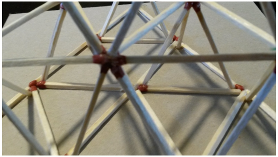
On the Parikh-de-Bruijn grid

Péter Burcsi Zsuzsanna Lipták W. F. Smyth

ELTE Budapest (Hungary), U of Verona (Italy),
McMaster U (Canada) & Murdoch U (Australia)



Words & Complexity 2018
Lyon, 19-23 Feb. 2018

Alphabet Σ: finite, ordered, constant size σ

- Given string s, the Parikh vector of s $\mathbf{pv}(s)$ is vector of multiplicities of characters
- Given a Pv p, its order is the sum of its entries = length of a string with Pv p

Ex. $s = \mathtt{aabaccba}$ over $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, then $\mathbf{pv}(s) = (4, 2, 2)$, order 8.

Alphabet Σ: finite, ordered, constant size σ

- Given string s, the Parikh vector of s $\mathbf{pv}(s)$ is vector of multiplicities of characters
- Given a Pv p, its order is the sum of its entries = length of a string with Pv p

Ex. $s = \texttt{aabaccba}$ over $\Sigma = \{\texttt{a}, \texttt{b}, \texttt{c}\}$, then $\mathbf{pv}(s) = (4, 2, 2)$, order 8.

- Two strings over the same alphabet are Parikh equivalent (a.k.a. abelian equivalent) if they have the same Parikh vector (i.e. if they are permutations of one another)

Ex. `aaaabbcc` and `aabcaabc` are both Parikh equivalent to s.

Alphabet Σ: finite, ordered, constant size σ

- Given string s, the Parikh vector of s $\mathbf{pv}(s)$ is vector of multiplicities of characters
- Given a Pv p, its order is the sum of its entries = length of a string with Pv p

Ex. $s = \texttt{aabaccba}$ over $\Sigma = \{\texttt{a}, \texttt{b}, \texttt{c}\}$, then $\mathbf{pv}(s) = (4, 2, 2)$, order 8.

- Two strings over the same alphabet are Parikh equivalent (a.k.a. abelian equivalent) if they have the same Parikh vector (i.e. if they are permutations of one another)

Ex. aaaabbcc and aabcaabc are both Parikh equivalent to s.

In Abelian stringology, equality is replaced by Parikh equivalence.

Abelian stringology

In Abelian stringology, equality is replaced by Parikh equivalence.

- Jumbled Pattern Matching
- abelian borders
- abelian periods
- abelian squares, repetitions, runs
- abelian pattern avoidance
- abelian reconstruction
- abelian problems on run-length encoded strings
- abelian complexity
- ...

Abelian stringology

In this talk, we introduce a new tool for attacking abelian problems.

Abelian stringology

In this talk, we introduce a new tool for attacking abelian problems.

But first: What's so different about abelian problems?

An example problem

Parikh-de-Bruijn strings

Recall: A de Bruijn sequence of order k over alphabet Σ is a string over Σ which contains every $u \in \Sigma^k$ exactly once as a substring.

Parikh-de-Bruijn strings

Recall: A de Bruijn sequence of order k over alphabet Σ is a string over Σ which contains every $u \in \Sigma^k$ exactly once as a substring.

Ex. $\Sigma = \{\mathtt{a}, \mathtt{b}\}$, order 2: `aabba`, order 3: `aaababbbaa`

Parikh-de-Bruijn strings

Recall: A de Bruijn sequence of order k over alphabet Σ is a string over Σ which contains every $u \in \Sigma^k$ exactly once as a substring.

Ex. $\Sigma = \{\mathtt{a}, \mathtt{b}\}$, order 2: `aabba`, order 3: `aaababbbaa`

We define the abelian analogue:

Def. A Parikh-de-Bruijn string of order k (a (k, σ)-PdB-string) is a string s over an alphabet of size σ s.t.

$$\forall\ p \text{ Parikh vector of order } k\ \exists!(i,j) \text{ s.t. } \mathbf{pv}(s_i \cdots s_j) = p$$

Ex. `aabbcca` is a $(\overset{k}{2}, \overset{\sigma}{3})$-PdB-string

Parikh-de-Bruijn strings

Classical case: De Bruijn sequences exist for every Σ and k.

Parikh-de-Bruijn strings

Classical case: De Bruijn sequences exist for every Σ and k.

