Discrete Mathematics and Theoretical Computer Science **(subm.)**, by the authors, 26–rev

On the Ehrenfeucht-Mycielski Balance Conjecture

John C. Kieffer[1†] and W. Szpankowski[2‡]

[1] *Dept. of Electrical & Computer Engr., University of Minnesota, 200 Union St. SE, Minneapolis, MN 55455, USA*
[2] *Dept. of Computer Science, Purdue University, 305 N. University St., West Lafayette, IN 47907, USA*

In 1992, A. Ehrenfeucht and J. Mycielski defined a seemingly pseuorandom binary sequence which has since been termed the EM-sequence. The balance conjecture for the EM-sequence, still open, is the conjecture that the sequence of EM-sequence initial segment averages converges to $1/2$. In this paper, we do not prove the balance conjecture but we do make some progress concerning it, namely, we prove that every limit point of the aforementioned sequence of averages lies in the interval $[1/4, 3/4]$, improving the best previous result that every such limit point belongs to the interval $[0.11, 0.89]$. Our approach is novel and exploits an analysis of the growth behavior as $n \to \infty$ of the rooted tree formed by the binary strings appearing at least twice as substrings of the length n initial segment of the EM-sequence.

1 Introduction

In the paper Ehrenfeucht and Mycielski (1992), an interesting binary sequence was defined, since termed the EM-sequence, which seems to possess pseudorandomness properties. The EM-sequence is sequence A038219 in the encyclopedia Sloane (2007), and is generated via an algorithm described in Sloane (2007) as follows: “The sequence starts 0,1,0 and continues according to the following rule: find the longest sequence at the end that has occurred at least once previously. If there are more than one previous occurrences select the last one. The next digit of the sequence is the opposite of the one following the previous occurrence.” For example, the first 50 terms of the EM-sequence are

$$01001101011100010000111101100101001001110100011000.$$

Despite the simplicity of this algorithm, not very much is known about the asymptotics of the EM-sequence. It is natural to conjecture that the EM-sequence behaves as a typical sequence generated by a binary IID process. In particular, we would expect that the averages of the initial segments of the EM-sequence converge to $1/2$; this is called the *balance conjecture*. The balance conjecture remains open, although various asymptotic properties of the EM-sequence, discussed in the following, have previously been established.

In Ehrenfeucht and Mycielski (1992), the following result concerning the EM-sequence was established.

[†] J. Kieffer’s research was supported in part by the NSF Grants NCR-9508282 and CCR-9902081.

[‡] W. Szpankowski’s research was supported in part by the NSF Grants CCR-0208709, CCF-0513636, and DMS-0503742, and the NIH Grant R01 GM068959-01.

Proposition 1. *Every binary string of finite length appears infinitely many times as a substring of the EM-sequence.*

This suggestive result motivated subsequent authors to try to prove the balance conjecture. In order to describe these efforts, let $\{x_i : i = 1,2,3,\cdots\}$ denote the EM-sequence, let x_i^j denote the segment $(x_i, x_{i+1}, \cdots, x_j)$, and let $N_n(0)$ ($N_n(1)$) be the number of zeroes (ones) in the initial segment x_1^n. The balance conjecture is equivalent to the statement

$$|N_n(0) - N_n(1)| = o(n).$$

A weaker result than the balance conjecture would be to show that

$$|N_n(0) - N_n(1)| \le \beta n + o(n) \tag{1}$$

for a specific real number β in the interval $[0,1]$.[§] The papers by McConnell (1996) and Sutner (2003) have established such a result. For each real number t in the interval $(0,1]$, let $\alpha(t)$ be the unique real number $u \in (0,1/2]$ such that

$$-u\log_2(u) - (1-u)\log_2(1-u) = t.$$

In the paper McConnell (1996), it was proved that statement (1) holds for

$$\beta = 1 - 2\alpha(1/7) \approx 0.96.$$

This result was subsequently improved in the paper Sutner (2003), where it was established that statement (1) holds for

$$\beta = 1 - 2\alpha(1/2) \approx 0.78.$$

In the present paper, we obtain an improvement, encapsulated in this our main result.

Theorem 1. $|N_n(0) - N_n(1)| \le n/2 + o(n)$.

