# The Ehrenfeucht-Mycielsky Sequence

Klaus Sutner

Carnegie Mellon University

May 2012

The Ehrenfeucht-Mycielsky Sequence

Disjunctiveness

Complexity of $U$

Density

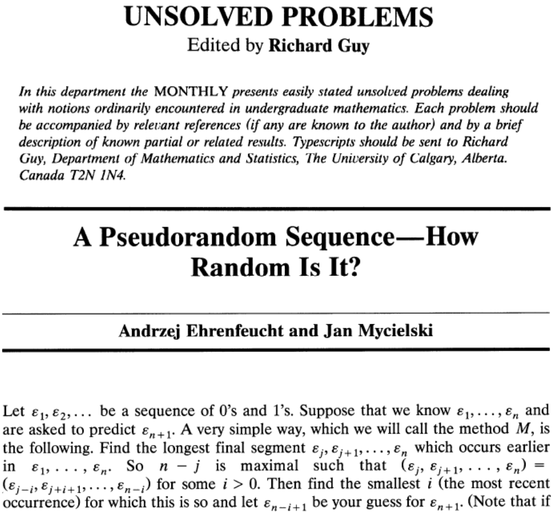
Open Problems

# Sage Advice

> Do not, more generally, publish your failures: I tried to prove so-and-so; I couldn't; here it is——see?!

Paul Halmos, *Four Panel Talks on Publishing*, 1975

# The Ehrenfeucht-Mycielsky Sequence



## UNSOLVED PROBLEMS

Edited by **Richard Guy**

*In this department the* MONTHLY *presents easily stated unsolved problems dealing with notions ordinarily encountered in undergraduate mathematics. Each problem should be accompanied by relevant references (if any are known to the author) and by a brief description of known partial or related results. Typescripts should be sent to Richard Guy, Department of Mathematics and Statistics, The University of Calgary, Alberta. Canada T2N 1N4.*

## A Pseudorandom Sequence—How Random Is It?

**Andrzej Ehrenfeucht and Jan Mycielski**

Let $\varepsilon_1, \varepsilon_2, \ldots$ be a sequence of 0's and 1's. Suppose that we know $\varepsilon_1, \ldots, \varepsilon_n$ and are asked to predict $\varepsilon_{n+1}$. A very simple way, which we will call the method $M$, is the following. Find the longest final segment $\varepsilon_j, \varepsilon_{j+1}, \ldots, \varepsilon_n$ which occurs earlier in $\varepsilon_1, \ldots, \varepsilon_n$. So $n - j$ is maximal such that $(\varepsilon_j, \varepsilon_{j+1}, \ldots, \varepsilon_n) = (\varepsilon_{j-i}, \varepsilon_{j+i+1}, \ldots, \varepsilon_{n-i})$ for some $i > 0$. Then find the smallest $i$ (the most recent occurrence) for which this is so and let $\varepsilon_{n-i+1}$ be your guess for $\varepsilon_{n+1}$. (Note that if

# Just say no . . .

Ehrenfeucht and Mycielsky construct an infinite binary sequence

$$U = u_1 u_2 u_3 \dots u_n \dots$$

based on the following simple idea:

When a situation arises that is similar to a previous one, do exactly the opposite of what you did last time.

# Formally . . .

Write

$$U_n = u_1 u_2 \dots u_{n-1} u_n$$

for the prefix of $U$ of length $n$.

## Definition

- $u_1 = 0$
- Find the longest suffix $v$ of $U_n$ that appears already in $U_{n-1}$.
  Let $b$ be the bit following the last occurrence of $v$ in $U_{n-1}$.
  Set $u_{n+1} = \overline{b}$.
- If no such suffix exists set $u_{n+1} = \overline{u}_n$.

# Less Formally

$$010\ldots a\,v\,b\ldots\overline{a}\,v \;\big|\; \overline{b}$$

$0\,1\,0$

$0\,1\,0\,0$

$0\,1\,0\,0\,1$

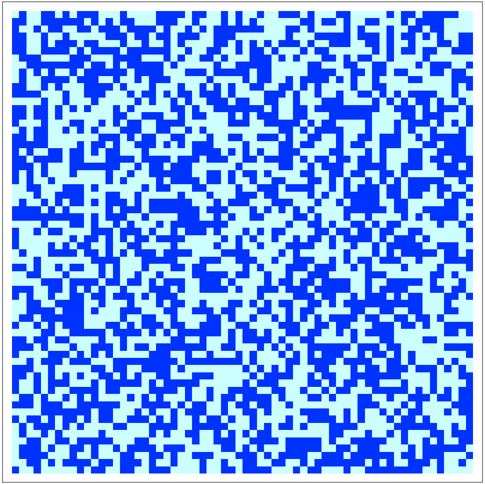
$0\,1\,0\,0\,1\,1$

$0\,1\,0\,0\,1\,1\,0$

$0\,1\,0\,0\,1\,1\,0\,1$

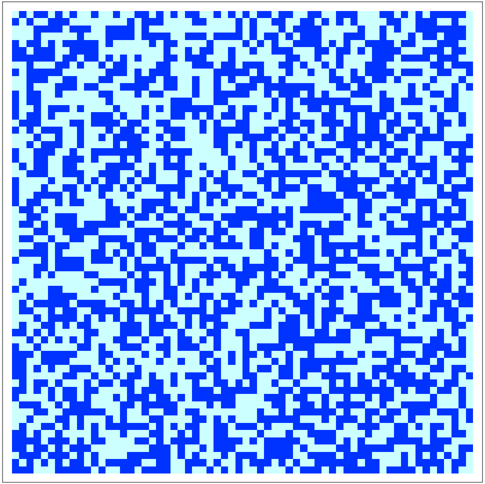
And so on . . .