- correspond to Hamiltonian paths in the de Bruijn graph of order k

Parikh-de-Bruijn strings

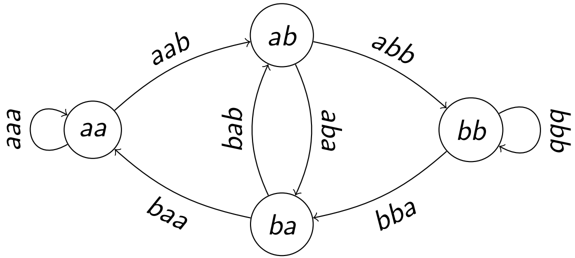
Classical case: De Bruijn sequences exist for every Σ and k.

- correspond to Hamiltonian paths in the de Bruijn graph of order k
- and to Euler-paths in the de Bruijn graph of order $k-1$

Parikh-de-Bruijn strings

Classical case: De Bruijn sequences exist for every Σ and k.

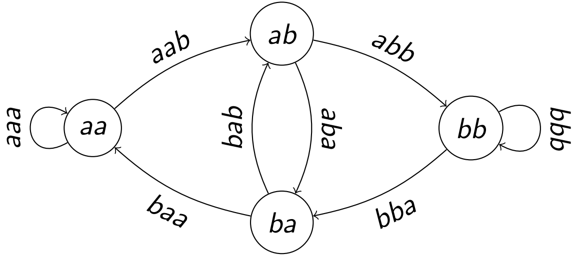
- correspond to Hamiltonian paths in the de Bruijn graph of order k
- and to Euler-paths in the de Bruijn graph of order $k-1$



Parikh-de-Bruijn strings

Classical case: De Bruijn sequences exist for every Σ and k.

- correspond to Hamiltonian paths in the de Bruijn graph of order k
- and to Euler-paths in the de Bruijn graph of order $k-1$

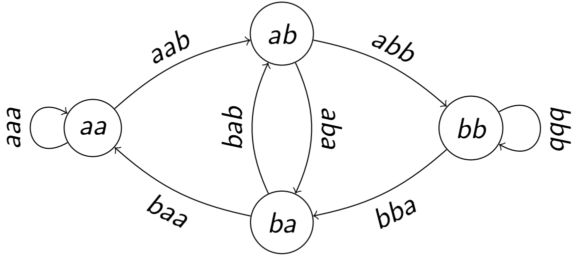


order 2: aabba

Parikh-de-Bruijn strings

Classical case: De Bruijn sequences exist for every Σ and k.

- correspond to Hamiltonian paths in the de Bruijn graph of order k
- and to Euler-paths in the de Bruijn graph of order $k-1$



order 2: aabba

order 3: aaababbbaa

Parikh-de-Bruijn strings

Abelian case:

- aabbcca is a $(\overset{k}{2}, \overset{\sigma}{3})$-PdB-string

Parikh-de-Bruijn strings

Abelian case:

- aabbcca is a $(\overset{k}{2}, \overset{\sigma}{3})$-PdB-string
- abbbcccaaabc is a $(3, 3)$-PdB-string

Parikh-de-Bruijn strings

Abelian case:

- aabbcca is a $(\overset{k}{2}, \overset{\sigma}{3})$-PdB-string
- abbbcccaaabc is a $(3, 3)$-PdB-string
- but no $(4, 3)$-PdB-string exists

Parikh-de-Bruijn strings

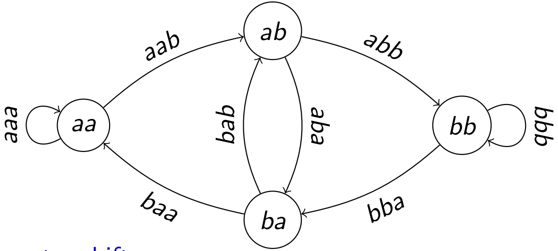
Abelian case:

- aabbcca is a $(\overset{k}{2}, \overset{\sigma}{3})$-PdB-string
- abbbcccaaabc is a $(3, 3)$-PdB-string
- but no $(4, 3)$-PdB-string exists
- and no $(2, 4)$-PdB-string exists

The Parikh-de-Bruijn grid

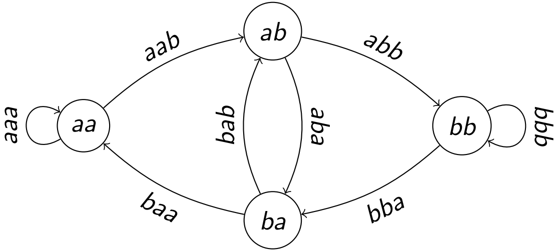
Recall: de Bruijn graph $dB_k = (V, E)$, where $V = \Sigma^k$ and $(\mathtt{x}u, u\mathtt{y}) \in E$ for all $\mathtt{x}, \mathtt{y} \in \Sigma$ and $u \in \Sigma^{k-1}$

