On unavoidable sets, universal cycles and word patterns

Sergey Kitaev
University of Kentucky

Joint work with Alexander Burstein
Iowa State University

- A. Burstein and S. Kitaev: On unavoidable sets of word patterns, preprint.

- F. Chung, P. Diaconis and R. Graham: Universal cycles for combinatorial structures, *Discrete Math.* **110** (1992), 43–60.

- A. Evdokimov: Complete sets of words and their numerical characteristics, *Metody Diskret. Analiz.*, Novosibirsk, IM SB RAS, **39** (1983), 7–32 (in Russian). See MR#86e:68087.

- G. Hurlbert: Universal Cycles: On Beyond de Bruijn, PhD thesis, Department of Mathematics, Rutgers University, 1990.

Let $\mathcal{A} = \{a_1, \ldots, a_n\}$ be an alphabet of n letters.

Any i consecutive letters of a word X constitute a *subword* of length i.

$\mathcal{A}^*$ is the set of all the words in the alphabet $\mathcal{A}$.

Let $\mathcal{S} \subseteq \mathcal{A}^*$. Then $\mathcal{S}$ is called a *set of prohibited words* or a *set of prohibitions*. A word that does not contain any words from $\mathcal{S}$ as its subwords is said to be *free from* $\mathcal{S}$. The set of all words that are free from $\mathcal{S}$ is denoted by $\widehat{\mathcal{S}}$.

Example. Let $\mathcal{A} = \{a,\ b\}$ and $\mathcal{S} = \{aa,\ ba\}$. The word $abbb$ is in $\widehat{\mathcal{S}}$.

If there exists a $k \in \mathcal{N}$ such that the length of any word in $\widehat{\mathcal{S}}$ is less than k, then $\mathcal{S}$ is called an *unavoidable* set.

Example 1. $\mathcal{A} = \{1,\ 2,\ 3,\ 4\}$. The set of prohibitions is

$$\mathcal{S} = \{123,\ 13,\ 14,\ 11,\ 22,\ 33,\ 44\ \}.$$

Then $\mathcal{S}$ is avoidable, since the word $\underbrace{124124\ldots124}_{3k}$ is in $\widehat{\mathcal{S}}$ for any k.

Example 2. $\mathcal{A} = \{1, 2, 3\}$. The set of prohibitions is

$$\mathcal{S} = \{12,\ 23,\ 31,\ 32,\ 11,\ 22,\ 33\}.$$

It is easy to check that $\mathcal{S}$ is unavoidable.

In general it is important to know criteria of avoidability for different classes of sets of prohibited words. An example of such a class can be the following:

$$\{\mathcal{S} : \mathcal{S} \subseteq \mathcal{A}^n\};$$

that is, we are interested in words of length n. It is interesting to know numerical characteristics that describe the boundary between unavoidable and avoidable sets of prohibitions. For instance, let

$$M(n) = \min|\mathcal{S}| \quad \text{and} \quad L(n) = \max L(\widehat{\mathcal{S}}),$$

where the extremum is taken with respect to all unavoidable $\mathcal{S} \subseteq \mathcal{A}^n$ and $L(\widehat{\mathcal{S}}) = \max\limits_{X \in \widehat{\mathcal{S}}} |X|$.

