On Combinatorial Generation of Prefix Normal Words

Péter Burcsi, Gabriele Fici, Zsuzsanna Lipták, Frank Ruskey, and Joe Sawada

CPM 2014
Moscow, 16-18 June 2014

How marrying two topics can lead to an explosion of results

Outline

- def 1: prefix normal words
- def 2: bubble languages
- the marriage
- generation algorithm
- $\rightsquigarrow$ enumeration results, testing algorithm, experimental results, new insights, and, and, and ...

Prefix Normal Words

Prefix normal words

Fici, Lipták (DLT 2011)

Definition

A binary word w is prefix normal (w.r.t. 1) if no substring has more 1s than the prefix of the same length.

Example

$$w = 101100010011011100 10$$

$$w' = 111010010110010100 10$$

Prefix normal words

Fici, Lipták (DLT 2011)

Definition

A binary word w is prefix normal (w.r.t. 1) if no substring has more 1s than the prefix of the same length.

Example

$$w = 10110001001101110010 \quad NO$$

$$w' = 11101001011001010010 \quad YES$$

Prefix normal words

Fici, Lipták (DLT 2011)

Definition

A binary word w is prefix normal (w.r.t. 1) if no substring has more 1s than the prefix of the same length.

Example

$$w = 10110001001101110010 \quad NO$$

$$w' = 11101001011001010010 \quad YES$$

$\mathcal{L}_{\mathrm{PN}}$ = all prefix normal words.
Exists canonical prefix normal form of every binary word w: $\mathrm{PNF}_1(w)$.

Binary Jumbled Pattern Matching (BJPM)

Does $\mathbf{w} = \mathbf{10100110110001110010}$ have a substring of length 11 containing exactly 5 ones?

Binary Jumbled Pattern Matching (BJPM)

Does $\mathbf{w} = \mathbf{10100110110001110010}$ have a substring of length 11 containing exactly 5 ones? (Online: easy, $\mathcal{O}(|w|)$. Indexed: ?)

Binary Jumbled Pattern Matching (BJPM)

Does $\mathbf{w} = \mathbf{10100110110001110010}$ have a substring of length 11 containing exactly 5 ones? (Online: easy, $\mathcal{O}(|w|)$. Indexed: ?)

Papers dealing with the indexed version of this problem appeared in:
PSC 2009, FUN 2010, IPL 2010, JDA 2012, ToCS 2012, IJFCS 2012, CPM 2012, SPIRE 2012, IPL 2013, IPL 2013, ESA 2013 x 2, SPIRE 2013, TCS 2014, PhTRS-A 2014, CPM 2014, CPM 2014, ISIT 2014, ICALP 2014, ...

(red ones have an intersection with the authors of this paper)

Binary Jumbled Pattern Matching (BJPM)

Does $\mathbf{w} = \mathbf{10100110110001110010}$ have a substring of length 11 containing exactly 5 ones? (Online: easy, $\mathcal{O}(|w|)$. Indexed: ?)

Interval property $\rightsquigarrow$ linear size index, $\mathcal{O}(1)$ query time:

Fix length k of substrings: # 1s builds an interval.

Ex: $k = 11 : 5, 6, 7$ ones.

For each k, store max and min no. of 1s.

k	1	2	3	4	5	...	11	...
max # 1s	1	2	3	3	4	...	7	...
min # 1s	0	0	0	1	2	...	5	...

Binary Jumbled Pattern Matching (BJPM)

Does $\mathbf{w} = \mathbf{10100110110001110010}$ have a substring of length 11 containing exactly 5 ones? (Online: easy, $\mathcal{O}(|w|)$. Indexed: ?)

Interval property $\rightsquigarrow$ linear size index, $\mathcal{O}(1)$ query time:

Fix length k of substrings: # 1s builds an interval.

Ex: $k = 11 : 5, 6, 7$ ones.

For each k, store max and min no. of 1s.

