple system, but such questions have practical ramifications on, for instance, data compression in the field of computer science—can we guarantee certain regularities in arbitrarily large data which we can exploit to compress such data?—and, indeed, combinatorics on words coalesced from results in numerous disparate fields of mathematics, demonstrating it is a field borne of its own utility.

2 Background Definitions and Results

The following definitions are summarized from [1, 2].

Words

Being a theorem about words, we must define what words *are*.

Consider a finite[1] set A. We call A an *alphabet*, with each element of A a *letter*. Any sequence of letters $a_i \in A$ is a *word* over A, of the form $a_1 a_2 \cdots a_n$. We define n in this example to be the *length*, or number of characters, in the word. The set of all words over A is denoted by A^*. This set includes the *empty word* ϵ, the unique word of length 0. We can also have words of infinite length, such that $u = a_1 a_2 a_3 \cdots$. Infinite words can also be thought of as a mapping from $\mathbb{N} \to A$.

We also define a binary operation *concatenation* such that for finite $u = a_1 a_2 \cdots a_n$ and arbitrary $v = b_1 b_2 \cdots b_m \cdots$ in A^*, the concatenation $uv \in A^*$ is

$$a_1 a_2 \cdots a_n b_1 b_2 \cdots b_m \cdots .$$

Note that the empty word ϵ is the unique word in A^* such that for any (finite) word $u \in A^*$, $u\epsilon = \epsilon u = u$. Note that though the result is symmetric in this case, concatenation is not so in general. However, concatenation *is* associative, a fact which will be of central importance in the proofs to come. Concatenation's property of associativity and the existence of ϵ classify A^*, in algebraic terms, as a monoid. Specifically, the set A^* of an alphabet A is called the *free monoid*.

We define a word $v \in A^*$ to be a *factor* of another word $u \in A^*$ if there exist $x, y \in A^*$ such that $u = xvy$. If $xy = \epsilon$, $u = v$. If $x = \epsilon$, v is a *prefix* of u, and if $y = \epsilon$, v is a *suffix* of u. A *subword* of a word u is a concatenation of distinct factors of u, in the order they appear in u.

We define the concatenation of a word with itself $uu = u^2$ to be a *square*, and a word of the form $uuu = u^3$ to be a *cube*. In general, words of the form $uu \cdots u = u^n$ are the nth power of u. A word is considered to be *square-free* if it contains no factors which are squares—equivalently, a word u is *square-free* if there are no words x, v, y such that $u = xvvy$ and $v \neq \epsilon$. The definition of *cube-free* is analogous.

A *morphism* $f : A^* \to B^*$ is a function such that

$$f(uv) = f(u)f(v), \quad u, v \in A^*$$

and (as can be easily verified) $f(\epsilon) = \epsilon$. An *isomorphism* is a morphism f with an inverse f^{-1}.

[1] Finititude is not technically necessary, but infinite alphabets are very much out of the scope of this paper.

Codes

A submonoid N of a monoid M is a subset N of M such that the binary operation of M is closed under N, and has the same neutral element ϵ. Given any subset X of our free monoid A^*, we can *generate* a submonoid of $X^* \subset A^*$ from the elements of X, through repeated concatenation of the elements of X. Conversely, for any submonoid Σ of A^* there is a unique set S which generates Σ and is minimal for set-inclusion—that is, no other set which generates Σ is a subset of S. In fact, S is computable from Σ, being

$$S = (\Sigma \setminus \epsilon) \setminus (\Sigma \setminus \epsilon)^2;$$

that is, the set of all nonempty words of Σ which cannot be written as the product of two nonempty words of Σ (the notation $(\Sigma \setminus \epsilon)^2$ referring to the product of any two nonempty words contained in Σ). This set S is called the *minimal generating set* of Σ.

A monoid M is said to be *free* if there is an alphabet B and an isomorphism from the free monoid B^* to M. The minimal generating set of a free monoid is called a *code*, or the *basis* of M.