De Brujin Graph

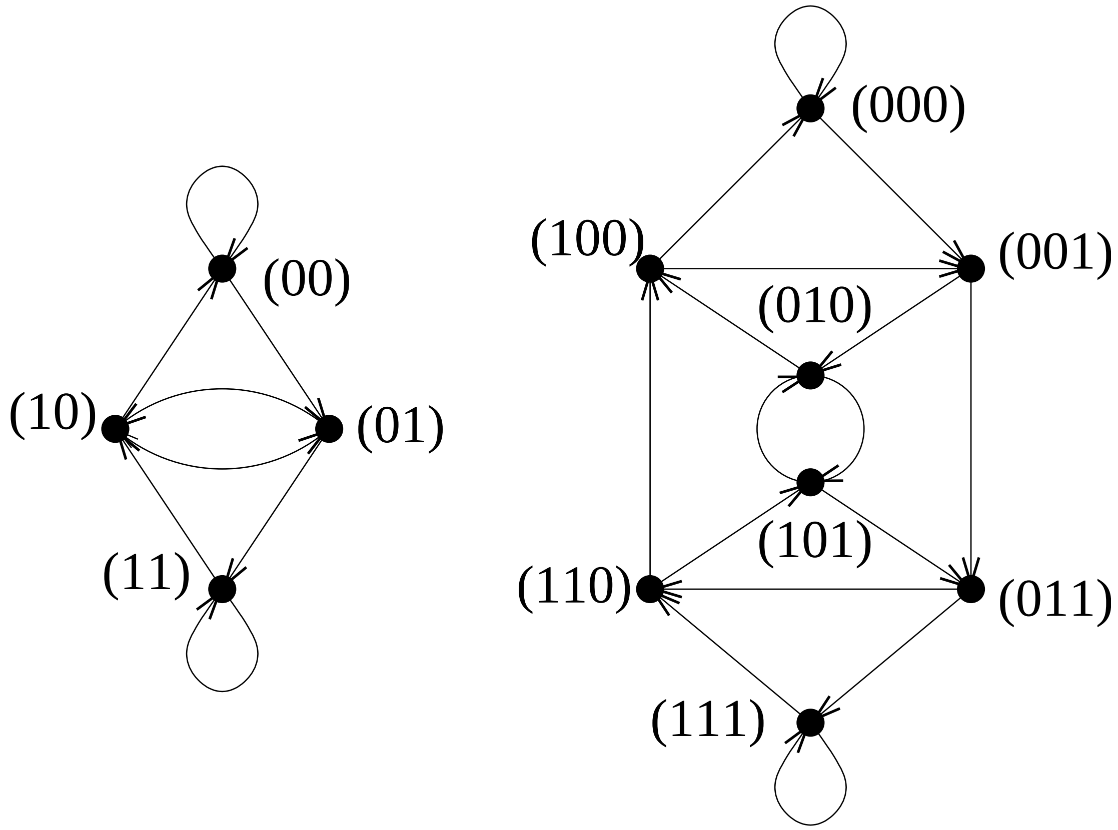
We define a directed graph $\vec{G}_n(V, E)$ by the following.

$V = \mathbf{A}^n$; There is an arc from $v_i = (v_{i_1}, \ldots, v_{i_n})$ to $v_j = (v_{j_1}, \ldots, v_{j_n})$ if

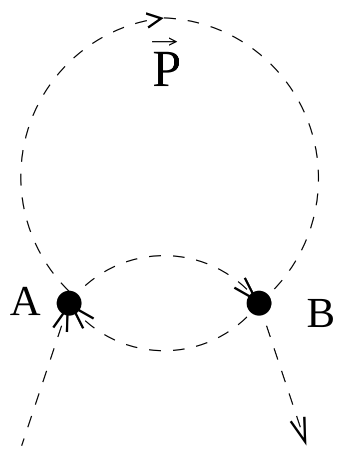
$$v_{i_2} = v_{j_1},\, v_{i_3} = v_{j_2}, \ldots, v_{i_n} = v_{j_{n-1}},$$

that is, when the words v_i and v_j overlap in $(n-1)$ letters.

The de Bruijn graphs for the alphabet $\mathcal{A} = \{0, 1\}$ and $n = 2, 3$.



A *chord* of a directed simple path $\vec{P}$ in $\vec{G}_n$ is an arc that does not belong to $\vec{P}$ but connects two of its vertices in a such way that there is a directed circuit generated by this arc and the part of the path between the ends of the arc. In the picture, the arc $\vec{BA}$ is a chord for the path $\vec{P}$, but $\vec{AB}$ is not.



$C(n)$ is the greatest length (the number of vertices) of a simple path in $\vec{G}_n$ that does not have chords and does not go through any vertex that has a loop.

Theorem 1. (A. Evdokimov) We have

$$L(n) = C(n) + n - 1 = |\mathbf{A}|^{n-1} + n - 2.$$

Theorem 2. (A. Evdokimov) We have

$$M(n) = \frac{1}{n} \sum_{d|n} \varphi(n/d) |\mathbf{A}|^d,$$

where $\varphi(n)$ is the number of integers among $1, 2, \ldots n-1$ that are relatively prime to n (Euler's φ-function).