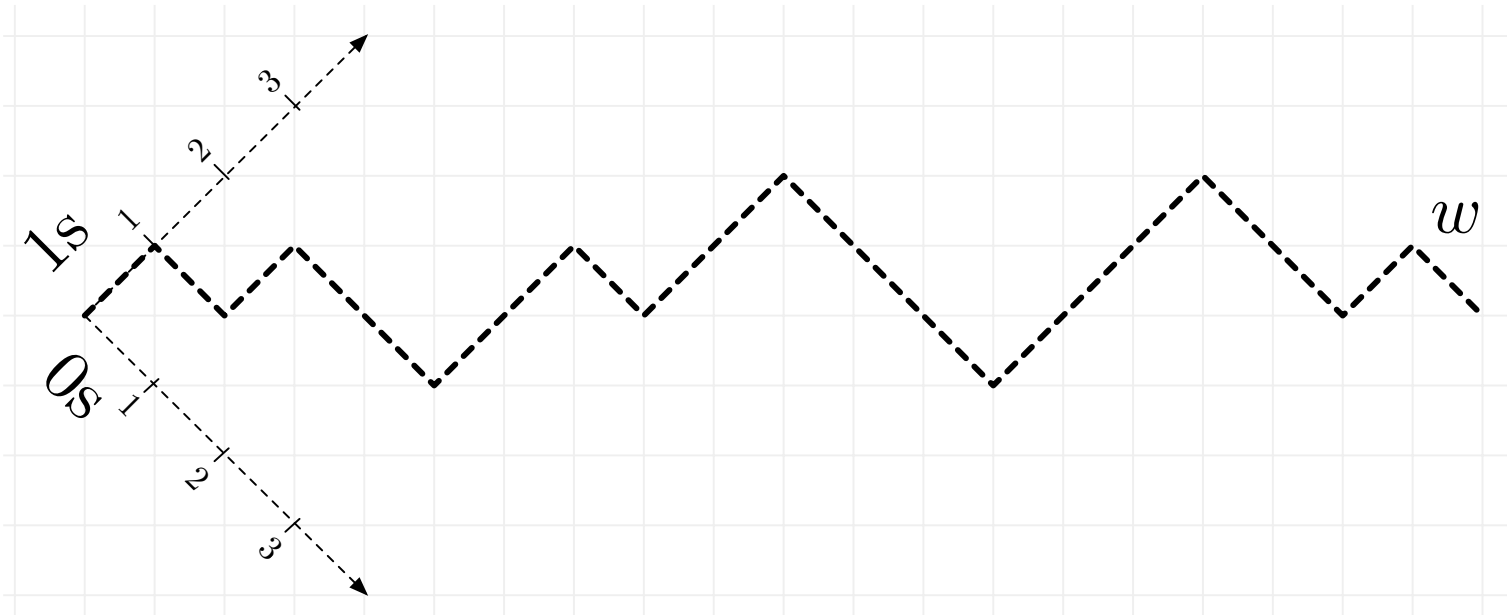
k	1	2	3	4	5	...	11	...
max # 1s	1	2	3	3	4	...	7	...
min # 1s	0	0	0	1	2	...	5	...

Research problem:

Compute this index efficiently.