# More Bits



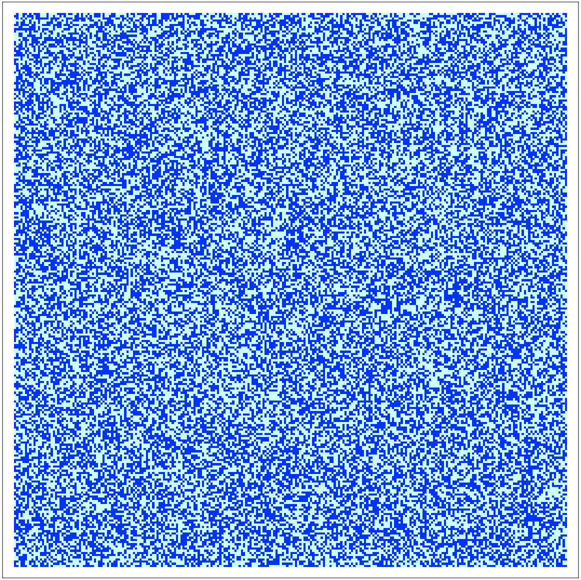
The first $2^{12}$ bits, in row-major order.

# Random Bits

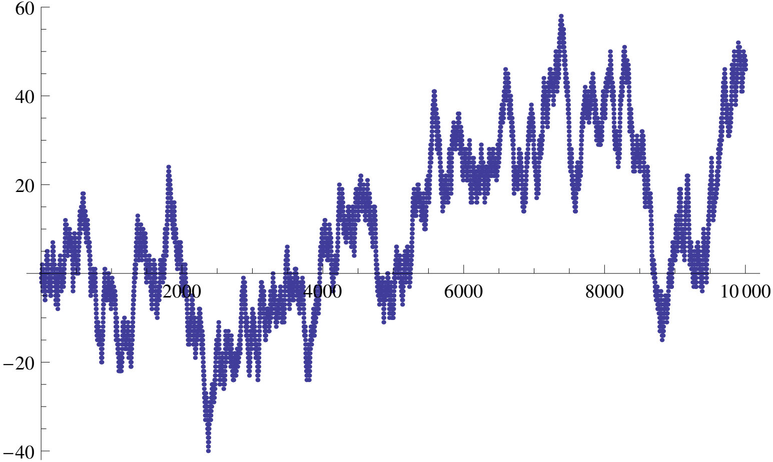


The first $2^{12}$ bits from a good pseudo-random source.

# 64K Bits



# EM Random Walk



# How Random Is It?

As an experiment one can try to compress the first million bits (actually, $2^{17} = 131072$ bytes).

- Lemple-Ziv-Welch gzip: 159,410 bytes.
- Burrows-Wheeler bzip2: 165,362 bytes.

Of course, the Kolmogorov complexity of $U$ is quite low; the sequence fails miserably as a random sequence in the sense of Martin-Löf.

# Golomb’s Postulates

Here are some classical, simple randomness criteria due to Golomb, initially used in the study of shift-register sequences.

**R1: Equidistribution**

The limiting density of 1’s should be $1/2$.

**R2: Blocks**

The limiting density of every block of length $k$ should be $2^{-k}$.

# Statistical Tests

```
PearsonChiSquareTest[ U, DiscreteUniformDistribution[{0, 1}] ]

==> 0.931466
```

This uses the first $10^6$ bits of $U$.

# EM and Density

Distribution of $0$'s and $1$'s in the first $i \cdot 10^8$ bits, for $i = 1, \ldots, 10$.

| $i \cdot 10^8$ | $\#0$ | $\Delta$ |
|---|---:|---:|
| 1 | 49996379 | 3621 |
| 2 | 99993568 | 6432 |
| 3 | 149998751 | 1249 |
| 4 | 199995036 | 4964 |
| 5 | 249995563 | 4437 |
| 6 | 299992953 | 7047 |
| 7 | 349998485 | 1515 |
| 8 | 400003768 | 3768 |
| 9 | 449989561 | 10439 |
| 10 | 499988410 | 11590 |

# Blocks

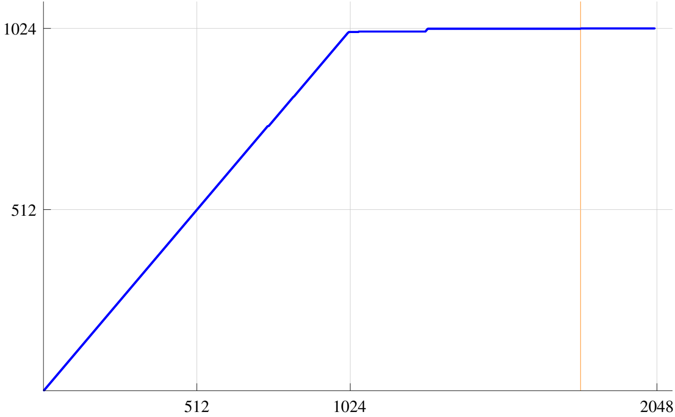
Here are the counts for all words of length 4 among the first $2^{20}$ bits.

| 0000 | 0001 | 0010 | 0011 | 0100 | 0101 | 0110 | 0111 |
|---|---|---|---|---|---|---|---|
| 96 | 58 | 24 | 12 | 41 | 28 | 50 | 15 |

