Some Variations on a Theme of Irina Mel'nichuk Concerning the Avoidability of Patterns in Strings of Symbols

George F. McNulty*

Department of Mathematics
University of South Carolina
Columbia, SC 29208, U.S.A.

mcnulty@math.sc.edu

Submitted: Jun 5, 2017; Accepted: Apr 10, 2018; Published: May 11, 2018

Abstract

The set of all doubled patterns on n or fewer letters can be avoided on an alphabet with k letters, where k is the least even integer strictly greater than $n+1$, with the exception of $n = 4$. The set of all doubled patterns on 4 or fewer letters can be avoided on the 8-letter alphabet. The set of all avoidable patterns on n or fewer letters can be avoided on an alphabet with $2(n+2)$ letters.

Mathematics Subject Classifications: 68R15

1 Introduction

By a *word* we understand here a finite sequence of elements, usually referred to as *letters*, drawn from some set A, usually referred to as an *alphabet*. We are interested in the combinatorial properties of words. Accounts, fairly recent, of the state of the combinatorial theory of words can be found in the monographs of Allouche and Shallit (2003) [1], of Lothaire (2005) [16], of Lothaire (2002) [15], and of Lothaire (1997) [14].

In this paper we will only deal with words of positive length. For a given alphabet A, we use A^+ to denote the set of all words on A that have positive length. Words can be concatenated to form other words. The operation of concatenation is associative, and when A^+ is endowed with this operation it becomes the semigroup freely generated by A. This entails that every map that assigns to each letter in A a word from B^+ can be extended to a unique morphism from A^+ to B^+. Let u and w be words. We say that u

*Supported by NSF Grant 1500216

is a *subword* of w provided either u is an initial segment of w or u is a final segment of w or there are words x and y so that $w = xuy$.

Let w be a word on the alphabet B and v be a word on the alphabet A. We way that v *encounters* w provided there is a morphism $h : B^+ \to A^+$ so that $h(w)$ is a subword of v. We think of w as a pattern or template and $h(w)$ has an instance of the pattern or template. So for v to encounter w means that an instance of the pattern w can be found among the subwords of v—that is within v itself. If v does not encounter w we say that v *avoids* w. We say that w is *avoidable on the k-letter alphabet* provided there are infinitely many words in A^+ that avoid w, where A is an alphabet with k letters. Similarly, we say that a set Σ of words is avoidable on the alphabet with k letters provided there are infinitely many words on the k-letter alphabet, each of which avoids every word belonging to Σ.

The notion of avoidable words was introduced independently by Bean, Ehrenfeucht, and McNulty [3] in 1979 and by Zimin [26] in 1982, and in these papers the notion of avoidability was given algorithmic characterizations. In §3 below, one of these characterizations will be more fully examined. The notion of avoidable words was inspired by work carried out by Axel Thue (see [24, 25]) in the early years of the 20^{th} century. Thue proved that the pattern xx is avoidable on the 3-letter alphabet and that xxx is avoidable on the 2-letter alphabet.

The word xx is the simplest example of a doubled word. In general, we say a word w is *doubled* provided every letter that occurs in w occurs at least twice. Suppose that w is doubled. We say that w is the *mesh* of w provided k is the smallest natural number such that whenever x is a letter and u is a word of length greater than k in which x does not occur, then xux is not a subword of w. In Bean, Ehrenfeucht, and McNulty [3] it was proved that

> For any positive natural number k, the set of all doubled words of mesh k on a countably infinite alphabet is avoidable on the alphabet with $8k+16$ letters.

This was the earliest global avoidability theorem for doubled words. It is easy to see by an inductive argument, that every word on an n-letter alphabet with length at least 2^n must have a subword that is doubled. Because the length of a word is an upper bound on its mesh, we also have

> The set of all doubled words on an alphabet with no more than n letters is avoidable on an alphabet with $8 \cdot 2^n + 16$ letters.

The method used by Zimin [26] is more efficient than that of Bean, Erhrenfeucht, and McNulty. The following is implicit in Zimin's work:

> The set of all doubled words on an alphabet with no more than n letters is avoidable on an alphabet with $6 \cdot 2^n + 14$ letters.

As an immediate corollary of the 1989 work of Baker, McNulty, and Taylor [2] we have

> The set of all doubled words on an alphabet with no more than n letters is avoidable on an alphabet with $9 \cdot n + 20$ letters,

if only because this bound was proven to hold for avoidable words generally.

In 1985 Irina Mel'nichuk [17] outlined a proof of the following theorem, which gives bounds sharper than any of those above.

Mel'nichuk's Global Avoidabilitiy Theorem for Doubled Patterns on n Letters. *Let n be a positive natural number. The set of all doubled patterns on the n-letter alphabet is avoidable on the alphabet with $3\lceil\frac{n+1}{2}\rceil$ letters.*

In 2015 Michael Lane [13] provides an exposition, in English, of Mel'nichuk's proof that fills in the all details not available in Mel'nichuk [17].

The first purpose of this paper is to establish a theorem that improves Mel'nichuk's theorem. Our second purpose is to offer a minor improvement to

Mel'nichuk's Global Avoidability Theorem for Avoidable Words on n Letters. *Let n be a positive natural number. The set of all avoidable patterns on the n-letter alphabet is avoidable on an alphabet with $4\left\lceil\frac{n+2}{2}\right\rceil$ letters.*[1]

Before taking up these tasks, it is interesting to observe that in contrast to the global or simultaneous avoidability results mentioned above, one can ask, more locally, of an individual word w, for the smallest k so that w is avoidable on the alphabet with k letters. Denote this smallest value by $\mu(w)$ and call it the *avoidability index* of w. In case the word w is doubled, the situation has been substantially clarified. In 1984 A. G. Dalalyan [12] proved that every doubled word is 4-avoidable and that any doubled word in which at least 6 distinct letters appear is 3-avoidable. Dalalyan's results were rediscovered by Bell and Goh [4] and the results of Bell and Goh were enhanced by Blanchet-Sadri and Woodhouse [5] in 2013. A substantial advance was obtained in 2015 by Michael Lane [13]. He proved

> Every doubled word on n or fewer letters of length at least $\min(2n+1, 12)$ is avoidable on the 3-letter alphabet.