A set X is called a *prefix* if no element of X is a prefix of any other. The definition for a *suffix* set is analogous. Any prefix or suffix set is a code.

3 Proof of Infinite Cube-Free Word

This section of the paper draws directly from the translation of Thue's paper. We begin by proving lemmas similar to our desired result.

Lemma 1. *Over a three-letter alphabet $\{a, b, c\}$, there exist arbitrarily long square-free words without factors aca or bcb.*

Proof. We begin with an arbitrary square-free word u over $\{a, b, c\}$, and then this construction is carried out in several steps:

First, we replace each occurrence in u of c preceded by a by the word $\beta\alpha$, and each occurrence of c preceded by b by $\alpha\beta$. The resulting word we call u'. We can show u' to be square-free by contraposition—first, note that by erasing each α and replacing β by c, we get u from u'. If u' has a square, then by this method, u has a square; thus u' is square-free.

Secondly, we insert γ after every letter of u' to get u''. Clearly, u'' is square-free.

Lastly, we replace any a in u'' with $\alpha\beta\alpha$ and any b by $\beta\alpha\beta$ to get w. This is our final word.

It remains to show w is square-free and does not contain $\alpha\gamma\alpha$ or $\beta\gamma\beta$. For the first fact, note that in u', all a or α alternate with all b or β. Thus, in u'', the three-letter factors with γ as the second letter are $a\gamma b$, $a\gamma\beta$, $\alpha\gamma b$, or $\alpha\gamma\beta$. Thus, in w the only such factors are $\alpha\gamma\beta$ and $\beta\gamma\alpha$.

Now we show w is square-free by contradiction. Suppose there is a square ss in w. Since the only factors between any two γ are α, β, $\alpha\beta\alpha$, and $\beta\alpha\beta$, there must in be at least one γ in ss and thus s. If there is only one γ and is not the last letter in s, assume without loss of generality that α follows γ (the argument for β is the same.) Then ss contains $\gamma\alpha\gamma\alpha$ or $\gamma\alpha\beta\alpha\gamma\alpha$, where in both cases w has a factor $\alpha\gamma\alpha$, contradiction.

Thus s must contain at least two γ. Then, for $X = \{\alpha, \beta, \alpha\beta\alpha, \beta\alpha\beta\}$, we express s in the form

$$s = p\gamma x_1\gamma\cdots\gamma x_m\gamma q, \quad x_j \in X, qp \in X, m \geq 1.$$

If $q = \epsilon$, then $p \in X$, and if we undo step three of our construction, we get a square in u''; the same holds for $p = \epsilon$. Thus $q, p \neq \epsilon$, and $qp = \alpha\beta\alpha$ or $\beta\alpha\beta$. Let us assume the first case (again, we have the same argument for the alternative). Then $(q, p) = (\alpha, \beta\alpha)$, or $(q, p) = (\alpha\beta, \alpha)$. These are symmetric—the following argument is generally the same for the end of u in the second case as it is for the beginning of u in the first case. Take the first case again without loss of generality. By construction, w cannot start with $\beta\alpha\gamma$, so we must have at least an α preceding ss in w, which means that w contains $qp\gamma x_1\gamma\cdots\gamma x_m\gamma$ as a factor, but this as before implies u'' has a square. Thus we have a contradiction, so in the end w must be square-free. □

Lemma 2. *There exists a sequence $(w_n)_{n\geq 0}$ of square-free words over three letters such that w_n is a prefix of w_{n+1}. Stated otherwise, there is an infinitely long square-free word over three letters.*

Proof. We use another construction as for Lemma 1, which is much more concise but much less illuminating. For any $u \in \{a, b, c\}^*$ with no factors aca or bcb, we create a new word $\sigma(u)$ by the function

$$\sigma : \begin{array}{ll} a \to abac & \\ b \to babc & \\ c \to bcac & \text{if } c \text{ is preceded by } a \\ c \to acbc & \text{if } c \text{ is preceded by } b. \end{array}$$

This produces the same result as the previous construction: however, with this, we can more easily apply σ an arbitrary number of times to u to get a square-free word on $\{a, b, c\}^*$ of arbitrarily great length, and therefore an infinitely long word.