Remark. Theorem 1 is equivalent to saying that any limit point of $\{N_n(1)/n\}$ belongs to the interval $[1/4, 3/4]$. The best previous result of which we are aware (Sutner (2003)) states that every such limit point belongs to the interval $[\alpha(1/2), 1-\alpha(1/2)]$; if we round to two decimal places, this best previous result tells us that every limit point of $\{N_n(1)/n\}$ belongs to the interval $[0.11, 0.89]$.

For any positive integer n, consider the rooted tree formed by the binary strings which appear as least twice as substrings of x_1^n. We obtain Theorem 1 via an analysis of the structure of this "recurrence" tree. This approach has not been used in previous work on the EM-sequence. It would be of interest to know whether this approach can lead to still further results about the EM-sequence in the future.

Notation and Terminology. We list the notation and terminology that will remain in force throughout the paper.

- $\{0,1\}^+$ denotes the set of all binary strings of finite nonzero length, λ denotes the empty string, and $\{0,1\}^*$ denotes the set of strings $\{0,1\}^+ \cup \{\lambda\}$. A string in $\{0,1\}^+$ is denoted in coordinate

[§] This is equivalent to saying that every limit point of the sequence $\{N_n(1)/n : n \ge 1\}$ belongs to the interval $[(1-\beta)/2, (1+\beta)/2]$.

form as $b_1b_2\cdots b_j$, where $b_1, b_2, \cdots, b_j$ belong to $\{0,1\}$ and j is the length of the string. If $B = b_1b_2,\cdots b_j$ and $C = c_1c_2\cdots c_k$ are two strings in $\{0,1\}^+$ expressed in coordinate form, then BC, the concatenation of string B with string C, is the string in $\{0,1\}^+$ expressed in coordinate form as $b_1b_2,\cdots, b_jc_1c_2\cdots c_k$. We make the obvious extension to the concatenation of more than two strings.

- $|b|$ denotes the length of string $b \in \{0,1\}^*$.
- If $a \in \{0,1\}$, then $\bar{a}$ is $1-a$, the complement of a.
- $\mathrm{card}(S)$ or $|S|$ denotes the cardinality of set S.
- If T is a tree, $|T|$ denotes the number of vertices.

2 Recurrent Substrings and Recurrence Trees

In this section, we introduce the concept of *recurrent substrings* of the EM-sequence and the concept of *recurrence trees* formed from the recurrent substrings. The concepts of recurrent substrings and recurrence trees are needed for proving Theorem 1.

Definitions. For each positive integer n, we define R_n to be the set consisting of those strings in $\{0,1\}^*$ which occur at least twice as substrings of the initial segment x_1^n of the EM-sequence. We call the elements of R_n the *recurrent substrings of* x_1^n. The *recurrence tree* T_n is the directed labelled graph specified as follows:

- The vertices of T_n are the elements of R_n.
- The edges of T_n are the pairs (aw, w) in which $w \in R_n$, $a \in \{0,1\}$, and $aw \in R_n$. aw is called the initial vertex of edge (aw, w) and w is called the final vertex of edge (aw, w).
- The direction along edge (aw, w) is taken to be $aw \rightarrow w$.
- Each edge (aw, w) carries the label a.

The children of vertex w of T_n are those members (if any) of the set $\{0w, 1w\}$ which belong to R_n. Each vertex of T_n which has no children is called a *leaf* of T_n. The vertex λ is called the *root* of T_n. A path in T_n is a finite nonempty sequence of edges $(e_1, e_2, \cdots, e_k)$ in which, for each i satisfying $1 \leq i \leq k-1$, the final vertex of edge e_i coincides with the initial vertex of edge e_{i+1}; k is called the length of path $(e_1, e_2, \cdots, e_k)$. The paths of length one in T_n are the edges of T_n. Given any vertex v of T_n which is not the root, there is a unique path $(e_1, e_2, \cdots, e_k)$ in T_n such that e_1 has initial vertex v and e_k has final vertex λ. Thus, if the recurrence tree T_n has j leaf vertices, there are j unique leaf-to-root paths in T_n. The *binary address* of a path $(e_1, e_2, \cdots, e_k)$ is defined to be the sequence of edge labels along the path. The set consisting of all the binary addresses of paths in T_n is precisely R_n.