BJPM with prefix normal forms

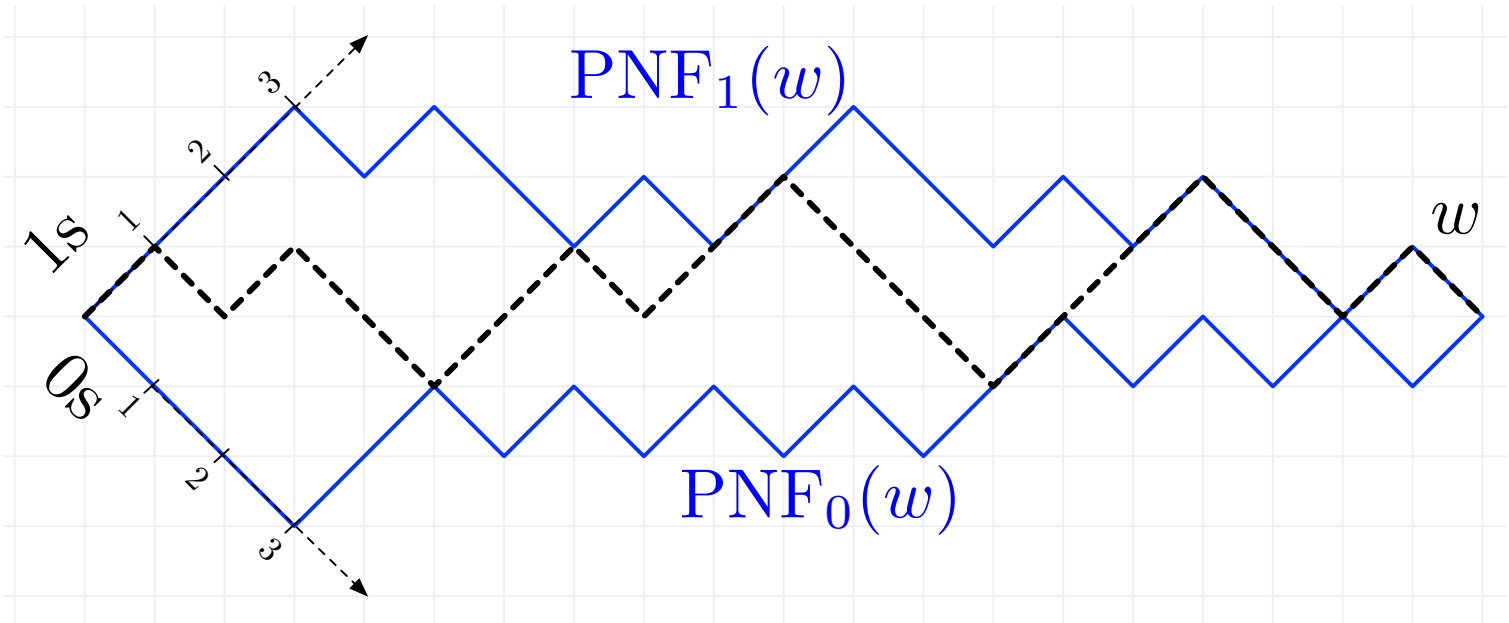
Does $\mathbf{w} = \mathbf{10100110110001110010}$ have a substring of length 11 containing exactly 5 ones?



$/ = 1$, $\backslash = 0$,

BJPM with prefix normal forms

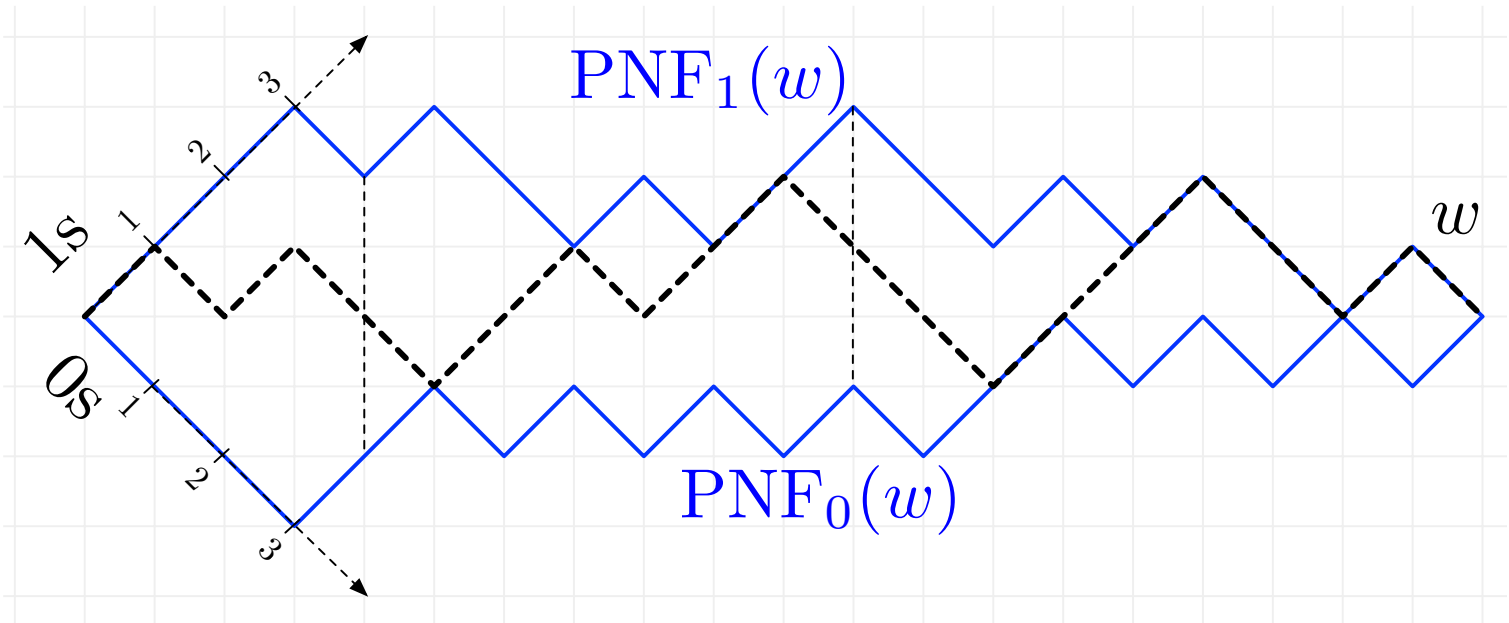
Does **w** = **10100110110001110010** have a substring of length 11 containing exactly 5 ones?