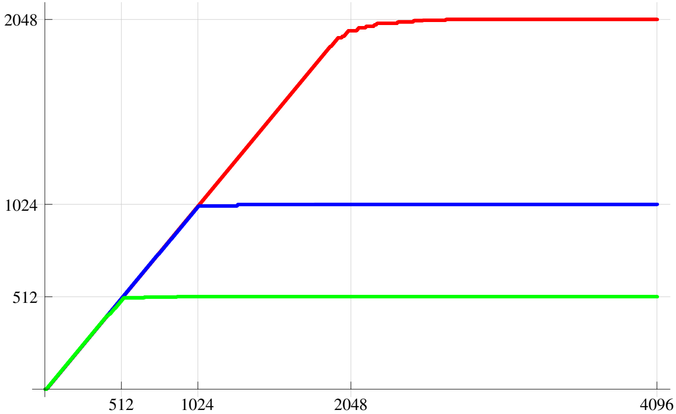
| 1000 | 1001 | 1010 | 1011 | 1100 | 1101 | 1110 | 1111 |
|---|---|---|---|---|---|---|---|
| 58 | 46 | 10 | 54 | 29 | 36 | 15 | 21 |

Difference to $2^{16} = 65536$.

# Census Length 10



# All Together Now



Words of lengths $k = 9, 10, 11$ (green, blue, red).

# Extremal Words at 2

This behavior is rather surprising.

Suppose one wants to construct a sequence $W \in \Sigma^\omega$ that maximizes the number of subwords of all lengths in its prefixes.

- For $|\Sigma| \geq 3$ one can construct an infinite de Bruijn word.
- For $|\Sigma| = 2$ there is no such word, though one can produce a limit of de Bruijn words at every other level.

The analogous problem for subsequences is easy (Flaxman, Harrow, Sorkin 2004).

The Ehrenfeucht-Mycielsky Sequence

Disjunctiveness

Complexity of $U$

Density

Open Problems

# All Finite Words

It looks like every finite word appears somewhere in $U$.

Definition

An infinite sequence is disjunctive if it contains all finite words as factors.

Ehrenfeucht-Mycielsky showed in 1992 that the sequence is disjunctive, using a combinatorial argument.

# Terminology

Recall the construction of $u_{n+1}$ from $U_n$:

$$\dots a\, v\, b \dots \overline{a}\, v \;\Big|\; \overline{b}$$

## Definition

- $v$ is said to match at time $n$, $v = \mu(n)$,
- $|v|$ is the match length $\lambda(n)$ at time $n$,
- the match position $\pi(n)$ at time $n$ is the location (position of the last bit) of the matching word $v$ in $U_{n-1}$.

# Single Matches

$\mu : \mathbb{N} \to 2^\star$ is almost injective: a word $v$ can can match at most once, except for inititial segments $v = U_k$; they can match twice.

$$\ldots avb \ldots \overline{a}v \mid \overline{b}$$

$$v \ldots avb \ldots \overline{a}v \mid \overline{b}$$

So, exactly one word of each length $k \geq 1$ matches at most twice, all others match at most once.

# Match Lengths

It follows from the definition that

$$\lambda(n+1) \leq \lambda(n) + 1$$

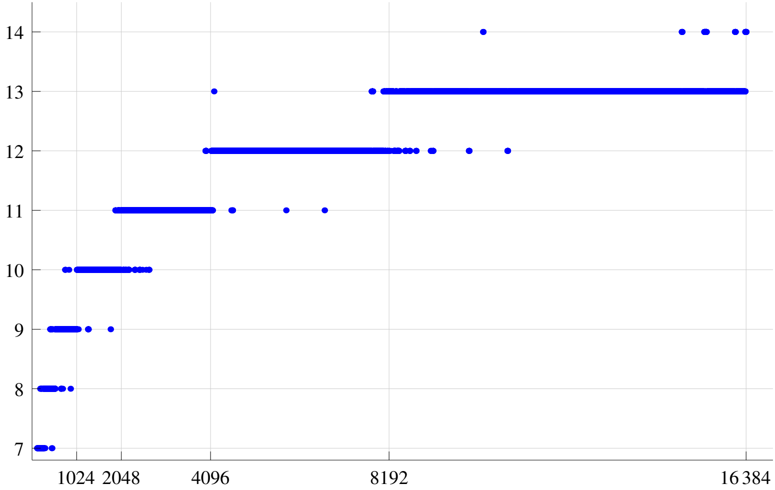
but it is perfectly conceivable that $\lambda(n+1)$ is much smaller than $\lambda(n)$.

Since there are infinitely many distinct matches, match lengths must be increasing in the sense that

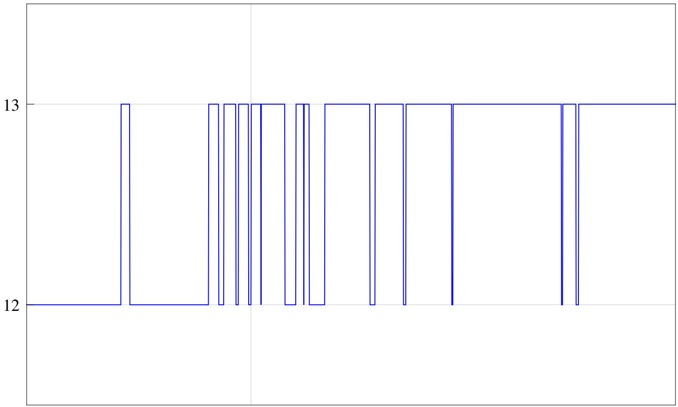
$$\liminf_n \lambda(n) = \infty.$$

Surprisingly, match lengths seem to increase very steadily.

