Self Referencing Sequences

Perry Kleinhenz, Fermi Ma, and Erik Waingarten

{pkleinhe,fermima,eaw}@mit.edu

1 Introduction

Written by Fermi Ma, edited by Perry Kleinhenz and Erik Waingarten

1.1 Problem Setup

Consider the sequence

$$1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 2, \ldots \tag{1}$$

We can break up the sequence into contiguous blocks, where each block is a stretch of repeated numbers. The block lengths, read from left to right, reproduce the original sequence (see Figure 1.1). Thus, this sequence is *self-referential*.

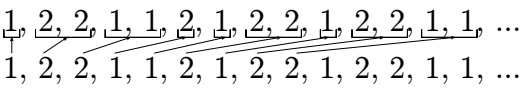

Fig. 1. Sequence broken up into contiguous blocks with lengths generating same sequence.

The sequence 1 contains the numbers 1 and 2. In general, we want to make self-refering sequences containing other numbers as well. Once there is more than two numbers in the sequence, generating the sequence becomes a bit ambiguous. Take the simple example of a self-referencing sequence with numbers 1, 2, and 3.

If the first number is 1, then we can make the second number either a 2, or a 3. In particular, both:

$$1, 3, 3, 3, 2, 2, 2, \ldots \tag{2}$$

$$1, 2, 2, 1, 1, 3, 1, \ldots \tag{3}$$

are in some sense "self-referencing" sequences. Each time we need to specify a new block, we have a choice of two elements. We would like to avoid having a large number of choices in the generation of the sequence, and so we require that the sequence contain elements from a *cycle*, and each time we start another block, the block will contain the next number in the cycle. In the case of sequence 1, the cycle is $(1, 2)$.

The self-referencing sequence with cycle $(1, 2, 3)$ that begins with 1 is

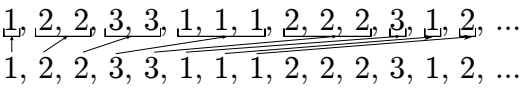

Fig. 2. Sequence generated by $(1, 2, 3)$ broken up into contiguous blocks.

Formally, we say that a sequence is *generated* by the set of numbers used in the sequence. So sequence 1 is generated by the set $\{1, 2\}$. We will refer to the cycle used in the sequence as the *generating cycle*.

Using the cycle, it is no longer ambiguous which number to use in the next block.

1.2 Overview

We address a variety of questions regarding these sequences. Our main results are:

- The self-referencing sequence of a given generating cycle is uniquely determined by its first number.
- Some characterizations of sequences with the same generating cycle.
- There exists an equivalent formulation of self-referencing sequences using substitution rules for some criteria of the generating cycles.
- One can compute the limiting density for the numbers in the sequence if a limiting density exists, for generating cycles that have substitution rules.
- The limiting densities of the self-referencing sequences of generating cycles $(1, 3)$ if it exists.

In Section 2 we show that if we are given a generating set and a starting number for the sequence, the sequence is uniquely determined. We consider how the starting number for the sequence affects the long term behavior of the sequence. In particular, in Section 3, we show that there exist limits on how much two sequences with the same generating set can resemble each other if they start with different numbers. In Section 4, we consider the limiting behavior of such sequences, and we conjecture that the limiting behavior of a sequence is determined only by the generating set, and not the starting number.

In Section 5, we take a different approach to analyzing these sequences by showing that there is an alternate way of generating them with an iterative process. In Section 6, we show how a large number of these sequences can actually be generated with a simple set of substitution rules. We introduce the density problem in Section 7, and use ideas developed in Section 5 and Section 6 to answer the question for certain types of generating sets. Unfortunately, we find that the open problem of determining the density for the sequence of 1's and 2's is hard to solve using our methods.

2 Determinism

Written by Fermi Ma, edited by Perry Kleinhenz and Erik Waingarten

We claim that a self-referencing sequence is uniquely determined by the generating set and the starting number.