Example 1. From the fact that

$$x_1^{16} = 0100110101110001,$$

one sees that

$$R_{16} = \{\lambda, 0, 01, 1, 10, 010, 101, 011, 11, 110, 100, 00, 001\}.$$

Since R_{16} consists of 13 strings, the recurrence tree T_{16} will therefore consist of 13 vertices. Fig. 1 gives us a pictorial representation of T_{16}. Our convention in Fig. 1 is that the root is to the right and one follows paths from left to right. Therefore, the address of a path goes from left to right, conforming to the appearance of that address as a recurrent substring of x_1^{16}. The reader can check that the addresses of the 13 vertex-to-root paths in Fig. 1 comprise the elements of the set R_{16} above.

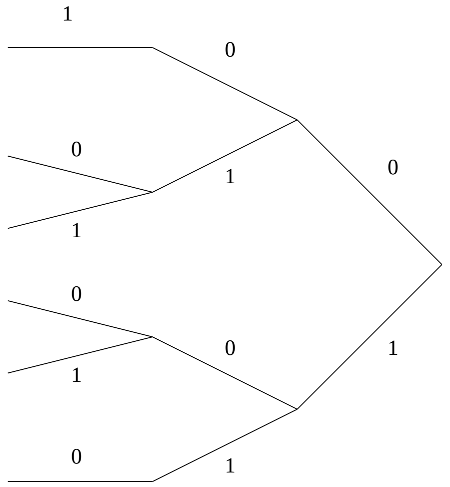

Fig. 1: The Tree T_{16}.

The sets $\{R_n : n \geq 1\}$ have various useful properties. We point out some of these properties which are easy to deduce. First of all, each set R_n is nonempty because it contains the empty string λ. We also have the obvious property

$$R_n \subset R_{n+1}, \quad n \geq 1.$$

By Proposition 1, we can deduce the property

$$\cup_{n=1}^{\infty} R_n = \{0,1\}^*.$$

Good strings. We define a string $B \in \{0,1\}^+$ to be *good* if its first two appearances in the EM-sequence are preceded by $0,1$ or $1,0$, respectively. If $B \in \{0,1\}^+$ is an initial segment of the EM-sequence, then B fails to be good (because the first appearance of B in the EM-sequence is preceded by the empty string). But there are also strings B which fail to be good which are not initial segments of the EM-sequence. For example, 1 is not good: it makes its first and second appearances in the initial segment 01001, but is preceded by 0 each time instead of being preceded by complementary bits.

We state the following result useful for proving Theorem 1, proved in Kieffer and Szpankowski (2007), the extended version of the present summary.

Proposition 2. *The sets $\{R_n\}$ obey the following asymptotic properties:*

- $\mathrm{card}(R_n) = n + o(n)$.
- $\mathrm{card}(\{b \in R_n : 0 \text{ is rightmost bit of } b\}) = N_n(0) + o(n)$.
- $\mathrm{card}(\{b \in R_n : 1 \text{ is rightmost bit of } b\}) = N_n(1) + o(n)$.
- $\mathrm{card}(\{b \in R_n : b \text{ is not good}\}) = o(n)$.

Definitions. We define $T_n(0)$ and $T_n(1)$ to be the subtrees of T_n which taken together give the tree T_n as indicated in Fig. 2. Define edge $e = (aw, w)$ of T_n to be *good* if and only if the string w is good. Suppose $e = (aw, w)$ is an edge of T_n, and let $(e_1, e_2, \cdots, e_k)$ be the path starting with edge $e_1 = e$ and ending at the root of T_n. Then w is the binary address of path $(e_2, \cdots, e_k)$. One concludes that e is good if and only if the address of the path which starts at the final vertex of e and ends at the root is good.

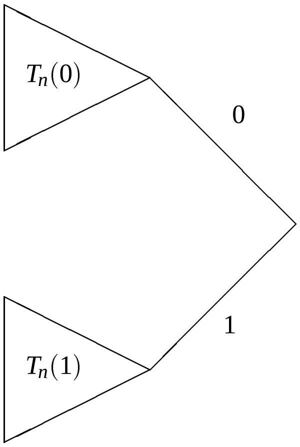

Fig. 2: Decomposition of T_n into subtrees $T_n(0)$ and $T_n(1)$.

Proposition 3. *The recurrence tree T_n has the following properties:*

- *T_n has $n + o(n)$ vertices.*
- *$T_n(0)$ has $N_n(0) + o(n)$ vertices.*
- *$T_n(1)$ has $N_n(1) + o(n)$ vertices.*
- *The cardinality of the set of edges of T_n which are not good is $o(n)$.*

The proof of this result is omitted because it follows straightforwardly from Proposition 2.