# Near Monotonicity

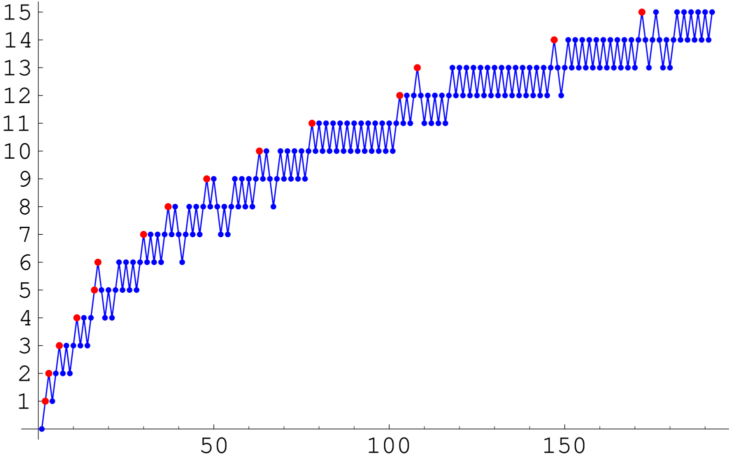


# Blowup



The region near $2^{13}$.

# No Drop



Condense constant runs of $\lambda$; red dots indicate a new maximum.

# Match Lengths

Note that the value of $\lambda$ never seems to drop by more than 1.

Theorem (No Drop)

*For all $n$:*

$$\lambda(n) - 1 \leq \lambda(n+1) \leq \lambda(n) + 1.$$

Proof is very tedious, one needs to worry about a prefixes of a matching word.

# Increasing Maxima

## Definition

The maximum match length and the critical time for $k$ are

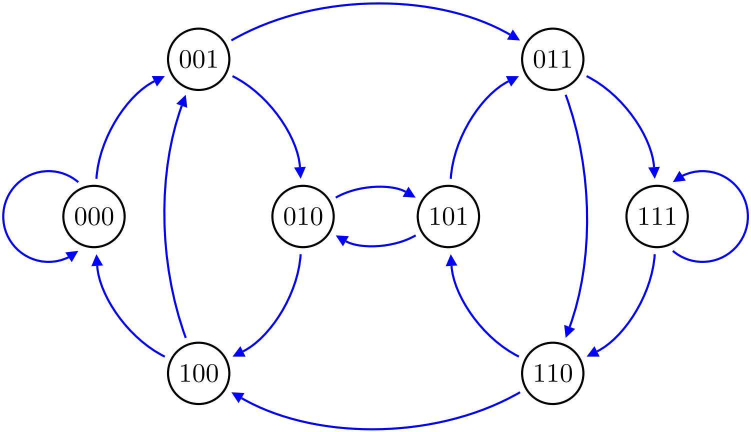
$$\Lambda(n) = \max\big(\,\lambda(m) \mid m \leq n\,\big)$$

$$\tau_k = \min\big(\,n \mid \Lambda(n) = k\,\big)$$

With a little bit of imagination one can see logarithmic growth for $\Lambda$.

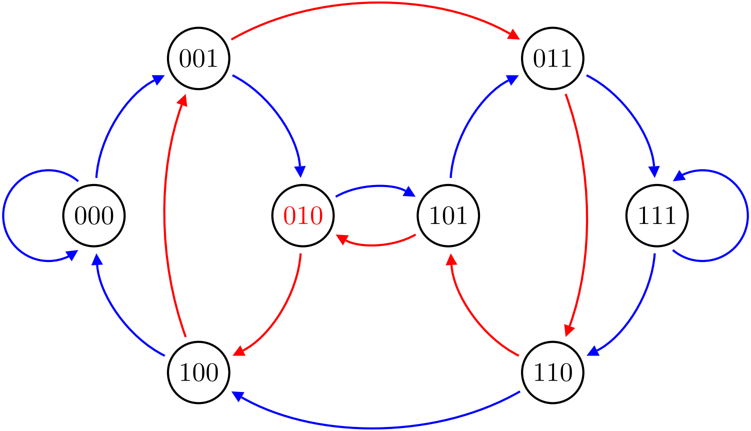
Understanding the $\tau_k$ is crucial for the analysis of the EM sequence.

# De Bruijn Graph of Order 3



Any infinite binary word traces a path in $\mathcal{B}_3$.

# De Bruijn Graph of Order 3



The first few edges of the path traced by $U$ in $\mathcal{B}_3$.

# Aside: Line Graphs

Recall that $\mathcal{B}_{k+1}$ is the line graph of $\mathcal{B}_k$.

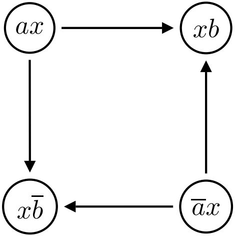
Hence a vertex-simple cycle in $\mathcal{B}_{k+1}$ gives rise to a cycle in $\mathcal{B}_k$, but not necessarily a vertex-simple one.

The thing may fold back onto itself, but it will remain edge-simple.