Lemma 1. *Let C be a generating cycle. Suppose (a_i) and (b_i) are both self-referencing sequences with $a_j \neq b_j$ and j is the smallest such number, then $j = 0$.*

Proof. Suppose $j > 0$, then $a_{j-1} = b_{j-1}$ since j was assumed to be the smallest index where (a_i) and (b_i) are different. Suppose $a_{j-1} \neq a_j$ and $b_{j-1} \neq b_j$, then we can ask what comes directly after a_{j-1} is the generating cycle; by looking at sequence (a_i), we get that a_j comes after a_{j-1}, but by looking at the sequence (b_i), we get that b_j comes after a_{j-1}. This is a contradiction since $a_j \neq b_j$. So lets assume without loss of generality that $a_{j-1} = a_j$, and also, we will assume that a_{j-1} belongs to the kth block. Since the sequences are the same up to index j, b_{j-1} belongs to the kth block. However, this means that $a_k \geq b_k + 1$. So $k \geq j$ and k designates the block length of the block containing a_j, and so $k \leq j$, so $k = j$. Figure 2 gives a visual representation of these sequences.

This implies that every block before the kth block just contains 1's since each block refered it itself. This means that either $j = 2$, or the sequence is not self-referential. But if $j = 2$ and $a_1 = b_1 = 1$, then $a_2 \neq b_2$, which means both a_2 and b_2 comes after 1 in the generating cycle. □

We present an algorithm that produces the unique self-referencing sequence given a generating cycle and a starting element. Assume the generating cycle is $(a_1, a_2, \ldots, a_m)$ and that the starting element is some a_i. For notational simplicity, if we ever write down a_k where $k > m$, we set this equal to $a_{k(\mathrm{mod}\ m)}$.

hb

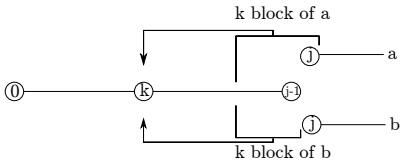

Fig. 3. Self-referencing sequences a and b which differ first at the jth element.

Algorithm:
Suppose the first number of the sequence is a_i. Then the first block of the sequence must be a block of a_i instances of a_i, so we extend the sequence to this block. If a_i is not 1, then there is a second element of that block. We read that second element, and then we add a block of a_i instances of a_{i+1} to the end of that sequence. If a_i is equal to 1, we write down a_{i+1} in the second spot, and use the value of a_{i+1} to dictate the length of the second block.

This procedure continues deterministically so long as there is an element to read. Every time we write down a block, we read 1 element and write down at least 1 element, thus the end of the sequence will always be at or after the spot we are reading.

From this point on, we will denote by $s(C, a)$ the unique sequence with generating cycle C and with starting number $a \in C$.

3 Prefixes

Written by Perry Kleinhenz, edited by Fermi Ma and Erik Waingarten

Definition 1. *We say that two sequences $\{a_i\}$ and $\{b_i\}$ differ by a prefix of length n if*

$$a_{n+i} = b_i,$$

for i any positive integer. We say that two sequences are independent if they do not differ by a prefix for any n, in either order.

Note that a pair of sequences can differ by multiple prefixes of different length. For example if we have

$$\{a_i\} = \{1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, \ldots\}$$
$$\{b_i\} = \{7, 8, 9, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, \ldots\},$$

then a_i and b_i differ by prefixes of length $3 + 6k$ for k a nonnegative integer. Because of this we make the following definition.

Definition 2. *We say that two sequences $\{a_i\}$ and $\{b_i\}$ differ by a minimal prefix of length n if they differ by a prefix of length n, but do not differ by a prefix of length m for all $0 < m < n$.*

We note that if a pair of sequences differs by a prefix they differ by a minimal prefix. In our above example a_i and b_i differ by a minimal prefix of length 3. The following lemma is a direct consequence of our algorithm for generating self-referential sequences:

Lemma 2. *Let $G = \{1, g_1, g_2, \ldots, g_n\}$, $s(G, 1)$ and $s(G, g_1)$ differ by a minimal prefix of length at most 1.*

Proof. Let a_i refer to the ith term of $s(G,1)$ and b_i refer to the ith term of $s(G,g_1)$. If we use our algorithm for generating the sequence, we note that $a_1 = 1$, and so $a_2 = g_1$, and that this is the beginning of a new block of length g_1. Furthermore, $b_1 = g_1$ and it indicated the beginning of a block of length g_1. Note that our algorithm is "memory-less" in the sense that it does not maintain any state about previous elements. Because the number and the begining of the block is the same. The algorithm will generate the same sequence, shifted by 1. Therefore, $s(G,1)$ and $s(G,g_1)$ differ by a minimal prefix of length at most 1. □

We now give a

Theorem 3. *Let $G = \{1, a, b\}$ where $1 < a < b$. $s(G,a)$ and $s(G,b)$ are independent.*