$/ = 1$, $\backslash = 0$, Blue: prefix normal forms of w

BJPM with prefix normal forms

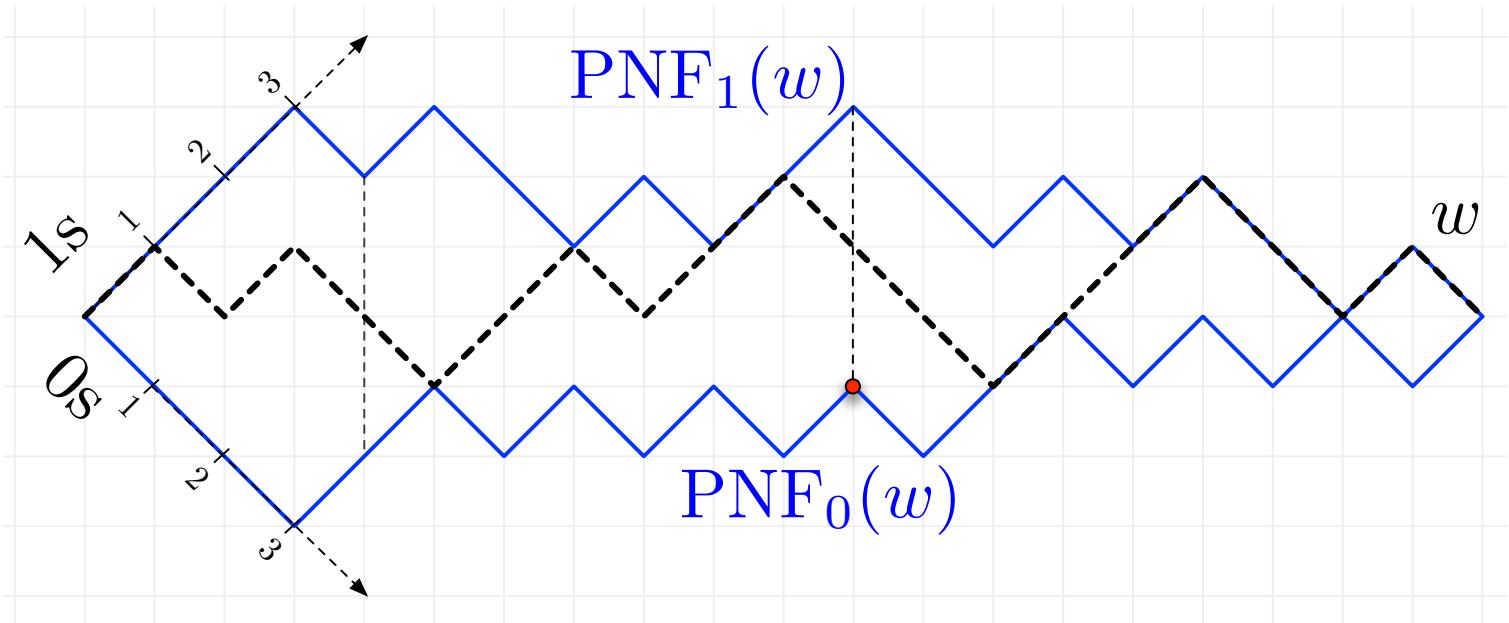
Does **w** = **10100110110001110010** have a substring of length 11 containing exactly 5 ones?