N.B. edges $\hat{=}$ $(k+1)$-length strings



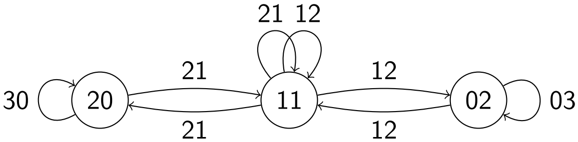
edges $\hat{=}$ one-character shifts

Recall: de Bruijn graph $dB_k = (V, E)$, where $V = \Sigma^k$ and $(\mathtt{x}u, u\mathtt{y}) \in E$ for all $\mathtt{x}, \mathtt{y} \in \Sigma$ and $u \in \Sigma^{k-1}$ **N.B.** edges $\hat{=}$ $(k+1)$-length strings



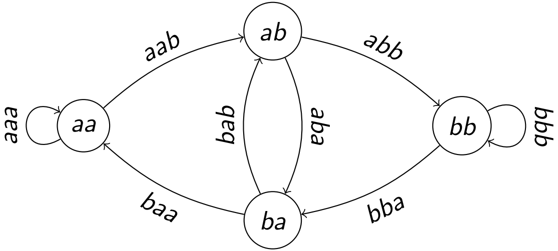
edges $\hat{=}$ one-character shifts

A straightforward generalization to Pv's gives:



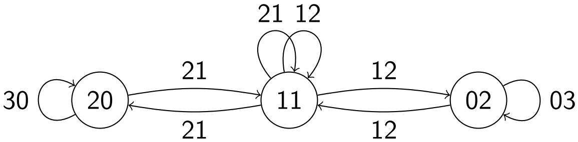
Recall: de Bruijn graph $dB_k = (V, E)$, where $V = \Sigma^k$ and $(\mathtt{x}u, u\mathtt{y}) \in E$ for all $\mathtt{x}, \mathtt{y} \in \Sigma$ and $u \in \Sigma^{k-1}$

N.B. edges $\triangleq$ $(k+1)$-length strings



edges $\triangleq$ one-character shifts

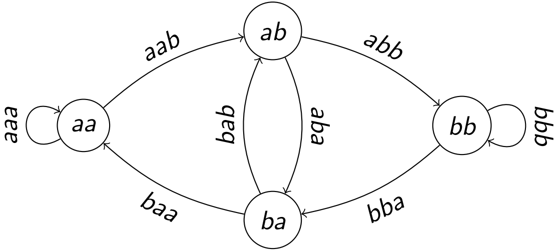
A straightforward generalization to Pv's gives:



edges $\triangleq$ one-character shifts,

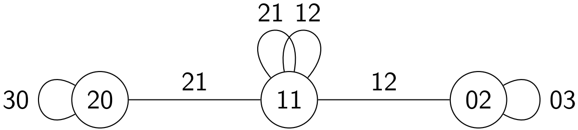
Recall: de Bruijn graph $dB_k = (V, E)$, where $V = \Sigma^k$ and $(\mathtt{x}u, u\mathtt{y}) \in E$ for all $\mathtt{x}, \mathtt{y} \in \Sigma$ and $u \in \Sigma^{k-1}$

N.B. edges $\hat{=}$ $(k+1)$-length strings



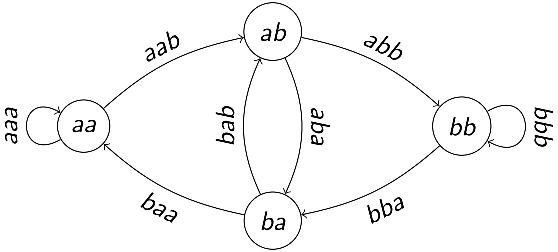
edges $\hat{=}$ one-character shifts

A straightforward generalization to Pv's gives:



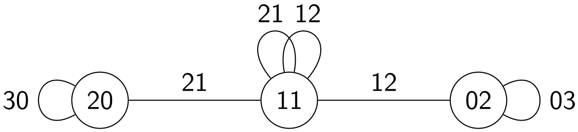
edges $\hat{=}$ one-character shifts, undirected edges

Recall: de Bruijn graph $dB_k = (V, E)$, where $V = \Sigma^k$ and $(\mathtt{x}u, u\mathtt{y}) \in E$ for all $\mathtt{x}, \mathtt{y} \in \Sigma$ and $u \in \Sigma^{k-1}$ **N.B.** edges $\hat{=}$ $(k+1)$-length strings