All the doubled words on these small alphabets are seen to be 3-avoidable (and even 2-avoidable in some cases). So Lane's result left as unsettled only certain doubled words of length 8 on 4 letters and certain doubled words on 5 letters that have length 10—this left a list of roughly 100 doubled patterns to check. In 2016 Ochem [21], working independently of Lane, but using largely similar methods, was finally able to prove that each doubled word is 3-avoidable. It appears that many doubled words are actually 2-avoidable. The problem of characterizing the doubled words that are 2-avoidable remains open.

For words w in general, not just doubled words, the situation is much less clear. Indeed, among the most vexing problems, first raised the the mid-1970's, are

Problem 0. Is the function μ on words over a countably infinite alphabet that returns the avoidability index of w (or ∞ when w is unavoidable) a computable function? If it is, what is its computational complexity?

[1] The bound stated by Lothaire is $4\lceil(n+1)/2\rceil$, but there appears to be a slight flaw in Lemma 3.2.7 there—the small adjustment of adding two letters to the alphabet remedies the matter.

Problem 1. Does the function μ have a finite upper bound?

The only progress on these problems has been in the investigation of the avoidability index on small alphabets. The avoidability of words on alphabets of size no more than 2 has been systematically investigated by Schmidt [23], Roth [22], and Cassaigne [8, 9] and is now completely understood. For alphabets of size 3 the task was begun by Cassaigne in his 1994 dissertation [9]. Clark in his 2001 dissertation [11] proved that the avoidability index of each avoidable word on a three letter alphabet is no more than 4, and, in 2006, Ochem [20] completed the work started by Cassaigne by verifying the avoidability index as 2 for the avoidable words not settled by Cassaigne. So the avoidability index of any avoidable word on the three letter alphabets is either 2 or 3. Finally, Clark [10, 11] has devised a word with avoidability index 5. No avoidable word of larger avoidability index is known.

2 Global avoidability of doubled words

The Global Avoidablitiy Theorem for Doubled Patterns on n Letters. *Let n be a positive natural number. The set of all doubled words on n or fewer letters is avoidable on*

- *the alphabet with 8 letters, if $n = 4$;*
- *the alphabet with $n + 2$ letters, if n is even and $n \neq 4$;*
- *the alphabet with $n + 3$ letters, if n is odd.*

Proof. The case when $n = 4$ is exceptional. We handle it at the end of the proof.

So for now, we stipulate that $n \neq 4$.

First, consider the case $n = 1$. The set of all doubled words on this alphabet is $\{x^2, x^3, x^4, \dots\}$, where x is the sole letter. Axel Thue showed that this set is avoidable on the alphabet with 3 letters. This settles the case since $3 \leqslant 4 = 1 + 3$.

So we take up the case when $n > 1$. Let k be the least even natural number so that

$$k - 1 > n.$$

So $k = n + 2$ if n is even and $k = n + 3$ if n is odd.

The plan of our proof is to start with an alphabet A with k letters. We take

$$A = \{a_0, a_1, \dots, a_{k-1}\}.$$

We will construct infinitely many words on A, each of them avoiding every doubled pattern on the n-letter alphabet. We will do this by a method introduced by Axel Thue [25], which has now become standard. Namely, we will define a map $\Psi\colon A^+ \to A^+$ so that the sequence $\Psi(a_0), \Psi^2(a_0), \Psi^3(a_0), \dots$ is an infinite list of words of increasing length such that each word on this list avoids every doubled word on the n-letter alphabet. The map Ψ will be easy to describe, but before doing this it helps to look ahead to see what this

map needs to be like. So for the moment suppose that Ψ is in hand. Now consider a doubled word w that, contrary to our hopes, is encountered by $\Psi^{\ell+1}(a_0)$. This means that there will be a morphism h such that $h(w)$ is a subword of $\Psi^{\ell+1}(a_0)$.

