3PCC

The structure of graphs representing all subwords of Thue-Morse sequences

Jakub Radoszewski, Wojciech Rytter

Instytut Informatyki,
Uniwersytet Warszawski,
Warszawa

The syntactic properties of Thue-Morse words are related to special type of automata/graphs.

The directed acyclic subword graph (*dawg*, in short) is a useful deterministic automaton accepting all suffixes of the word. Its compacted version (resulted by compressing chains of states) is denoted by *cdawg*.

The cdawgs of Thue-Morse words have regular and very simple structure, in particular they offer a powerful (exponential) compression of the set of all subwords in case of finite Thue-Morse words.

Using the special structure of cdawgs several unknown properties of Thue-Morse words as well as new (graph-based) proofs of some well-known properties are demonstrated.

A slight modification of the compact cdawg of the infinite Thue-Morse word yields an infinite graph with "2-counting property". Also a new numeration system is introduced and its relation to cdawgs of Thue-Morse words is discussed.

Denote:

$$infinite\ \ graph: \quad G \ = \ cdawg(\tau)$$

$$finite\ \ graph: \quad G_n \ = \ cdawg(\tau_n)$$

Let $\bar{x}$ be the sequence resulting by negating the bits of x. The finite TM words are defined as follows:

$$\tau_0 \;=\; 0; \quad \tau_n \;=\; \tau_{n-1}\bar{\tau}_{n-1} \quad \text{for} \;\; n > 0.$$

The infinite TM word τ is the *limit* of τ_n words, the limit in the sense that each τ_n is a prefix of τ.

Also let φ be the TM morphism:

$$\varphi(0) = 01, \quad \varphi(1) = 10.$$

Then

$$\tau_n \;=\; \varphi^n(\tau_0).$$

$$\tau_0 = 0,$$

$$\tau_1 = 01,$$

$$\tau_2 = 0110,$$

$$\tau_3 = 01101001$$

$$\tau_4 = 0110100110010110$$

$$\tau \;=\; 0110100110010110100101100110\ldots\ldots$$

Denote by $Sub(u)$ the set of all finite subwords (factors) of u.

The word $v \in \{0,1\}^*$ is a **right special factor** of the word u iff $v0$, $v1 \in Sub(u)$.

$v \in \{0,1\}^*$ is a **left special factor** of the word u iff $0v$, $1v \in Sub(u)$.

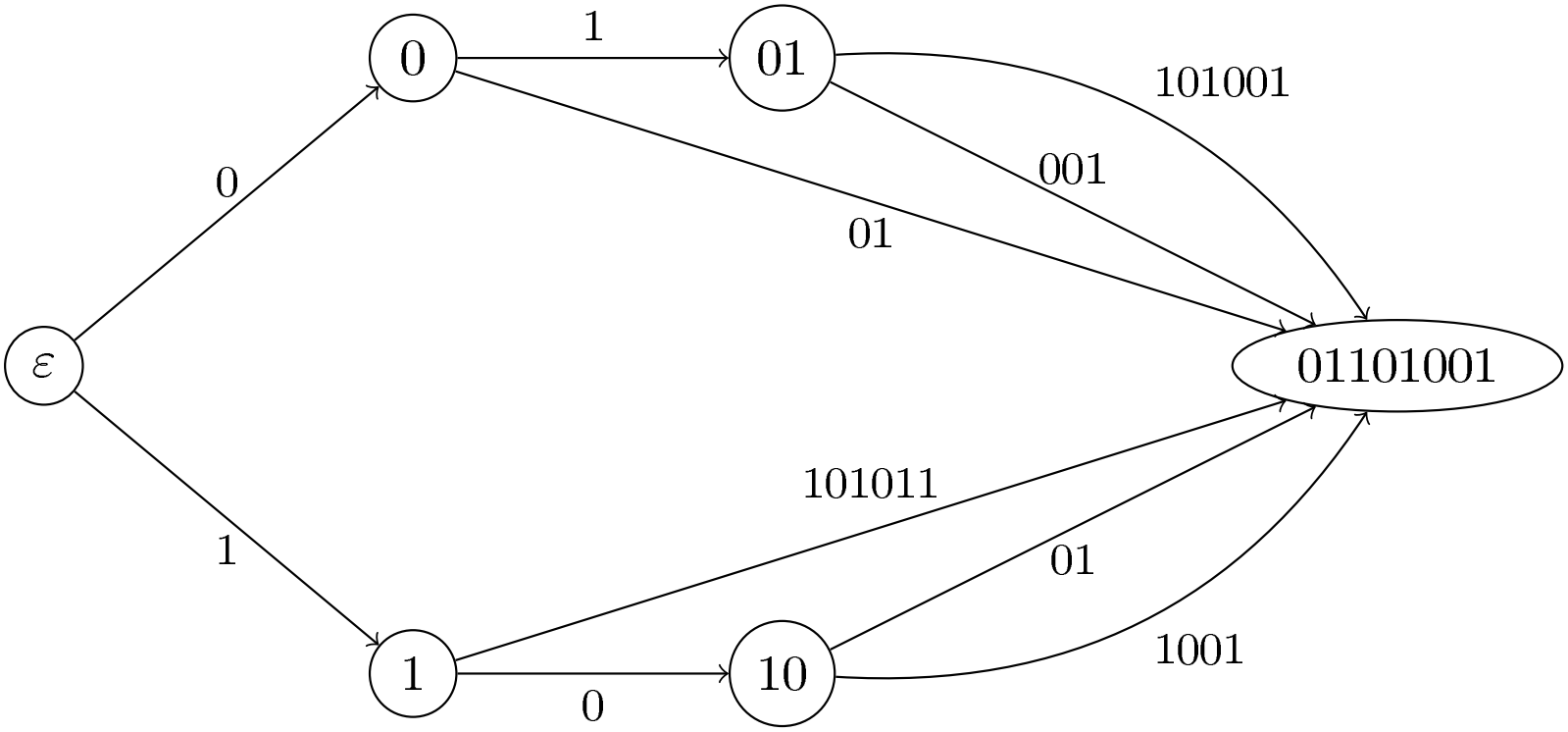
The word is a **bispecial factor** iff it is both left and right special.