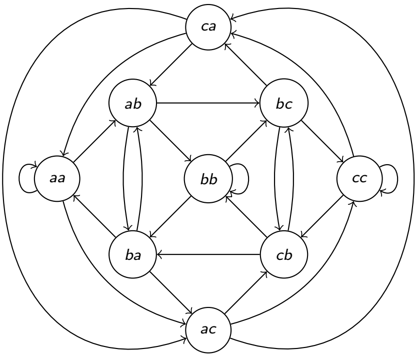
edges $\hat{=}$ one-character shifts

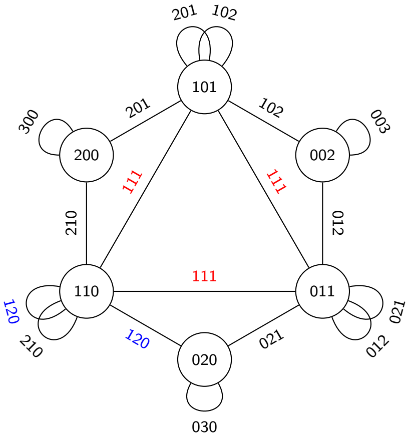
A straightforward generalization to Pv's gives: **NO:** edges $\hat{=}$ $(k+1)$-order Pv's!



edges $\hat{=}$ one-character shifts, undirected edges

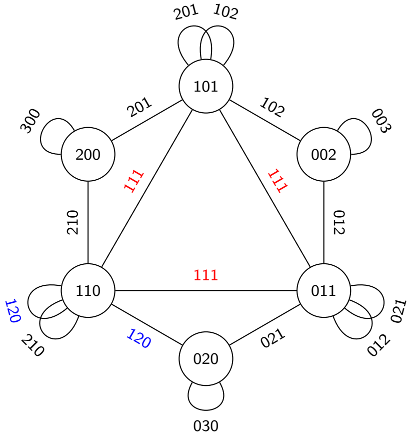
Let's look at another example: Here, $\sigma = 3, k = 2$.





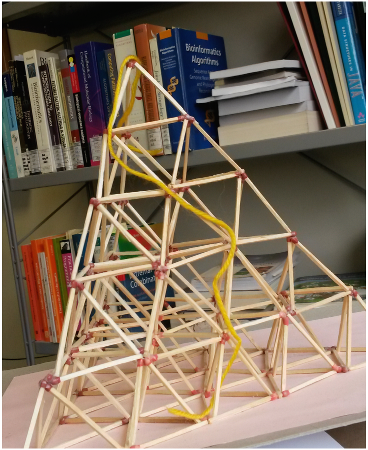
In the abelian version, several edges have the same label (i.e. here: 3-order Pv).

Turns out the right way to generalize de Bruijn graphs is the Parikh-de-Bruijn grid:



?

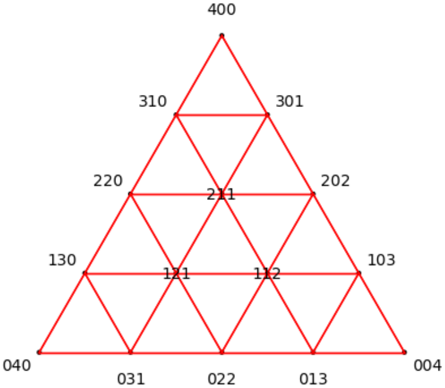
For $\sigma = 4, k = 5$, the Parikh-de-Bruijn grid looks like this:



The Parikh-de-Bruijn grid

PdB-grid:

- $V = k$-order Pv's

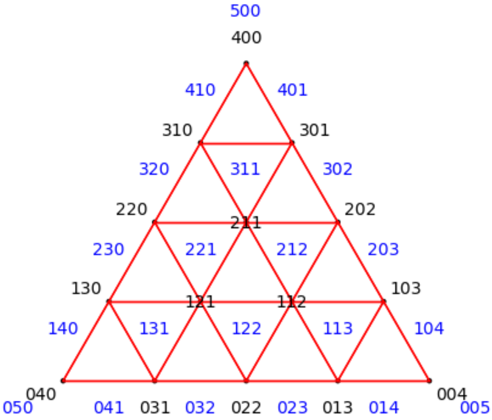


(4, 3)-grid; loops not included in figure

The Parikh-de-Bruijn grid

PdB-grid:

- $V = k$-order Pv's
- $pq \in E$ iff exist $\mathtt{x}, \mathtt{y} \in \Sigma$ s.t. $p = q - \mathtt{x} + \mathtt{y}$

