Journal of Integer Sequences, Vol. 16 (2013),
Article 13.4.3

On Curling Numbers of Integer Sequences

Benjamin Chaffin
Intel Processor Architecture
2111 NE 25th Avenue
Hillsboro, OR 97124
USA
chaffin@gmail.com

John P. Linderman
1028 Prospect Street
Westfield, NJ 07090
USA
jpl.jpl@gmail.com

N. J. A. Sloane[1]
The OEIS Foundation Inc.
11 South Adelaide Avenue
Highland Park, NJ 08904
USA
njasloane@gmail.com

Allan R. Wilks
425 Ridgeview Avenue
Scotch Plains, NJ 07076
USA
allan.wilks@gmail.com

[1]Corresponding author.

Abstract

Given a finite nonempty sequence S of integers, write it as XY^k, where Y^k is a power of greatest exponent that is a suffix of S: this k is the *curling number* of S. The *curling number conjecture* is that if one starts with *any* initial sequence S, and extends it by repeatedly appending the curling number of the current sequence, the sequence will eventually reach 1. The conjecture remains open. In this paper we discuss the special case when S consists just of 2's and 3's. Even this case remains open, but we determine how far a sequence consisting of n 2's and 3's can extend before reaching a 1, conjecturally for $n \leq 80$. We investigate several related combinatorial problems, such as finding $c(n, k)$, the number of binary sequences of length n and curling number k, and $t(n, i)$, the number of sequences of length n which extend for i steps before reaching a 1. A number of interesting combinatorial problems remain unsolved.

1 The curling number conjecture

Given a finite nonempty sequence S of integers, write it as $S = XY^k$, where X and Y are sequences of integers and Y^k is a power of greatest exponent that is a suffix of S: this k is the *curling number* of S, denoted by $\mathrm{cn}(S)$. X may be the empty sequence ϵ; there may be several choices for Y, although the shortest such Y which achieves k (which as we shall see in §3.1 is primitive) is unique.

For example, if $S = 0\,1\,2\,2\,1\,2\,2\,1\,2\,2$, we could write it as XY^2, where $X = 0\,1\,2\,2\,1\,2\,2\,1$ and $Y = 2$, or as XY^3, where $X = 0$ and $Y = 1\,2\,2$. The latter representation is to be preferred, since it has $k = 3$, and as $k = 4$ is impossible, the curling number of this S is 3.

The following conjecture was stated by van de Bult et al. [2]:

Conjecture 1. The curling number conjecture. If one starts with any initial sequence of integers S, and extends it by repeatedly appending the curling number of the current sequence, the sequence will eventually reach 1.

In other words, if $S_0 = S$ is any finite nonempty sequence of integers, and we define S_{m+1} to be the concatenation

$$S_{m+1} \ := \ S_m \ \mathrm{cn}(S_m) \ \text{ for } m \geq 0\,, \tag{1}$$

then the conjecture is that for some $t \geq 0$ we will have $\mathrm{cn}(S_t) = 1$. The smallest such t is the *tail length* of S_0, denoted by $\tau(S_0)$ (and we set $\tau(S_0) = \infty$ if the conjecture is false).

For example, suppose we start with $S_0 = 2\,3\,2\,3$. By taking $X = \epsilon$, $Y = 2\,3$, we have $S_0 = Y^2$, so $\mathrm{cn}(S_0) = 2$, and we get $S_1 = 2\,3\,2\,3\,2$. By taking $X = 2$, $Y = 3\,2$ we get $\mathrm{cn}(S_1) = 2$, $S_2 = 2\,3\,2\,3\,2\,2$. By taking $X = 2\,3\,2\,3$, $Y = 2$ we get $\mathrm{cn}(S_2) = 2$, $S_3 = 2\,3\,2\,3\,2\,2\,2$. Again taking $X = 2\,3\,2\,3$, $Y = 2$ we get $\mathrm{cn}(S_3) = 3$, $S_4 = 2\,3\,2\,3\,2\,2\,2\,3$. Now, unfortunately, it is impossible to write $S_4 = XY^k$ with $k > 1$, so $\mathrm{cn}(S_4) = 1$, $S_5 = 2\,3\,2\,3\,2\,2\,2\,3\,1$, and we have reached a 1, as predicted by the conjecture. For this example, $\tau(S_0) = 4$. (If we continue the sequence from this point, it joins Gijswijt's sequence, discussed in §5.)

Some of the proofs in van de Bult et al. [2] could be shortened and the results strengthened if the conjecture were known to be true. All the available evidence suggests that the conjecture *is* true, but it has so far resisted all attempts to prove it.

In this paper we report on some extensive investigations into the case when the starting sequence consists of 2's and 3's (although even in this special case the conjecture remains open).

In Section 2 we study how far a starting sequence consisting of n 2's and 3's can extend before reaching a 1. Call the maximum such length $\Omega(n)$. That is, $\Omega(n)$ is the maximal value of the tail length $\tau(S_0)$ taken over all sequences S_0 of 2's and 3's of length n. We determine $\Omega(n)$ for all $n \le 48$, and conjecturally for all $n \le 80$ (Table 1 and Figure 1). The data suggests some properties that should be possessed by especially good starting sequences (Properties P2, P3, P4 in §2.2). Although we have not found any algebraic construction for good starting sequences, Section 2.3 describes a method which sometimes succeeds in building starting sequences of greater length. The algorithm which allowed us to extend the search to length 80 is discussed in §2.4. We would not be surprised if the conjecture in this special case turns out to be a consequence of known results on the unavoidability of patterns in long binary sequences—we discuss this briefly in §2.5.

Section 3 is devoted to the combinatorial question: what is the number $c(n,k)$ of binary sequences of length n and curling number k? This seems to be a surprisingly difficult problem, and we have succeeded only in relating $c(n,k)$ to two subsidiary quantities: $p(n,k)$, the number of such sequences that are primitive, and $p'(n,k)$, the number that are both primitive and robust (see §3.1). The main results of this section are the formulas for $c(n,k)$ in Theorems 8 and 20. With their help we are able to enumerate the curling numbers of all binary sequences of length $n \le 104$. The resulting table can be seen in entry A216955[2] in [9]. The number of binary sequences with curling number 1, $c(n,1)$ (A122536), is especially interesting and is discussed in §3.4. Some further recurrences given there enable us to compute $c(n,1)$ for $n \le 200$ (although we still do not know an explicit formula). We make frequent use of the classical Fine-Wilf theorem, and it and two other preliminary results are given in §3.2. The differences $d(n,k) := 2\,c(n-1,k) - c(n,k)$ show the structure of the $c(n,k)$ table more clearly than the numbers $c(n,k)$ themselves, and are the subject of §3.6.

In Section 4, we study the number $t(n,i)$ of sequences of length n with tail length i, where $0 \le i \le \Omega(n)$. By direct search we have determined $t(n,i)$ for $n \le 48$ (A217209), although without finding any recurrences (except for $t(n,0)$, which is the same as $c(n,1)$). The terms in each row of the $t(n,i)$ table occur in clumps, at least for $n \le 48$. In §4.1 and §4.2 we investigate some statistics of the $t(n,i)$ table, although we are a long way from finding a model which explains the clumps. Sections 4.3, 4.4, 4.5 discuss some combinatorial questions related to tail lengths. If the starting sequence S_0 is sufficiently long, it seems plausible that prefixing S_0 with a 2 or 3 is unlikely to *decrease* the tail length. If one of these prefixes decreases the tail length, we call S_0 *rotten*, and if both prefixes 2 and 3 decrease the tail length we call it *doubly rotten*. Rotten sequences certainly exist, but up to length 34

[2]Throughout this article, six-digit numbers prefixed by A refer to entries in the OEIS [9].

there are no doubly rotten sequences, and we conjecture than none exist of any length (see Conjecture 22). If this conjecture were true, it would explain a certain phenomenon that we observed in §2.2, and it would also imply that $\Omega(n+1) \geq \Omega(n)$ for all n, something that we do not know at present.

In Section 5 we briefly describe Gijswijt's sequence (A090822), which was the starting point for this investigation. The last section summarizes the open problems mentioned in the paper.

Notation Since the starting sequence S can be any sequence of integers, it seems appropriate in this paper to speak about "sequences" rather than "words" over some alphabet. However, we will make use of certain terminology (such as "prefix", "suffix") from formal language theory (cf. [7]).

Sequences will be denoted by upper case Latin letters. S^k means $SS \cdots S$, where S is repeated k times. The length of S is denoted by $|S|$. ϵ denotes the empty sequence.

Sets of sequences will be denoted by script letters (e.g., $\mathcal{C}(n,k)$) and their cardinalities by the corresponding lower case Latin letters (e.g., $c(n,k)$). Greek letters and other lower case Latin letters will also denote numbers. The symbol # denotes the cardinality of a set.

The curling number of S is denoted by $\mathrm{cn}(S)$. For a starting sequence $S_0 := s_1\, s_2 \cdots s_n$ of length n, where the s_i are arbitrary integers, we define S_{m+1} to be the concatenation $S_m\ \mathrm{cn}(S_m) = s_1 \cdots s_{n+m+1}$ for $m \geq 0$. If $\mathrm{cn}(S_t) = 1$ for some $t \geq 0$, then we call the smallest such t the *tail length* of S_0, denoted by $\tau(S_0)$, and the corresponding sequence $S^{(e)} := S_t = s_1 \cdots s_{n+t}$ is the *extension* of S_0. If no such t exists, then we set $\tau(S_0) = \infty$, $S^{(e)} = S_\infty$ (and the curling number conjecture would be false).

2 Sequences of 2's and 3's

A preliminary report on the work in this section was given in [3].

2.1 Maximal tail length $\Omega(n)$

One way to approach the conjecture is to consider the simplest nontrivial case, where the initial sequence S_0 contains only 2's and 3's, and see how far such a sequence can extend using the rule (1) before reaching a 1. Perhaps if one were sufficiently clever, one could invent a starting sequence that would never reach 1, which would disprove the conjecture. Of course it cannot reach a number greater than 3, either, for the first time this happens the next term will be 1. So the sequence must remain bounded between 2 and 3. Unfortunately, even this apparently simple case has resisted our attempts to solve it. At the end of this section (see §2.5) we will mention some slight evidence that suggests the conjecture is true. First we report on our numerical experiments.