The internal vertices of the subword graph correspond to left special factors.

The internal vertices of the compacted subword graph correspond to bispecial factors.

G_n is exponentially compressed representation of subwords of τ_n. For $n \geq 3$ the number of vertices of G_n is:

$$4n - 6 \;=\; 4 \;\log|\tau_n| - 6$$



The cdawg for $\tau_3 = 01101001$. The internal vertices are the bispecial factors of 01101001. $\mathsf{cdawg}(\tau_3)$ is not extremely regular, the regular structure starts from τ_4. Labels of edges outgoing from the same vertex start with different symbols, they correspond to factors of τ.

Define a family of operations, $DelQuart_i$, which remove from the word w the i-th quarter, assuming $|w|$ is divisible by 4:

$$w \;=\; w_1w_2w_3w_4 \;\;\&\;\; |w_1| = |w_2| = |w_3| = |w_4|$$

$$\Rightarrow\; DelQuart_1(w) = w_2w_3w_4 \;\&\; DelQuart_3(w) = w_1w_2w_4.$$

We introduce special factors of *rank* n:

$$\sigma_n \;=\; DelQuart_3(\tau_n), \quad \tau'_n \;=\; DelQuart_1(\tau_n).$$

Example 1

$$DelQuart_3: \quad \tau_3 \;=\; 0110\,\underline{10}\,01 \quad\Rightarrow\quad 011001 \;=\; \sigma_3$$

$$DelQuart_1: \quad \tau_3 \;=\; \underline{01}\,101001 \quad\Rightarrow\quad 101001 \;=\; \tau_3'$$

Equivalently

$$\sigma_n \;=\; \varphi^{n-2}(\sigma_2), \quad \text{where } \sigma_2 = 010.$$

We also have:

$$\sigma_n \;=\; \tau_{n-2}\bar{\tau}_{n-1} \;=\; \tau_{n-1}\tau_{n-2}, \quad \tau_n' \;=\; \bar{\tau}_{n-2}\bar{\tau}_{n-1}.$$

Lemma 1 *A word $u \in Sub(\tau)$ starting with the letter 0 is a left special factor of the infinite Thue-Morse word if and only if it is a prefix of $\varphi^n(\sigma_2)$ for some n.*

Lemma 2 *A word $u \in Sub(\tau)$, $|u| > 4$, is a bispecial factor of the infinite Thue-Morse word if and only if it is a morphic image (under φ) of a shorter bispecial word. Moreover, τ_0, τ_1, σ_2 and their negations are bispecial factors of τ.*

We have exactly 8 bispecial words of length at most 4 in τ. They are

$$\tau_0 = 0,\ \sigma_2 = 010,\ \tau_1 = 01,\ \tau_2 = 0110$$

and their negations.