# Zig-Zags

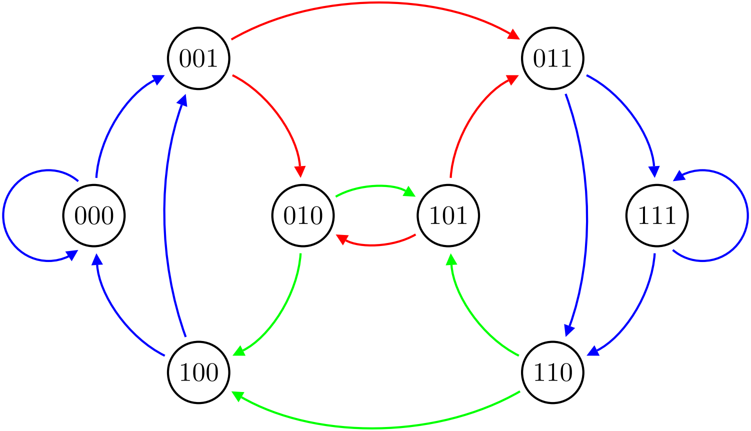
A zig-zag is an alternating path of length 4

$$v_1 \to v_2 \leftarrow v_3 \to v_4 \leftarrow v_1$$



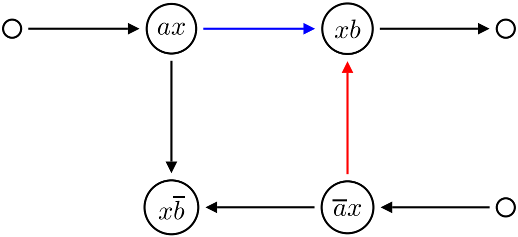
Binary de Bruijn graphs are the edge-disjoint union of zig-zags.

# Zig-Zag Decomposition



# No Merge

The first two encounters of $U$ with a zig-zag **cannot** look like this:



Blue: first hit, red: second hit.

# The Second Coming

Consider the path traced by $U$ in the de Bruijn graph $\mathcal{B}_k$ of order $k$.

## Proposition

*This path begins with a vertex-simple cycle returning to $U_k$.*
*Thus, $U_k$ is the first word of length $k$ that matches.*

In other words, $\tau_k$ is the time when $U$ has traced a first vertex-simple cycle, the principal cycle, in the de Bruijn graph $\mathcal{B}_k$.

# Recurrence

## Corollary

*Every finite word in $U$ must appear at least twice (and, therefore, infinitely often).*

*Proof.*

If $w$ appears at all it must appear in some prefix $U_k$.

□

# Disjunctiveness

Lemma

*The Ehrenfeucht-Mycielsky sequence is disjunctive.*

*Proof.*

Consider all nodes in $\mathcal{B}_k$ that are hit by $U$.

They must all be hit at least twice, so everybody has out-degree 2.

But the only subgraph with this property is $\mathcal{B}_k$ itself. □

Unfortunately, this proof does not give any reasonable bound on when all words of length $k$ must already have appeared in $U$.

It seems that $n \approx 2^{k+2}$ suffices, but that is an open conjecture.

# The Internal Clock

The inner life of $U$ seems to unfold like so:

$k$ Prefix $u = U_k$.
Start tracing a simple cycle $C_0$ in $\mathcal{B}_k$.

$\tau_k$ $u$ is first match of length $k$.
Start tracing secondary cycle $C_1$ in $\mathcal{B}_k$.

$\tau_{k+1}$ All zig-zags in $\mathcal{B}_k$ touched, all degree 4 cycles in the residual are bordered by degree 2 points.

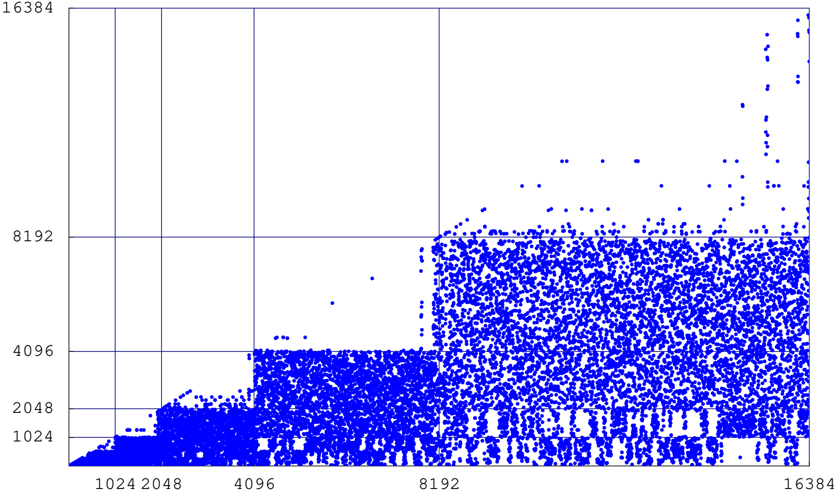
$\tau_{k+2}$ No residual points in $\mathcal{B}_k$ left, edges may remain.