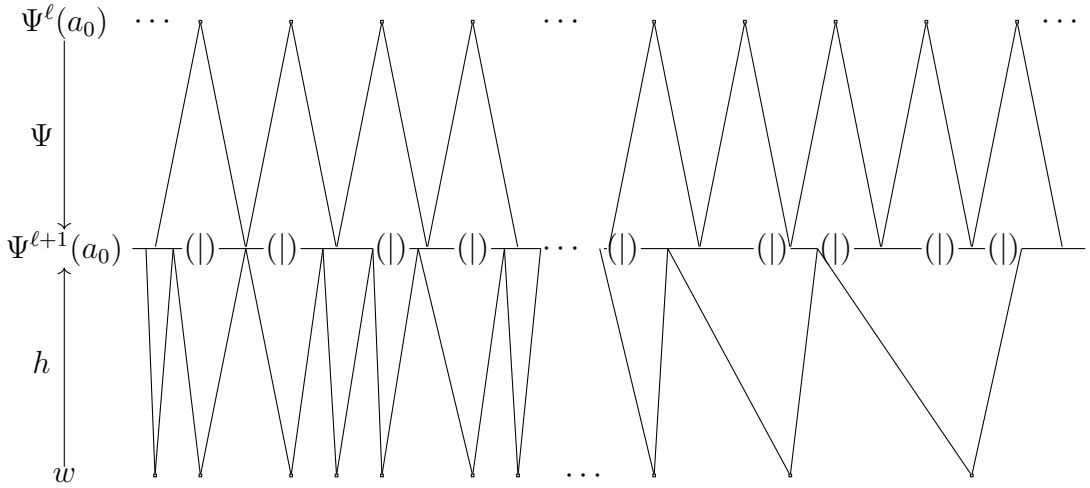

Figure 1: A Visualization of the Proof

The situation at hand is illustrated in Figure 1. In this figure, points at the top are (some of) the letters of $\Psi^{\ell}(a_0)$ arranged as they are in $\Psi^{\ell}(a_0)$. The points at the bottom are the letters of w; they comprise a copy of w. Of course Ψ takes each letter at the top to a word (indicated here as a line segment) in the middle of the diagram. Together they constitute a subword of $\Psi^{\ell}(a_0)$. Likewise, h takes each letter at the bottom to a word in the middle. The diagram is drawn in a way to suggest that all the Ψ images of letters have the same length—this will indeed be an attribute of Ψ when we finally define it. The (|)'s that appear once in the Ψ-image of each letter from A indicate specially chosen subwords of length 2 that we will call *Mel'nichuk decisive representatives* of the letter they come from by way of Ψ. We will rig matters so that these decisive representative never straddle a border between h-images of adjacent letters of w. As seen in the diagram, some letters of w have images that may engulf at least one Mel'nichuk decisive representative, while others do not. Now obtain w' from w by deleting all the letters whose images under h engulf no decisive representative. The diagram suggests how to construct a morphism h' so that $h'(w')$ is a subword of $\Psi^{\ell}(a_0)$. Then an appeal to induction will finish the proof.

Rather than an appeal to induction, it is convenient to prove the theorem indirectly. So, in pursuit of a contradiction, we assume the theorem fails. We pick a doubled word w, as short as possible, on no more the n letters, that is encountered by $\Psi^{t}(a_0)$ for some natural number t. For this w, we pick ℓ as small as possible so that $\Psi^{\ell+1}$ encounters w. Finally, we pick a morphism h so that $h(w)$ is a subword of $\Psi^{\ell+1}(a_0)$.

The essential requirement placed on Ψ by the idea sketched above is that the image of each letter x in A should have a large number of distinct subwords of length 2 that will permit us to identity a decisive representative of x. Since at most n letters occur in w, there will only be at most n possibilities for the first letter of $h(\xi)$, as ξ runs through

the letters occurring in w. So we must have at more than n suitable choices for the decisive representatives. The main difficulty, is that the k decisive representatives must be pairwise distinct.

We can regard words on the alphabet A as right-directed paths where the vertices have been labeled with letters for A. What we require is a graceful labeling. In 1982 Bloom and Hsu [6] introduced the notion of graceful labelings of directed graphs. In 1985 Bloom and Hsu [7] observed that one-way directed paths of even length have a graceful labelings. A single example suffices to see the general case. Consider the graceful numberings of the directed path with 10 vertices displayed in Figure 2.

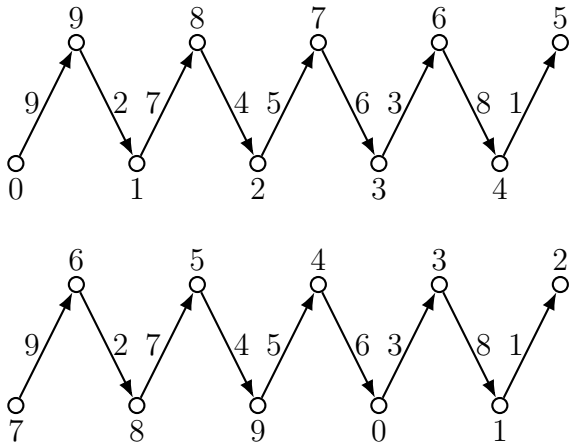

Figure 2: Two Graceful Labellings of Left-directed Path on 10 Vertices

Observe that the edge labels are constructed from the vertex labeling as follows: if $x \to y$, then the label of the edge from x to y is the natural number $r < 10$ so that $y - x \equiv r \pmod{10}$. Another way to say this is that the label is $y - x$ as computed in cyclic group $\mathbb{Z}_{10}$. The numberings shown in Figure 2 are graceful in the sense that all the edge labels are distinct. More is true. The labeling of the path at the bottom can be obtained from the labeling of the path at the top. Actually, we give two ways to see how these two labellings are related. In the first way, to obtain the bottom labeling simple add 7 (in $\mathbb{Z}_{10}$) to the label of each vertex in the top labeling. It is evident that the labels on the edges will not change when this is done. In this way, we can obtain 10 different graceful vertex labellings with the same edge labeling. The second way to see the connection is to realize that the vertex labeling occurs around an oblong. Each of these labellings amount to numbering the vertices counterclockwise around the oblong, starting at some vertex with 0. In the numbering at the top, we started with the leftmost of the lower vertices, while in the numbering at the bottom, we started at the fourth vertex from the left among the lower vertices. With this second viewpoint in mind, it is easy to see that if 0 is on the lower level, but not the leftmost vertex, then 9 will be somewhere on the directed path preceding 0. On the other hand, if 0 is on the upper level, then 1 will precede 0 on the directed path.