□

This, along with certain algebraic features of words, leads directly to our result.

Theorem 3. *There exists an infinite cube-free word over two letters.*

Proof. Our proof begins with a square-free infinite word $u \in \{x, y, z\}^*$. We create a morphism

$$f : \begin{array}{l} x \to a \\ y \to ab \\ z \to abb, \end{array}$$

and assert that $f(u) \in \{a, b\}^*$ is cube-free.

Consider $X = \{a, ab, abb\}$. This is a suffix code, giving us the following chain of results.

1. If u and v are words on $\{x, y\}$ such that $f(u) = f(v)$, then $u = v$.

2. The morphism f is injective.

   Proof. This holds because X is a code. □

   Now let $\mathbf{x}$ be an infinite square-free word over the letters x, y, and z, and let $\mathbf{y} = f(\mathbf{x})$.

3. If $\mathbf{y}$ contains the factor uuu (that is, a cube,) then u does not start with the letter a.

   Proof. If u starts with the letter a, then there is a (unique) factor v of $\mathbf{x}$ such that $f(v) = u$. But then $\mathbf{x}$ contains the square vv, contradiction. □

4. If $\mathbf{y}$ contains a factor uuu, then u does not begin with bb.

Proof. If u begins with bb, then any occurence of u is preceded by an a, and therefore also u ends with an a. Thus, setting $u = u'a$, the word $\mathbf{y}$ has a factor $au'au'au'$, contrary to the preceding lemma. □

5. If y contains the factor uuu, then u does not end with b.

   Proof. By the preceding lemmas, u must begin with ba, and suppose u ends with b, such that $u = bau'b$. Then $\mathbf{y}$ contains a factor $bau'bbau'bbau'b$, which means $\mathbf{y}$ contains the factor $au'bbau'bb$, meaning (by the fact f is injective) $\mathbf{x}$ contains a square, contradiction. □

This is sufficient to prove the theorem. If $\mathbf{y}$ includes uuu, then u must start with ba and end with a. If $u = ba$, then $\mathbf{y}$ contains $bababa$, which means $\mathbf{x}$ contains the square yy. Otherwise, if $u = bau'a$ for some u', then $\mathbf{y}$ contains $bau'abau'abau'a$ and thus contains $abau'abau'$, meaning $\mathbf{x}$ has a square, contradiction.

Thus $\mathbf{y}$ contains no factors of the form uuu, and is therefore square-free, which is what we wanted.

□

4 Further Applications

In a direct sense, the proof of infinite square-free (and/or cube-free in cases) words has applications in the Burnside problem in abstract algebra (Is every finitely generated torsion semigroup finite?)[2]

Further, the square-free word constructed in Lemma 2 as well as the cube-free word constructed in the final theorem were both independently demonstrated and proven by Morse in later decades; the latter word is known now as the Thue-Morse sequence and has had applications in several fields of number theory and combinatorics even before Thue first codified the sequence. Investigations of square-free words and generalizations thereof (often kth-power-free words, or theorems on the distance between repeated factors of infinite words) have been further developed throughout the 20th century.

References

[1] J. Berstel, *Axel Thue's papers on repetitions in words: a translation.* Laboratoire de combinatoire et d'informatique mathématique, Montreal, Qc., 1995.

[2] M. Lothaire, *Combinatorics on Words.* Encyclopedia of Mathematics and Its Applications. Addison-Wesley, Reading, Massachusetts, 1983

[3] A. Thue, Über unendliche Zeichenreihen, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl.,* Christianna 1906, Nr. 7.

[4] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeighenreihen, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl.,* Christiana 1912, Nr. 10.