Let $\Omega(n)$ denote the maximal tail length that can be achieved before a 1 appears, for any starting sequence S_0 consisting of n 2's and 3's. If a 1 is never reached, we set $\Omega(n) = \infty$.

n	1	2	3	4	5	6	7	8	9	10	11	12
$\Omega(n)$	0	2	2	4	4	8	8	58	59	60	112	112
n	13	14	15	16	17	18	19	20	21	22	23	24
$\Omega(n)$	112	118	118	118	118	118	119	119	119	120	120	120
n	25	26	27	28	29	30	31	32	33	34	35	36
$\Omega(n)$	120	120	120	120	120	120	120	120	120	120	120	120
n	37	38	39	40	41	42	43	44	45	46	47	48
$\Omega(n)$	120	120	120	120	120	120	120	120	120	120	120	131
n	49	50	51	52	53	54	55	56	57	58	59	60
$\Omega(n)$	131	131	131	131	131	131	131	131	131	131	131	131
n	61	62	63	64	65	66	67	68	69	70	71	72
$\Omega(n)$	131	131	131	131	131	131	131	132	132	132	132	132
n	73	74	75	76	77	78	79	80				
$\Omega(n)$	132	132	132	133	173	173	173	173				

Table 1: Lower bounds on $\Omega(n)$, the maximal tail length that can be achieved before a 1 appears, for any starting sequence S_0 consisting of n 2's and 3's. Entries for $n \le 48$ are known to be exact; the other entries are conjectured to be exact.

The curling number conjecture would imply $\Omega(n) < \infty$ for all n.

By direct search, we have found $\Omega(n)$ for all $n \le 48$. (The values for $n \le 30$ were given in [2].) The results are shown in Table 1 and Figure 1, together with lower bounds (which we conjecture are in fact equal to $\Omega(n)$) for $49 \le n \le 80$. The values of $\Omega(n)$ also form sequence A217208 in [9].

In [2], before we began computing $\Omega(n)$, we did not know how fast it would grow—would it be a polynomial, exponential, or other function of n? Even now we still do not know, since we have only limited data. But up to $n = 48$, and probably up to $n = 80$, $\Omega(n)$ is a piecewise constant function of n. There are occasional *jump points*, where $\Omega(n) > \Omega(n-1)$, but in between jump points $\Omega(n)$ does not change. Of course this piecewise constant behavior is not incompatible with polynomial or exponential growth, if the jump points are close enough together, but up to $n = 80$ this seems not to be the case. There are long stretches where $\Omega(n)$ is flat. A probabilistic argument will be given in §4 which suggests (not very convincingly) that, on the average, $\Omega(n)$ may be roughly $c_1\, n$, for a constant $c_1 \approx 1.34$. Up to $n = 49$, $\Omega(n)$ never decreases, although we cannot prove that this is always true (see §4.3).

The jump points are at $n = 1, 2, 4, 6, 8, 9, 10, 11, 14, 19, 22, 48$ and we believe the next three values are 68, 76 and 77 (A160766).

2.2 Properties of good starting sequences

From $n = 2$ through 48 (and probably through $n = 80$) the starting sequences S_0 which achieve $\Omega(n)$ at the jump points are unique. These especially good starting sequences are

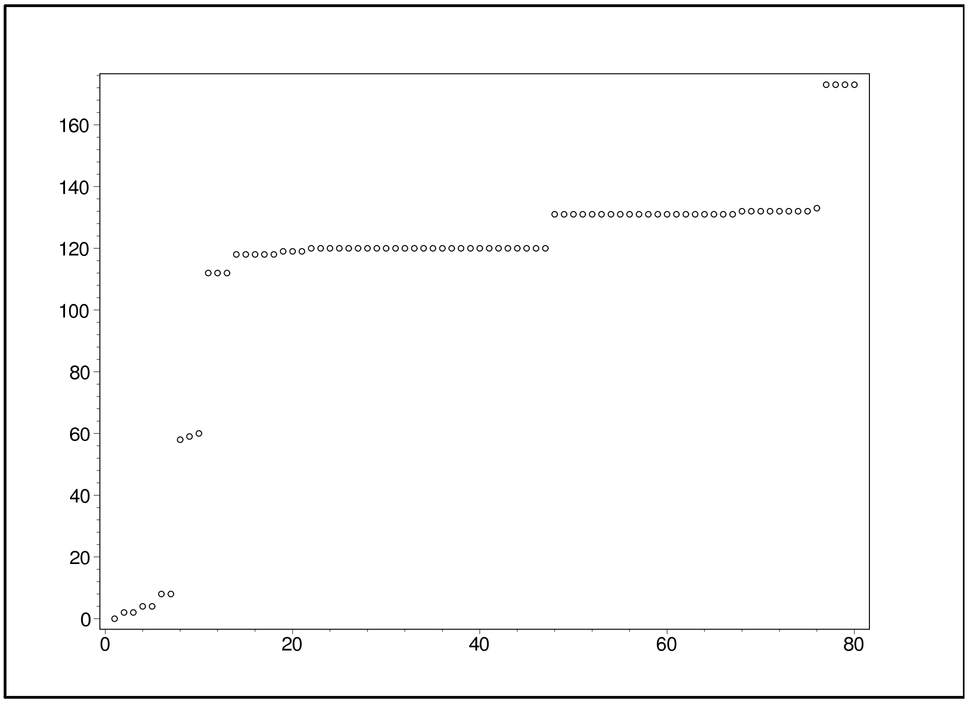

Figure 1: Scatter-plot of lower bounds on $\Omega(n)$, the maximal tail length that can be achieved before a 1 appears, for any starting sequence S_0 consisting of n 2's and 3's. Entries for $n \leq 48$ are known to be exact; the other entries are conjectured to be exact.

listed in Tables 2 and 3. For $2 \leq n \leq 48$ (and probably for $2 \leq n \leq 80$) these sequences S_0 also have the following properties:

(P2) S_0 begins with 2.

(P3) S_0 does not contain the subword 3 3.

(P4) S_0 contains no nonempty subword of the form V^4 (and in particular does not contain 2 2 2 2).

These are empirical observations. However, since they certainly hold for the first $2^{49} - 1$ choices for S_0, we venture to make the following conjecture:

Conjecture 2. If a starting sequence S_0 of 2's and 3's of length $n \geq 2$ achieves $\Omega(n)$ with $\Omega(n) > \Omega(n-1)$, then S_0 is unique and has properties P2, P3 and P4.

We can at least prove one result about these especially good starting sequences. Let $S_0 = s_1 \, s_2 \cdots s_n$ be any sequence of integers with extension $S^{(e)} = S_t = s_1 \, \cdots \, s_{n+t}$, where $\mathrm{cn}(S_t) = 1$. Call S_0 *weak* if each S_r $(r = 0, \ldots, t-1)$ can be written as $XY^{s_{n+r+1}}$ with $X \neq \epsilon$. In other words, S_0 is weak if the initial term s_1 is not necessary for the computation of the curling numbers $s_{n+1}, \ldots, s_{n+t}$. This implies that $\tau(S_0) = \tau(s_2 \, \cdots \, s_n)$, and establishes

n	Starting sequence
1	2
2	2 2
4	2 3 2 3
6	2 2 2 3 2 2
8	2 3 2 2 2 3 2 3
9	2 2 3 2 2 2 3 2 3
10	2 3 2 3 2 2 2 3 2 2
11	2 2 3 2 3 2 2 2 3 2 2
14	2 2 3 2 3 2 2 2 3 2 2 3 2 3
19	2 2 3 2 2 3 2 3 2 2 2 3 2 2 3 2 2 3 2
22	2 3 2 2 3 2 2 3 2 3 2 2 2 3 2 3 2 2 3 2 2 3
48	2 2 3 2 2 3 2 3 2 2 2 3 2 2 2 3 2 2 3 2 2 2 3 2 2 3 2 3 2 2 2 3 2 2 2 3 2 2 3 2 2 2 3 2 2 3 2 3

Table 2: Starting sequences consisting of n 2's and 3's for which $\Omega(n) > \Omega(n-1)$, complete for $1 \leq n \leq 48$.

n	Starting sequence
68	2 2 3 2 2 3 2 2 2 3 2 2 3 2 3 2 2 2 3 2 2 2 3 2 2 3 2 2 2 3 2 2 3 2 2 2 3 2 2 3 2 3 2 2 2 3 2 2 2 3 2 2 3 2 2 2 3 2 2 3 2 2 2 3 2 2 3 2
76	2 3 2 2 2 3 2 2 3 2 2 2 3 2 2 3 2 3 2 2 2 3 2 2 2 3 2 2 3 2 2 3 2 2 2 3 2 2 3 2 2 2 3 2 2 3 2 3 2 2 2 3 2 2 2 3 2 2 3 2 2 3 2 2 2 3 2 2 3 2 2 2 3 2 2 3
77	2 2 3 2 2 2 3 2 3 2 2 2 3 2 2 2 3 2 2 3 2 2 2 3 2 2 2 3 2 3 2 2 2 3 2 2 2 3 2 2 3 2 2 2 3 2 2 2 3 2 3 2 2 2 3 2 2 2 3 2 2 3 2 3 2 2 2 3 2 2 3 2 2 2 3 2 3

Table 3: Conjectured to be the complete list of starting sequences of n 2's and 3's for which $\Omega(n) > \Omega(n-1)$, for the range $49 \leq n \leq 80$.

Lemma 3. *If a starting sequence S_0 of length $n \geq 2$ achieves $\Omega(n) > \Omega(n-1)$, then S_0 is not weak.*

One further empirical observation is worth recording, concerning the starting sequences between jump points. Suppose n_0, n_1 are consecutive jump points, so that

$$\Omega(n) \;=\; \Omega(n-1) \quad \text{ for } n_0 < n < n_1\,,$$

and $\Omega(n) > \Omega(n-1)$ at $n = n_0$ and n_1. Then for $n \leq 48$ and conjecturally for $n \leq 80$, if $n_0 < n < n_1$, one can obtain a starting sequence that achieves $\Omega(n)$ by taking the starting sequence of length n_0 and prefixing it by a "neutral" string of $n - n_0$ 2's and 3's that do not get used in the computation of $\Omega(n)$. Although this is not surprising, we are unable to prove that such neutral prefixes must always exist. We return to this topic in §4.3.

The large gaps between the jump points at 22 and 48 and between 48 and 68 are especially noteworthy. In particular, we have

$$\Omega(n) = 120 \text{ for } 22 \le n \le 47\,, \tag{2}$$

and, conjecturally,

$$\Omega(n) = 131 \text{ for } 48 \le n \le 67\,. \tag{3}$$

The data shown in Tables 1, 2, 3 and Figure 1 for n in the range 49 to 80 were obtained by computer search under the assumption that the starting sequence has the properties P3 and P4 mentioned above, although without making any assumption about uniqueness. As it turned out, assuming P3 and P4, the best starting sequences at the jump points are indeed unique and start with 2. Assuming P3 and P4 greatly reduces the number of starting sequences that must be considered. For example, simply excluding sequences that contain four consecutive 2's or four consecutive 3's reduces the number of candidates of length n from 2^n to a constant times c_2^n, where $c_2 = 1.839\cdots$ (cf. A135491). However, this by itself is not enough to enable us to reach $n = 80$. We discuss the algorithm that we used in more detail in §2.4.