# Matches

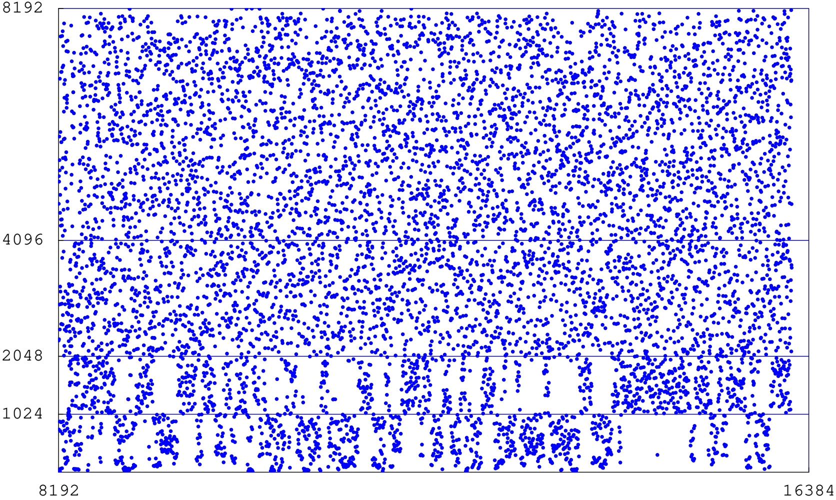
As a consequence of disjunctiveness we know that there are exactly $2^k + 1$ matches of length $k$:

- Each word of length $k$ other than $U_k$ appears exactly once as match.
- $U_k$ appears exactly twice as match.

# EM Staircase (Match Positions)



# Fine Structure



The Ehrenfeucht-Mycielsky Sequence

Disjunctiveness

Complexity of $U$

Density

Open Problems

# Prefixes

The language $\{ U_n \mid n \geq 0 \}$ of all prefixes of $U$ fails to be regular. It follows from the "Gap Theorem" by Calude and Yu (1995) that the prefix language cannot be context-free (it is trivially context-sensitive).

If one accounts for space the right way, we have:

- Prefixes of $U$ can be recognized in logarithmic space and quadratic time using Knuth-Morris-Pratt.

- A linear space lookup algorithm can generate one bit of the sequence in amortized constant time, assuming near-monotonicity.

# Computing a Billion Bits of $U$

From the No-Drop lemma we know that if the previous match length was $k$ one of the following three suffixes must work at the next step:

$$\begin{aligned}
v &= u_{n-k}u_{n-k+1}u_{n-k+2}\ldots u_{n-1}u_n\\
v &= \phantom{u_{n-k}}u_{n-k+1}u_{n-k+2}\ldots u_{n-1}u_n\\
v &= \phantom{u_{n-k}u_{n-k+1}}u_{n-k+2}\ldots u_{n-1}u_n
\end{aligned}$$

The obvious brute-force implementation would require three searches of length $O(n)$.

# Simultaneous Search

One can modify the classical Knuth-Morris-Pratt string search algorithm to perform all three searches at once, in $O(n)$ steps.

The KMP machine has states $Q = \{0, 1, \ldots, k, k+1\}$ and we feed $U_{n-1}$ to it, in reverse order.

State $p$ means: we have seen a match of length $p$.

Record the first time $p = k - 1$ and $p = k$, stop when $p = k + 1$.

Hence we can compute the next bit in $O(n + k)$ steps.

# Position Tables

All we really need to know to get bit $u_{n+1}$ is the last occurrence of the three candidates

$$u_{n-k} \dots u_n \qquad u_{n-k+1} \dots u_n \qquad u_{n-k+2} \dots u_n.$$

So we could keep a hash table for all words up to length $k+1$ that have already been encountered.

In fact, hashing is not necessary: the table will grow to size $2^k$ and fill up, so we might as well use a simple array.

# Maintaining the Lookup Table

Crucial: if all matches so far have length at most $k$ we only need $P_l$ for $l \leq k$.

In fact, we can even delete $P_1$, $P_2$ etc. once all words have matched.

At time $\tau_{k+1}$ the max match length increases.

At that point, allocate and initialize $P_{k+1}$.

This costs $\Theta(2^{k+1})$ steps, but the cost can be amortized over the following steps until the max match-length increases again.

The Ehrenfeucht-Mycielsky Sequence

Disjunctiveness

Complexity of $U$

Density

Open Problems

# Density

For any word $w \in \mathbf{2}^k$ we write

$$\Delta(w) = (\#\,1 \text{ in } w)/k$$

for the density of $w$, and $\Delta(W)$ for the average density of a set of words $W \subseteq \mathbf{2}^k$.

For an infinite word $V \in \mathbf{2}^\omega$ let

$$\Delta(V) = \lim_{n \to \infty} \Delta(V_n)$$

We suspect strongly that $\Delta(U) = 1/2$.

Of course, the limit might fail to exist or might be different from $1/2$.

# Density Basics

Clearly $\Delta(\mathbf{2}^k) = 1/2$.

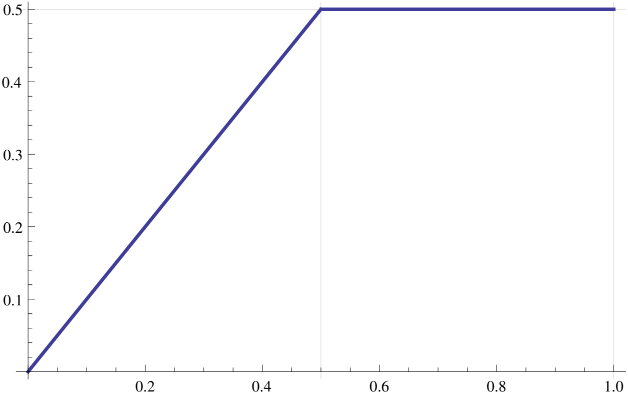
Write $\mathbf{2}^{k,p} = \{\, x \in \mathbf{2}^k \mid \#1 \text{ in } x = p \,\}$.

