On the Structure of Compacted Subword Graphs of Thue-Morse Words and Their Applications[*]

Jakub Radoszewski[†] and Wojciech Rytter
[jrad,rytter]@mimuw.edu.pl

Institute of Informatics,
University of Warsaw,
ul. Banacha 2, 02-097 Warsaw, Poland

Abstract

We investigate how 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 we present several unknown properties of Thue-Morse words as well as new (graph-based) proofs of some well-known properties. In particular we show a simple algorithm that checks, for a given string w, if w is a subword of a Thue-Morse word and computes its number of occurrences in nth Thue-Morse word in $O(|w| + \log n)$ time and $O(1)$ space. Additionally, a slight modification of the compact dawg of the infinite Thue-Morse word yields an infinite graph with 2-counting property.

Keywords: Thue-Morse word, compacted subword graph, graph counting property.

1 Introduction

Thue-Morse words (*TM* words, in short) form a famous family of words, due to many interesting properties related not only to text algorithms and combinatorics on words, but also to other disciplines, see [1]. In particular they do not contain factors of type $axaxa$, where a is a single letter (overlaps), consequently they do not contain cubes. A very good source for properties of these words is for example the book [4]. We rediscover/discover several known/unknown properties of *TM* words in a novel way: analyzing the compacted subword graphs (cdawgs) of finite and infinite *TM* words. This approach was already

[*] The authors thank anonymous referees for their insightful remarks which helped significantly improve the presentation of the paper.

[†] corresponding author, e-mail: jrad@mimuw.edu.pl, tel. +48-22-55-44-117, fax +48-22-55-44-400

successfully applied by one of the authors to another well-known family of words, the Fibonacci words [15]. We also study how the cdawg of the infinite *TM* word is related to an infinite graph with 2-counting property, similar analysis for Fibonacci words and, in general, Sturmian words can be found in [13].

The structure of cdawg of a word w is closely related to right special factors of w (defined later on in the text). Such factors of *TM* words were already studied thoroughly in relation to the subword complexity function of the infinite *TM* word (i.e., the number of distinct factors of the word of a given length), see [6, 12, 16]. On the other hand, the vertices of cdawg of w can be seen as bispecial factors of w; bispecial factors of the infinite *TM* word are characterized in [3, 11].

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

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

We say that τ_n is of *rank* n. 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.

Let φ be the *TM* morphism, defined as:

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

A well known property (alternative definition) of *TM* words is:

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

We have:

$$\tau_0 = 0,\; \tau_1 = 01,\; \tau_2 = 0110,\; \tau_3 = 01101001, \ldots$$

$$\tau \;=\; 01101001100101101001011001 10\ldots\ldots$$

We consider words u over the alphabet $\{0, 1\}$, $u \in \{0, 1\}^*$. The positions are numbered from 0 to $|u| - 1$. By $P = \{p_0, p_1, \ldots, p_{|u|-2}\}$ we denote the set of inter-positions that are located *between* pairs of consecutive letters of u. The empty word is denoted by ε. If $u, v \in \{0, 1\}^*$ then by $u \cdot v = uv$ we denote the concatenation of words u and v.

For $u = u_0u_1 \ldots u_{m-1}$, denote by $u[i\,..\,j]$ a *factor* (subword) of u equal to $u_i \ldots u_j$ (in particular $u[i] = u[i\,..\,i]$). Words $u[0\,..\,i]$ are called prefixes of u, and words $u[i\,..\,m-1]$ — suffixes of u. Similarly, one can define factors, prefixes and suffixes (resulting by cutting off an initial prefix) of an infinite word $u_0u_1u_2\ldots$ By $\#occ(x, u)$ we denote the number of occurrences of a factor x in u.

Denote by $Sub(u)$ the set of all finite subwords (factors) of u. We say that 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. In particular, for each word containing at least 2 different letters, the empty word is one of its bispecial factors.

We say that an integer i is an *end-occurrence* of the word u in the (finite or infinite) word w if $u = w[i - |u| + 1\,..\,i]$. Let $Fin(u)$ be the set of end-occurrences of the word u

in τ. From the point of view of the dawg two words u_1, u_2 correspond to the same vertex if and only if $Fin(u_1) = Fin(u_2)$.

A *dawg* (directed acyclic subword graph) of a finite word u (notation: $\mathsf{dawg}(u)$) is the minimal automaton accepting all suffixes of u [8, 9]. In this paper we deal with compacted dawgs (*cdawgs*). Cdawgs were first introduced by Blumer et al [5], for references on cdawgs see also [7, 10, 14]. The cdawg for τ_3 is illustrated in Fig. 1.