We know that:

$$\varphi(\tau_i) = \tau_{i+1},\ \varphi(\sigma_i) = \sigma_{i+1}$$

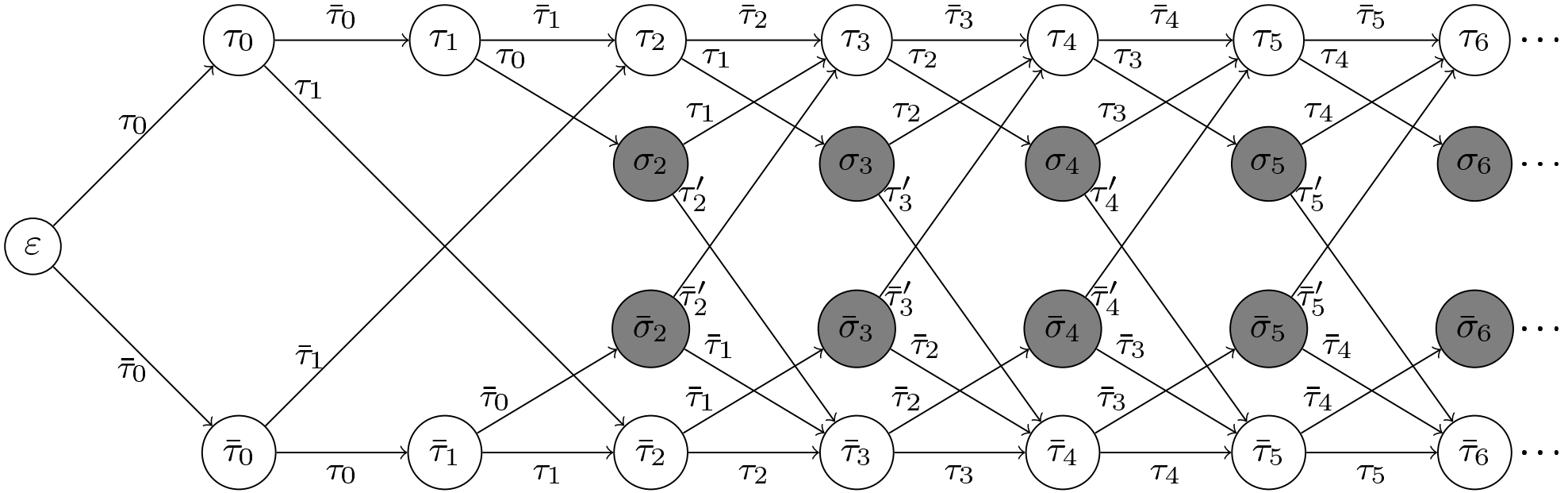
Hence the internal vertices of the compacted subword graph correspond precisely to words τ_i, σ_i and their negations for $i \geq 0$.

Theorem 1 *The edges of G are as follows (other edges are symmetric copies by bitwise negation):*

$$\varepsilon \xrightarrow{\tau_0} \tau_0, \qquad \tau_0 \xrightarrow{\tau_1} \bar{\tau}_2, \tag{1}$$

$$\tau_i \xrightarrow{\bar{\tau}_i} \tau_{i+1} \ \ for \ \ i \geq 0, \qquad \tau_i \xrightarrow{\tau_{i-1}} \sigma_{i+1} \ \ for \ \ i \geq 1, \tag{2}$$

$$\sigma_i \xrightarrow{\tau_{i-1}} \tau_{i+1}, \qquad \sigma_i \xrightarrow{\tau_i'} \bar{\tau}_{i+1} \ \ for \ \ i \geq 2. \tag{3}$$



The cdawg for τ — an initial part.

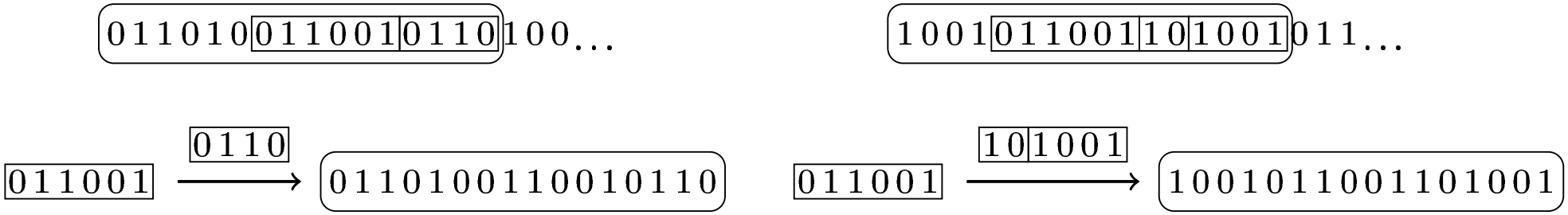
The vertices of $G_n = cdawg(\tau_n)$ are vertices of G with the sink τ_n and other indices smaller thean $n-1$.