Proof. We proceed by contradiction. If the two sequences are not independent then they must differ by some minimal prefix of length n. If we let a_i be the sequence beginning with a and b_i be the sequence beginning with b then

$$a_{n+k} = b_k.$$

We know that the first b terms of b_i are b's. Because $b > a > 1$ no block can be longer then b. Thus the terms immediately preceding and following this block in a_i cannot be b's. Because of this the first block of b_i is the $m+1$th block of a_i, for some m a positive integer. Since the value of the sequence at position k is the length of the kth block this means that $a_{m+1} = b_1$. In fact the kth block of b_i is the $(m+k)$th block of a_i and so

$$a_{m+k} = b_k$$

We now claim that the prefix has length strictly larger than m. Assume otherwise so $n \leq m$. This means that a_{m+1} which gives the length of the $m+1$th block occurs at or after the beginning of the $m+1$th block, which begins at position a_{n+1}. By our proof that self-referential sequences are well defined we know that the second case cannot occur. Therefore we must have the $m+1$th block start at position $m+1$. This would require every block to have length exactly 1, but this is a contradiction because the first block has length $a_1 = a > 1$. Therefore the prefix must have length strictly larger than m.

We note that based on a previous step the two sequences differ by a prefix starting at m. Since $m < n$ we have produced a prefix which starts before n which is a contradiction. Therefore the two sequences are independent. □

4 Equivalence

Written by Fermi Ma and Perry Kleinhenz, edited by Erik Waingarten

In this section, we consider a slightly weaker definition of equivalence. Consider, for example, sequences generated by $G = \{1, 2, 3\}$. Theorem 3 shows that the sequence starting with 2

$$2, 2, 3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \ldots$$

and the sequence starting with 3

$$3, 3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \ldots$$

are independent in the sense that neither sequence is exactly contained in the other. However, if we delete the first two numbers of the sequence starting with 2, and the first number of the sequence starting with 3, we get

$$3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \ldots$$

in both cases. Lets look at the state of the algorithm that generates both sequences. After the first block, the execution of the algorithm of $s(G,2)$ right after placing the first 3's (in the 3rd and 4th position), the algorithm is in state corresponding to:

– block size 3
– no elements in the current block
– starting number is 1

. If we look at the execution of the algorithm that generates $s(G, 3)$ after placing the first three 3's, the algorithm is in state corresponding to:

– block size 3
– no elements in the current block
– starting number is 1

Therefore, since the states are the same, the sequence generated after the 4th position of $s(G, 2)$ and after the $3rd$ position of $s(G, 3)$ will be the same. By inspection, we see we can also include two of the 3's.

Thus, since both sequences are essentially the same if a prefix is removed from each of them, we call these sequences *equivalent*. To formalize, this, we say that two sequences $\{a_i\}$ and $\{b_i\}$ are equivalent if there exists some integers $n, k \geq 0$ such that

$$a_{n+i} = b_{k+i}$$

We now give the following conjecture

Conjecture 1. Any sequence generated by a set S is equivalent to any other sequence generated by the same set S.

This conjecture is partly motivated by the fact that the property holds for $\{1, 2, 3\}$. For generating sets with larger numbers, we have been unable to verify or refute the conjecture. However, in Section 7, we show that for some generating sets the starting number of a sequence does not affect the densities of each number, which provides some evidence for this conjecture.

We now introduce a notion which will help us categorize sequences as equivalent.

Definition 3. *Consider a sequence a. We define the read distance for an element a_i as the number of terms in between it and the beginning of the ith block of numbers.*

The read distance can be any nonnegative integer. For example if we consider the sequence a_i generated by $\{1, 2, 3\}$ beginning with a 2

$$2, 2, 3, 3, 1, 1, 1, 2, 2, 2, 3, 1, 2, 3, 3, 1, 1, 2, 2, 3, 3, 3, \ldots$$

the read distance for a_1 (the first two) is 0 as the first block of numbers starts at the first number, the read distance for a_2 (the second two) is 1 as the second block of numbers starts at the third number. We can continue to calculate the read distances for this sequence and see that they form a monotonically increasing sequence $0, 1, 2, 4, 6, 6, 6, 6, 7, 8, 9, \ldots$. From this we might suspect that the read distance for any self referencing sequence will also be monotonically increasing. We can in fact compute the read distance exactly.