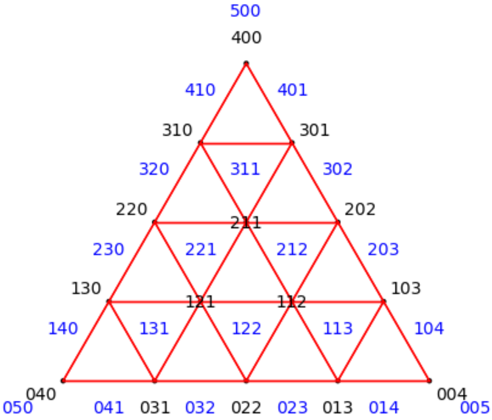


(4, 3)-grid; loops not included in figure

The Parikh-de-Bruijn grid

PdB-grid:

- $V = k$-order Pv's
- $pq \in E$ iff exist $\mathtt{x}, \mathtt{y} \in \Sigma$ s.t. $p = q - \mathtt{x} + \mathtt{y}$
- undirected edges (or: bidirectional edges)

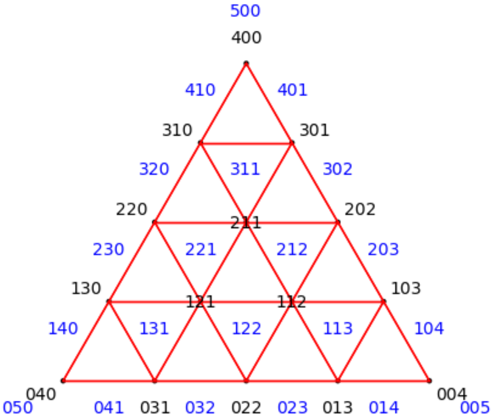


$(4, 3)$-grid; loops not included in figure

The Parikh-de-Bruijn grid

PdB-grid:

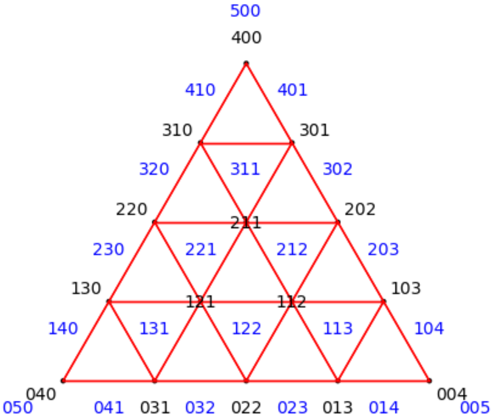
- V = k-order Pv's
- $pq \in E$ iff exist $\mathtt{x}, \mathtt{y} \in \Sigma$ s.t. $p = q - \mathtt{x} + \mathtt{y}$
- undirected edges (or: bidirectional edges)
- loops at every node, different one for each non-zero character



(4, 3)-grid; loops not included in figure

PdB-grid:

- V = k-order Pv's
- $pq \in E$ iff exist $\mathtt{x}, \mathtt{y} \in \Sigma$ s.t. $p = q - \mathtt{x} + \mathtt{y}$
- undirected edges (or: bidirectional edges)
- loops at every node, different one for each non-zero character
- $(k+1)$- and $(k-1)$-order Pv's $\hat{=}$ $(\sigma - 1)$-simplices (triangles for $\sigma = 3$, tetrahedra for $\sigma = 4$ etc.)

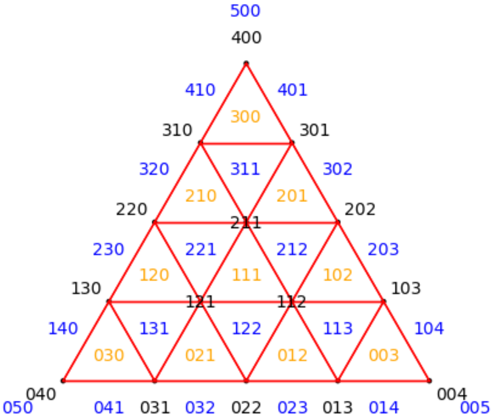


$(4, 3)$-grid; loops not included in figure

The Parikh-de-Bruijn grid

PdB-grid:

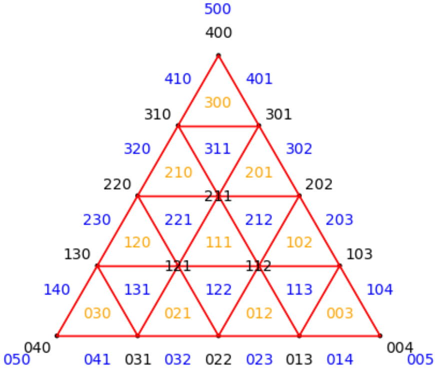
- V = k-order Pv's
- $pq \in E$ iff exist $\mathtt{x}, \mathtt{y} \in \Sigma$ s.t. $p = q - \mathtt{x} + \mathtt{y}$
- undirected edges (or: bidirectional edges)
- loops at every node, different one for each non-zero character
- $(k+1)$- and $(k-1)$-order Pv's $\hat{=}$ $(\sigma - 1)$-simplices (triangles for $\sigma = 3$, tetrahedra for $\sigma = 4$ etc.)