The remarks above hold in general. A graceful numbering of a directed path with k

vertices, where k is even, can be made by labeling the leftmost vertex with 0 and running counterclockwise around the oblong. The last vertex on the path will get the label $k/2$. Using this graceful labeling, we can devise others as described above. There will be k of them and they will all generate the same edge labels. Moreover, for any one of these graceful labelings, if 0 is on the lower level (but not the left end of the path) then $k-1$ precedes 0 on the path, whereas, if 0 is on the upper level, then 1 precedes 0 on the path.

In the above constructions, if we replace the label i by a_i, for each $i < k$, we get k graceful words from the k gracefully labeled left-directed paths. In case $k = 10$, here is what they look like:

$$a_0a_9a_1a_8a_2a_7a_3a_6a_4a_5$$
$$a_1a_0a_2a_9a_3a_8a_4a_7a_5a_6$$
$$a_2a_1a_3a_0a_4a_9a_5a_8a_6a_7$$
$$a_3a_2a_4a_1a_5a_0a_6a_9a_7a_8$$
$$a_4a_3a_5a_2a_6a_1a_7a_0a_8a_9$$
$$a_5a_4a_6a_3a_7a_2a_8a_1a_9a_0$$
$$a_6a_5a_7a_3a_8a_3a_9a_2a_0a_1$$
$$a_7a_6a_8a_3a_9a_4a_0a_3a_1a_2$$
$$a_8a_7a_9a_4a_0a_5a_1a_4a_2a_3$$
$$a_9a_8a_0a_5a_1a_6a_2a_5a_3a_4$$

Observe that the word on each row begins with one of our k letters. We denote the word on the i^{th}-row by $\vec{a}_i$. It is important to observe that the word a_pa_q, where $p \neq q$, is a subword of exactly one $\vec{a}_i$. Indeed, $q - p$ (calculated in $\mathbb{Z}_k$) is the edge-label associated with two adjacent columns in our array of words. In the left of these adjacent columns a_p occurs exactly once. What this will mean is that each of the $k-1$ the subwords of $\vec{a}_i$ that have length 2 are available as potential decisive representatives of a_i.

We define our map Ψ so that

$$\Psi(a_i) = \vec{a}_i a_i, \qquad \text{for each } i < k.$$

So $\Psi(a_i)$ is a word of length $k+1$ that begins and ends with the letter a_i and in which every letter a_j with $j \neq i$ occurs exactly once. Also, there are $k-1$ subwords of length 2 that are available to be chosen as decisive representatives of a_i. Since $k - 1 > n$, chose one of these available words of length 2 so that its right letter is not the leftmost letter in $h(\xi)$, for any letter ξ occurring in w.

At this point, our proof would be essentially complete, except for the possibility that w' is empty. To prevent this, we have show that there is some letter ξ of w so that some decisive representative is a subword of $h(\xi)$. But since decisive representative cannot straddle the h-image of adjacent letters in w, it will be enough to prove the next lemma.

Lemma A.
Some decisive representative is a subword of $h(w)$.

Proof. It follows easily by induction on t that no two adjacent letters in $\Psi^t(a_0)$ are identical.

Since $h(w)$ is a subword of $\Psi^{\ell+1}(a_0)$ we have three alternatives to consider.

Alternative: $\Psi(a_i)$ is a subword of $h(w)$, for some $i < k$.
In this case the decisive representative of a_i will be a subword of $h(w)$.

Alternative: $h(w)$ is a subword of $\Psi(a_i)$, for some $i < k$.
We reject this alternative, since $\Psi(a_i)$ has no doubled subwords, but $h(w)$ must be doubled since w is doubled.

Alternative: $h(w)$ is a subword of $\Psi(a_ia_j)$, for some $i, j < k$ with $i \neq j$.
In view of the second alternative, here we know that $h(w)$ straddles the boundary between $\Psi(a_i)$ and $\Psi(a_j)$. It is harmless to suppose that $i = 0$. So $\Psi(a_0a_j)$ is

$$a_0a_{k-1}a_1\cdots a_{\frac{k}{2}}a_0a_ja_{j-1}a_{j+1}\cdots a_{\frac{k}{2}+j}a_j.$$

To keep out of the first Alternative, we need only consider that $h(w)$ is a subword of the following word:

$$a_{k-1}a_1\cdots a_{\frac{k}{2}}a_0\overbrace{a_ja_{j-1}a_{j+1}\cdots a_{\frac{k}{2}+j}}^{\vec{a}_j}.$$

In this word every letter occurs exactly 2 times. Since w is doubled, $h(w)$ is doubled. So any letter that occurs in $h(w)$ occurs at least twice. Since $h(w)$ straddles the middle of the displayed word, we know that both a_0 and a_j occur in $h(w)$. The other occurrence of a_0 is in $\vec{a}_j$. If a_{k-1} precedes a_0 in $\vec{a}_j$, then a_{k-1} must occur (twice) in $h(w)$. But it follows that every letter that occurs in w, must occur exactly twice in w. This will force all k of the letters to occur in $h(w)$. But then $\vec{a}_j$ will be a subword of $h(w)$. Hence the decisive representative of a_j will be a subword of $h(w)$.

So it remains to consider the case when a_{k-1} does not occur in $h(w)$. In this case, we know that a_1 must precede a_0 in $\vec{a}_j$. This forces all the letters except a_{k-1} to occur in $h(w)$. So we must have the following situation:

$$a_{k-1}\underbrace{a_1\cdots a_{\frac{k}{2}}a_0a_ja_{j-1}a_{j+1}\cdots}_{h(w)}a_{\frac{k}{2}+j}.$$

This forces $\frac{k}{2} + j \equiv k - 1 \pmod{k}$. That is $j = \frac{k}{2} - 1$. So we are reduced to the situation when

$$h(w) = a_1\cdots a_{\frac{k}{2}}a_0a_{\frac{k}{2}-1}a_{\frac{k}{2}-2}a_{\frac{k}{2}}a_{\frac{k}{2}-3}\cdots a_{k-2}.$$

We see that the length of $h(w)$ is $2(k-1)$.