$/ = 1$, $\backslash = 0$, Blue: prefix normal forms of w
verticals: fixed length substrings $k = 4, 11$.

BJPM with prefix normal forms

Does **w** = **10100110110001110010** have a substring of length 11 containing exactly 5 ones? YES



$/ = 1$, $\backslash = 0$, Blue: prefix normal forms of w
verticals: fixed length substrings $k = 4, 11$.

BJPM with prefix normal forms

Does $\mathbf{w} = \mathbf{10100110110001110010}$ have a substring of length 11 containing exactly 5 ones?

$\#$ 1s in pref$(\mathrm{PNF}_1(w), 11) = 7$

$\#$ 1s in pref$(\mathrm{PNF}_0(w), 11) = 5$

$$7 \geq 5 \geq 5 \quad \rightsquigarrow \text{YES}$$

Thus, fast computation of PNFs yields fast solution to BJPM.

Bubble Languages

Bubble languages

Ruskey, Sawada, Williams (JCombTh.A, 2012)

Sawada, Williams (ElJComb. 2012)

Definition

A language $\mathcal{L} \subset \{0,1\}^*$ is called bubble if, for all $w \in \mathcal{L}$, exchanging the first 01 with 10 (if any) results in another word in $\mathcal{L}$.

Example

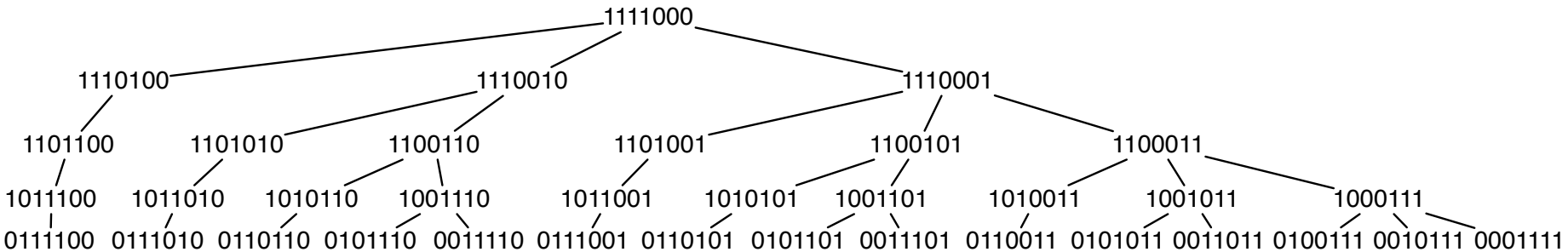
- $\{1001, 1010, 1100, 1000, 0000\}$ – YES
- $\{1001, 1010\}$ – NO

Theorem

$\mathcal{L}_{PN}$ is a bubble language.

An alternative characterization of bubble languages

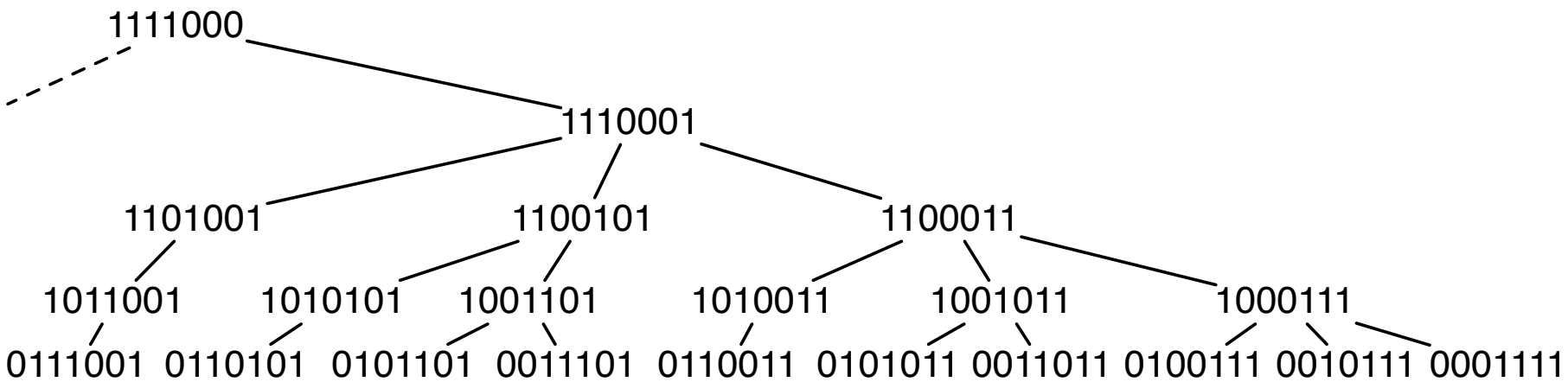
The bubble tree T_d^n on all strings of length n with d ones:



$w = 1^s0^t\gamma$, children of w: $1^{s-1}0^i10^{t-i}\gamma$, for $i = 1, \ldots, t$.

An alternative characterization of bubble languages

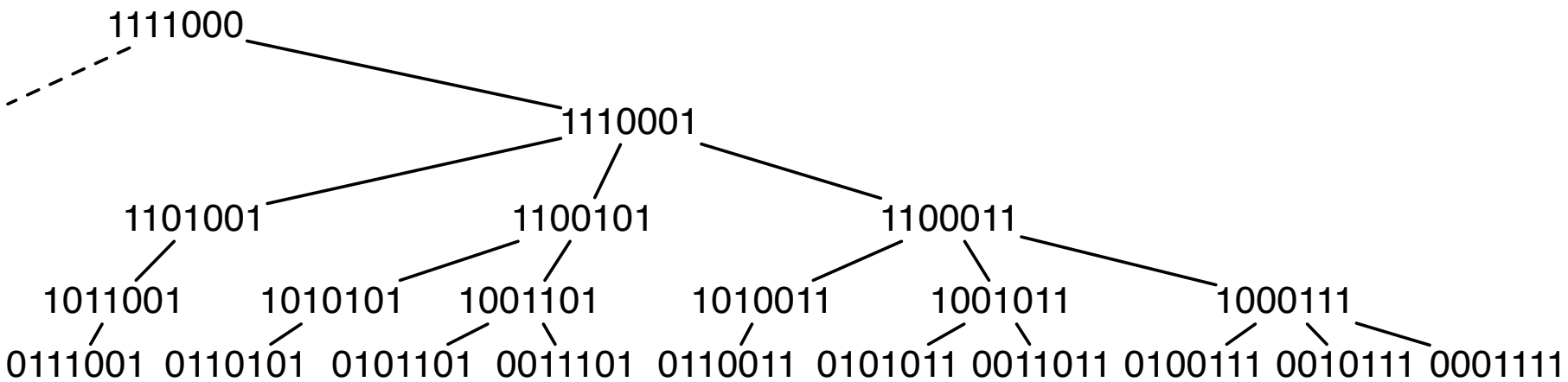
The bubble tree T_d^n on all strings of length n with d ones:



$w = 1^s0^t\gamma$, children of w: $1^{s-1}0^i10^{t-i}\gamma$, for $i = 1, \ldots, t$.

An alternative characterization of bubble languages

The bubble tree T_d^n on all strings of length n with d ones:



$w = 1^s0^t\gamma$, children of w: $1^{s-1}0^i10^{t-i}\gamma$, for $i = 1, \ldots, t$.

Observation

A language is bubble iff it is left- and up-closed in T_d^n, for all n, d.