Lemma 4. *Let a_i be any self referencing sequence generated by some set of positive integers. Then the read distance for the element a_i, for $i > 1$, is*

$$\sum_{k=1}^{i-1} (a_k - 1)$$

Proof. We will prove this using induction. We know that for $i = 1$ the read distance is 0, as the first block always starts at the first number. When $i = 2$ we must show that the read distance is $a_1 - 1$. The first block will be of length a_1 since the sequence is self referencing, therefore the second block will begin at position a_1. The number of elements between position a_1 and 2 is exactly $a_1 - 1$.

Now if we assume that our result holds for $i = l$ we would like to show that it holds for $i = l + 1$. We know that the distance between the start of the lth block and position l is

$$\sum_{k=1}^{l-1}(a_k - 1).$$

This lth block has length a_l and so the $l + 1$th block will begin a_l positions to the right of the beginning of the lth block. Also note that a_{l+1} is one position to the right of a_l. Therefore the beginning of the $(l + 1)$th block is

$$a_l - 1 + \sum_{k=1}^{l-1}(a_k - 1) = \sum_{k=1}^{l}(a_k - 1)$$

positions to the right of a_{l+1}. But this is exactly the reading distance for a_{l+1} and so we have our result. □

Note that because a_i is generated by positive integers we always have $a_k - 1 \geq 0$ and so the read distance is monotonically increasing. Although this shows us that read distance never decreases it does not tell us how quickly it increases, our next result gives a simple bound on the number of consecutive times the read distance can stay constant.

Lemma 5. *Consider a self referencing sequence x generated by $\{a_1, a_2, \ldots, a_n\}$ with $a = \max(a_1, a_2, \ldots, a_n)$. Suppose the read distance for x_i is d, then the read distance for x_{i+a} is at least $d + 1$.*

Proof. Assume otherwise. By Lemma 4 we know that the read distance for x_{i+a} is at least d. Thus it must be exactly d. Lemma 4 also tells us that if this is the case then $x_{i+k} = 1$ for $k = 0, 1, 2, \ldots, a_n$. But this means that we have a block of length $a + 1$, which is a contradiction of the self referencing property of the sequence. Therefore the read distance for x_i is at least $d + 1$. □

Theorem 6. *If two sequences are equivalent then the first element of the shared sequence must have the same read distance in both sequences.*

Proof. Assume otherwise. So if the two sequences are a_j and b_j and we have

$$a_{i+l} = b_{i+k} = c_i$$

such that if m or l were made smaller this relation would not hold. Suppose the read distance for b_{1+k} is x and the read distance for a_{1+l} is y. Without loss of generality let us assume that $x > y$. Therefore we know there are two separate blocks of length c_1 in the sequence, one which corresponds to a_{i+l} and another which corresponds to b_{i+k}. Because the two sequences are identical there exists some index n such that $c_{n+1} = a_m$ is the element of sequence a which corresponds to the second block of length c_1. Because the sequences are self referential we know that $c_1, c_2, \ldots, c_n$ must appears in b prior to element $i + k$. In fact the self referential nature of the sequence guarantees that $b_{i+k-1} = c_n$, that is this subsequence must end just before the two sequences become the same. Thus our sequence b is

$$\cdots c_1 c_2 c_3 \cdots c_n c_1 c_2 \cdots c_n \cdots$$

where the prefix begins at the second c_1. We now claim that the common sequence cycles through

$$c_1 c_2 \cdots c_n.$$

Well by Lemma 4 since the read distance for b_{1+k} is greater than the read distance for a_{1+l}, and the sequences are identical for subsequent indices, we know that the read distance for b_{i+k} is longer than the read distance for a_{i+l} by exactly $x - y$. Thus b_{k+i} corresponds to the same block as $a_{l+(x-y)+i}$. But $b_{k+i} = c_i$ and $a_{l+(x-y)+i} = c_{x-y+i}$ and so $c_i = c_{x-y+i}$ and so the common sequence cycles. Because the common sequence cycles and b has the same cycle, but one which begins before the common sequence l could be made smaller and have the equivalence relation hold. This is a contradiction and so the first element of the shared sequence must have the same read distance in both sequences. □

We also have a sufficient condition for equivalence

Lemma 7. *If two sequences a_i, b_i are such that for some integers m, l the read distance for a_m and b_l are the same and*

$$a_{m+k} = b_{l+k}$$

where k ranges from 0 to the read distance for a_m then a_i and b_i are equivalent.