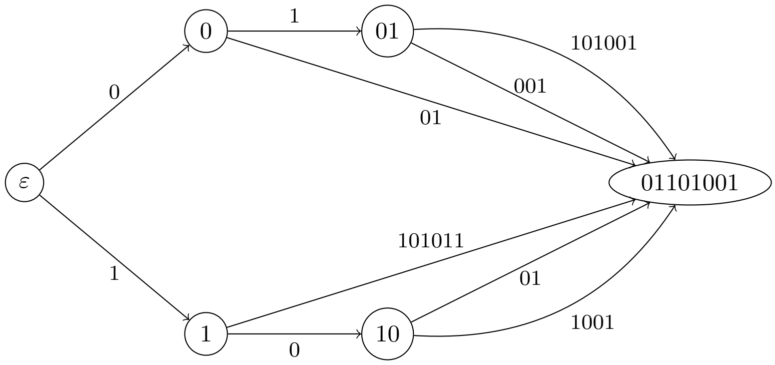

Figure 1: The cdawg for $\tau_3 = 01101001$. The set of vertices is the set of bispecial factors and the sink. $\mathsf{cdawg}(\tau_3)$ does not show a regular structure of general case, such a regular structure starts from τ_4. Labels of edges outgoing from the same vertex start with different symbols, these labels also have compact representations as factors of τ.

A cdawg of a word u, denoted as $\mathsf{cdawg}(u)$, represents all (finite or infinite) suffixes of u. In the finite case let $G_n = \mathsf{cdawg}(\tau_n)$, the set $V(G_n)$ of vertices is the set of bispecial factors of τ_n including τ_n as a sink node. For the infinite word, $G = \mathsf{cdawg}(\tau)$, the only difference is lack of a sink node, in case of *TM* words this simplifies the construction considerably.

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.

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

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

Equivalently

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

We also have:

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

2 Useful Syntactic-Combinatorial Properties of *TM* Words

In this section we recall several already known facts about *TM* words and use them to prove new properties which we will use to classify vertices and edges of cdawgs of finite and infinite *TM* words.

The next two (already known) lemmas show that the words σ_n defined in the previous section are strongly related to special factors of τ.

Lemma 2. [Proposition 2.15 in [4]]
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 3. [3, 11]
A word $u \in Sub(\tau)$, $|u| > 4$, is a bispecial factor of the infinite Thue-Morse word if and only if $\varphi^{-1}(u)$ is a (shorter) bispecial word. Moreover, τ_0, τ_1, σ_2 and their negations are bispecial factors of τ.

Let us also recall the following observation, its proof can be found in [4].

Observation 4.

(a) If u is a factor of τ such that $|u| \geq 4$ then all positions in $Fin(u)$ are even or all of them are odd.

(b) For any $n \geq 2$, $2^n - 1 \in Fin(\tau_n)$ and $2^{n+1} - 1 \in Fin(\bar{\tau}_n)$.

The technical Observation 5 provides a characterization of bispecial factors of τ_n for $n \geq 4$ and also a useful tool for the analysis of edges of G_n and G.

Observation 5. *For $n \geq 2$, the word τ_{n+2} contains:*

(1) three end-occurrences of the factor τ_n: $a_n = 2^n - 1$ followed by the letter 1, $b_n = 2^{n+1} + 2^{n-1} - 1$ preceded by letter x and followed by the letter 0, and $t_n = 2^{n+2} - 1$ preceded by the letter $\bar{x}$, where $x \in \{0, 1\}$

(2) two end-occurrences of the factor $\bar{\tau}_n$: $c_n = 2^{n+1} - 1$ preceded by letter y and followed by the letter 1, and $d_n = 2^{n+1} + 2^n - 1$ preceded by the letter $\bar{y}$ and followed by the letter 0, where $y \in \{0, 1\}$

(3) two end-occurrences of the factor σ_n: $e_n = 2^n + 2^{n-1} - 1$ preceded by letter z and followed by the letter 0, and $f_n = 2^{n+1} + 2^n + 2^{n-2} - 1$ preceded by the letter $\bar{z}$ and followed by the letter 1, where $z \in \{0, 1\}$

(4) two end-occurrences of the factor $\bar{\sigma}_n$: $g_n = 2^n + 2^{n-2} - 1$ preceded by letter w and followed by the letter 0, and $h_n = 2^{n+1} + 2^n + 2^{n-1} - 1$ preceded by the letter $\bar{w}$ and followed by the letter 1, where $w \in \{0, 1\}$.

Moreover, the words τ_0, τ_1 and their negations are bispecial factors of τ_3.

Proof. The proof goes by induction on n. The inductive basis ($n = 2$) can be verified by hand for the words

$$\tau_2 = 0110,\ \bar{\tau}_2 = 1001,\ \sigma_2 = 010,\ \bar{\sigma}_2 = 101$$

within $\tau_4 = 0110100110010110$.

As for the inductive step ($n > 2$), let us note that τ_n (or $\bar{\tau}_n$) has an end-occurrence in τ at position j if and only if τ_{n-1} ($\bar{\tau}_{n-1}$ resp.) has an end-occurrence at position $(j-1)/2$ in τ. Indeed, this is due to Observation 4 and the fact that τ is a fixed point of the morphism φ. In such a case, the letters immediately following the considered occurrences of τ_n ($\bar{\tau}_n$ resp.) and τ_{n-1} ($\bar{\tau}_{n-1}$ resp.) are the same, while the letters preceding them are bitwise negations of each other (in both cases, if the considered letters exist).