Definitions. A subtree $\tilde{T}$ of rooted tree T shall be called a *principal subtree* of T if $\tilde{T}$ is a rooted tree whose root coincides with the root of T. Fig. 3 indicates the principal subtree of T_n in which the subtree T_n^* (appearing in two places as indicated) is uniquely specified by requiring that $|T_n^*|$ be maximized. We call this principal subtree of T_n the *principal symmetric subtree* of T_n.

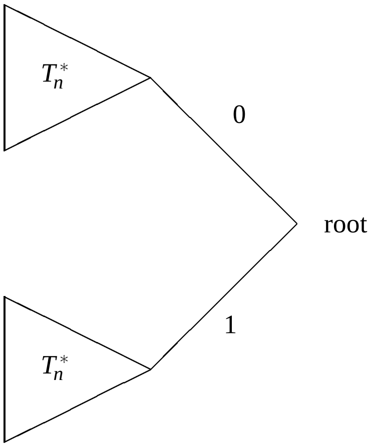

Fig. 3: Principal symmetric subtree of T_n.

We can specify the principal symmetric subtree of T_n and the tree T_n^* without referring to a figure: Let R_n^* be the set of all $b \in \{0,1\}^*$ such that both $b0$ and $b1$ belong to R_n; then T_n^* is the tree generated by R_n^* and the principal symmetric subtree of T_n is the tree generated by $R_n^*0 \cup R_n^*1$.

Example 2. Fig. 4 gives the tree T_{16}^*, easily extracted from the tree T_{16} in Fig. 1.

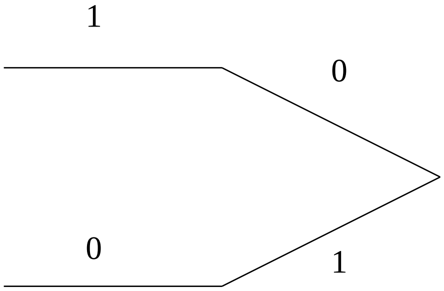

Fig. 4: The Tree T_{16}^*.

Let V_n be the set of all leaves of T_n which do not belong to the principal symmetric subtree of T_n. For each $v \in V_n$, let $\pi(v)$ be the unique path in T_n which starts at v and ends at the first vertex of the principal symmetric subtree of T_n which is encountered. Suppose we remove the principal symmetric subtree of T_n from T_n. Then what remains is a forest of trees, which is the union of the paths $\pi(v)$ for $v \in V_n$.

The following auxiliary result is easy to prove.

Lemma 1. *For each n, no two paths in $\{\pi(v) : v \in V_n\}$ have an edge in common.*

Proof. Let v_0, v_1 be distinct vertices in V_n. Assume that $\pi(v_0)$ and $\pi(v_1)$ have an edge in common. The proof will be complete once we show that this assumption leads to a contradiction. Let e_0 be the last edge along path $\pi(v_0)$ which does not belong to path $\pi(v_1)$, and let e_1 be the last edge along path $\pi(v_1)$ which does not belong to path $\pi(v_0)$. Then the remainder of path $\pi(v_0)$ after edge e_0 (which is a nonempty path) coincides with the remainder of path $\pi(v_1)$ after edge e_1. e_0, e_1 are thus "sibling edges" terminating at the same vertex, and their binary labels must therefore be distinct; relabelling v_0, v_1 if necessary, we may assume that e_0 carries label 0 and e_1 carries label 1. Let π be the path in T_n and let e be the edge in T_n such that (e_1, π, e) and (e_2, π, e) are the paths starting at e_1, e_2, respectively, and going back to the root of T_n. Let $a \in \{0,1\}$ be the label of e and let $b \in \{0,1\}^+$ be the address of π. The path π is not a path in the principal symmetric subtree of T_n, since the first edge of π belongs to both paths $\pi(v_0)$, $\pi(v_1)$ and these two paths contain no edges in the principal symmetric subtree of T_n. Therefore, b does not belong to R_n^*. Since ba belongs to R_n, we conclude that $b\bar{a}$ does not belong to R_n. Each of the strings $0ba$, $1ba$ belongs to R_n and therefore each of these strings appears at least twice in x_1^n. It follows that $0b$ appears at least twice in x_1^{n-1}, and so does $1b$. The first two appearances of $0b$ in x_1^{n-1} are followed by $c, \bar{c}$, respectively, where $c \in \{0,1\}$. The first two appearances of $1b$ in x_1^{n-1} are followed by $d, \bar{d}$, respectively, where $d \in \{0,1\}$. Consequently, all of the following strings appear in x_1^n: $0bc, 0b\bar{c}, 1bd, 1b\bar{d}$. It follows that $b\bar{a}$ appears at least twice in x_1^n, a contradiction.