Proof. We will prove this using a simple induction. Suppose the read distance for a_m is d then know that $a_{m+d} = b_{l+d}$ because $a_{m+d-1} = b_{l+d-1}$ and a_{m+d} is the beginning of a new block. Furthermore we have that the read distance for a_{m+d} equals the read distance for b_{l+d}. Now suppose that we have for $k \leq j$ that $a_{m+k} = b_{l+k}$ and the read distance for a_{m+k} equals the read distance for b_{l+k} and we would like to show that this holds for $k = j + 1$. By Lemma 4 we actually already know that the read distance for the two is the same, as it is the same for the preceding terms, which are equal.

If a_{m+k} is the last term in a block then b_{m+k} must also be the last term in its block, because the length of blocks is determined by earlier terms which we know to be equal and to have the same read distance. Since $a_{m+k} = b_{l+k}$ in this case we have $a_{m+k+1} = b_{l+k+1}$.

Otherwise a_{m+k} and a_{m+k+1} are in the same block, then we only need to show that b_{l+k} and b_{l+k+1} are in the same block as well. But as we argued above the length of blocks is determined by terms earlier in the sequence which we know to be equal and have the same read distance. Therefore we have $a_{m+k+1} = b_{l+k+1}$ for this case as well. This completes our induction and so a_i and b_i are equivalent. □

These conditions for equivalence lead us to make the following conjecture on a sufficient condition for nonequivalence.

Conjecture 2. If two sequences a_i and b_i have elements $a_k \neq b_l$, for $k, l > 1$ with the same read distance then they are not equivalent.

5 An Iterative Process

Written by Fermi Ma, edited by Perry Kleinhenz and Erik Waingarten

We change focus in this section and consider how to generate self-referential sequences with an iterative process. We first look at the sequences generated by $\{1, 2\}$.

One possible iterative process is as follows. Let the sequence s^0 be 2, and let the sequence $s^{(i)}$ for $i \geq 1$ be the unique sequence that starts with 2 and is such that its block lengths, read from left to right, reproduce the sequence $s^{(i-1)}$. This gives:

$$\begin{aligned} s^{(0)} &= 2 \\ s^{(1)} &= 22 \\ s^{(2)} &= 2211 \\ s^{(3)} &= 221121 \\ s^{(4)} &= 221121221 \\ &\vdots \end{aligned}$$

Let s^∞ be the limit of this process. By limit, we mean the limit with respect to finitely many terms $s_j^{(i)} \to s_j^\infty$ as $i \to \infty$. We claim that s^∞ is the unique self-referential sequence $s(\{1, 2\}, 2)$ generated by $\{1, 2\}$ with starting number 2. In fact we can prove a more general result about this iterative process.

Theorem 8. *Let $s^0 = a$ for some $a \in A$ where $a \neq 1$, and let $s^{(i)}$ for $i \geq 1$ be the unique sequence that starts with a and is such that its block lengths, read from left to right, reproduce the sequence $s^{(i-1)}$. Then the limit of this process, s^∞, is the sequence $s(A, a)$.*

Proof. We prove this with induction. We show that if $s^{(i)}$ is a prefix of $s(A,a)$, then $s^{(i+1)}$ is a longer prefix of $s(A,a)$. The base case of the induction is satisfied, as $s^{(0)}$ is a prefix of $s(A,a)$.

For the inductive step, we observe that a prefix of the sequence describes the block lengths of a *longer* prefix of the sequence. We know that the sequence it describes is longer because the sequence it describes has length equal to the sum of the original prefix. In other words, the length of $s^{(i+1)}$ equals the sum of of the numbers in $s^{(i)}$, which is strictly greater than the length of $s^{(i)}$ (since there are numbers that are greater than 1).

The resultant sequence is a prefix of the self referencing sequence because if we apply our process to any prefix of the self referencing sequence it will produce a prefix of the self referencing sequence. □

Note that we must specify that $a \neq 1$, because that is the only case in which the length of $s^{(i+1)}$ is not strictly greater than the length of $s^{(i)}$.

It turns out that we can be less restrictive about what our starting sequence is. Going back to the case of $\{1,2\}$, we can let $s^{(0)}$ be *any* sequence of 1's and 2's that starts with a 2. For example, suppose $s^{(0)} = 2,1,1,1,2$. The rules for generating $s^{(i)}$ give:

$$\begin{aligned}
s^{(0)} &= \mathbf{2},1,1,1,2 \\
s^{(1)} &= \mathbf{2},\mathbf{2},\mathbf{1},2,1,2,2 \\
s^{(2)} &= \mathbf{2},\mathbf{2},\mathbf{1},\mathbf{1},\mathbf{2},\mathbf{1},1,2,1,1,2,2 \\
s^{(3)} &= \mathbf{2},\mathbf{2},\mathbf{1},\mathbf{1},\mathbf{2},\mathbf{1},\mathbf{2},\mathbf{2},\mathbf{1},\mathbf{2},1,1,2,1,2,2,1,1 \\
&\vdots
\end{aligned}$$