We should emphasize that in the (we believe unlikely) event that there are starting sequences of length n with $49 \le n \le 80$ that achieve $\Omega(n)$ but do not satisfy properties P3 and P4, it is possible our conjecture that there are jump points at lengths 68, 76, and 77 may be wrong, and there may be better starting sequences than those shown in Table 3.

2.3 A construction for larger n

We have not succeeded in finding any algebraic constructions for good starting sequences. However, one simple construction enables us to obtain lower bounds on $\Omega(n)$ for some larger values of n. Let S_0 be a sequence of length n that achieves $\Omega(n)$, and let $S^{(e)}$ be its extension of length $n+\Omega(n)$. Then in some cases the starting sequence $S^{(e)}S_0$ will extend to $S^{(e)}S^{(e)}2$ and beyond before reaching a 1. For example, taking S_0 to be the length 48 sequence in Table 2, the sequence $S^{(e)}S_0$ has length $179 + 48 = 227$ and extends to a total length of 596 before reaching a 1, showing that $\Omega(227) \ge 369$.

2.4 Computational details

Our results are complete for $n \le 48$ and are probably complete through $n = 80$. In order to extend the search this far, the algorithms used were specifically tuned to the case of sequences of 2's and 3's. There is no easy way (as far as we know) to avoid the basic process of computing the extension of S (compute $\mathrm{cn}(S)$, append it to S, and repeat until $\mathrm{cn}(S) = 1$), and so the focus is on computing $\mathrm{cn}(S)$ quickly. In the following discussion we assume that S has length at least 36. The first step is brute force: look up the curling number $\mathrm{cn}(s_{n-35} \cdots s_n)$ in a table. Two bits are sufficient to record cn, since we only care about whether it is 1, 2, 3 or ≥ 4; at two bits per entry, this table occupies 16 gigabytes.

This provides a lower bound on $\text{cn}(S)$, and also gives a lower bound for the length of the repeated substring Y which maximizes $\text{cn}(S)$. For example, if $\text{cn}(s_{n-35} \cdots s_n) = 1$, then any Y which gives $\text{cn}(S) > 1$ must be at least 19 digits long, or there would have been two copies within the last 36 digits of S.

There is also an upper bound on the length of Y. Since we are looking for that Y which maximizes $\text{cn}(S)$, we are only interested in Y's which could be repeated more times than the current best known value of $\text{cn}(S)$. For example, if we know $\text{cn}(S) \geq 3$, then we only want a Y which is repeated four times, and so we only need consider lengths up to the length of S divided by 4.

We now consider the last n digits of S as a candidate for Y, for all values of n between the lower and upper bounds. The sequences are represented as 128-bit binary numbers, and so looking for repetitions of Y can be done with bit manipulation. A few shifts and OR's generate 4 copies of Y, or as many as will fit in 128 bits. Then an XOR finds digits in which this differs from S, a bit scan locates the index of the first difference, and we can divide by the length of Y to find how many times this Y is repeated. (In fact, all divisions are done with precomputed tables.) If some Y increases the best known value for $\text{cn}(S)$, then the upper bound on the length of Y can be revised downwards. If we reach 128 digits and are still going, we resort to a slow string-based routine. In practice this slow routine accounts for less than 1% of the program's execution time.

To compute the conjectured values up to length 80, we exclude (most) strings containing $3\,3$ or a subword V^4. Obviously we cannot check all 2^{80} strings to see if they violate one of these conditions, so we need an efficient way to avoid considering them at all. To do this, we compute a 256×256 table which lists, for every string of length 8, all the length-8 strings which could legally follow it. We then construct S recursively in 8-digit blocks, ensuring that the rules are not broken within any two consecutive blocks. This is not perfect (it will allow a V^4 to slip by if V is 9 digits long, for example), but it efficiently eliminates the vast majority of undesirable cases.

2.5 Unavoidable regularities

One reason we think the curling number conjecture may be true, at least in the special case of sequences of 2's and 3's, is that there are several theorems in formal language theory about the inevitability of regularities in long binary strings. A classical example is Shirshov's theorem [7, Theorem 7.1.4], [8, Theorem 2.4.3]. Unfortunately that does not quite do what we need, but it does offer hope that a proof along these lines may exist. Lyndon's theorem [7, p. 67] is another example. Suppose we have a very long sequence of 2's and 3's generated by (1), and consider its canonical decomposition into Lyndon words. There are relatively few Lyndon words that are possible (e.g., $2\,2\,2\,2$ is forbidden), but since this attack has not yet led to a contradiction we shall say no more about it.

3 Number of binary sequences with given curling number

In this section we study the number $c(n,k)$ of binary sequences of length n and curling number k. For consistency with the other sections, we continue to consider sequences of 2's and 3's, although for this question any alphabet of size 2 (such as {0,1}) would do equally well.

3.1 Primitive and robust sequences

A sequence S is *imprimitive* (or *periodic*) if it is equal to T^i for some sequence T and an integer $i \geq 2$. Otherwise, S is *primitive* [7, p. 7].

Lemma 4. *Suppose S has curling number k. Then S can be written as XY^k, possibly in several ways. The shortest such Y is primitive and unique, and has curling number $< k$ if $k > 1$, curling number 1 if $k = 1$.*

Proof. Consider all possible ways of writing $S = XY^k$, and let $\mathcal{Y}$ denote the set of such Y's of minimal length. Every $Y \in \mathcal{Y}$ is primitive, for if $Y = T^i$, $i \geq 2$, then $S = XT^{ik}$, and $\mathrm{cn}(S) \geq ik > k$, contradicting the definition of $\mathcal{Y}$. To establish uniqueness, we observe that $S = X_1{Y_1}^k = X_2{Y_2}^k$ with $|Y_1| = |Y_2|$ implies $Y_1 = Y_2$. If $k > 1$ and $Y \in \mathcal{Y}$ has curling number $c \geq k > 1$, say $Y = UV^c$, $|V| \geq 1$, then $S = X(UV^c)^k = X'V^c$ with $c \geq k$, $|V| < |Y|$, a contradiction. Finally, if $k = 1$, certainly Y cannot have curling number greater than 1, or S would too. $\square$

We denote the length of this shortest Y by π. We let $\mathcal{C}(n,k,\pi)$ (for $n \geq 1$, $1 \leq k \leq n$, $1 \leq \pi \leq n$) denote the set of all S with the given values of n, k, and π, $c(n,k,\pi) := \#\mathcal{C}(n,k,\pi)$, $\mathcal{C}(n,k) := \bigcup_{\pi=1}^{\lfloor n/k \rfloor} \mathcal{C}(n,k,\pi)$, and $c(n,k) := \#\mathcal{C}(n,k) = \sum_{\pi=1}^{\lfloor n/k \rfloor} c(n,k,\pi)$.

If S has curling number 1 then the shortest Y for which $S = XY$ is simply the last term of S, so $\pi = 1$ and $S \in \mathcal{C}(n,1,1)$. The sets $\mathcal{C}(n,1,\pi)$ for $\pi > 1$ are empty.

We let $\mathcal{P}(n,k)$ (for $1 \leq k \leq n$) denote the subset of primitive $S \in \mathcal{C}(n,k)$, and $p(n,k) := \#\mathcal{P}(n,k)$. Note that $\mathcal{C}(n,1) = \mathcal{P}(n,1)$, since curling number 1 implies primitive.

Also let $\mathcal{Q}(n,k) := \bigcup_{i=1}^{k} \mathcal{P}(n,i)$ (for $1 \leq k \leq n$) denote the set of primitive sequences with curling number at most k, and $q(n,k) := \#\mathcal{Q}(n,k) = \sum_{i=1}^{k} p(n,i)$. We also set $q(n,0) := 0$ and $q(n,k) := q(n,n)$ for $k > n$. By definition, $q(n,n)$ is the total number of aperiodic binary sequences of length n, and it is well known ([5]; see also entry A217943 in [9]) that

$$q(n,n) \; = \; \sum_{d|n} \mu\left(\frac{n}{d}\right) 2^d \,, \tag{4}$$

where μ is the Möbius function ($q(n,n)$ is sequence A027375).

Call $S \in \mathcal{P}(n,k)$ *robust* if no proper suffix of S^{k+1} has curling number $\geq k+1$. Examples of non-robust sequences first appear at length 5, where $S = 3\,2\,2\,3\,2 \in \mathcal{C}(5,1)$ is not robust since

$$S^2 = 3\,2\,2\,3\,2\,3\,2\,2\,3\,2$$

has the suffix $(2\,3\,2)^2$. At length 8 there are examples with $k = 2$, such as $S = 3\,2\,2\,3\,2\,2\,3\,2$, for which S^3 has the suffix $(2\,3\,2)^3$. Let $\mathcal{P}'(n,k)$ denote the subset of robust $S \in \mathcal{P}(n,k)$, and let $p'(n,k) := \#\mathcal{P}'(n,k)$.

Tables 4, 5, 6, and 7 show the initial values of $c(n,k)$, $p(n,k)$, $q(n,k)$, and $p'(n,k)$, respectively. There are far fewer non-robust sequences than robust, and their numbers are shown in Table 8.

$n\backslash k$	1	2	3	4	5	6	7	8	9	10	11	12
1	2											
2	2	2										
3	4	2	2									
4	6	6	2	2								
5	12	12	4	2	2							
6	20	26	10	4	2	2						
7	40	52	20	8	4	2	2					
8	74	110	38	18	8	4	2	2				
9	148	214	82	36	16	8	4	2	2			
10	286	438	164	70	34	16	8	4	2	2		
11	572	876	328	140	68	32	16	8	4	2	2	
12	1124	1762	660	286	134	66	32	16	8	4	2	2

Table 4: Table of $c(n,k)$, the number of binary sequences of length n and curling number k, for $1 \leq k \leq n$ and $n \leq 12$ (for an extended table see A216955).

3.2 Three preliminary theorems

The classical Fine-Wilf theorem ([4]; [1, p. 13], [6], [7, p. 10]) turns out to be very useful for studying curling numbers.

Theorem 5. (*Fine and Wilf*) *If sequences $S = X^i$ and $T = Y^j$ have a common suffix U of length*

$$|U| \;\geq\; |X| \;+\; |Y| \;-\; \gcd(|X|,|Y|)\,, \tag{5}$$

then, for some sequence Z and integers g, h, we have $X = Z^g$, $Y = Z^h$, $|Z| = \gcd(|X|,|Y|)$.

In most applications all we will need is $|U| \geq |X| + |Y| - 1$, rather than (5) itself.

There is an equivalent definition of robustness that is easier to check.

$n\backslash k$	1	2	3	4	5	6	7	8	9	10	11	12
1	2											
2	2	0										
3	4	2	0									
4	6	4	2	0								
5	12	12	4	2	0							
6	20	20	8	4	2	0						
7	40	52	20	8	4	2	0					
8	74	100	36	16	8	4	2	0				
9	148	214	76	36	16	8	4	2	0			
10	286	414	160	68	32	16	8	4	2	0		
11	572	876	328	140	68	32	16	8	4	2	0	
12	1124	1722	640	276	132	64	32	16	8	4	2	0

Table 5: Table of $p(n, k)$, the number of primitive binary sequences of length n and curling number k, for $1 \leq k \leq n$ and $n \leq 12$ (for an extended table see A218869).