Consider any letter a_i with i different from each of $0, a_j$, and $a_{\frac{k}{2}}$. Pick a letter ξ of w so that a_i occurs in $h(\xi)$. Now a_i occurs once in $\vec{a}_0$ and once in $\vec{a}_j$. Because $\vec{a}_0$ and $\vec{a}_j$ have no subwords of length 2 in common, it must be that $h(\xi)$ has length 1. The same also applies when $i = j$. So consider a_0. Pick the letter ξ of w so that a_0 occurs in $h(\xi)$. Now the word a_0a_j can occur only once in $h(w)$. However, there is one situation when the

word $a_{\frac{k}{2}}a_0$ can occur twice. This happens when $\frac{k}{2} \equiv 3 \pmod{k}$. This means that $k = 6$. In turn, this means that n is either 3 or 4. But recall that $n \neq 4$. So if n is 3, it might be that $h(\xi)$ has length 2, where ξ is the letter so that a_0 occurs in $h(\xi)$. In this event, the length of $h(\zeta)$ is 1 for each letter $\zeta \neq \xi$ of w. It follows that the length of $h(w)$ is no greater than $2(3+1)$. So we have $2(k-1) \leqslant 2(3+1)$. This means that $k \leqslant 5$. But in this case, $k = 3+3 = 6$. So we must reject this case. It follows that $h(\xi)$ has length 1 for each letter ξ of w. But then the length of $h(w)$ cannot be greater than $2n$. So we would have $2(k-1) \leqslant 2n$. This is impossible since we have chosen k so that $k - 1 > n$.

In this way, the lemma is established. □

With the Lemma A in hand, we see that not all the letters in w are deletable. So w' is not empty. By the minimality in the choice of w we see that the length of w and the length of w' must be the same. This means that no letters have been deleted. But then $w = w'$ and we see that $\Psi^\ell(a_0)$ encounters w (with the help of h'). This violates the minimality in the choice of ℓ. At last, this is the contradiction needed to complete our indirect proof of the theorem, when $n \neq 4$.

For the exceptional case $n = 4$, observe that each doubled word on no more than 4 letters is also a word on no more than 5 letters. Now $n = 5$ is an unexceptional instance of the theorem and $5 + 3 = 8$. So the set of all doubled words on no more than 4 letters is avoidable on an alphabet with 8 letters. □

There is only one case where Mel'nichuk's bound is sharper than the bound proved here: $n = 1$. The bound given in the present theorem is 4, whereas Mel'nichuk's bound is 3, which is actually the bound established by Axel Thue [25]. Table 1 provides a comparison of the bounds.

The present bound	Mel'nichuk's bound: $3\left\lceil\frac{n+1}{2}\right\rceil$	Parity of n
$n+2$	$\frac{3}{2}(n+2)$	n is even and $n \neq 4$
$n+3$	$\frac{3}{2}(n+1)$	n is odd
$8^\dagger$	9	$n = 4$

†The referee has sketched an argument that, when $n = 4$, the true value is 5.

As Mel'nichuk observed, the bound must be at least $n + 1$. When $n = 1$, the bound must be at least $n + 2 = 3$ and this bound can be achieved, as Axel Thue showed in 1906. It is conceivable that $n + 2$ is a sharp bound, but no proof is known that this bound can be acheived when n is odd; when $n = 4$ the referee contends that the bound is $4 + 1 = 5$. Neither is it known that $n + 1$ will not suffice in every case except $n = 1$.

3 Global Avoidability of Avoidable Words

After her 1985 work on avoiding doubled patterns, Irina Mel'nichuk turned to avoiding patterns generally. It appears she has never put her methods in this direction into

the literature. However, in 1991 Mikhail Volkov gave a presentation of her methods at Marquette University and Pavel Goralčik brought her methods to Paris. An account of Mel'nichuk's methods can be found in Chapter 3 of Lothaire [15], where the following theorem is credited to Irina Mel'nichuk.

Mel'nichuk's Global Avoidability Theorem for Avoidable Words on n Letters. *Let n be a positive natural number. The set of all avoidable patterns on the n-letter alphabet is avoidable on an alphabet with $4\left\lceil\frac{n+2}{2}\right\rceil$ letters.*

The bound $4\lceil\frac{n+2}{2}\rceil$ is equal to $2(n+2)$ when n is even and to $2(n+3)$ when n is odd. Here we can only make a small improvement: Mel'nichuk's bound in the even case works for the odd case as well. She submitted an abstract to the Colloquium on Universal Algebra held in Szeged in August 1989 in which she asserted that the bound $n+6$ would serve, but included no proof.

The Global Avoidability Theorem for Avoidable Words on n Letters. *Let n be a positive natural number. The set of all avoidable patterns on the n-letter alphabet is avoidable of an alphabet with $2(n+2)$ letters.*

We need the notion of reducibility in the proof of this theorem. We associate with each pattern w a bipartite graph, called the *adjacency graph* of w, as follows. The two parts of this graph are called the left alphabet and the right alphabet. In the adjacency graph their is an edge joining the letter x in the left alphabet with y is the right alphabet provided the length 2 word xy is a subword of w. For example, for the word $a_0b_0a_1b_1a_2b_2a_3b_3a_4b_4$ has the adjacency graph displayed in Figure 3.