Here, we bold the numbers that form a prefix of $s(\{1,2\},2)$. We note that $s^{\infty} = s(A,a)$ because because $s^{(0)}$ contains a 2 at the beginning, which by the above theorem will produce $s(A,a)$ in the limit. We call the rest of each $s^{(i)}$ that is not a prefix of $s(1,2,2)$ the *tail*.

We now consider two different forms of convergence. We know that $s^{(i)}$ in some sense approaches $s(A,a)$, as arbitrarily many elements at the beginning of the sequence will match up with the beginning of $s(A,a)$ as i increases. However, if the tail of the sequence never entirely disappears, then there does not exist an i where $s^{(i)}$ is exactly a prefix of $s(A,a)$. Thus, we formalize two types of convergence

Definition 4. *We say that a sequence s^i is weakly convergent to $s(A,a)$ if $s^{\infty} = s(A,a)$. We say that the sequence s^i is strongly convergent if $s^{(i)}$ is a prefix of $s(A,a)$ for all $i \geq k$ for some k.*

It is unclear what conditions must be placed on the starting sequence in order to ensure strong convergence, although based on numerical simulation we make the following conjecture.

Conjecture 3. A sequence s^i is strongly convergent to $s(\{1,2\},2)$ if the starting sequence begins with a 2 and contains no more than 3 elements. A sequence s^i is strongly convergent to $s(\{1,2\},1)$ if the starting sequence begins with a 1 and contains between 2 and 3 elements. A sequence s^i is weakly convergent but not strongly convergent to $s(\{1,2\},a)$ if the starting sequence begins with an $a \in \{1,2,\}$ and contains 4 or more elements.

We make this conjecture on the basis on computer simulations which executes the iterative process to at least 20 iterations. More iterations were not completed because after this point the simulations began to take impractically long amounts of time to complete.

6 Substitution Rules

Written by Fermi Ma and Erik Waingarten, edited by Perry Kleinhenz

The iterative process given in Section 5 is one way to generate self-referencing sequences. However, for certain generating sets A, there is a simpler ways to perform this iteration using a

fixed set of substitution rules. Take $A = \{1, 3\}$ as an example, and consider the sequence that starts with a 3:

$$3331113331313331111333\ldots$$

We can use the iterative process from Section 5, and suppose that $s^{(0)} = 33$. We get:

$$\begin{aligned}
\{s^{(0)}\} &= 3, 3 \\
\{s^{(1)}\} &= 3, 3, 3, 1, 1, 1 \\
\{s^{(2)}\} &= 3, 3, 3, 1, 1, 1, 3, 3, 3, 1, 3, 1 \\
\{s^{(3)}\} &= 3, 3, 3, 1, 1, 1, 3, 3, 3, 1, 3, 1, 3, 3, 3, 1, 1, 1, 3, 3, 3, 1, 3, 3, 3, 1 \\
&\ldots
\end{aligned}$$

It turns out that we can model this evolution with the following substitution rules:

$$\begin{aligned}
&\text{Rule } A : 3, 3 \to 3, 3, 3, 1, 1, 1 \\
&\text{Rule } B : 3, 1 \to 3, 3, 3, 1 \\
&\text{Rule } C : 1, 1 \to 3, 1
\end{aligned}$$

This is to be interpreted as follows. Starting from 3,3, the only possible rule to apply is Rule A, which gives $3, 3, 3, 1, 1, 1$ at the following iteration. We then apply Rule A to $3, 3$, Rule B to $3, 1$ and Rule C to $1, 1$. Note that at every step, we break up the sequence into chunks of 2 and then simultaneously apply *all* the rules. This will always be possible, since the rules preserve the fact that these sequences have even length. So at the following step, we have 3,3,3,1,1,1,3,3,3,1,3,1, which is $s^{(2)}$. Notice that in general, after the ith application of the substitution rules, the sequence is equal to $s^{(i)}$.

Why does this work? Well, first of all, notice that the rules are the natural applications of the self-referential sequence. Also, note that we have decided to make the rules of length 2; if we had made the self-referential rules $1 \to 1$ and $3 \to 333$, this would not have worked. In addition, we know that each block length preserves parity.