The characterization of edges of G_n:

Theorem 2 *The edges of G_n (for $n \geq 4$) are of the form (1), (2), (3) for $i \leq n-2$ (and their negations), and additionally the following edges pointing to the sink:*

$$\tau_{n-2} \overset{\bar{\tau}'_{n-1},\ \tau'_n}{\longrightarrow} \tau_n, \qquad \bar{\tau}_{n-2} \overset{\bar{\tau}_{n-1},\ \tau_{n-2}}{\longrightarrow} \tau_n, \tag{4}$$

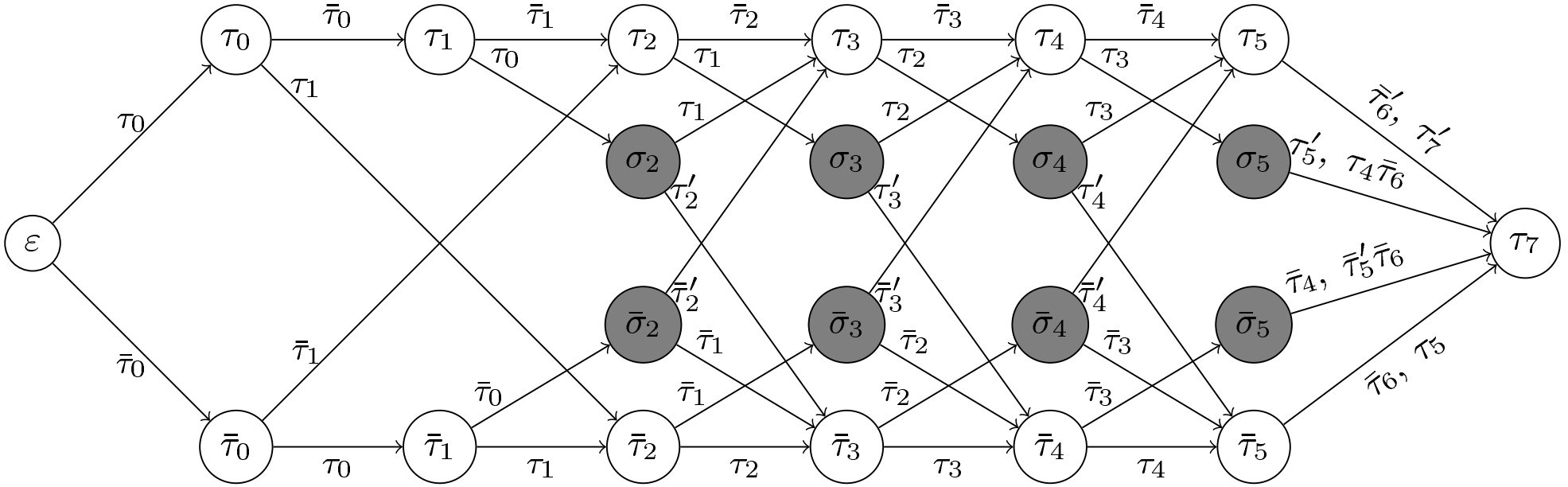
$$\sigma_{n-2} \overset{\tau'_{n-2},\ \tau_{n-3}\bar{\tau}_{n-1}}{\longrightarrow} \tau_n, \qquad \bar{\sigma}_{n-2} \overset{\bar{\tau}_{n-3},\ \bar{\tau}'_{n-2}\bar{\tau}_{n-1}}{\longrightarrow} \tau_n. \tag{5}$$



The edges

$$\sigma_3 \xrightarrow{\tau_2} \tau_4$$

$$\sigma_3 \xrightarrow{\bar{\tau}_1\bar{\tau}_2} \bar{\tau}_4$$



The cdawg for τ_7. Note the power of compaction:
$|\tau_7| = 128$ and it contains 6 232 distinct factors, however
$|V(G_7)| = 22$ and $|E(G_7)| = 42$.