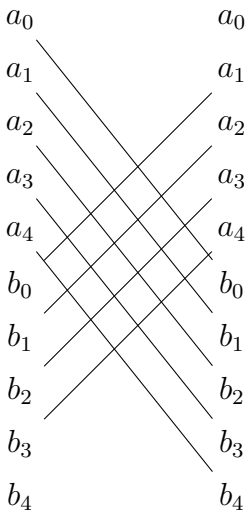

Figure 3: The Adjacency Graph of $a_0b_0a_1b_1a_2b_2a_3b_3a_4b_4$

This word has a simple adjacency graph, mostly due to the fact the each letter appears in the order only once. A subset F of the alphabet is *free for w* provided whenever ξ and ζ are letters occurring in w there is no path in the adjacency graph connecting ξ in the

left alphabet with ζ in the right alphabet. The set $\{a_0, a_1, a_2, a_3, a_4\}$ is free for the word used in Figure 3.

We say that a word w *reduces in one step* to a word u if and only if u can be obtained from w be deleting all occurrence of the letters belonging to some set free from w. We way that w *reduces* to u if and only if u can be reached from w by a series of finitely many one-step reductions.

The following theorem was proved in 1979 by Bean, Ehrenfeucht, and McNulty [3] and independently in 1982 by Zimin [26].

The Characterization Theorem for Unavoidable Patterns. *A word w is unavoidable if and only if w is reducible to the empty pattern.*

Zimin [26] provided an additional equivalent condition.

The proof of the Global Avoidability Theorem for Words on n Letters amounts to a proof that unavoidable words are reducible—one direction of the proof of the Characterization Theorem.

Proof of the Global Avoidability Theorem for Words on n Letters. This proof follows the plan used in the proof of the Global Avoidability Theorem for Doubled Patterns.

Let us take $k = n + 2$. Let $A = \{a_0, \ldots, a_{k-1}\}$ and let $B = \{b_0, \ldots, b_{k-1}\}$. Our alphabet will be $A \cup B$. This time, we give define the morphism Ψ as follows.

Case: k is odd.

$$\begin{aligned}
\Psi(a_0) &= a_0 b_0 a_1 \ldots b_{(k-1)/2} a_{(k+1)/2} & \Psi(b_0) &= b_{(k+1)/2} a_{(k+1)/2} \ldots a_{k-1} b_{k-1} \\
\Psi(a_1) &= a_1 b_0 a_2 \ldots b_{(k-1)/2} a_{(k+1)/2+1} & \Psi(b_1) &= b_{(k+1)/2} a_{(k+1)/2+1} \ldots\ a_0\ b_{k-1} \\
\Psi(a_2) &= a_2 b_0 a_3 \ldots b_{(k-1)/2} a_{(k+1)/2+2} & \Psi(b_2) &= b_{(k+1)/2} a_{(k+1)/2+2} \ldots\ a_1\ b_{k-1} \\
&\vdots & &\vdots \\
\Psi(a_{k-1}) &= a_{k-1} b_0 a_0 \ldots b_{(k-1)/2} a_{(k+1)/2+k-1} & \Psi(b_{k-1}) &= b_{(k+1)/2} a_{(k+1)/2-1} \ldots a_{k-2} b_{k-1}
\end{aligned}$$

Case: k is even.

$$\begin{aligned}
\Psi(a_0) &= a_0 b_0 \ldots a_{k/2-1} b_{k/2-1} & \Psi(b_0) &= \ a_{k/2}\ b_{k/2} \ldots a_{k-1} b_{k-1} \\
\Psi(a_1) &= a_1 b_0 \ldots\ a_{k/2}\ b_{k/2-1} & \Psi(b_1) &= a_{k/2+1} b_{k/2} \ldots\ a_0\ b_{k-1} \\
\Psi(a_2) &= a_2 b_0 \ldots a_{k/2+1} b_{k/2-1} & \Psi(b_2) &= a_{k/2+2} b_{k/2} \ldots\ a_1\ b_{k-1} \\
&\vdots & &\vdots \\
\Psi(a_{k-1}) &= a_{k-1} b_0 \ldots a_{k/2-2} b_{k/2-1} & \Psi(b_{k-1}) &= a_{k/2-1} b_{k/2} \ldots a_{k-2} b_{k-1}
\end{aligned}$$

In each display, the subscripts on the a's on each successive row can be obtained by adding 1 modulo k to the a in preceding row. One the other hand, each b_i occurs in only one

column, in each case. Every Ψ-image of a letter has length k. In each case, any word of length 2 can occur in at most one Ψ-image of a letter. It will be useful below to note here that reading across the Ψ-image of any letter one observes that the a's and b's alternate. This last property still holds when reading across $\Psi^t(a_0)$, for any natural number t.

In case k is odd, $\Psi(a_i)$ begins and ends with a's with some indices and b_0 is the second letter of the image, whereas $\Psi(b_i)$ begins and ends with a b's with some indices and the beginning is $b_{(k+1)/2}$. In case k is even, $\Psi(a_i)$ always begins with some a and ends with some b and the second letter is always b_0, whereas $\Psi(b_i)$ begins with an a and ends with a b and its second letter is always $b_{k/2}$.

It is convenient to let b_* be $b_{(k+1)/2}$ in the odd case and to let it be $b_{k/2}$ is the even case.

The even case differs in no important way from the method described in Lothaire [15].

We will prove that the set comprised of $\Psi^t(a_0)$ as t runs through the natural numbers avoids every pattern on the alphabet with n letters that is avoidable. This is the same as proving that every pattern on the alphabet with n letters that is encountered by some $\Psi^t(a_0)$ is unavoidable—or, what is the same, is reducible to the empty word.