$n\backslash k$	1	2	3	4	5	6	7	8	9	10	11	12
1	2											
2	2	2										
3	4	6	6									
4	6	10	12	12								
5	12	24	28	30	30							
6	20	40	48	52	54	54						
7	40	92	112	120	124	126	126					
8	74	174	210	226	234	238	240	240				
9	148	362	438	474	490	498	502	504	504			
1	286	700	860	928	960	976	984	988	990	990		
11	572	1448	1776	1916	1984	2016	2032	2040	2044	2046	2046	
12	1124	2846	3486	3762	3894	3958	3990	4006	4014	4018	4020	4020

Table 6: Table of $q(n, k)$, the number of primitive binary sequences of length n and curling number at most k, for $1 \leq k \leq n$ and $n \leq 12$ (for an extended table see A218870).

$n\backslash k$	1	2	3	4	5	6	7	8	9	10	11	12
1	2											
2	2	0										
3	4	2	0									
4	6	4	2	0								
5	10	12	4	2	0							
6	20	20	8	4	2	0						
7	36	52	20	8	4	2	0					
8	72	98	36	16	8	4	2	0				
9	142	214	76	36	16	8	4	2	0			
10	280	414	160	68	32	16	8	4	2	0		
11	560	870	326	140	68	32	16	8	4	2	0	
12	1114	1720	640	276	132	64	32	16	8	4	2	0

Table 7: Table of $p'(n,k)$, the number of robust primitive binary sequences of length n and curling number k, for $1 \le k \le n$ and $n \le 12$ (for an extended table see A218875).

$n\backslash k$	1	2	3	4	5	6	7	8	9	10	11	12
1	0											
2	0	0										
3	0	0	0									
4	0	0	0	0								
5	2	0	0	0	0							
6	0	0	0	0	0	0						
7	4	0	0	0	0	0	0					
8	2	2	0	0	0	0	0	0				
9	6	0	0	0	0	0	0	0	0			
10	6	0	0	0	0	0	0	0	0	0		
11	12	6	2	0	0	0	0	0	0	0	0	
12	10	2	0	0	0	0	0	0	0	0	0	0

Table 8: The numbers $p(n,k) - p'(n,k)$ of non-robust primitive binary sequences of length n and curling number k, for $1 \le k \le n$ and $n \le 12$ (for an extended table see A218876).

Theorem 6. *If $S \in \mathcal{P}(n,k)$ is not robust, implying that S^{k+1} has a proper suffix T^{k+1} for some T, then T^{k+1} is in fact a proper suffix of S^2.*

Proof. The assertion is trivially true if $k = 1$, so we assume $k \ge 2$. The hypotheses imply $t := |T| < n$. Now S^{k+1} and T^{k+1} have a common suffix of length $(k+1)t$. If it were the case that $(k+1)t \ge n + t - 1$, by Theorem 5 we would have $S = Z^g$, $T = Z^h$, for some Z, g, h with $g > h$, implying $g \ge 2$ and so S would be imprimitive, a contradiction. So $(k+1)t < n + t - 1 < 2n$, as required. $\square$

It follows that $S \in \mathcal{P}(n,k)$ is robust if and only if no proper suffix of S^2 has curling number $k+1$. This greatly simplifies the computation of the numbers $p(n,k)$.

A trivial but useful observation is that prefixing a sequence with a single number cannot increase the curling number by more than 1:

Theorem 7. *If $S \in \mathcal{C}(n,k)$ then $2S$ (and equally $3S$) is in either $\mathcal{C}(n+1,k)$ or $\mathcal{C}(n+1,k+1)$.*

Proof. If, for example, $2S \in \mathcal{C}(n+1,l)$ with $l \ge k+2$, then $2S = U\,V^l$ for some U, V, l, and V^{l-1} (at least) is a suffix of S, contradicting the fact that S has curling number k. $\square$

3.3 A recurrence for $c(n,k)$

The first main theorem of this section expresses the n-th row of the $c(n,k)$ table in terms of the $(n-1)$st row and much earlier rows of the $p(n,k)$ and $p'(n,k)$ tables.

Theorem 8. *The numbers $c(n,k)$ have the following properties: $c(n,k) = 0$ for $n \le k-1$, $c(n,k) = 2$ for $n = k$ and $k+1$, and, for $n \ge k+2$,*

$$\begin{aligned} c(n,k) \;=\; & 2\,c(n-1,k) \\ & + \; [k \mid n] \; \left(p'\left(\frac{n}{k}, k-1\right) \;+\; q\left(\frac{n}{k}, k-2\right)\right) \\ & - \; [k+1 \mid n] \; \left(p'\left(\frac{n}{k+1}, k\right) \;+\; q\left(\frac{n}{k+1}, k-1\right)\right), \end{aligned} \tag{6}$$

where the Iverson bracket $[R]$ is 1 if the relation R is true, 0 otherwise.

Proof. We assume $k \ge 1$ and $n \ge k+2$. Suppose $S \in \mathcal{C}(n,k)$ and let T denote S with its left-most term deleted. We consider the cases $\mathrm{cn}(T) = k$ and $\mathrm{cn}(T) < k$ separately.

In the first case, if T is any sequence in $\mathcal{C}(n-1,k)$, and S is $2T$ or $3T$, then, by Theorem 7, S is in either $\mathcal{C}(n,k)$ or $\mathcal{C}(n,k+1)$. So we will obtain $2c(n-1,k)$ sequences in $\mathcal{C}(n,k)$, except that we must exclude from the count those $T \in \mathcal{C}(n-1,k)$ with the property that $2T$ or $3T = V^{k+1}$ for some primitive V of length $n/(k+1)$. This can only happen when n is a multiple of $k+1$. These V's are primitive sequences of length $n/(k+1)$, with curling number $l \le k$, and are such that no proper suffix of V^{k+1} has curling number greater than

k. If $l = k$, the number of such V's is (by definition) $p'(n/(k+1), k)$. On the other hand, if $1 \le l \le k-1$, any $V \in \mathcal{P}(n/(k+1), l)$ has the property that no proper suffix of V^{k+1} has curling number greater than k (and the number of these is $p(n/(k+1), l)$). This follows from the Fine-Wilf theorem (Theorem 5). For if V^{k+1} has a proper suffix of the form U^{k+1}, then these two sequences overlap in the last $(k+1)u$ terms, where $u = |U|$, and also $u < v$, where $v = |V| = n/(k+1)$. Since V has curling number $l < k$, the right-most k copies of U are not a suffix of V, and so $ku > v$. This implies

$$(k+1)u \;\ge\; v + u - 1\,, \tag{7}$$

and so by Theorem 5, $V = Z^g$, $U = Z^h$, $h < g$, $g \ge 2$. But $V^k = Z^{2g}$ is a suffix of T, so $\mathrm{cn}(T) \ge 2k > k$, a contradiction. (Further applications of the Fine-Wilf theorem will follow this same pattern, and we will not give as much detail.)

In the second case we must consider sequences $S = V^k$ where $\mathrm{cn}(T) < k$. Now n must be a multiple of k, and $V \in \mathcal{P}(n/k, l)$ for $1 \le l \le k-1$ is such that no proper suffix of V^k has curling number k. If $l = k-1$, the number of such V's is (by definition) $p'(n/k, k-1)$. On the other hand, if $1 \le l \le k-2$, the condition that no proper suffix of V^k has curling number k follows from the Fine-Wilf theorem by an argument similar to that given above (except that $k+1$ is replaced by k), and the number is $p(n/k, l)$. This completes the proof of the theorem. $\square$

3.4 Sequences with curling number 1

For the purpose of investigating the curling number conjecture, we are particularly interested in the first three columns of the $c(n,k)$ table, since they determine the probabilities that a random sequence of 2's and 3's has curling number 1, 2, 3, or ≥ 4 (see §4.2). The values of $c(n,1)$ are especially intriguing, as this is a combinatorial problem of independent interest. The first 30 terms of $c(n,1)$ were contributed to [9] by G. P. Srinivasan in 2006, who described it as the "number of binary sequences of length n with no initial repeats", which is equivalent to our definition (see A122536). However, we have been unable to find a formula for $c(n,1)$[3], or even a recurrence that expresses $c(n,1)$ in terms of the values of $c(m,1)$ for $m < n$. Theorem 8 says only that

$$c(n,1) \;=\; 2c(n-1,1) \;-\; [2 \mid n]\; p'(n/2,1), \tag{8}$$

$p'(n/2,1)$ being the number of robust primitive binary sequences of length $n/2$ and curling number 1.

Use of (8) enables n terms of the $c(\cdot,1)$ sequence to be obtained from $n/2$ terms of the $p'(\cdot,1)$ sequence. In practice, this limits us to about 100 terms of the former sequence. In order to obtain more terms, we introduce some further terminology (which will be used only in this section).

[3] Apart from the conjectured asymptotic estimate (27).

If S has length n, let $S^{[i]}$ denote its length-i suffix, for $1 \leq i < n$. Then we define

$$\begin{aligned}
\mathcal{A}(n,i) &:= \{S \in \mathcal{C}(n,1) \mid \mathrm{cn}(S^{[i]}\, S) = 1\}, && 1 \leq i < n\,,\\
\mathcal{B}(n,i) &:= \{S \in \mathcal{C}(n,1) \mid \mathrm{cn}(S\, S^{[i]}\, S) = 1\}, && 1 \leq i < n\,,\\
\mathcal{E}(n,i,j) &:= \{S \in \mathcal{C}(n,1) \mid S^{[i]}\, S \in \mathcal{B}(n+i,j)\}, && 1 \leq i < n, 1 \leq j < n+i\,,
\end{aligned}$$

and let $a(n,i) = \#\mathcal{A}(n,i)$, $b(n,i) = \#\mathcal{B}(n,i)$, $e(n,i,j) = \#\mathcal{E}(n,i,j)$. $S \succeq T$ will mean that T is a suffix of S, and $S \succ T$ that T is a proper suffix of S.

The following two theorems give a canonical form (see (9)) for non-robust sequences with curling number 1.

Theorem 9. *If* $\mathrm{cn}(S) = 1$ *but* $\mathrm{cn}(TS) > 1$ *for some* T *with* $S \succ T$*, then there exist* $X \neq \epsilon$*,* $Y \neq \epsilon$ *with*

$$S = XYX, \text{ where } \mathrm{cn}(X) = 1,\ T \succeq Y, \text{ and } X \succ Y\,. \tag{9}$$

Proof. Since $\mathrm{cn}(TS) > 1$, $TS \succeq ZZ$ for some Z with $|Z| < |S|$, and therefore $S \succ Z$. We write $S = XZ$ and observe that $TXZ \succeq ZZ$, so $TX \succeq Z$. Therefore either $X \succeq Z$ or $Z \succ X$. The former implies $\mathrm{cn}(S) > 1$, a contradiction. So $Z \succ X$, say $Z = YX$, and $S = XYX$.