How about words of density up to $\alpha$:

$$\lim \Delta(\mathbf{2}^{k,\leq \alpha k}) = ???$$

As a function of $\alpha$ this is non-decreasing, 0 at 0 and 1/2 at 1.

# A Surprise



# The Proof

$$\mathit{\Delta}(\mathbf{2}^{k,\leq p}) = \frac{\sum_{i \leq p} \binom{k}{i} i/k}{\binom{k}{\leq p}} = 1/2 - \left( 4 \frac{\binom{k-1}{<p}}{\binom{k-1}{p}} + 2 \right)^{-1}$$

Let $0 \leq \varepsilon < 1/2$ and $p = \lfloor \varepsilon k + c \rfloor$ where $c$ is constant. Then

$$\lim_{k \to \infty} \frac{\binom{k}{<p}}{\binom{k}{p}} = \frac{\varepsilon}{(1 - 2\varepsilon)}.$$

# Near Monotonicity

There is a connection between the density of $U$ and match lengths.

## Conjecture (2-Monotonicity)

$$m \geq n \quad \textit{implies} \quad \lambda(m) \geq \lambda(n) - 2.$$

This is true for the first billion bits, but the conjecture is still open.

# Monotonicity Implies Balance

We have for $0 \leq \alpha \leq 1/2$: $\lim_{k\to\infty} \mathit{\Delta}(\mathbf{2}^{k,\leq \alpha k}) = \alpha$.

Applying this to $\tau_{k+c} \leq t < \tau_{k+c+1}$ and the set of $k$-factors of $U_t$ we get the following:

## Theorem

*$c$-monotonicity of $\lambda$ for any constant $c$ implies balance.*

▸ Proof

# Bounding Density

Alas, the best result known so far is

**Theorem**

*The density of 1's in $U$ is at least* $0.11$.

The proof combines a counting method by McConnell 2000 with a detailed analysis of the behavior of $U$ in de Bruijn graphs.

Here is an outline.

# Another Cycle

Recall the principal cycle in $\mathcal{B}_k$. Upon completion of the principal cycle, $U$ traces another cycle, also anchored at $U_k$.

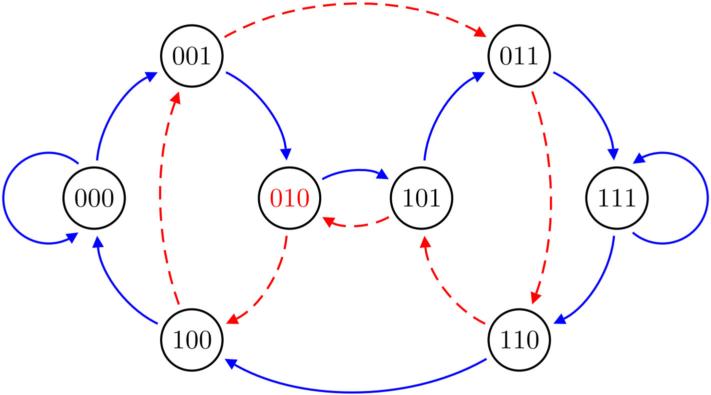
Up to time $t = \tau_{k+1} - 1$ we have two cycles $C_0$ and $C_1$ in $\mathcal{B}_k$, both anchored at $u = U_k$:

$$\overbrace{u\,a\ldots b\,u}^{C_0}\underbrace{\overline{a}\ldots\bar{b}\,u}_{C_1}\,a\ldots$$

$C_0$ is a vertex-simple cycle, and the two cycles are edge-disjoint.

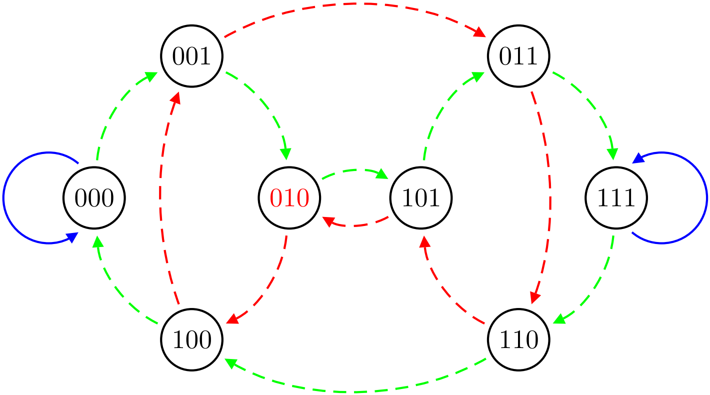
Doubly hit vertices correspond to matches of length $k$ up to time $t$.

# Principal Cycle in $\mathcal{B}_3$



Note how one can read off the secondary cycle (up to degree 4 points).

# Both Cycles in $\mathcal{B}_3$



Note how the match length drops at $000$ and $111$.

# Residual Graph

Define the residual graph to be

$$\overline{\mathcal{B}}_k(t) = \mathcal{B}_k - C_0 - C_1$$

$\overline{\mathcal{B}}_k(t)$ consists only of degree 2 and, possibly, degree 4 points.

The strongly connected components of $\overline{\mathcal{B}}_k(t)$ are all Eulerian.

# Irregular Words

By disjunctiveness, $U$ must later touches the components in the residual graph. We have the following situation:

$$\ldots a\,v\,b \ldots a\,v\,\bar{b} \ldots$$

The first two occurrences of $v$ are preceded by the same bit, $v$ is irregular.

Taxonomy

- initial
- regular
- irregular

# Counting Irregulars

The number of irregular words seems to be small.