We would like to know if we can do this for any generating set G and starting number g. In particular we would like to be able to simulate our original algorithm with a substitution rule. Previously we established that our algorithm had a state corresponding to

- current block size
- number of elements already written down in the current block
- what number the current block is composed of

When we apply our substitution rules we should not need to know what happened to blocks other than the one we are currently substituting for. Therefore, each rule must somehow encode the length of blocks, the number of elements already written in the current block, and what number the current block is composed of. The first condition is simple to implement, as the rules will specify the length of blocks. In order to keep track of the numbers already written, each rule should be self-contained. That is each rule should write complete blocks. In order to comply with the third condition it is enough for the size of the generating set to divide the length of the rules. This is so that at the start of each rule, we always know which numbers to write.

Therefore, we can formulate a set of sufficient conditions to make rules to generate the self-referential sequence with generating set G. We want to have a set of rules R where each rule $r \in R$ has length $|r| = l$, such that all $|G|^l$ strings are in the rule set and for each $r \in R$

- each rule writes blocks in the same order of G
- $\sum_{i \in r} i$ is divisible by l
- l is divisible by $|G|$

Given these conditions on every rule, we can guarantee that the self-referential sequence is generated. Note that in this case, the above example does not satisfy our rules until we say that $1, 3 \to 3, 1, 1, 1$.

Another example, we can have $G = \{2, 4\}$, and we can have that the rules be

$$\text{Rule } A : 2, 2 \to 2, 2, 4, 4$$
$$\text{Rule } B : 2, 4 \to 2, 2, 4, 4, 4, 4$$
$$\text{Rule } C : 4, 2 \to 2, 2, 2, 2, 4, 4$$
$$\text{Rule } D : 4, 4 \to 2, 2, 2, 2, 4, 4, 4, 4$$

One of the reasons why generating a rule set for $\{1, 2\}$ is complicated is that since $1, ..., 1$ must be in the rule set, the length of the rules must be even (since $|G| = |\{1, 2\}| = 2$). On the other hand, $2, 1, ..., 1$ must also be in the set and so the rule length must be odd.

7 Density

Written by Erik Waingarten, edited by Perry Kleinhenz and Fermi Ma

A topic of interest when considering a self referencing sequence is the relative density of the various elements of the generating set. It is unclear if such a density even exists and if so how to compute it for arbitrary generating sets. In this section, we show how to use substitution rules to compute relative densities for certain sequences if we assume such densities exists. We explain the method and compute the relative density of 3's for $G = \{1, 3\}$. We will also produce bounds on the density of 2's $G = \{1, 2\}$ in a way that suggests that the density converges to 2; however, this method is computationally difficult for smaller bounds since the number of cases to check increases exponentially.

To begin we will discuss a technique that can be used to compute the densities for some generating sets. In particular, suppose we have a generating set G with a set of rules R that satisfy the conditions from the section above. For each rule $r \in R$, we will say denote the rule as $r_l \to r_r$ as the inputs and the outputs.

Now we can count the fraction of times r_r appears in the sequence at each step. Since we know the rules, these will become a set of $|R|$ equalities that reference each other. We can take their ratios and solve for the fraction of times that each r_r appears and compute the density by assuming a stable state.

For example, lets look at the rules we have determined for $G = \{1, 3\}$.

$$\text{Rule } A : 3, 3 \to 3, 3, 3, 1, 1, 1$$
$$\text{Rule } B : 3, 1 \to 3, 3, 3, 1$$
$$\text{Rule } C : 1, 1 \to 3, 1$$

Then we can let $3, 3, 3, 1, 1, 1 = A$, $3, 3, 3, 1 = B$ and $3, 1 = C$. If A_i is the number of occurrences of A in $s^{(i)}$, then

$$A_i = B_{i-1} + A_{i-1}$$

Likewise, we have

$$B_i = A_{i-1} + B_{i-1} + C_{i-1}$$
$$C_i = A_{i-1}$$

So now we can take the fraction by dividing by the total number of them:

$$A'_i = \frac{B'_{i-1} + A'_{i-1}}{3A'_{i-1} + 2B'_{i-1} + C'_{i-1}}$$

$$B'_i = \frac{A'_{i-1} + B'_{i-1} + C'_{i-1}}{3A'_{i-1} + 2B'_{i-1} + C'_{i-1}}$$

$$C'_i = \frac{A'_{i-1}}{3A'_{i-1} + 2B'_{i-1} + C'_{i-1}}$$

So if we assume a density, there will be a stable state which will be the density, so in the case where $A'_i = A'_{i-1}, B'_i = B'_{i-1}, C'_i = C'_{i-1}$, we can solve for the fractional appearance of each and get the density of the numbers.