Since $S \succeq X$, $\mathrm{cn}(X) = 1$. Also $TX \succeq Z = YX$, so $T \succeq Y$. It remains to show that $X \succ Y$. Now $S \succ T \succeq Y$ and $S \succeq X$, so either $Y \succeq X$ or $X \succ Y$. The former implies $Y = WX$ for some W, and then $S = XYX = XWXX$, contradicting $\mathrm{cn}(S) = 1$. So $X \succ Y$. □

For example, $S = 33223\,22333223$ has curling number 1 (the conspicuous substring 333 makes it easy to check this). If $T = 2333\,223$, then $TS = 2333\,Z^2$, where $Z = 223\,33223$ (the spaces in these strings are for legibility). Then, following the steps of the proof, we write $S = XZ$, which defines $X = 33223$, and then write $Z = YX$, which defines $Y = 223$, and so finally we have

$$S \;=\; XYX \;=\; 33223\,223\,33223\,,$$

as claimed.

Theorem 10. *If* $S = XYX = UVU$*, with* $X \succ Y \neq \epsilon$*,* $U \succ V \neq \epsilon$*,* $X \neq U$*, then* $\mathrm{cn}(S) > 1$.

Proof. Without loss of generality, $U \succ X$. Since both X and U are *prefixes* of S, we have $U = XZ$ for some $Z \neq \epsilon$, and $S \succ XZ$. Now $2|Z| = |S| - 2|X| - |V| < |S| - 2|X| = |Y| < |X|$, so $|X| > |Z|$. This implies $X \succ Z$ (they are both suffixes of S), say $X = AZ$, so $S = UVU = UVXZ = UVAZZ$, contradicting $\mathrm{cn}(S) = 1$. □

Theorems 9 and 10 say that a non-robust sequence S with curling number 1 can be written in a unique way as $S = XYX$, where Y is a suffix of X.

Corollary 11. *(i) For $1 \le i < n/3$, there is a bijection between the sets $\mathcal{C}(n,1) \setminus \mathcal{A}(n,i)$ and*

$$\bigcup_{m=\lceil (n-i)/2 \rceil}^{\lfloor (n-1)/2 \rfloor} \mathcal{B}(m, n-2m)\,.$$

(ii) For $n/3 \le i < n$, there is a bijection between the sets $\mathcal{C}(n,1) \setminus \mathcal{A}(n,i)$ and

$$\bigcup_{m=1+\lfloor n/3 \rfloor}^{\lfloor (n-1)/2 \rfloor} \mathcal{B}(m, n-2m)\,.$$

Proof. Fix i, where $1 \le i < n$. First, suppose that S is in $\mathcal{C}(n,1) \setminus \mathcal{A}(n,i)$. Taking $T = S^{[i]}$ in Theorem 9 we may write $S = XYX$ where $m := |X| \ge 1$, $X \in \mathcal{C}(m,1)$, $n - 2m = |Y| \le |S^{[i]}| = i$ and $m = |X| > |Y| = n - 2m \ge 1$. Hence $X \in \mathcal{B}(m, n-2m)$ for some m satisfying the three conditions: $m \ge (n-i)/2$, $m > n/3$ and $m \le (n-1)/2$. By Theorem 10, X belongs to only one such $\mathcal{B}(m, n-2m)$.

Conversely, if m satisfies these three conditions and $X \in \mathcal{B}(m, n-2m)$ then let $S = XX^{[n-2m]}X$. By the definition of $\mathcal{B}(m, n-2m)$, S must be in $\mathcal{C}(n,1)$ and since $m \ge (n-i)/2$, we have $S^{[i]}S \succeq S^{[n-2m]}S = (YX)^2$, so that S is not in $\mathcal{A}(n,i)$.

This establishes a bijection between $\mathcal{C}(n,1) \setminus \mathcal{A}(n,i)$ and the union of $\mathcal{B}(m, n-2m)$ for m satisfying the three earlier conditions. The proof is completed by observing that $(n-i)/2 > n/3$ if and only if $n > 3i$, which is the condition that separates cases (i) and (ii) of the Corollary. ∎

Since the unions in Corollary 11 are clearly disjoint, we immediately obtain the following formulas for $a(n,i)$.

Corollary 12. *(i) For $1 \le i < n/3$,*

$$a(n,i) \;=\; c(n,1) \;-\; \sum_{m=\lceil (n-i)/2 \rceil}^{\lfloor (n-1)/2 \rfloor} b(m, n-2m)\,. \tag{10}$$

(ii) For $n/3 \le i < n$,

$$a(n,i) \;=\; c(n,1) \;-\; \sum_{m=1+\lfloor n/3 \rfloor}^{\lfloor (n-1)/2 \rfloor} b(m, n-2m)\,. \tag{11}$$

The next three theorems give a further refinement of non-robust sequences, and lead to the set bijections and formulas in Corollaries 16 and 17. We postpone their proofs to the Appendix. The proof that Corollary 16 follows from Theorems 13 through 15 is similar to the proof of Corollary 11 and is omitted.

Theorem 13. *If $X \succ Y$, $\mathrm{cn}(YX) = 1$, and $\mathrm{cn}(XYX) > 1$, then there exist S and T such that $YX = STS$ with $X \succeq T$, $S \succ T$, $\mathrm{cn}(S) = 1$. Furthermore, either $|S| = |Y|$ or $|S| > 2|Y|$.*

Theorem 14. *If $n/2 < i < n$ then there is a bijection between $\mathcal{A}(n,i)\backslash\mathcal{B}(n,i)$ and $\mathcal{B}(i, n-i)$.*

Theorem 15. *If $1 \le i < n/3$ then $\mathcal{A}(n,i)\backslash\mathcal{B}(n,i)$ is a disjoint union of $\mathcal{E}(m-i, i, n+i-2m)$, where $\max(2i, 1 + \lfloor (n+i)/3 \rfloor) < m \le \lfloor (n+i-1)/2 \rfloor$.*

Corollary 16. *(i) For $1 \le i < n/5$, there is a bijection between $\mathcal{A}(n,i) \setminus \mathcal{B}(n,i)$ and the disjoint union of $\mathcal{E}(m-i, i, n+i-2m)$, where $2i < m \le (n+i-1)/2$.*

(ii) For $n/5 \le i < n/3$, there is a bijection between $\mathcal{A}(n,i) \setminus \mathcal{B}(n,i)$ and the disjoint union of $\mathcal{E}(m-i, i, n+i-2m)$, where $(n+i)/3 < m \le (n+i-1)/2$.

(iii) For $n/3 \le i \le n/2$, $\mathcal{B}(n,i)$ is empty.

(iv) For $n/2 < i < n$, there is a bijection between $\mathcal{A}(n,i) \setminus \mathcal{B}(n,i)$ and $\mathcal{B}(i, n-i)$.

Corollary 17. *(i) For $1 \le i < n/3$,*

$$b(n,i) \;=\; a(n,i) \;-\; \sum_{m=\max(2i, 1+\lfloor (n+i)/3 \rfloor)}^{\lfloor (n+i-1)/2 \rfloor} e(m-i, i, n+i-2m)\,. \tag{12}$$

(ii)

$$b(n,i) \;=\; 0 \quad \text{for } \; n/3 \le i \le n/2\,. \tag{13}$$

(iii)

$$b(n,i) = a(n,i) - b(i, n-i) \quad \text{for } \; n/2 < i < n\,. \tag{14}$$

Observing that $p'(n/2, 1) = a(n/2, n/2-1)$, equations (8) and (10) through (14) can be used recursively to compute values of $c(n,1)$, using brute force to determine $e(m,i,j)$ only for relatively small values of m (see §3.7).

We have briefly investigated the possibility of generalizing the approach in this section to deal with curling numbers k greater than one. The following theorems replace Theorems 9 and 10:

Theorem 18. *Suppose $S \in \mathcal{P}(n,k) \setminus \mathcal{P}'(n,k)$, where $k > 1$. Then there exist X and T with $S = X(TX)^k$ and $S \succ T$.*

Proof. By Theorem 6, $S^2 = PQ^{k+1}$ with $P \ne \epsilon$. If $(k+1)|Q| \ge n + |Q| - 1$, then Theorem 5 would imply that S is periodic. So $k|Q| < n-1$, and k copies of Q lie properly inside S, say $S = XQ^k$ with $|X| < |Q|$, $X \ne \epsilon$. Define T by $Q = TX$ and we have $S = X(TX)^k$. Also $PQ^{k+1} = SXQ^k$, so $PQ = PTX = SX$ and $S \succ T$. □

Theorem 19. *The representation $S = X(TX)^k$ obtained in Theorem 18 is unique.*

Since we will not make any use of Theorem 19, we omit the somewhat tedious proof.

Because $S \in \mathcal{P}(n,k)$, we know that S can be written as XY^k, where Y is primitive, possibly in several ways. Theorems 18 and 19 say that if S is not robust, then exactly one of these Y's has the corresponding X as a suffix. We have not pursued the generalizations of Theorems 13–15 and Corollaries 16–17 to this case.

3.5 The values of $c(n,k)$ for $k \geq \lfloor \sqrt{n} \rfloor$

The second main theorem of this section gives an expression for $c(n,k)$ in the range $k \geq \lfloor \sqrt{n} \rfloor$ that involves the partial sum function $q(m,k)$.

Theorem 20. *We have $c(n,n) = 2$ for all n, $c(n,n-1) = 2$ for $n \geq 2$, and, for $n \geq 4$ and $k \geq \lfloor \sqrt{n} \rfloor$,*

$$c(n,k,\pi) \;=\; \begin{cases} 2^{n-(k+1)\pi}\,(2^{\pi}-1)\,q(\pi,k-1), & \text{if } 1 \leq \pi \leq \lfloor \frac{n}{k+1} \rfloor\,, \\ 2^{n-k\pi}\,q(\pi,k-1), & \text{if } \lceil \frac{n+1}{k+1} \rceil \leq \pi \leq \lfloor \frac{n}{k} \rfloor\,, \end{cases} \tag{15}$$

and $c(n,k) = \sum_{\pi=1}^{\lfloor n/k \rfloor} c(n,k,\pi)$.

Proof. We assume $n \geq 4$ and $k \geq \lfloor \sqrt{n} \rfloor \geq 2$. Note that $k \geq \lfloor \sqrt{n} \rfloor$ is equivalent to $k+1 > \sqrt{n}$.

We consider the cases $n \leq \pi(k+1)-1$ and $n \geq (k+1)\pi$ separately.