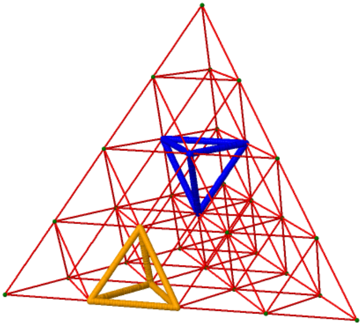
(4, 3)-grid; loops not included in figure

The Parikh-de-Bruijn grid

The $(4, 3)$-PdB-grid

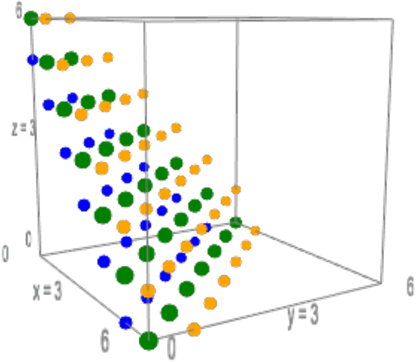


The $(4, 4)$-PdB-grid



vertices: k-order Pv's (vertices), downward triangles/tetrahedra: $(k+1)$-order Pv's, (upward triangles/tetrahedra: $(k-1)$-order Pv's.

The Parikh-de-Bruijn grid



The diagonal section of the integer grid with the hyperplanes $\mathcal{H}_k$ (green), $\mathcal{H}_{k+1}$ (blue), and $\mathcal{H}_{k-1}$ (yellow), for $k = 6$ and $\sigma = 3$.

Back to the example problem

More differences

Classical case: (dB_k) One-to-one correspondence: walks and strings.
Abelian case: Every string corresponds to a walk in the PdB-grid, but not every walk corresponds to a string.

More differences

Classical case: (dB_k) One-to-one correspondence: walks and strings.
Abelian case: Every string corresponds to a walk in the PdB-grid, but not every walk corresponds to a string.

Classical case: De Bruijn sequences exist for every k and σ.
Abelian case: PdB-strings do not exist for every k and σ.

(**N.B.** Not all PdB-strings come from circular strings = universal cycles!)

Back to Parikh-de-Bruijn strings

Theorem 1

No $(k, 3)$-PdB strings exist for $k \geq 4$.

Theorem 2

A $(2, \sigma)$-PdB string exists if and only if σ is odd.

Theorem 3

A $(3, \sigma)$-PdB string exists if and only if $\sigma = 3$ or σ not a multiple of 3.

Theorem 1

No $(k, 3)$-PdB strings exist for $k \geq 4$.

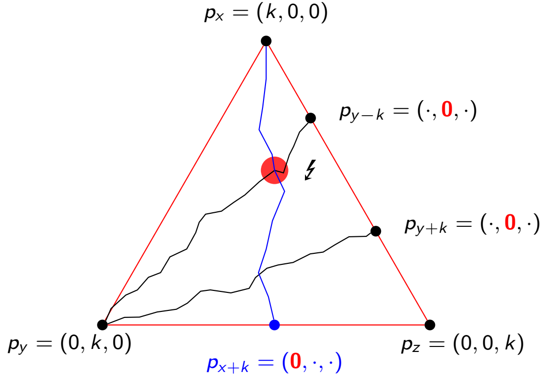
Lemma 1

If the walk induced by string w does not use any loops, then for all i: $w_i \neq w_{i+k}$.

Proof

Otherwise we would have two consecutive occurrences of the same Pv p, thus using a loop at p.

Theorem 1: No $(k,3)$-PdB strings exists for $k \geq 4$. (Proof uses Lemma 1.)



$$w = \cdots \quad \underbrace{aaa\cdots a}_{k}\cdots \quad \underbrace{bbb\cdots b}_{k}\cdots \quad \underbrace{ccc\cdots c}_{k}\cdots$$

$\uparrow$ position x

$\uparrow$ position y

$\uparrow$ position z

Example 2: Covering strings

Classical case: If s is a (classical) de Bruijn sequence of order k, then it also contains all $(k-1)$-length strings as substrings.

Example 2: Covering strings

Classical case: If s is a (classical) de Bruijn sequence of order k, then it also contains all $(k-1)$-length strings as substrings.