Theorem 3 *We can test if a word w is a factor of a TM word τ_n in time $O(|w|)$ and $O(1)$ space. The number of occurrences can be found within the same complexity.*

Proof: It is easy to see that we can test if a specified factor of w is a TM word in linear time and constant space. Then we can traverse $\mathsf{cdawg}(\tau_n)$ without remembering it explicitly, just keeping track of the current position within w and the current vertex of the cdawg. Traversing an edge of the cdawg reduces to one or several tests if a factor of w is a TM word, which take $O(|w|)$ time in total. □

Let $\#occ(x, w)$ be the number of occurrences of factor x in w.
The node of G_n is accepting iff the correspoding word is a suffix of τ_n. We can show:

- The accepting nodes are

$$\tau_n,\ \tau_{n-2},\ \bar{\tau}_{n-3},\ \tau_{n-4},\ ,\ \bar{\tau}_{n-5}\cdots$$

- The number of occurrences of x equals the number of paths from the implicit node corresponding to x to accepting node

- The number of such path equals the number of such paths from the closest explicit node to accepting node

The numbers $\#occ(\tau_k, \tau_7)$

k =	1	2	3	4	5	6
$\#occ$ =	1	3	5	11	21	43

The numbers $\#occ(\bar{\tau}_k, \tau_7)$

k =	1	2	3	4	5	6
$\#occ$ =	1	2	5	10	21	42

What about simple formulas ?

For $1 \leq i < n$ we have:

- $\#occ(\tau_i, \tau_n) = \lceil (2^{n-i+1} - 1)/3 \rceil$
- $\#occ(\bar{\tau}_i, \tau_n) = \lfloor (2^{n-i+1} - 1)/3 \rfloor$

For $2 \leq i < n - 2$ we have:

- $\#occ(\sigma_i) = \#occ(\tau_{i+1}) + \#occ(\bar{\tau}_{i+1})$

This implies simple calculation of $\#occ(x, \tau_n)$ for any x.

The following theorem provides a compact formula for the number of distinct factors of a TM word.

Theorem 4

The number of different factors of τ_n *for* $n \geq 4$ *equals*

$$\frac{73}{192}|\tau_n|^2 + \frac{8}{3}.$$

Denote by $\mathsf{mult}(v)$ the number of paths from initial node ε to v.

For $2 \leq i \leq n-2$, by simple induction we obtain

$$\mathsf{mult}(\tau_i) = \mathsf{mult}(\bar{\tau}_i) = 2^{i-1}$$

$$\mathsf{mult}(\sigma_i) = \mathsf{mult}(\bar{\sigma}_i) = 2^{i-2}.$$

The total number of different factors of τ_n equals

$$S(\tau_n) = \sum_{e=(u,v) \in E(G_n)} \mathsf{mult}(u) \cdot |e|.$$

We compute $S(\tau_n)$ level by level, using the fact that

$$|\tau_i| = |\bar{\tau}_i| = 2^i \quad \text{and} \quad |\tau'_i| = |\bar{\tau}'_i| = 3 \cdot 2^{i-2}.$$

The local period at inter-position p_i as $|v|$, where vv is the shortest square centered at this inter-position. The critical factorization point of a word u is an inter-position of u for which the local period equals the (global) period of u.

Theorem 5 *The critical factorization point of the TM word τ_n, for $n \geq 4$, is the inter-position p_i for $i = 2^{n-1} + 1$.*

It is a known fact that the critical factorization point of a word u corresponds to the first letter of the shorter of the following two suffixes:

1. the lexicographically largest suffix of u under the standard order of letters: $0 < 1$
2. the lexicographically largest suffix of u under the reversed order of letters: $1 \prec 0$.

To get suffix (1) we traverse G_n, starting from ε, along the lexicographically largest path, i.e.:

$$\varepsilon \xrightarrow{\bar{\tau}_0} \bar{\tau}_0 \xrightarrow{\bar{\tau}_1} \tau_2 \xrightarrow{\bar{\tau}_2} \tau_3 \xrightarrow{\bar{\tau}_3} \ldots \xrightarrow{\bar{\tau}_{n-3}} \tau_{n-2} \xrightarrow{\tau'_n} \tau_n.$$