First, if $n \leq \pi(k+1)-1$, we have

$$\left\lceil \frac{n+1}{k+1} \right\rceil \leq \pi \leq \left\lfloor \frac{n}{k} \right\rfloor \,.$$

Let us write $S = X\,Y^k$, where Y is minimal and has length π. Then $n \leq \pi(k+1)-1$ implies $|X| < \pi$. By Lemma 4, $Y \in \mathcal{Q}(\pi,k-1)$. There are $2^{n-\pi k}$ choices for X, and $q(\pi,k-1)$ choices for Y, and we claim that the resulting sequence $X\,Y^k$ always has curling number k. For suppose it has curling number $> k$, so that we have $X\,Y^k = U\,V^{k+1}$, with $u = |U|$, $v = |V|$. There are two sub-cases. If $(k+1)v \geq k\pi$, then we have $(k+1)\pi > |S| \geq (k+1)v$, implying $\pi > v$. The two different representations of S have a common suffix Y^k of length $k\pi$, which, since $k \geq 2$, satisfies

$$k\pi \;\geq\; v+\pi-1\,. \tag{16}$$

By Theorem 5, $Y = Z^g$, $V = Z^h$, with $g > h$, so $g \geq 2$, and Y is imprimitive, a contradiction. On the other hand, suppose $(k+1)v < k\pi$. Again $\pi > v$. Since $\mathrm{cn}(Y) < k$, $kv > \pi$. Now the common suffix has length $(k+1)v$, our inequalities imply

$$(k+1)v \;\geq\; v+\pi-1\,, \tag{17}$$

and, again by Theorem 5, Y is imprimitive, a contradiction. So the number of sequences S of this type is $2^{n-k\pi}q(\pi,k-1)$, as claimed.

Second, if $n \geq (k+1)\pi$, we have

$$1 \leq \pi \leq \left\lfloor \frac{n}{k+1} \right\rfloor \,.$$

Let us write

$$S \;=\; XBY^k\,, \tag{18}$$

where X has length $n-(k+1)\pi$, B has length π, and $Y \in \mathcal{Q}(\pi, k-1)$. Certainly $B \neq Y$ (B stands for "blocker", the purpose of which is to ensure that Y is repeated only k times). There are potentially $2^{n-(k+1)\pi}$ choices for X, $2^{\pi}-1$ choices for B, and $q(\pi, k-1)$ choices for Y. We claim that the assumption $k \geq \lfloor\sqrt{n}\rfloor$ guarantees that all choices result in a sequence with curling number k. For suppose on the contrary that S (in (18)) is also equal to $U\,V^{k+1}$, with $u=|U|$, $v=|V|$. Again there are two sub-cases. If $(k+1)v \geq k\pi$, then we have

$$(k+1)^2 \;>\; n \;\geq\; (k+1)v \;\geq\; k\pi\,,$$

so $k+1>v$, $k \geq v$. The two different representations of S have a common suffix of length $k\pi$, and our inequalities imply (16). On the other hand, suppose $(k+1)v<k\pi$. Again we have $kv>\pi$, and the common suffix satisfies (17). In both cases Theorem 5 now leads to a contradiction. This complete the proof of the theorem. $\square$

The formulas in Theorem 20 cover a large portion of the $c(n,k)$ table. However, although with more work they could be extended so as to apply to slightly smaller values of k, it seems unlikely that this approach will lead to a formula for $c(n,k,\pi)$ for small values of k.

3.6 The difference table $d(n,k)$

$n\backslash k$	1	2	3	4	5	6	7	8	9	10	11	12
2	2	−2										
3	0	2	−2									
4	2	−2	2	−2								
5	0	0	0	2	−2							
6	4	−2	−2	0	2	−2						
7	0	0	0	0	0	2	−2					
8	6	−6	2	−2	0	0	2	−2				
9	0	6	−6	0	0	0	0	2	−2			
10	10	−10	0	2	−2	0	0	0	2	−2		
11	0	0	0	0	0	0	0	0	0	2	−2	
12	20	−10	−4	−6	2	−2	0	0	0	0	2	−2

Table 9: The difference table $d(n,k)$ defined by (19) (for an extended table see A217943).

In the $c(n,k)$ table (Table 4), if we look at the difference between each row and twice the previous row, we obtain a much simpler table.[4] We define

$$d(n,k) \;:=\; 2\,c(n-1,k) \;-\; c(n,k)\,, \tag{19}$$

[4]It was by studying the $d(n,k)$ table that we were led to Theorems 8 and 20.

for $n \geq 2$, $1 \leq k \leq n-1$, with $d(n,n) = -2$. The initial values are shown in Table 9. We see that if one ignores the initial entries in each row, most of the remaining entries are zero, except for diagonal lines of pairs of nonzero entries. More precisely, it appears that

$$\begin{aligned} d(2k, k-1) &= -d(2k, k) = 2, \quad k \geq 4, \\ d(3k, k-1) &= -d(3k, k) = 6, \quad k \geq 5, \\ d(4k, k-1) &= -d(4k, k) = 12, \quad k \geq 6, \\ d(5k, k-1) &= -d(5k, k) = 30, \quad k \geq 7, \end{aligned} \tag{20}$$

and so on. Only the first of these diagonal lines can be seen in Table 9, but they are all visible in the extended table that is given in entry A217943 in [9]. These expressions all follow from Theorem 20:

Theorem 21. *In the range $k \geq \lfloor \sqrt{n} \rfloor$, the only nonzero entries in the $d(n,k)$ table are*

$$d(mk, k-1) = -d(mk, k) = q(m,m), \text{ for } m \geq 1, \ k \geq m+2. \tag{21}$$

Proof. This follows easily from Theorem 20. We prove the second assertion in (21) as an illustration. We have

$$d(mk, k) = 2c(mk-1, k) - c(mk, k). \tag{22}$$

From (15),

$$c(mk, k) = \sum_{\pi=1}^{m-1} c(mk, k, \pi) + c(mk, k, m), \tag{23}$$

$$c(mk-1, k) = \sum_{\pi=1}^{m-1} c(mk-1, k, \pi). \tag{24}$$

Each summand in (23) (see (15)) is exactly twice the corresponding term in (24), and $c(mk, k, m) = q(m, k) = q(m, m)$, so $d(mk, k) = -q(m, m)$. $\square$

Note that whereas the expression for $c(n,k)$ in Theorem 20 involves the general function $q(\pi, k)$, the expression for $d(n,k)$ in the range $k \geq \lfloor \sqrt{n} \rfloor$ is fully explicit, since $q(m,m)$ is given by (4).

Theorem 8 gives another formula for $d(n,k)$:

$$d(n,k) = [k{+}1 \mid n] \left(p' \left(\frac{n}{k+1}, k \right) + q \left(\frac{n}{k+1}, k-1 \right) \right) - [k \mid n] \left(p' \left(\frac{n}{k}, k-1 \right) + q \left(\frac{n}{k}, k-2 \right) \right), \tag{25}$$

and in particular,

$$\begin{aligned} d(n,1) &= [2 \mid n] \, p'(n/2, 1), \\ d(n,2) &= [3 \mid n] \, (p'(n/3, 2) + p(n/3, 1)) - [2 \mid n] \, p'(n/2, 1). \end{aligned} \tag{26}$$

The first of these is nicely checked by noticing that the nonzero entries in the first column of the $d(n,k)$ table, namely $2, 2, 4, 6, 10, 20, \cdots$ are also the entries in the first column of the $p'(n,k)$ table (Table 7). It is also worth mentioning that if p is prime then $c(p,k) = 2c(p-1,k)$ for all k (see (8)) and so $d(p,k) = 0$.

3.7 Computation of $c(n,k)$

We constructed an extensive table of values of $c(n,k)$, hoping that it would lead to additional insight into these numbers. First, by direct enumeration, using a number of different programs and different computers (including a four-day computation on a cluster of 64 SPARC processors), we calculated $c(n,k)$ for $n \leq 51$.

Second, we tabulated $e(n,i,j)$ for $n \leq 23$. This was sufficient for the recurrences (8) and (10)–(14) to give $c(n,1)$ for $n \leq 200$. These values suggest the conjecture that

$$\lim_{n\to\infty} \frac{c(n,1)}{2^n} \;=\; 0.27004339525895354325\cdots\,. \tag{27}$$

From Equation (8) we have

$$c(n,1) \;\geq\; 2c(n-1,1) \;-\; [2 \mid n]\, c(n/2,1)\,,$$

which implies, using the known values of $c(n,1)$, that

$$c(n,1) \;>\; 0.27 \cdot 2^n \quad \text{for } n \geq 200\,. \tag{28}$$

We omit the proof. But we have no comparable upper bound for $c(n,1)$ (other than 2^n), nor a proof that the limit (27) exists.

Third, we used a different approach, which enabled us to take a table of the curling numbers of all sequences of length $n \leq n_0$, and from this produce a table of $c(n,k)$ for all $n \leq 2n_0$, without having to compute the curling numbers of all 2^{2n_0} sequences of length $2n_0$. The idea underlying this approach is the following. Consider a sequence S of length n with $n_0 \leq n \leq 2n_0$, and let M be its length-n_0 suffix. As a first approximation, we set $\mathrm{cn}(S) = \mathrm{cn}(M) = l$ (say). This approximation will be wrong if for some suffix T of M it should happen that T^{l+1} is a suffix of S. If so, we must increase $\mathrm{cn}(S)$ by 1 for all S having suffix T^{l+1}. There are complications if there is more than one such T to be considered, but the Fine-Wilf theorem (Theorem 5) shows that this can only happen when $l = 1$. We omit discussion of the details. Using this approach (with $n_0 = 32$) we were able to extend the table of values of $c(n,k)$ and $p(n,k)$ to $n = 64$.

Finally, we tabulated $p'(n,k)$ for $n \leq 36$. This, together with the 200 terms of $c(n,1)$, was sufficient for the recurrence in Theorem 8 to give the first 104 rows of the $c(n,k)$ table. These results can be seen in A216955 and A122536.

4 Tail lengths of {2,3}-sequences

4.1 Distribution of tail lengths

Let $t(n,i)$ denote the number of starting sequences S_0 of n 2's and 3's which have tail length i, where i ranges from 0 to $\Omega(n)$. The initial values are shown in Table 10. Since the rows rapidly increase in length (cf. Table 1), we end this table at $n = 9$. Note that the entries for $i = 9$ through 55 (which are all zero) have been compressed into a single column. Rows $n = 22$ and 32 are shown in Tables 11 and 12. Entry A217209 in [9] gives the first 48 rows in full. The first column is the same as the first column of the $c(n,k)$ table, and contains the numbers $c(n,1)$ that are the subject of §3.4.

$n \backslash i$	0	1	2	3	4	5	6	7	8	9-55	56	57	58	59
1	2													
2	2	1	1											
3	4	2	2											
4	6	5	3	1	1									
5	12	9	6	2	3									
6	20	18	12	6	7	0	0	0	1					
7	40	34	25	11	14	1	0	1	2					
8	74	71	47	24	28	1	3	2	3	0	0	2	1	
9	148	139	95	48	56	6	4	3	6	0	2	3	1	1

Table 10: Table of $t(n,i)$, the number of sequences of n 2's and 3's with tail length i, for $0 \le i \le \Omega(n)$ (A217209).