Definition. We call the paths belonging to $\{\pi(v) : v \in V_n\}$ *spaghetti strands* (of the tree T_n).

Exploiting Lemma 1, we now have a decomposition of T_n as the principal symmetric subtree of T_n with spaghetti strands adjoined to it, as conceptualized in Fig. 5. There may not be any spaghetti strands, in which case $|T_n(0)| = |T_n(1)|$; if this happens for infinitely many n one could conclude that $1/2$ is a limit point of the sequence $\{N_n(1)/n\}$. Our approach to proving Theorem 1 in the next section involves showing that only a limited portion of recurrence tree T_n can be occupied by spaghetti strands as $n \to \infty$.

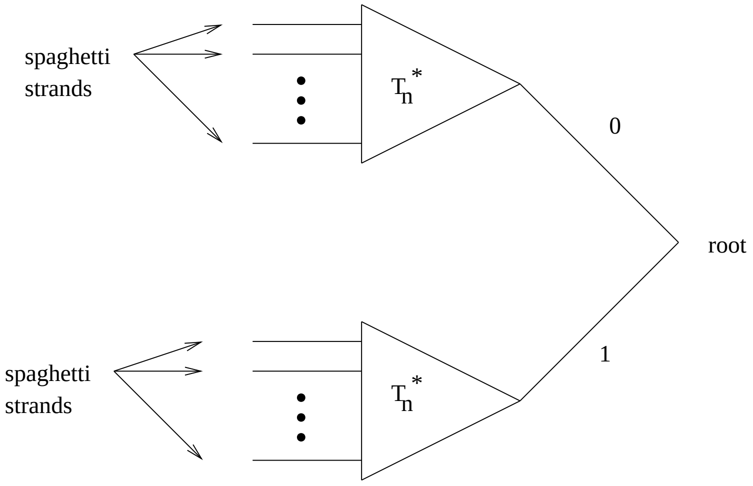

Fig. 5: Decomposition of T_n showing the spaghetti strands.

Example 3. Examining Fig. 1, we see by inspection that T_{16} has exactly two spaghetti strands, each consisting of one edge:

$$110 \to 10, \quad 001 \to 01.$$

3 Proof of Theorem 1

The following two results provide the machinery needed to prove Theorem 1.

Proposition 4. *Let $B \in \{0,1\}^+$ be a good string. Let $a < b < c < d < e$ be the positive integers at which the first five appearances of B in the EM-sequence $\{x_i : i \geq 1\}$ end. Let u, v be the strings*

$$\begin{aligned} u &= x_{a+1}x_{b+1}x_{c+1}x_{d+1}x_{e+1}, \\ v &= x_{a+2}x_{b+2}x_{c+2}x_{d+2}x_{e+2}. \end{aligned}$$

Then at least one of the following statements must be true:

(a): *u is a permutation of* 00011 *or* 11100.

(b): *v is a permutation of* 00011 *or* 11100.

Proposition 5. *For each n, the set of all edges of T_n which belong to spaghetti strands may be partitioned into two subsets $E_n(1), E_n(2)$ satisfying the following properties:*

- *For each n, $E_n(1)$ contains at most 2 edges from each spaghetti strand of T_n.*
- $|E_n(2)| = o(n)$.

Example 4. If we look at the first five appearances of 11000 in the EM-sequence, together with the following bit, we obtain:

$$\begin{aligned} x_{11}^{16} &= 110001 \\ x_{46}^{51} &= 110000 \\ x_{80}^{85} &= 110001 \\ x_{114}^{119} &= 110001 \\ x_{123}^{128} &= 110000 \end{aligned}$$

Note that the first and third appearances of 11000 are followed by 1, whereas the second and fifth appearances are followed by 0. Thus, property(a) of Proposition 4 holds for the string $B = 11000$.

The proofs of Propositions 4-5 are given in the paper Kieffer and Szpankowski (2007), the extended version of the present summary. We remark that in our development in Kieffer and Szpankowski (2007), we obtain Proposition 5 as a consequence of Proposition 4.