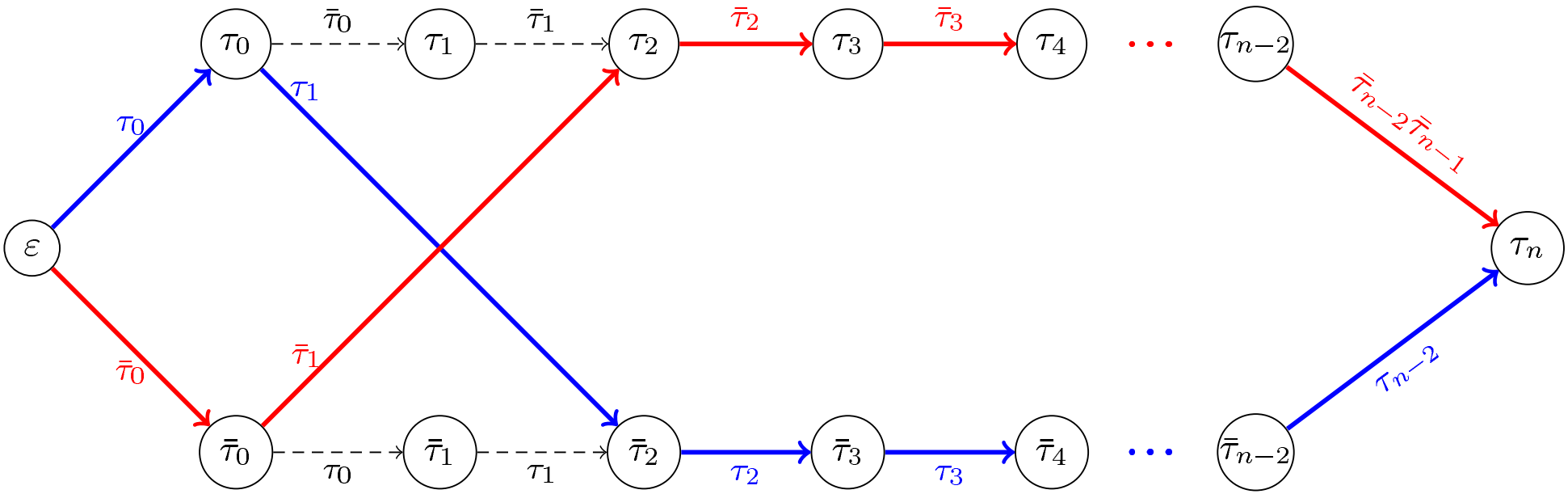
The length of this path is the length of τ_n minus 1, hence it is $2^n - 1$.

On the other hand, the suffix (2) corresponds to the maximal path starting from ε that always prefers 0:

$$\varepsilon \xrightarrow{\tau_0} \tau_0 \xrightarrow{\tau_1} \bar{\tau}_2 \xrightarrow{\tau_2} \bar{\tau}_3 \xrightarrow{\tau_3} \ldots \xrightarrow{\tau_{n-3}} \bar{\tau}_{n-2} \xrightarrow{\tau_{n-2}} \tau_n.$$

The length of this path is $|\bar{\tau}_{n-1}| - 1$, since the label of this path equals $\bar{\tau}_{n-1}$ with the first letter removed. Thus the suffix (2) has length $2^{n-1} - 1$ and is always shorter than (1) and using it we obtain the critical factorization point:

$$|\tau_n| - (2^{n-1} - 1) = 2^{n-1} + 1$$



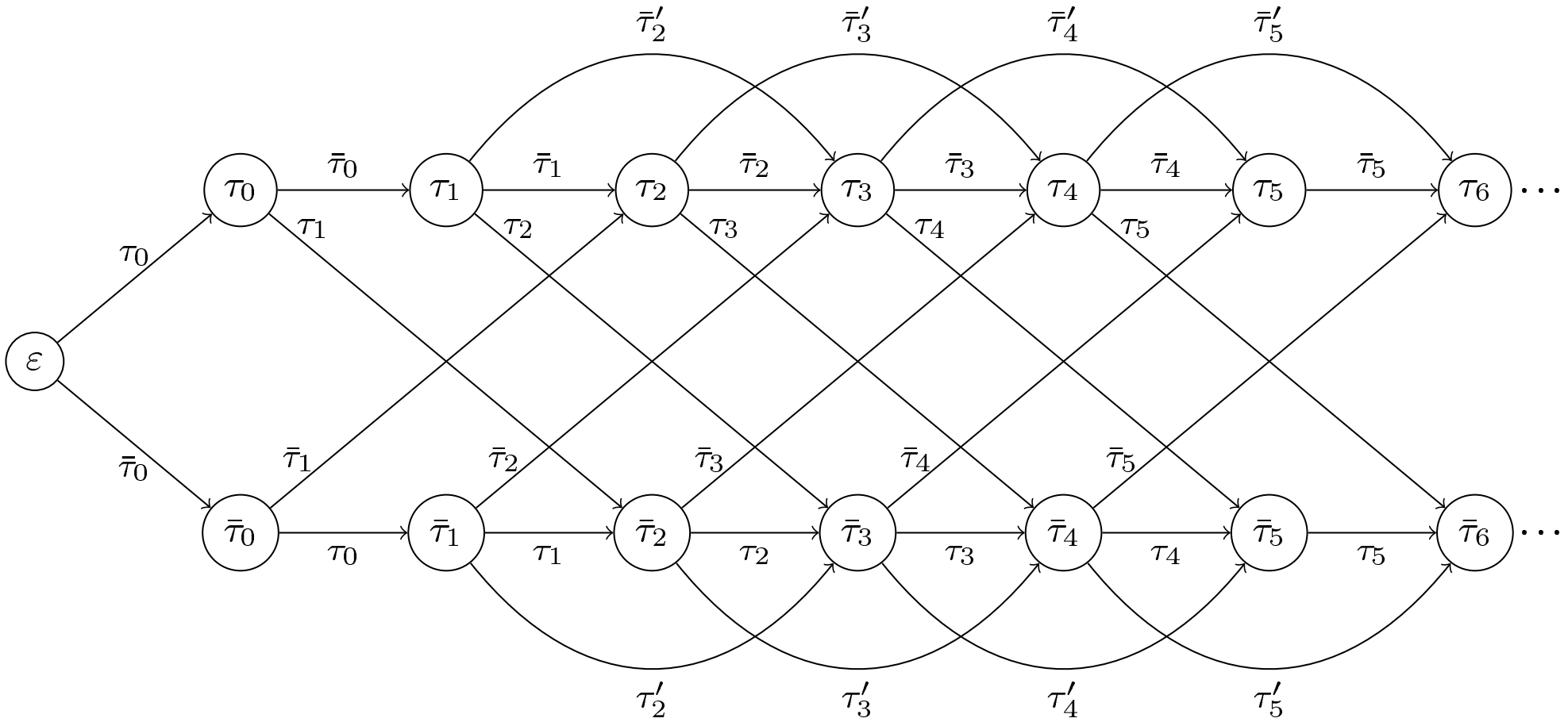
The red path = maximal suffix (with $0 < 1$)
The (completely) upper path = τ_n.

The blue path = maximal suffix (with $1 < 0$)
The (completely) lower path = $\bar{\tau}_{n-1}$.

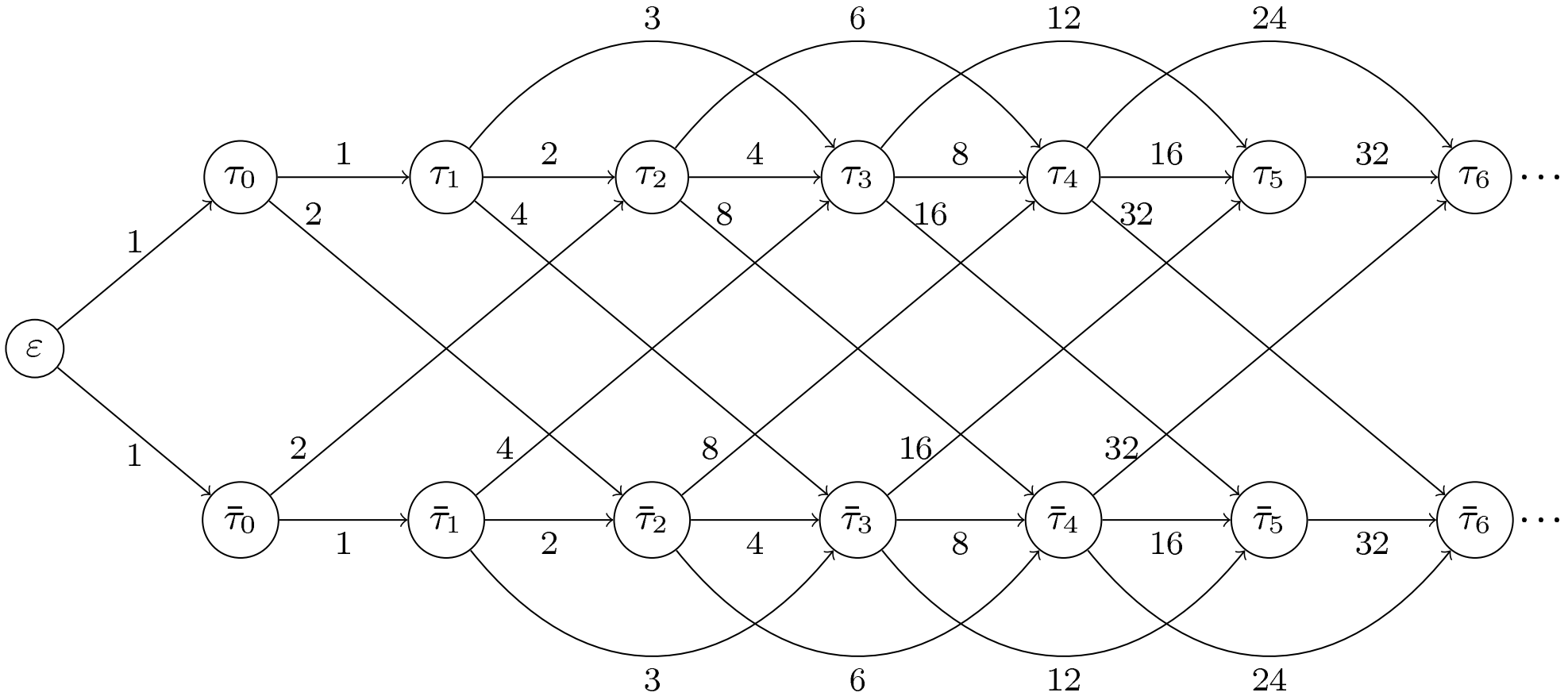
Let G' be an infinite labeled graph obtained from G by removing all vertices σ_n, $\bar{\sigma}_n$ and replacing pairs of edges traversing them with single edges with concatenated labels. We call G' the *pseudo-cdawg* of τ.