Abelian case: not always true, e.g.

`aaaaabbbbbcaaaadbbbcccccddddaaaccdbcbaccaccddbddbadacddbbbb`

is a $(5,4)$-PdB-string but is not $(4,4)$-covering: no substring with Pv $(1,1,1,1)$.

Example 2: Covering strings

Classical case: If s is a (classical) de Bruijn sequence of order k, then it also contains all $(k-1)$-length strings as substrings.

Abelian case: not always true, e.g.

`aaaaabbbbbcaaaadbbbcccccddddaaaccdbcbaccaccddbddbadacddbbbb`

is a $(5,4)$-PdB-string but is not $(4,4)$-covering: no substring with Pv $(1,1,1,1)$.

Theorem 4

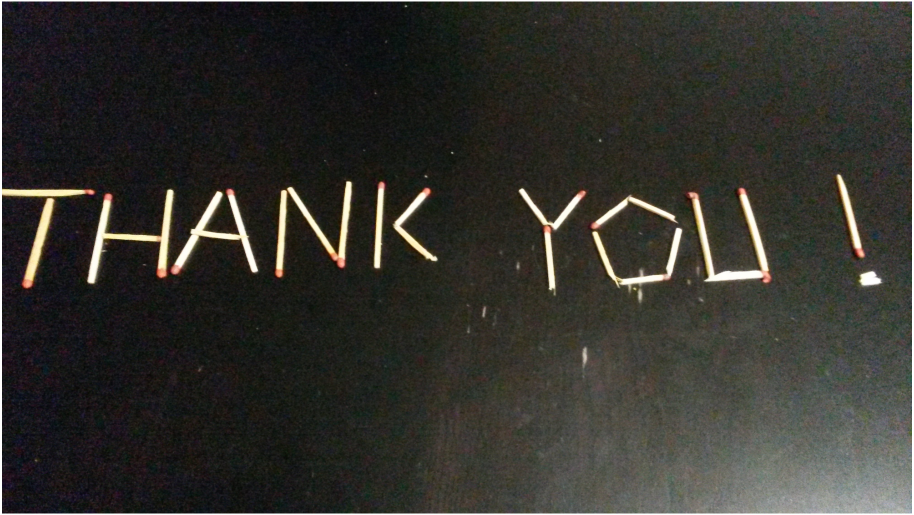
For every $\sigma \geq 3$ and $k \geq 4$, there exist (k,σ)-covering strings which are not $(k-1,\sigma)$-covering.

Experimental results

k	σ	*string*	*length (excess)*
2	3	aabbcca	7 (0)
3	3	abbbcccaaabc	12 (0)
4	3	aaaabbbbccccaacabcb	19 (1)
5	3	aaaaabbbacccccbbbbbaacaaccb	27 (2)
6	3	aaaabccccccaaaaaabbbbbbcccbbcabbaca	35 (2)
7	3	aabbbccbbcccabacaaabcbbbbbbbaaaaaaaccccccccba	44 (2)
2	4	aabbcadbccdd	12 (1)
3	4	aaabbbcaadbdbccadddccc	22 (0)
4	4	aabbbbcaacadbddbccacddddaaaabdbbccccdd	38 (0)
5	4	aaaaabbbbbcaaaadbbbcccccddddaaaccdbcbaccaccddbddbadacddbbbb	60 (0)
2	5	aabbcadbeccddeea	16 (0)
3	5	aaabbbcaadbbeaccbdddcccebededadceeeaa	37 (0)
4	5	aaaabbbbcaaadbbbeaaccbbddaaeaebcccadbeeeadddcccceeeedddd...	73 (0)

Conclusion and open problems

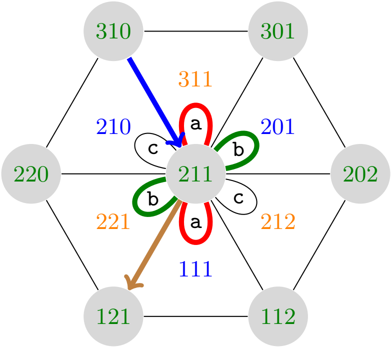
- new tool for modeling and solving abelian problems
- find good characterization for walks which correspond to strings
- many open problems on PdB- and covering strings, e.g.
  - For which σ and k do *PdB*-strings exist? (We answered this question only for some special cases.)
  - What is the length of a shortest covering string when no PdB-string exists, e.g. $k = 3$, $\sigma = 6$?
  - What is the minimum proportion of $(k - 1)$-order Pv's covered by a k-covering string?
- apply PdB-grid to other abelian problems
- paper on Arxiv


THANK YOU !