We prove the theorem indirectly. So, in pursuit of a contradiction, we assume the theorem fails. We pick a word w on no more than n letters that cannot be reduced to the empty word, with w as short as possible, that is encountered by $\Psi^t(a_0)$ for some natural number t. For this w, we pick ℓ as small as possible so that $\Psi^{\ell+1}$ encounters w. (Notice that $\Psi^0(a_0) = a_0$ and a_0 reduces to the empty word.) Finally, we pick a morphism h so that $h(w)$ is a subword of $\Psi^{\ell+1}(a_0)$.

With h in hand, we pick a system of decisive representatives of the letters in $A \cup B$. These decisive representative can be of only two kinds: a_ib_j and b_ia_j. Just has in the earlier proof, we have a pattern w' and a morphism h' so that w' is obtained by deleting all occurrences of some letters from w, and so that $h'(w')$ is a subword of $\Psi^\ell(a_0)$.

In the earlier proof the difficulty was that w' might have been empty, but here the difficulty is that the set of letters deleted from w to obtain w' might not be free for w. So we need the following lemma.

Lemma B. *The pattern w reduces to w'.*

Proof. Zimin [26] proved, as Lemma 8 in his paper, a statement that, in the context at hand, reads

> If there is a pattern v and a morphism f so that
>
> - v reduces to w';
> - $f(w)$ is a subword of v;
> - ξ is a letter of w not in w' if and only if no letters of w' occur in $f(\xi)$,
>
> then w is reducible to w'.

Our task, then, is to provide an appropriate pattern v and an appropriate morphism f.

Consider any letter ξ that occurs in w. To obtain $f(\xi)$ we modify $h(\xi)$. In $h(\xi)$ we might find any number of decisive representatives. If some decisive representative is a subword of $h(\xi)$, consider the leftmost one: xy. To get $f(\xi)$ our first step is to replace xy by $x\xi y$, if x is a b_i, and by $xy\xi$ if y is a b_j. Now say there are r decisive representatives remaining in $h(\xi)$. We introduce new letters ξ_i for each $i < r$ and insert them from left to right within or after each of remaining decisive representatives in $h(\xi)$, just as we did with ξ itself.

Of course, if ξ was a letter deleted from w, then no decisive representative is a subword of $h(\xi)$ and in this case $f(\xi) = h(\xi)$. We take $v = f(w)$. At this point, the last two stipulations in Zimin's Lemma 8 are fulfilled.

It remains to show that v reduces to w'. Observe that in v

- each occurrence on an a_i is followed, if at all, by a b_j,
- each occurrence of a b_i is followed, if at all, by either an a_j or by some ξ, perhaps with a subscript.
- each occurrence of a ξ (even those with subscripts) is followed, if at all, by an a_j.

It follows that in the adjacency graph of v, the only edges from vertices in the set A of the left alphabet have their other vertices in the set B of the right alphabet. Also, the only edges from vertices in the set B of the right alphabet have their other vertices in the set A of the left alphabet. Therefore, the set A is free for v. So delete this free set to obtain the word v'.

Now v' is made up of b's with certain subscripts and the various ξ's that we inserted as well as their subscripted versions. The unsubscripted ξ's are exactly the letters of w' and they occur in v' exactly as they occur in w'. To obtain w', as we desire, all we have to do is delete the b's and the subscripted versions of the ξ's from v'. There are three points that make this possible. First, the b's occur in order, as their indices cycle modulo k. Second, between any occurrence of a ξ (perhaps with subscripts) and the next one (perhaps a different one) to the right, either b_0 or b_* must occur and if one of these occurs then the other does not. Third, the subscripted ξ's occur in the order of their subscripts. This means that in the adjacency graph of v' we have edges of the following kinds: edges from b_i in the left alphabet to b_{i+1} in the right alphabet, where the $+1$ in the subscripts is computed modulo k; we also have some edges between b's in the left alphabet and ξ's in the right, as well as edges from ξ's in the be left alphabet to b's in the right. This means that each singleton $\{b_i\}$ is free for v'. Let us delete b_1, resulting in v''. The adjacency graph of v'' has the properties just mentioned (unless $b_i = b_*$) so again all the singletons of the remaining b's are free. So deleting the singletons one after another we can reduce v' to v^*, where the only b's remaining are b_0 and b_*. In v^* these now alternate with the ξ's. So in the adjacency graph of v^* there are only edges from the b's in the left alphabet to certain ξ's in the right alphabet and edges from certain ξ's in the left alphabet to b's in the right alphabet. This means $\{b_0, b_*\}$ is free from v^*. Deleting this free set results in v°. It remains to delete from v° the subscripted versions of the ξ's. But just as the with

the b_i's above, each singleton is a free set and we can delete these one at a time, since the freeness of the remaining singletons will persist from step to step. .

Putting the reductions together, we see that v reduces to w'. According to Zimin, w must also reduce to w', proving our lemma. □

With the Lemma B in hand, we see that w reduces to w'. By the minimality in the choice of w we see that the length of w and the length of w' must be the same. This means that no letters have been deleted. But then $w = w'$ and we see that $\Psi^\ell(a_0)$ encounters w (with the help of h'). This violates the minimality in the choice of ℓ. At last, this is the contradiction needed to complete our indirect proof of the theorem. □

What was actually accomplished in this proof was that every word w of n or few letters that is encountered by any $\Psi^T(a_0)$ is reducible to the empty word. Every unavoidable w must be encountered in this way. So every unavoidable word reduces to the empty word. So this is also a proof of one direction of the Characterization Theorem.

The bound $2(n + 2)$ given in this theorem might not be the best. That coefficient 2 at the front can be attributed to the way in which the reductions were constructed in the proof of Lemma B. At least the most transparent attempts to use graceful words lead to adjacency graphs that have large portions that are complete bipartite graphs. Such graphs simply do not provide enough free sets to support the needed reductions. But it is conceivable that some other part of graph theory will provide the means to improve this bound.