0-8	1133200	1140102	768386	417081	479224	47190	33440	32283	51035
9-17	6388	6096	1031	2074	516	807	67	0	0
18-26	1	1	3	6	12	7	0	0	0
27-35	0	0	0	0	0	0	0	0	0
36-44	0	0	0	0	0	0	1	7	16
45-53	24	50	98	198	394	786	1316	2633	5121
54-62	5891	7687	9230	14622	12983	6486	2659	642	1099
63-71	463	299	32	0	0	0	0	0	0
72-80	1	0	0	0	0	0	0	0	0
81-89	0	0	0	0	0	0	0	0	0
90-98	0	0	0	0	0	0	0	0	0
99-107	0	0	1	1	2	4	8	16	32
108-116	64	128	256	512	1024	17	17	34	70
117-120	139	282	8	1					

Table 11: Distribution of tail lengths $t(22,i)$, $0 \le i \le 120$, for all starting sequences of length 22 (22 is the first time a tail of length 120 is reached). Note the three "clumps."

As can be seen from Tables 10–12, the values in each row are distributed into clumps, with each clump gradually thickening as n increases. Table 11 shows the distribution of

0-8	1159845258	1167273283	786757853	427198253	490970976	48399112	34266983	33065461	52260747
9-17	6585936	6286710	1088875	2157877	553922	848516	69469	519	1038
18-26	836	1547	3092	6184	11843	7303	206	28	57
27-35	99	194	0	0	0	2	9	21	34
36-44	72	130	198	394	788	1576	3153	6305	12610
45-53	25219	50438	100876	201752	403504	804960	1347868	2695736	5244019
54-62	6034490	7874728	9455010	14977616	13308516	6658834	2742615	676305	1153446
63-71	487704	309650	32814	28	24	48	96	193	385
72-80	770	0	0	0	0	0	0	0	0
81-89	0	0	0	0	1	2	5	10	20
90-98	0	1	1	2	4	8	16	32	64
99-107	128	256	512	1024	2048	4096	8192	16384	32768
108-116	65536	131072	262144	524288	1048544	18331	18265	36530	73119
117-120	146237	292601	8798	1144					

Table 12: Distribution of tail lengths $t(32, i)$, $0 \le i \le 120$, for all starting sequences of length 32. The clumps have thickened.

tail lengths at length 22, the first time that a tail of length 120 is reached (note the final "1", indicating that the starting sequence was unique). By length 32 (Table 12), the clumps have thickened but still end at 120. A tail of length greater than 120 does not appear until length 48, when the greatest tail length jumps to 131. The powers of 2 in Tables 11 and 12 suggest that the clumps tend to grow by prefixing good starting sequences of shorter length by random strings of 2's and 3's. However, we do not have a satisfactory model which explains this distribution.

The mean value of the nth row,

$$\frac{1}{2^n} \sum_{i=0}^{\Omega(n)} i\, t(n, i)\,,$$

at least for $n \le 48$, is converging to a value around $2.741\cdots$ (see A216813). That is, if a starting sequence consisting of n 2's and 3's is chosen at random, it will reach a 1 on average after only $2.741\cdots$ steps. This is in sharp contrast to the behavior of the best starting sequences, as we see from Table 1. Of course if the curling number conjecture is false for sequences of 2's and 3's, the mean will be infinite beyond some point.

4.2 A probabilistic model

Let $\theta_k^{(n)} := c(n, k)/2^n$ denote the probability that a randomly chosen sequence consisting of n 2's and 3's has curling number k. The available data ($n \le 200$ for $k = 1$, $n \le 104$ for $k > 1$) suggests that as n increases these probabilities are converging to the values

$$\theta_1 \approx .270, \ \theta_2 \approx .434, \ \theta_3 \approx .162, \ \sum_{k \ge 4} \theta_k \approx .134\,.$$

When we extend a sequence S by appending the curling number $k = \text{cn}(S)$, *if* it were the case that the concatenation Sk were independent of S, we could model this process as a two-state Markov chain with states "curling number is 2 or 3" and "curling number is 1 or ≥ 4." The probability of staying in the "2 or 3" state would be $\theta_2 + \theta_3 \approx .596\cdots$ and the probability of leaving that state would be $.404\cdots$. If the starting sequence is randomly chosen from all 2^n possibilities, this model would imply that the maximal number of steps before reaching the "1 or ≥ 4" state for the first time would be about

$$t \;\approx\; n \frac{\log 2}{\log(1/.596)} \;\approx\; 1.34\, n \,.$$

This Markov model certainly does not apply at the beginning of the appending process, but it could conceivably be valid once the sequence has been extended for a while, so we think it is worth mentioning.

4.3 "Rotten" sequences: prefix decreases tail

Let S_0 be an arbitrary sequence of 2's and 3's of length n, with tail length $\tau(S_0) = i$, say. It seems plausible that if n is large, then prefixing S_0 by a single 2 or 3 will not change $\tau(S_0)$, i.e., that $\tau(2\,S_0) = \tau(3\,S_0) = \tau(S_0)$. But could doing this actually *decrease* the tail length? Choosing an adjective not normally used in mathematics, we will call S_0 *rotten* if either $\tau(2\,S_0) < \tau(S_0)$ or $\tau(3\,S_0) < \tau(S_0)$, and *doubly rotten* if both $\tau(2\,S_0) < \tau(S_0)$ and $\tau(3\,S_0) < \tau(S_0)$ hold. There are surprisingly few rotten sequences of length up through 34. The first few examples are shown in Table 13, and the numbers of rotten sequences of lengths 1 through 34 are given in Table 14. If $S_0 = 3\,2\,3\,2\,3$, for example, then $S_0^{(e)} = 3\,2\,3\,2\,3\,2\,3\,3\,2$, and $\tau(S_0) = 4$. But if we prefix S_0 with a 2, so the starting sequence is $2\,S_0 = 2\,3\,2\,3\,2\,3$, the extension is $2\,3\,2\,3\,2\,3\,3\,2$, so $\tau(2\,S_0) = 2$, and S_0 is rotten.

22	333	32323	323232	2323232	3232323	22322232
23222322	23223223	33233233	223222322	223222323	232223222	332332332
2232223222	2232223223	2232223232	2322232223	2322322322	2332332332	3322332233
3323323323	22322232223	22322232232	22322232322	22322322232	22322322322	22323222322

Table 13: The first 28 rotten sequences (A216730).

0	1	1	0	1	1	2	4	4	8
14	11	18	30	26	24	40	35	58	69
48	84	158	67	139	287	215	242	490	323
624	919	516	1072						

Table 14: Number of rotten sequences of lengths 1 through 34 (A216950).

However, up to length 34 there are no doubly rotten sequences.

Conjecture 22. Doubly rotten sequences do not exist.

If this conjecture were true, it would imply that one can always prefix a starting sequence S_0 by one of $\{2, 3\}$ without decreasing the tail length. This would explain the observation made in §2.2 about the behavior of $\Omega(n)$ between jump points. It would also imply that $\Omega(n+1) \geq \Omega(n)$ for all n, something that we do not know at present.

4.4 Sequences in which first term is essential

A statistic that is relevant to the study of rotten sequences is the following. If a starting sequence S_0 of length n is chosen at random, and has curling number k, this means we can write $S_0 = XY^k$ for suitable sequences X, Y. What is the probability that we must necessarily take X to be the empty sequence, i.e., that the only such representation goes all the way back to the beginning of S_0 (and so the first term is essential for the computation of the curling number)? The sequence $2\,2\,3\,2\,2\,3$ is an example, since here $k = 2$, and $X = \epsilon$, $Y = 2\,2\,3$ is the only representation. But $2\,3\,3\,2\,3\,3$ is not, since $k = 2$ and we can either take $X = \epsilon$, $Y = 2\,3\,3$ or $X = 2\,3\,3\,2$, $Y = 3$, and the latter representation avoids using $X = \epsilon$. The number of such sequences of length n for $1 \leq n \leq 35$ is given in Table 15. If n is prime, the number is 2, but the limit supremum of these numbers appears to grow exponentially.

2	2	2	4	2	8	2	10	8	14
2	40	2	40	32	88	2	192	2	324
100	564	2	1356	32	2226	370	4564	2	9656
2	17944	1450	35424	152					

Table 15: Number of sequences of lengths 1 through 35 whose curling number representation $X\,Y^k$ requires $X = \epsilon$ (A216951).

4.5 Sequences where prefix increases tail

In contrast to "rotten" sequences, we also investigated starting sequences S_0 for which either $\tau(2\,S_0) > \tau(S_0)$ or $\tau(3\,S_0) > \tau(S_0)$. The sequence $S_0 = 2\,2\,3\,2\,2$ is an example, since $\tau(S_0) = 2$, $\tau(2\,S_0) = 8$, $\tau(3\,S_0) = 2$. The numbers of such sequences of lengths 1 through 30 are shown in Table 16. There are rather more of these than there are rotten sequences, although we found no example where both $\tau(2\,S_0) > \tau(S_0)$ and $\tau(3\,S_0) > \tau(S_0)$ hold.

5 Gijswijt's sequence

If we simply start with $S_0 = 1$, and generate an infinite sequence by continually appending the curling number of the current sequence, as in (1), we obtain

$$G := 1\,1\,2\,1\,1\,2\,2\,2\,3\,1\,1\,2\,1\,1\,2\,2\,2\,3\,2\,1\,1\,2\,1\,1\,2\,2\,2\,3\,1\,1\,2\,1\,1 \cdots .$$

2	1	2	1	5	3	12	9	19	16
38	20	59	42	104	65	213	111	400	245
765	439	1563	820	3046	1731	5955	3292	12078	6343

Table 16: Number of sequences S_0 of lengths 1 through 30 such that $\tau(2\,S_0) > \tau(S_0)$ or $\tau(3\,S_0) > \tau(S_0)$ (A217437).

This is *Gijswijt's sequence*, A090822, invented by D. Gijswijt in 2004, and analyzed by van de Bult et al. [2].

The first time a 4 appears in G is at term 220. One can calculate quite a few million terms without finding a 5 (as the authors of [2] discovered), but in [2] it was shown that a 5 eventually appears for the first time at about term

$$10^{10^{23}}\,.$$

Van de Bult et al. [2] also show that G is in fact unbounded, and conjecture that the first time that a number $m\ \geq 6$ appears is at about term number

$$2^{2^{3^{4^{\cdot^{\cdot^{\cdot^{m-1}}}}}}}\,,$$

a tower of height $m-1$. The fairly complicated arguments used in [2] could be considerably simplified and extended if the curling number conjecture were known to be true.