We say that a weighted directed graph with a designated source vertex has *k-counting property* if there are exactly k different paths for each length $n > 0$ starting from the source and ending in some arbitrary vertex.

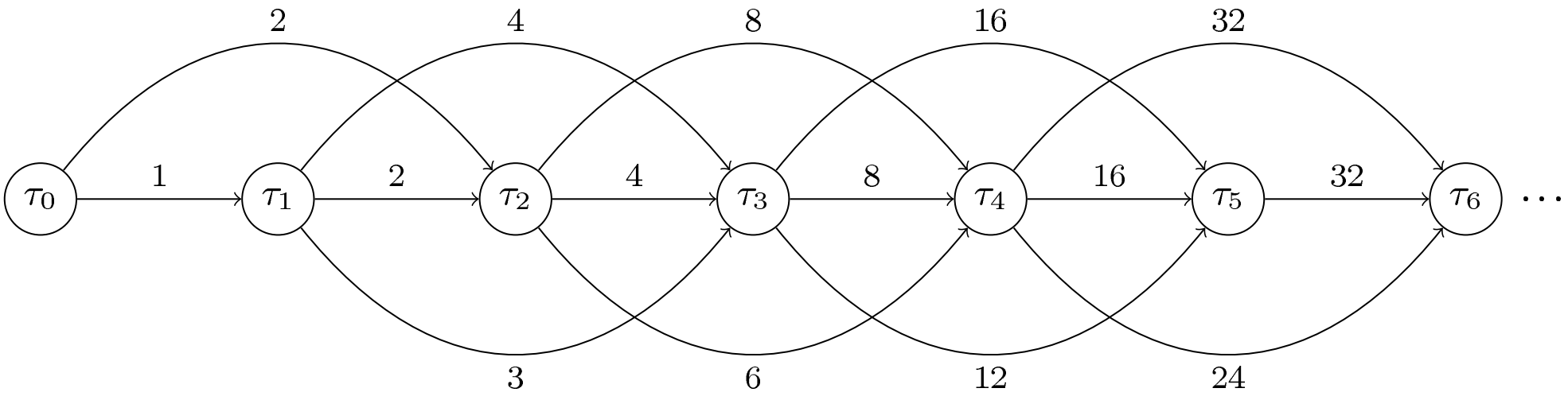
Let H be a directed weighted graph obtained from G' by replacing labels of edges with their lengths. Note that the edges of H can be divided into three groups: the backbone (two series of edges $1, 1, 2, 4, 8, \ldots$), in-branch edges (two series $3, 6, 12, 24, \ldots$) and inter-branch edges



The pseudo-cdawg for τ — an initial part.



The initial part of the weighted graph H obtained from G' by replacing labels with their lengths. This graph has 2-counting property.



The initial part of the weighted graph H' obtained from H by removing the vertex ε and gluing the upper and lower branch. This graph has 1-counting property.