Since any set of prohibitions $\mathcal{S}$ for $|\mathcal{S}| \leq M(n)$ is avoidable, it is helpful to have a table for $M(n)$. If $|\mathcal{A}| = 2$ and $1 < n \leq 10$ then we have

n	2	3	4	5	6	7	8	9	10
M(n)	3	4	6	8	14	20	36	60	108

In particular, any set of binary words of length 9 that has less than 60 words is avoidable.

$M(n) \sim |\mathcal{A}|^n/n$, when $n \to \infty$.

From now on the alphabets are totally ordered.

Word patterns without internal dashes = *patterns* in this talk

We define a pattern to be a word that contains each of the letters $1, 2, \ldots, k$ at least once for some k, and no other letters.

E.g. 2613235 contains an occurrence of the pattern 1323, but its subword 2613 is not a pattern.

An n*-pattern word* is a word in which each subword of length n is a pattern.

We consider *unavoidable sets of prohibited patterns*, or simply *sets of prohibitions*. We assume that all prohibitions are of length n.

Suppose our alphabet is $[m] = \{1, 2, \ldots, m\}$.

By analogy with the case of words, dealing with patterns and n-pattern words, we define

$$M(n, m) = \min |\mathcal{S}| \quad \text{and} \quad L(n, m) = \max L(\widehat{\mathcal{S}}),$$

where the extremum is taken with respect to all unavoidable sets of patterns $\mathcal{S}$ and $L(\widehat{\mathcal{S}}) = \max\limits_{X \in \widehat{\mathcal{S}}} |X|$.

Recall that the Möbius function is defined by

$$\mu(n) = \begin{cases} 0, & \text{if } n \text{ has a repeated prime factor,} \\ 1, & \text{if } n = 1, \\ (-1)^k, & \text{if } n \text{ is a product of } k \text{ distinct primes,} \end{cases}$$

so $\mu(n) \neq 0$ indicates that n is square-free.

Theorem. For n-pattern words over $[m]$, $M(n,m)$ is given by

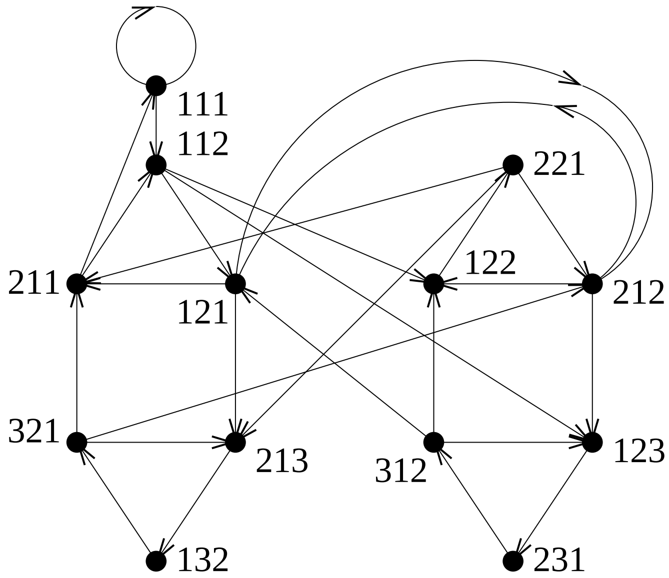
$$\sum_{i|n} \sum_{j=0}^{\min(i,m)-1} (-1)^j \binom{\min(i,m)-1}{j} \frac{1}{i} \sum_{d|i} \mu(d)(\min(i,m)-j)^{i/d}.$$

Remark 1. In the theorem, we can assume that $n \geq m$, since if $n < m$ we can only use the first n letters in $[m]$ to construct n-pattern words, which reduces to the case $n = m$.

Remark 2. For $n = m$, we have $\min(i,m) = i$ in the formula of the theorem.