We can compute solutions to this set of quadratic equations and find the density of 3's is approximately 0.602.

As discussed in the previous section, it is unclear how to establish a set of rules for $G = \{1, 2\}$; however, there is some structure that we can exploit. In particular, we know that the sequences $1, 1, 1$ and $2, 2, 2$ will never appear in the self-referential sequence. This is because such sequences would be part of a block of size at least 3.

What does this mean? Well, we know that if there is a density of 1's and 2's, then that density d_2 is bounded between

$$\frac{1}{3} \leq d_2 \leq \frac{2}{3}$$

In general, this gives a procedure for computing the limiting density. For example, now we know that there must be at least one 2 for every three numbers. So now we can look at how many 2's are in any given string of length 9. We know there must be at least three 2's. There are only three cases where there are three 2's in a segment of 9 numbers in the sequence, namely

$$2, 1, 1, 2, 1, 1, 2, 1, 1$$

$$1, 2, 1, 1, 2, 1, 1, 2, 1$$

$$1, 1, 2, 1, 1, 2, 1, 1, 2$$

In each of these cases, we can read the subsequence that must have generated them part of them. In order, they are

$$2, 1, 2, 1, 2$$

$$1, 2, 1, 2, 1$$

$$2, 1, 2, 1, 2$$

but each of these are generated by $1, 1, 1$. This means that none of the sequences of length 9 are possible, so in addition to having one 2 for every group of 3, there is another 2 for every group of 9. Therefore, we have shown that the density must be bounded by

$$\frac{1}{3} + \frac{1}{9} \leq d_2 \leq 1 - \frac{1}{3} - \frac{1}{9}$$

In general, we can continue this process, and we did, to show that

$$\frac{1}{3} + \frac{1}{9} + \frac{1}{27} \leq d_2 \leq 1 - \frac{1}{3} - \frac{1}{9} - \frac{1}{27}$$

Of course, the number of cases to check increases very quickly: there are 3 choices to put the 2 in the groups of 3, 3 choices to put the 3 groups of 2's and 2 positions within these groups for the groups of 9, and so on. Its not hard to see that this number grows exponentially fast, making it computationally difficult to analyze more cases.

We believe that if there is a way to show this, we will be able to show that the density is $\frac{1}{2}$, since the density will be bounded below by $\sum_{n=1}^{\infty} \frac{1}{3^n} = \frac{1}{2}$ and bounded above by the same number. However, generalizing the procedure to improve the bound seems difficult.

8 Unresolved Problems

Written by Fermi Ma, edited by Perry Kleinhenz and Erik Waingarten

Our work leaves a number of unresolved problems.

- In Section 4, we propose the idea that in some sense, the number that we start a sequence with does not matter. In other words, any sequence that uses a certain generating set is equivalent to any other sequence created by the same generating set.

- In the same section, we propose a condition for determining when two sequences are not equivalent. We conjecture that if two sequences have unequal numbers with the same read distances, then they are not equivalent. It seems from trying cases that this condition is sufficient, but we have been unable to prove it.
- In Section 5 we make a conjecture on when sequences are strongly and weakly convergent. We have not yet been able to investigate strong and weak convergence in great detail, as we have been unable to make precise mathematical statements about these types of convergence. Most of what we know about them comes from analyzing numerical data.
- In Section 6, we look at substitution rules for certain "nice" generating sets of numbers, such as $\{1, 3\}$ and $\{2, 4\}$. These sets allow for substitution rules because they satisfy the trio of properties we outline in that section. However, we do not know that these are the only properties. We leave unresolved the question of determining precisely which sets allow for substitution rules and which ones ones do not. The major difficulty here seems to come in proving that a certain generating set does *not* allow for substitution rules.
- In Section 7 we were able to compute the density of any sequence with generating set $\{1, 3\}$, and we were able to give bounds on the density of any sequence with generating set $\{1, 2\}$. However, we leave unresolved the question of what the density for $\{1, 2\}$ actually is, or even if we can obtain tighter bounds.

9 Haikus

Written by Fermi Ma

One, Two, Two, One, One,
Two, One, Two, Two, One, Two, Two,
One, One, Two, One, One...

For the final draft,
we will improve the haikus
(and the whole paper).