Appendix

The Parikh-de-Bruijn grid

$k = 4$, $\sigma = 3$



$(k+1)$	a	3	3	2	2				
	b	1	1	2	2				
	c	1	1	1	1				
		a	a	b	a	c	a	b	b
k	a	3	2	2	2	1			
	b	1	1	1	1	2			
	c	0	1	1	1	1			
$(k-1)$	a		2	1	2	1			
	b		1	1	0	1			
	c		0	1	1	1			

Walk corresponding to aabacabb. $(k+1)$- and $(k-1)$-order Pv's: triangles incident to the edges traversed by the walk. The $(k+1)$ and $(k-1)$-order Pv's for loops (same k-order Pv twice) lie in opposite direction, hence the name bow.

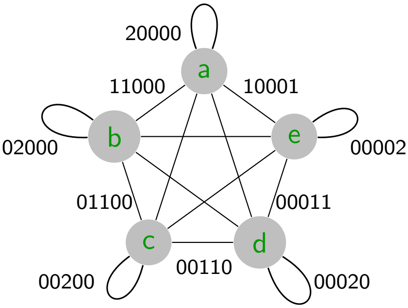
Parikh-de-Bruijn and covering strings

Theorem 2

A $(2, \sigma)$-PdB string exists if and only if σ is odd.

Proof

Pv's of order 2 either form $(0...0, \overset{i}{2}, 0..0)$ or $(0...0, \overset{i}{1}, 0...0, \overset{j}{1}, 0..0)$. So we need exactly one substring aa for all $\mathtt{a} \in \Sigma$, and either ab or ba for all $\mathtt{a}, \mathtt{b} \in \Sigma$. Consider the undirected complete graph $G = (V, E)$ with loops where $V = \Sigma$ (N.B.: not the PdB-grid!): an Euler path exists iff σ is odd.



Example 2: Covering strings

Next best thing: covering strings.

Def.

- We call a string s (k,σ)-covering if

$$\forall\ p \text{ Parikh vector of order } k\ \exists(i,j) \text{ s.t. } \mathbf{pv}(s_i \cdots s_j) = p$$

  (There is at least one substring in s which has Pv p.)
- The excess of s is: $|s| - \underbrace{\binom{\sigma+k-1}{k} + k - 1}_{\text{length of a PdB-string}}$.

Ex.

- `aaaabbbbccccaacabcb` is a shortest $(4,3)$-covering string, with excess 1.
- `aabbcadbccdd` is a shortest $(2,4)$-covering string, with excess 1.

Example 2: Covering strings

Classical case: If s is a (classical) de Bruijn sequence of order k, then it also contains all $(k-1)$-length strings as substrings.

For PdB-strings, this is not always true, e.g.

`aaaaabbbbbcaaaadbbbcccccddddaaaccdbcbaccaccddbddbadacddbbbb`

is a $(5,4)$-PdB-string but is not $(4,4)$-covering: no substring with Pv $(1,1,1,1)$.

Theorem 4

For every $\sigma \geq 3$ and $k \geq 4$, there exist (k,σ)-covering strings which are not $(k-1,\sigma)$-covering.

The Parikh-de-Bruijn grid

Lemma 2

A set of k-order Parikh vectors is realizable if and only if the induced subgraph in the k-PdB-grid is connected.

realizable = exists string with exactly these k-order sub-Pv's.

Proof sketch

$\Rightarrow$: clear.

$\Leftarrow$: Use loops until undesired character $\mathtt{x}$ exits, replace by new character $\mathtt{y}$.

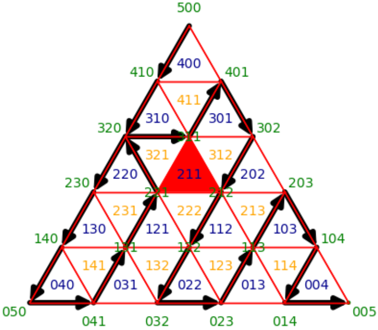
Parikh-de-Bruijn and covering strings

Theorem 4

For every $\sigma \geq 3$ and $k \geq 4$, there exist (k, σ)-covering strings which are not $(k-1, \sigma)$-covering.

Proof

$w =$ `aaaaabbbbbcabbaaacacbbcbccacaccccbccccc`



General construction:

- remove $(k-1)$-order Pv $p = (k-3, 1, 1, 0, \ldots, 0)$ with incident edges and vertices
- the rest is connected, hence a string exists (Lemma 2)
- add vertices of p without traversing edges incident to p
- can be done by detours from corners of PdB-grid