The *graph of pattern overlaps* $\vec{P}_n = \vec{P}_n(V, E)$ is a subgraph of the de Bruijn graph $\vec{G}_n$, where the set of vertices V contains all n-letter patterns over the underlying alphabet $\mathcal{A}$.

The graph of pattern overlaps for $\mathcal{A} = \{1, 2, 3\}$ and $n = 3$:



Let $T(n,m)$ denote the number of conjugacy classes of patterns of length n over the alphabet $[m]$ with respect to a cyclic shift.

Lemma 1. $M(n,m) = T(n,m)$.

Lemma 2. $T(n,m)$ is given by

$$\sum_{i|n} \sum_{j=0}^{\min(i,m)-1} (-1)^j \binom{\min(i,m)-1}{j} \frac{1}{i} \sum_{d|i} \mu(d)(\min(i,m)-j)^{i/d}.$$

Let $C(n, m)$ denote the greatest length (the number of vertices) of a simple path in $\vec{P}_n$ that does not have chords and does not pass through any vertex incident with a loop.

Theorem 1. $L(n, m) = C(n, m) + n - 1.$

Theorem 2. $L(n, 2) = 2^{n-1} + n - 2.$

Proposition 1. $C(n,m) \leq$

$$\sum_{i|n}(i-1) \sum_{j=0}^{\min(i,m)-1} (-1)^j \binom{\min(i,m)-1}{j} \frac{1}{i} \sum_{d|i} \mu(d)(\min(i,m)-j)^{i/d}.$$

Proposition 2. $C(n,m) \geq \displaystyle\sum_{i=1}^{m} \sum_{\substack{a_1+\cdots+a_i=n-1 \\ a_1\geq 1,\ldots,a_i\geq 1}} \binom{n-1}{a_1,\ldots,a_i}.$

Theorem. The graph of pattern overlaps $\vec{P}_n$ contains a Hamiltonian circuit.

Universal cycles (U-cycles) for combinatorial structures

00011101 – includes every binary triple exactly once, if the end of the sequence is thought of as looping back to the beginning.

0000100110101111 – every 0-1 word of length 4 occurs uniquely under the same conditions as above.

These are examples of (binary) *de Bruijn cycles*.

A k*-ary de Bruijn cycle of order* n is a cyclic listing of length k^n of letters from the alphabet $\{0, 1, \ldots, k-1\}$ with the property that every n-tuple over the alphabet appears exactly once as a subword on the cycle.

Flye-Sainte Marie (1894) – existence of such cycles for any n and k in response to the question by Riviere.

de Bruijn (1946) – rediscovery $+$ the number of such cycles for $k = 2$.