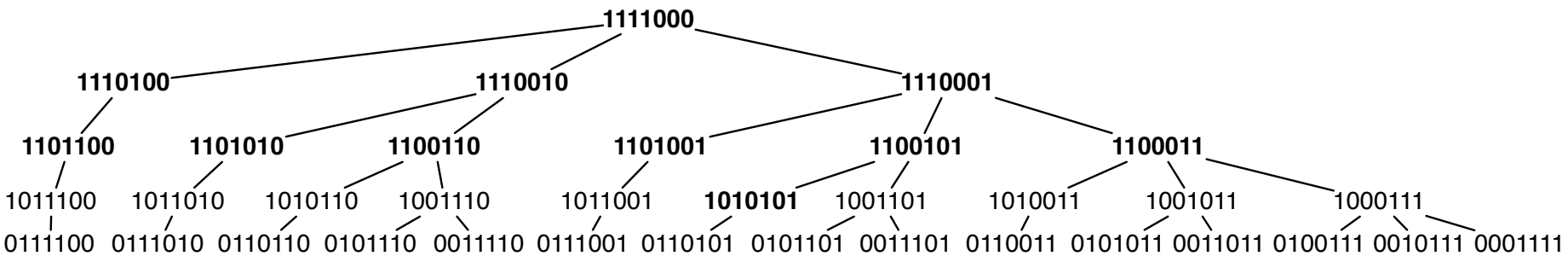
Bubble miracles

Let $\mathcal{L}$ be a bubble language.

- (Fixed-density subsets of) $\mathcal{L}$ are subtrees in the T_d^n's
- Traversal of these subtrees = generation algorithm for $\mathcal{L}$. (enumeration, listing)
- post-order yields a Gray code for $\mathcal{L}$ (cool-lex order)
- Need only: For $w \in \mathcal{L}$, which is the rightmost child still in $\mathcal{L}$? (Oracle for $\mathcal{L}$)
- If Oracle in time $\mathcal{O}(f(n) \cdot k)$, where $k =$ rightmost child, then generation algorithm in $\mathcal{O}(f(n))$ amortized time per word.

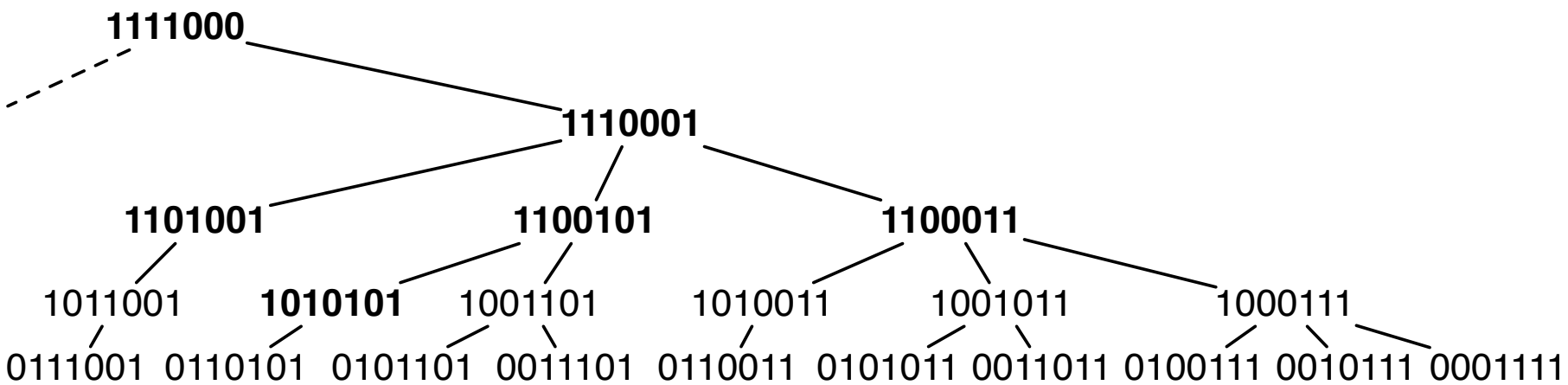
Prefix Normal Words and Bubble Languages

$\mathcal{L}_{\mathrm{PN}}$ in the bubble tree



For every node in $\mathcal{L}_{\mathrm{PN}}$, we need to decide which is rightmost child in $\mathcal{L}_{\mathrm{PN}}$.

$\mathcal{L}_{\mathrm{PN}}$ in the bubble tree



For every node in $\mathcal{L}_{\mathrm{PN}}$, we need to decide which is rightmost child in $\mathcal{L}_{\mathrm{PN}}$.