We now embark upon the proof of Theorem 1. Let $t_n = |T_n^*|$. Let $k_0(n)$ be the total number of spaghetti strands of T_n whose paths, continued back to the root, end in the edge $(0, \lambda)$, and let $j_0(n)$ be the total number of edges in these $k_0(n)$ spaghetti strands. Let $k_1(n)$ be the total number of spaghetti strands of T_n

whose paths, continued back to the root, end in the edge $(1,\lambda)$, and let $j_1(n)$ be the total number of edges in these $k_1(n)$ spaghetti strands. Then

$$\begin{aligned} |T_n(0)| &= t_n + j_0(n) \\ |T_n(1)| &= t_n + j_1(n) \\ |T_n(0)| + |T_n(1)| &= 2t_n + j_0(n) + j_1(n) \end{aligned}$$

Let $L(T_n^*)$ be the number of leaf vertices of T_n^* and let $U(T_n^*)$ be the number unary vertices of T_n^*. Then

$$t_n = 2L(T_n^*) + U(T_n^*) - 1.$$

Since each spaghetti strand terminates at either a leaf vertex or unary vertex of T_n^*, we have

$$\max(k_0(n), k_1(n)) \leq 2L(T_n^*) + U(T_n^*) = t_n + 1.$$

By Proposition 5, we have

$$\begin{aligned} j_0(n) &\leq 2k_0(n) + o(n) \\ j_1(n) &\leq 2k_1(n) + o(n) \end{aligned}$$

Therefore,

$$\max(j_0(n), j_1(n)) \leq 2t_n + o(n). \tag{2}$$

We will argue that

$$\limsup_{n\to\infty} n^{-1} N_n(0) \leq 3/4. \tag{3}$$

A similar argument will give

$$\limsup_{n\to\infty}\ n^{-1} N_n(1) \leq 3/4. \tag{4}$$

Together, (3) and (4) yield Theorem 1. Since by Proposition 3 we have

$$|T_n(0)| + |T_n(1)| = n + o(n)$$

and

$$|T_n(0)| = N_n(0) + o(n),$$

it follows that

$$\limsup_{n\to\infty} n^{-1} N_n(0) = \limsup_{n\to\infty} \left[\frac{|T_n(0)|}{|T_n(0)| + |T_n(1)|} \right].$$

Thus, to establish (3), we can prove that

$$\limsup_{n\to\infty} \left[\frac{|T_n(0)|}{|T_n(0)| + |T_n(1)|} \right] \leq 3/4. \tag{5}$$

By (2), we may pick a sequence of positive numbers $\{\varepsilon_n\}$ tending to zero such that

$$j_n(0) \leq 2t_n + n\varepsilon_n, \ \ n = 1, 2, \cdots.$$

We then obtain

$$\frac{|T_n(0)|}{|T_n(0)|+|T_n(1)|} = \frac{t_n + j_0(n)}{2t_n + j_0(n) + j_1(n)} \leq \frac{t_n + j_0(n)}{2t_n + j_0(n)} \leq \frac{3t_n + n\varepsilon_n}{4t_n + n\varepsilon_n} \leq (3/4) + \left(\frac{n}{4t_n}\right)\varepsilon_n.$$

To finish our proof of (5), we can simply show that $n/t_n = O(1)$. To see this, first note that

$$n = |T_n(0)| + |T_n(1)| + o(n) = 2t_n + j_0(n) + j_1(n) + o(n) \leq 6t_n + o(n).$$

The inequality

$$n \leq 6t_n + o(n)$$

implies $n/t_n = O(1)$.

References

A. Ehrenfeucht and J. Mycielski. A pseudorandom sequence—how random is it? *Amer. Math. Monthly*, 99:373–375, 1992.

J. Kieffer and W. Szpankowski. *On the Ehrenfeucht-Mycielski Balance Conjecture (Extended Version)*, 2007. `http://www.ece.umn.edu/users/kieffer/presentations.html`.

T. McConnell. Laws of large numbers for some non-repetitive sequences. Technical report, Syracuse University Department of Mathematics, 1996.

N. Sloane. *On-Line Encyclopedia of Integer Sequences*, 2007. `http://www.research.att.com/˜njas/sequences/`.

K. Sutner. The Ehrenfeucht-Mycielski sequence. *Lecture Notes in Computer Science*, 2759:282–293, 2003.