Aardenne-Ehrenfest and de Bruijn (1951) – the number of such cycles is

$$(k!)^{k^{n-1}}/k^n.$$

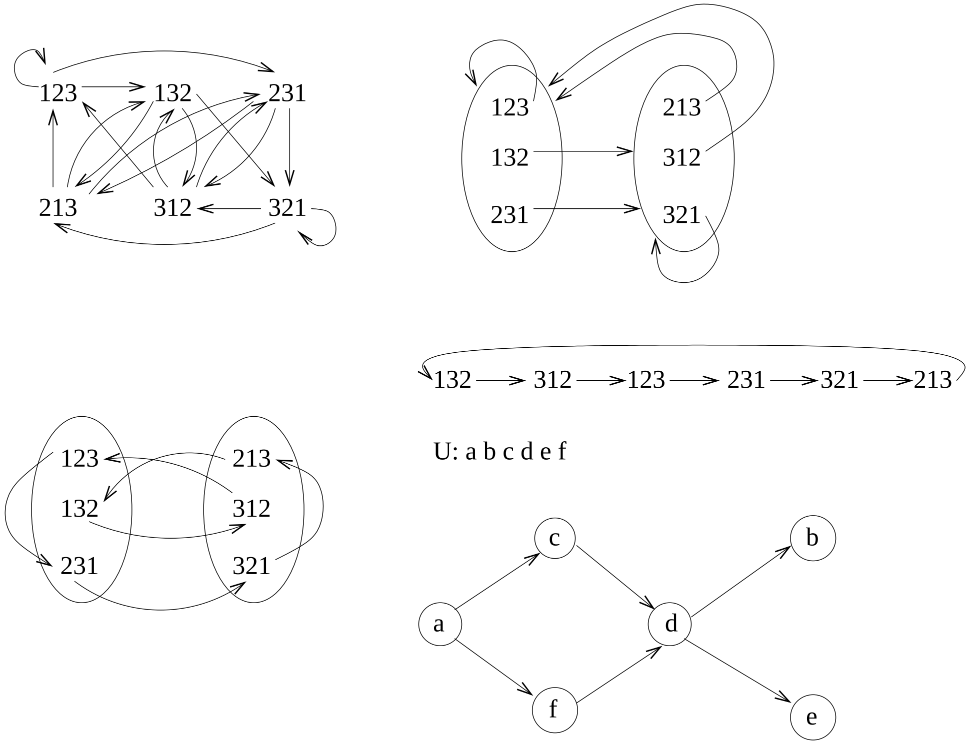
Universal cycles for permutations

$\mathcal{S}_n$ – the set of all $n!$ permutations of $[n] = \{1, 2, \ldots, n\}$.

$a_1 a_2 \ldots a_n$ is *order-isomorphic* to $b_1 b_2 \ldots b_n$ if $a_i < a_j \Leftrightarrow b_i < b_j$.

A U-cycle $a_0 a_1 \ldots a_{n!-1}$, $a_i \in \{1, 2, \ldots, N\}$, for $\mathcal{S}_n$ is an $n!$-tuple such that each $\sigma \in \mathcal{S}_n$ is order-isomorphic to exactly one block $a_{i+1} \ldots a_{i+n}$, where index addition is performed modulo $n!$. Clearly, we need $N > n$.

Theorem. (B. Jackson) U-cycles for permutations of $[n]$ exist for all n.



We need a mapping of $\{a, b, \ldots, f\}$ into $\{1, 2, \ldots, N\}$ which *preserves order*, i.e. a *linear extension* λ for suitable N.

$N = 4$, $\lambda(a) = 1$, $\lambda(c) = \lambda(f) = 2$, $\lambda(d) = 3$, $\lambda(b) = \lambda(e) = 4$, which gives the U-cycle 142342 for $\mathcal{S}_3$.

A U-cycle for $\mathcal{S}_4$: 123412534153214532413254.

Let $N(n)$ be the minimal number of different symbols any U-cycle for $\mathcal{S}_n$ must use.

$N(2) = 2,\ N(3) = 4,\ N(4) = 5,$ and $n + 1 \leq N(n) \leq 6n$ for $n \geq 5$

Conjecture. (F. Chung, P. Diaconis, and R. Graham) $N(n) = n + 1$, for $n \geq 3$.

Universal cycles for set partitions

The number of *partitions* of $[n]$ into k nonempty parts is denoted by $S(n,k)$ and called the *Stirling number of the second kind*.

$$S(n,k) = kS(n-1,k) + S(n-1,k-1)$$

$13|25|467|8$ corresponds to $abacbccd$.

Theorem 1. (F. Chung, P. Diaconis, and R. Graham) U-cycles for partitions of $[n]$ exist for all $n \geq 4$.

Theorem 2. (G. Hurlbert) U-cycles for $\leq k$-partitions of $[n]$ exist for all n and $4 \leq k \leq n$.

Theorem 3. (G. Hurlbert) U-cycles for k-partitions of $[n]$ exist for $k = n-1$ and all odd $n \geq 3$.

Theorem. For any m and n, there exists a U-cycle for word patterns of length n over an m-letter alphabet.

Considerations of the conjugacy classes of word patterns with respect to a cyclic shift provide a combinatorial proof of the following identity.

$$\sum_{i=1}^{m} \sum_{\substack{a_1+\cdots+a_i=n \\ a_1\geq 1,\ldots,a_i\geq 1}} \binom{n}{a_1,\ldots,a_i} =$$

$$\sum_{i|n} \sum_{j=0}^{\min(i,m)-1} (-1)^j \binom{\min(i,m)-1}{j} \sum_{d|i} \mu(d)(\min(i,m)-j)^{i/d}.$$