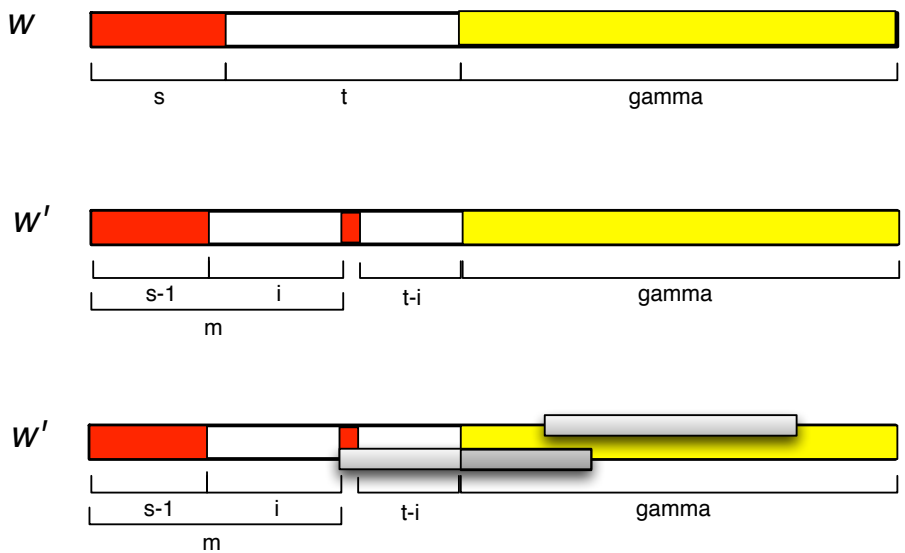
Oracle for $\mathcal{L}_{\mathrm{PN}}$

Theorem

Let $w = 1^s0^t\gamma \in \mathcal{L}_{PN}$ and $w' = 1^{s-1}0^i10^{t-i}\gamma$ one of its children. Then it can be decided in $\mathcal{O}(s+t)$ time whether $w' \in \mathcal{L}_{PN}$.

Proof

One has to compute the max # 1s in a window of size $s + i$ only.



Bubble miracles for prefix normal words

- Efficient generation algorithm for $\mathcal{L}_{\mathrm{PN}}$: amortized linear time per word
  conjectured $\mathcal{O}(\log n)$ = conjectured average length of $s + t$ for a pnw
- Best previous: $\mathcal{O}(2^n n^2)$ time; very substantial improvement
  (no. pn-words grows much slower than 2^n)
- Gray code for $\mathcal{L}_{\mathrm{PN}}$
- enumeration results (experiments)—not possible before!
- many new insights from the bubble property, the generation algorithm, the new representation of prefix normal words
- and, and, and ...

Related work:

- F. Cicalese, G. Fici, Zs. Lipták:
  Searching for Jumbled Patterns in Strings.
  PSC '09, pp. 105–117 (2009).
- P. Burcsi, F. Cicalese, G. Fici, Zs. Lipták:
  On Table Arrangements, Scrabble Freaks, and Jumbled Pattern Matching.
  FUN '10, LNCS 6099: 89–101 (2010).
- G. Fici, Zs. Lipták:
  On Prefix Normal Words.
  DLT '11, LNCS 6795: 228–238 (2011).
- P. Burcsi, F. Cicalese, G. Fici, Zs. Lipták:
  Algorithms for Jumbled Pattern Matching in Strings.
  Internat. J. Found. Comput. Sci., 23: 357–374 (2012).
- P. Burcsi, F. Cicalese, G. Fici, Zs. Lipták:
  On Approximate Jumbled Pattern Matching in Strings.
  Theory Comput. Syst., 50: 35–51 (2012).
- G. Badkobeh, G. Fici, S. Kroon, Zs. Lipták:
  Binary Jumbled String Matching for Highly Run-Length Compressible Texts.
  Inform. Process. Lett., 113: 604-608 (2013).
- P. Burcsi, G. Fici, Zs. Lipták, F. Ruskey, J. Sawada:
  Normal, Abnormal, Prefix Normal.
  FUN '14, LNCS 8496: 79–93 (2014).

THANK YOU!