Our final theorem shows that if the curling number conjecture is true, any starting sequence S that does not contain a 1 must eventually merge with G.

Theorem 23. *Assume the curling number conjecture is true. Let S be an initial sequence not containing a* 1*, let $S^{(e)}$ be its "extension" (defined in §1), and let $S^{(\infty)}$ be its infinite continuation. Then $S^{(\infty)} = S^{(e)}G$.*

Proof. By definition, $S^{(e)}$ does not contain a 1 but is immediately followed by a 1. Suppose $S^{(\infty)} \neq S^{(e)}G$, and suppose they first differ at a position where $S^{(\infty)}$ is n, say, whereas $S^{(e)}G$ is $m < n$. This n must be the curling number of some portion of $S^{(\infty)}$ that begins with a suffix X, say, of $S^{(e)}$. Let $S^{(e)} = WX$. Then $S^{(\infty)} = W(XT)^n\,n\cdots$ for some prefix T of G, whereas $G = T(XT)^{n-1}\,m\ \cdots$. If $n=2$, $m=1$, we have $G = TXT1\cdots$. The curling number of the first copy of T is the first term of X, which is not 1, but the curling number of the second T is 1, a contradiction. On the other hand, if $n \geq 3$, $G = TXTXT\cdots XTm\cdots$, and the initial $TXTX$ has curling number at least 2 and cannot be followed by T (which begins with 1), again a contradiction. ∎

We do not know if the theorem is still true if S is allowed to contain a 1 but does not end with 1.

6 Open questions and topics for future research

1. Is the curling number conjecture (even just for the case of sequences of 2's and 3's) true?

2. It would be nice to have some further exact values of $\Omega(n)$, beyond $n = 48$, even though they will require extensive computations.

3. What is the asymptotic behavior of $\Omega(n)$?

4. Can the especially good starting sequences shown in Tables 2 and 3 (in particular those of lengths 22, 48 and 77) be generalized? What makes them so special?

5. Can the properties of good starting sequences mentioned in Conjecture 2 be justified?

6. Can Shirshov's theorem (see §2.5) be modified so as to apply to our problem?

7. Are there analogs of Theorems 8 and 20 for $p(n, k)$ (the number of primitive sequences) or $p'(n, k)$ (the number of primitive and robust sequences) ?

8. Are there formulas for $c(n, k)$ that are more explicit than those given in Theorems 8 and 20? Is there a formula that matches the 200 known terms of the $c(n, 1)$ sequence?

9. Are there formulas or recurrences for the numbers $t(n, i)$ of starting sequences with tail length i?

10. Is there a probabilistic model that better explains the distribution of values of $t(n, i)$ visible in Tables 10–12 and A217209? The model presented in §4.2 is certainly inadequate.

11. Do "doubly rotten" sequence exist? (See Conjecture 22.)

12. The question implicit in the last sentence of §5.

7 Appendix: Proofs of Theorems 13, 14, 15

7.1 Theorem 13

Proof. The first statement follows immediately from Theorem 9, taking S and T in that theorem to be YX and X respectively. To prove the second statement, let $x := |X|$, $y := |Y|$, $s := |S|$, $t := |T|$, and note that $YX = STS$ implies $x + y = 2s + t$. Also $X = BY$ (say), with $B \neq \epsilon$.

We are to show that $s = y$ or $s > 2y$. First, suppose that $y < s \leq 2y$. Since $s > y$, there exists a sequence U with $|U| = s - y$ such that $S = YU$. Then we have the following chains of implications [the successive assertions are enclosed in square brackets]: $[s > y] \Rightarrow [s > y-$

$t] \Rightarrow [x = 2s + t - y > s] \Rightarrow [X \succ S \succ U]$, and $[s \le 2y] \Rightarrow [s - y \le y] \Rightarrow [|Y| \le |U|] \Rightarrow [Y \succeq U$ (since $YX = YBY = STYU)] \Rightarrow [Y = CU$ (say)$] \Rightarrow [X \succ S = CUU] \Rightarrow [\mathrm{cn}(X) > 1]$, a contradiction.

Second, suppose that $s < y$. Then there exists $U \neq \epsilon$ with $|U| = y - s$ such that $Y = SU$. But $x > y > s$ and $YX = STS$ imply $X \succ S$, and since $X \succ Y$ then $X \succ U$ also. If $S \succeq U$ then $YX = YBSUX \succ UU$, which contradicts $\mathrm{cn}(YX) = 1$. Hence $[U \succ S] \Rightarrow [s < y - s] \Rightarrow [2s < y] \Rightarrow [x + y = 2s + t < y + t \le y + x]$, since $X \succeq T$. Since this is impossible, $s < y$ is also impossible. $\square$

Note that the condition $|S| = |Y|$ is equivalent to $2|Y| > |X|$: if $s = y$ then $x = y + t$, which implies $2y = s + y > t + y$ (since $t > s$). Conversely, if $s \neq y$ then $s > 2y$, which implies $x + y = 2s + t > 4y + t$, $x > 3y + t$, so $x \ge 2y$. Similar reasoning shows that the condition $|S| > 2|Y|$ is equivalent to $3|Y| < |X|$.

7.2 Theorem 14

Proof. If $X \in \mathcal{A}(n, i) \setminus \mathcal{B}(n, i)$ then we may apply Theorem 13 to X, taking $Y = X^{[i]}$, with $|X| = n$, $|Y| = i$, where $n/2 < i < n$. So there exist S, T with $YX = STS$, $Y \succeq T$, $S \succ T$, and either $|S| = |Y|$ or $|S| > 2|Y|$. We cannot have $|S| > 2|Y|$, since that implies $|S| > n$, $2|S| > 2n > |YX|$, which contradicts $YX = STS$. So $|S| = |Y|$, $Y = S$, $X = TY$, and $|T| = n - i$. Also $\mathrm{cn}(YX) = 1$ by definition of $\mathcal{A}(n, i)$, i.e., $\mathrm{cn}(YTY) = 1$, so $Y \in \mathcal{B}(i, n - i)$.

The map from X to Y is one-to-one, since X determines Y. To show it is onto, take $Y \in \mathcal{B}(i, n - i)$, let $Q = Y^{[n-i]}$, and define P by $Y = PQ$ and set $X := QY = QPQ$. Then we have $\mathrm{cn}(YQY) = \mathrm{cn}(YX) = 1$, so $X \in \mathcal{A}(n, i)$. Also $XYX = QPQ\,PQ\,QPQ$ has curling number at least 2, so $X \notin \mathcal{B}(n, i)$. Hence $X \in \mathcal{A}(n, i) \setminus \mathcal{B}(n, i)$. $\square$

7.3 Theorem 15

Proof. Since the sets $\mathcal{E}$ in the sum are clearly disjoint, we just need to establish a bijection between the elements of $\mathcal{A}(n, i) \setminus \mathcal{B}(n, i)$ and the disjoint union of the $\mathcal{E}$ sets defined by the range of m.

As in the previous proof, if $X \in \mathcal{A}(n, i) \setminus \mathcal{B}(n, i)$, then we may apply Theorem 13 to X, taking $Y = X^{[i]}$, with $|X| = n$, $|Y| = i$, where now $1 \le i < n/3$. There exist S, T with $YX = STS$, $Y \succeq T$, $S \succ T$, and either $|S| = |Y|$ or $|S| > 2|Y|$. Let $|S| = m$, $|T| = n + i - 2m$. As before, $S \in \mathcal{B}(m, |T|)$. There are three conditions that m must satisfy: (i) $|T| \ge 1$ implies $m \le (n+i-1)/2$; (ii) $|S| > |T|$ implies $m > \lceil (n+i)/3 \rceil$; (iii) $m = i < n/3$ is incompatible with $YX = STS$, so $m > 2i$.

Since $m > i$, we may write $S = YU$, with $|U| = m - i$. Since $m > 2i$, $m - i > i$ and $|U| > |Y|$. Now $X \succ S$, so $X \succ U$ and therefore $U \succ Y$. Since $S = YU \in \mathcal{B}(m, |T|)$, $U \in \mathcal{E}(m - i, i, |T|)$. The mapping $X \mapsto U$ is one-to-one since X determines $Y = X^{[i]}$, S and T are unique by Theorem 10, $m = |S|$, and $S = YU$ determines U.

To show the map is onto, suppose $U \in \mathcal{E}(m-i, i, n+i-2m)$ for some m satisfying conditions (i)-(iii) above. Then set $Y = U^{[i]}$, $S = YU$, $T = S^{[n+i-2m]}$, and $X = UTS$. Then $YX = STS$ so that $U \in \mathcal{E}(m-i, i, n+i-2m)$ implies $YX \in \mathcal{A}(n, i)$. But $XYX = XSTS \succ TSTS$, so $\mathrm{cn}(XYX) > 1$ and therefore $XYX \notin \mathcal{B}(n, i)$. $\square$

8 Acknowledgments

We thank the referees for several helpful comments.

References

[1] J.-P. Allouche and J. Shallit, *Automatic Sequences: Theory, Applications, Generalizations*, Cambridge Univ. Press, 2003.

[2] F. J. van de Bult, D. C. Gijswijt, J. P. Linderman, N. J. A. Sloane, and A. R. Wilks, A slow-growing sequence defined by an unusual recurrence, *J. Integer Sequences*, **10** (2007), Article 07.1.2.

[3] B. Chaffin and N. J. A. Sloane, The curling number conjecture, http://arxiv.org/abs/0912.2382.

[4] N. J. Fine and H. S. Wilf, Uniqueness theorems for periodic sequences, *Proc. Amer. Math. Soc.*, **16** (1965), 109–114.

[5] E. N. Gilbert and J. Riordan, Symmetry types of periodic sequences, *Illinois J. Math.*, **5** (1961), 657–665.

[6] A. Lentin and M. P. Schützenberger, A combinatorial problem in the theory of free monoids, in R. C. Bose and T. A. Dowling, eds., *Combinatorial Mathematics and its Applications*, Univ. North Carolina Press, Chapel Hill, 1969, pp. 128–144.

[7] M. Lothaire, *Combinatorics on Words*, Addison-Wesley, 1983.

[8] A. de Luca and S. Varricchio, *Finiteness and Regularity in Semigroups and Formal Languages*, Springer, 1998.

[9] The OEIS Foundation Inc., *The On-Line Encyclopedia of Integer Sequences*, http://oeis.org.

2010 *Mathematics Subject Classification*: Primary 68R15; Secondary 11B37.
Keywords: curling number, Gijswijt sequence, Fine-Wilf theorem.

(Concerned with sequences A027375, A090822, A122536, A135491, A160766, A216730, A216813, A216950, A216951, A216955, A216955, A217208, A217209, A217437, A217943, A218869, A218870, A218875, and A218876.)

Received December 25 2012; revised version received March 12 2013. Published in *Journal of Integer Sequences*, March 16 2013.

Return to Journal of Integer Sequences home page.