4 Concluding Comment

Loosely speaking the first proof in this paper was obtained by combining Mel'nichuk's method with the graceful numbering of left-directed paths with an even number of vertices. The second proof amounts to a very small alteration in her method. So the imprint of Irina Mel'nichuk's thinking is heavy in the paper before you.

Note Added in Proof

Since this article was accepted two things happened that deserve mention.

In connection with the Global Avoidability Theorem for Doubled Patterns on n Letters, Pascal Ochem has kindly shared with me his method to handle the exceptional case $n = 4$. The upshot of his method is the best possile bound in this case: $n+1$. His method relies on an application of the Perron-Frobenius Theorem and Moulin Ollagnier's 1992 article [19] that the set of words on the 5-letter alphabet that have repetition index $\frac{5}{4}$ is infinite. This is an adaptation of ideas in Ochem's 2016 article [21].

In connection with the Global Avoidability Theorem for Avoidable Words on n Letters, after several years I was finally able to make contact with Irina Mel'nichuk. She very kindly shared with me a manuscript that she had prepared in 1996 but had never brought to final publication. In that mansucript, Mel'nichuk gave exactly the same theorem. While our

two proofs share something in common, they differ in important ways—indeed, Mel'nichuk actually proves a smaller bound in the case n is even. Her article is now available on the arXiv as [18].

References

[1] Jean-Paul Allouche and Jeffrey Shallit. *Automatic sequences*. Cambridge University Press, Cambridge, 2003. Theory, applications, generalizations.

[2] Kirby A. Baker, George F. McNulty, and Walter Taylor. Growth problems for avoidable words. *Theoret. Comput. Sci.*, 69(3):319–345, 1989.

[3] Dwight R. Bean, Andrzej Ehrenfeucht, and George F. McNulty. Avoidable patterns in strings of symbols. *Pacific J. Math.*, 85(2):261–294, 1979.

[4] Jason P. Bell and Teow Lim Goh. Exponential lower bounds for the number of words of uniform length avoiding a pattern. *Inform. and Comput.*, 205(9):1295–1306, 2007.

[5] Francine Blanchet-Sadri and Brent Woodhouse. Strict bounds for pattern avoidance. In *Developments in language theory*, volume 7907 of *Lecture Notes in Comput. Sci.*, pages 106–117. Springer, Heidelberg, 2013.

[6] Gary S. Bloom and D. Frank Hsu. On graceful digraphs and a problem in network addressing. In *Proceedings of the thirteenth Southeastern conference on combinatorics, graph theory and computing (Boca Raton, Fla., 1982)*, volume 35, pages 91–103, 1982.

[7] Gary S. Bloom and D. Frank Hsu. On graceful directed graphs. *SIAM J. Algebraic Discrete Methods*, 6(3):519–536, 1985.

[8] Julien Cassaigne. Unavoidable binary patterns. *Acta Inform.*, 30(4):385–395, 1993.

[9] Julien Cassaigne. *Motifs évitables et régularité dans les mots*. 1994. Ph.D. Thesis–Univeristé Pierre-et-Marie-Curie - Paris VI.

[10] Ronald J. Clark. The existence of a pattern which is 5-avoidable but 4-unavoidable. *Internat. J. Algebra Comput.*, 16(2):351–367, 2006.

[11] Ronald James Clark. *Avoidable formulas in combinatorics on words*. ProQuest LLC, Ann Arbor, MI, 2001. Thesis (Ph.D.)–University of California, Los Angeles.

[12] A. G. Dalalyan. Word eliminability. *Akad. Nauk Armyan. SSR Dokl.*, 78(4):156–158, 1984.

[13] Michael Lane. *Avoiding Doubled Words in Strings of Symbols*. 2015. Ph.D. Thesis—University of South Carolina.

[14] M. Lothaire. *Combinatorics on words*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1997.

[15] M. Lothaire. *Algebraic combinatorics on words*, volume 90 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2002.

[16] M. Lothaire. *Applied combinatorics on words*, volume 105 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2005.

[17] I. L. Mel'nichuk. Existence of infinite finitely generated free semigroups in certain varieties of semigroups. In *Algebraic systems with one action and relation*, pages 74–83. Leningrad. Gos. Ped. Inst., Leningrad, 1985.

[18] I. L. Mel'nichuk. Avoidable Words. arXiv:1801.07458, 2018.

[19] Jean Moulin Ollagnier. Proof of Dejean's conjecture for alphabets with $5, 6, 7, 8, 9, 10$ and 11 letters. *Theoretical Computer Science*, 95 (2): 187–205, 1992.

[20] Pascal Ochem. A generator of morphisms for infinite words. *Theor. Inform. Appl.*, 40(3):427–441, 2006.

[21] Pascal Ochem. Doubled patterns are 3-avoidable. *Electron. J. Combin.*, 23(1), 2016. # P1.19.

[22] Peter Roth. Every binary pattern of length six is avoidable on the two-letter alphabet. *Acta Inform.*, 29(1):95–107, 1992.

[23] Ursula Schmidt. Avoidable patterns on two letters. *Theoret. Comput. Sci.*, 63(1):1–17, 1989.

[24] Axel Thue. Über die gegenseitige lage gleicher teile gewisser zeichreihen. 10:1–67.

[25] Axel Thue. Über in endlishe zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl. Christiana*, 7:1–22.

[26] A. I. Zimin. Blocking sets of terms. *Mat. Sb. (N.S.)*, 119(161)(3):363–375, 447, 1982.