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $I_k$ | 1 | 2 | 2 | 2 | 4 | 4 | 6 | 6 | 6 | 12 | 6 | 16 |

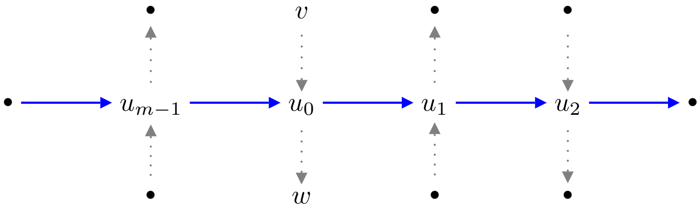
There are 12 irregular words of length 10:

$$\begin{aligned}
&0000000000, 0010010010, 0010110101, 0011000000,\\
&\qquad 0011001100, 0011100000, 0111100001, 1001110010,\\
&\qquad\qquad 1010110000, 1110100111, 1111000111, 1111111111.
\end{aligned}$$

# Who Cares?

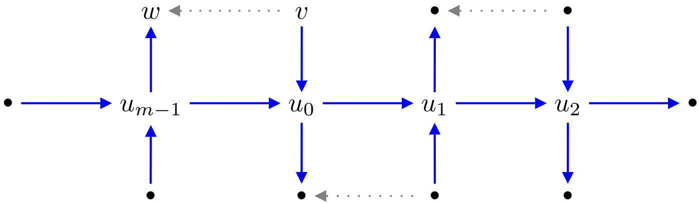
$\lambda$ can decrease only when an irregular word is encountered for the second time, and will then correspondingly increase when the same word is encountered for the third time, at which point it appears as a match.

The easy case: the SCC in the residual graph is a cycle (blue).



# The Hard Case: Degree 4

The $u_i$ are degree 4 in the residual graph. Neighbors are degree 2.



The Ehrenfeucht-Mycielsky Sequence

Disjunctiveness

Complexity of $U$

Density

Open Problems

# Density

Conjecture

*The limiting density of 1's in $U$ is $1/2$.*

Conjecture

*The limiting density of any word of length $k$ in $U$ is $2^{-k}$.*

Conjecture

*The last conjecture holds even if we start with an arbitrary finite word as seed.*

# Entropy

For two words $z, x \in \mathbf{2}^\star$ define

$$F_z(x) = \text{ number of occurrences of } z \text{ in } x$$

$$H_k(x) = \sum_{|z|=k} -F_z(x) \log_2(F_z(x))$$

## Conjecture

$$\limsup_m H_k(U_m) = 1$$

This can be handled for the *Linus sequence*: minimize length of last double block $\ldots vv$, has 0 entropy (Balister, Kalikow, Sarkar 2008).

# Matches

### Conjecture

*All words of length $k$ match by time $2^{k+2}$.*

### Conjecture

*The match length function is 2-monotonic.*

**Let me know if you want to work on this.**

# $c$-Monotonicity Implies Balance: Proof

Assume otherwise; say for infinitely many $t$ we have $\Delta(U_t) < \alpha_0 < 1/2$.

Consider the prinicipal round for $k + c$ and pick some time $t$ in the interval $[\tau_{k+c}, \tau_{k+c+1})$.

Let $W$ be the multiset of all $k$-factors of $U_t$, so $\Delta(W) < \alpha_0$.

We must have $\mathbf{2}^k \subseteq W$: all matches after $t$ have length at least $k$ by our assumption.

# Proof, ctnd.

All words of length $k+c+1$ on $U_t$ are unique, so there is a constant bounding the multiplicities of $x \in \mathbf{2}^k$ in $W$.

Split $W$ into $\mathbf{2}^k$ and a multiset: $W = \mathbf{2}^k + V$ where $\forall\, x \in \mathbf{2}^k\, (V(x) \leq d)$.

Let $\delta = \mathit{\Delta}(V)$ and $m = |V|$, so that

$$\alpha_0 > \mathit{\Delta}(W) = \frac{2^k \cdot 1/2 + m \cdot \delta}{2^k + m}$$

Hence $m = \mathit{\Omega}(2^k)$.

# Proof, ctnd.

On the other hand, for some $p$ we have

$$\alpha_0 \geq \Delta(V) \geq \Delta(d \cdot \mathbf{2}^{k,\leq p}) = \Delta(\mathbf{2}^{k,\leq p}).$$

If for some $x \in \mathbf{2}^k$, $q/k = \Delta(x) < \Delta(\mathbf{2}^k + d \cdot \mathbf{2}^{k,<q})$ then $\mathbf{2}^k + d \cdot \mathbf{2}^{k,\leq q}$ minimizes the density of all multisets with multiplicities bounded by $d$ that include $x$.

From a previous observation, $p \leq \alpha_0 k$.

# Proof, ctnd.

Using Sterling approximation we see that the cardinality $m$ is bounded by

$$d\binom{k}{\leq \alpha_0 k} \leq d + d\alpha_0 k\binom{k}{\alpha_0 k} \approx d + d\sqrt{\frac{\alpha_0 k}{2\pi(1-\alpha_0)}} < 2^{kH(\alpha_0)}$$

where $H(x) = -x \lg x - (1-x)\lg(1-x)$ is the binary entropy function.

$H$ is symmetric about $x = 1/2$ and concave, with maximum $H(1/2) = 1$.

Hence $2^{H(\alpha_0)} < 2$, contradicting our previous lower bound. □

◂ Return