A similar condition can be stated for the factors σ_n and $\bar{\sigma}_n$: end-occurrence of one of them in τ at position j corresponds to an end-occurrence of σ_{n-1} ($\bar{\sigma}_{n-1}$ resp.) at position $(j-1)/2$ in τ. This is, again, due to Observation 4, since $\bar{\tau}_{n-1}$ (τ_{n-1}) is a suffix of σ_n ($\bar{\sigma}_n$ resp.).

Hence, to conclude this part of the proof, it suffices to note that $(a_n - 1)/2 = a_{n-1}$ and same conditions hold for $b_n, c_n, \ldots, h_n, t_n$.

The "moreover" part of the observation can easily be verified by hand. □

The following observation provides an analogical result regarding the words σ_n.

Observation 6. *The word σ_{n+1} (for $n \geq 2$) contains only a single occurrence of the factor $\bar{\tau}_{n-1}$, which is followed by the letter 0, and two occurrences of τ_{n-1}, one followed by the letter 1 and the other being a suffix of σ_{n+1}.*

Proof. We prove the observation by induction on n. For $n = 2$ the verification of the conclusion of the observation ($\tau_1 = 01$, $\bar{\tau}_1 = 10$, $\sigma_3 = 011001$) is trivial.

Now let $n > 2$. Note that we can use a similar machinery as in the proof of Observation 5. Indeed, σ_{n+1} is a factor of τ of even length having the suffix $\bar{\tau}_n$. Therefore, by Observation 4, all its end-occurrences in τ are odd. This concludes, by the same Observation, that any end-occurrence of τ_{n-1} (or $\bar{\tau}_{n-1}$) in σ_{n+1} is odd and thus corresponds to an end-occurrence of τ_{n-2} ($\bar{\tau}_{n-2}$ resp.) in σ_n. Note that the letters immediately following the considered factors in σ_{n+1} and in σ_n are the same (provided that they exist). Using the inductive hypothesis, we conclude the proof. □

3 Structure of Cdawg of Infinite *TM* Word

We start the description of $\mathsf{cdawg}(\tau)$ by showing the structure of its vertices. The following fact is a simple consequence of previous work related to combinatorics of *TM* words.

Fact 7. *The vertices of G are all words τ_i, σ_i and their bitwise negations for $i = 0, 1, \ldots,$ together with the source vertex corresponding to ε.*

Proof. The vertices of G correspond to bispecial factors of τ. According to Lemma 3 and [3, 11], the (non-empty) bispecial factors of τ are exactly τ_i, σ_i and their negations. □

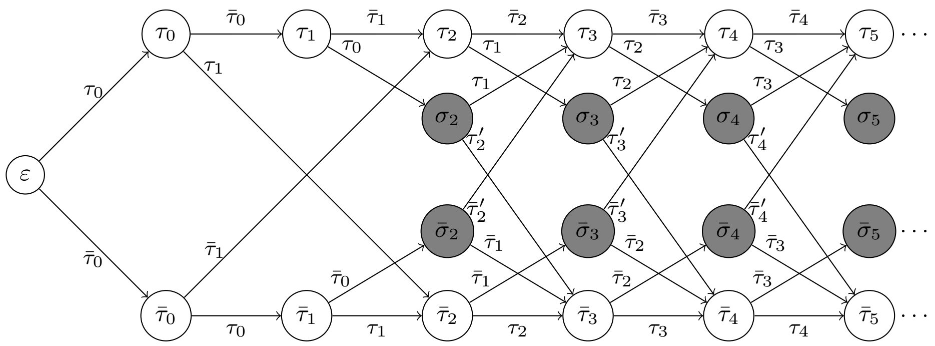

Figure 2: The cdawg for τ — an initial part.

We know the vertices of G, however the main component of the structure of this graph is characterization of its edges. Now we proceed to the analysis of the set of edges $E(G)$, since we wish to represent the labels of edges in a compact way using the factors σ_n, τ_n, τ'_n and their bitwise negations.

Each vertex $v \in V(G)$ has exactly two outgoing edges, one with label starting with the letter 0 and the other with the letter 1. It is a well known fact that:

Observation 8. *The ending vertex of an edge outgoing from u in a (finite or infinite) cdawg H with the label starting with the letter $c \in \{0,1\}$ is the shortest $v \in V(H)$ such that $uc \in Sub(v)$. Then the label of this edge is $c\alpha$, such that $uc\alpha$ is a suffix of v.*

Using this observation, we can provide the characterization of edges of $\mathsf{cdawg}(\tau)$.

Theorem 9. *The edges of G are as follows (other edges are symmetric copies by bitwise negation):*

$$\varepsilon \xrightarrow{\tau_0} \tau_0, \qquad \tau_0 \xrightarrow{\tau_1} \bar\tau_2, \tag{2}$$

$$\tau_i \xrightarrow{\bar\tau_i} \tau_{i+1} \;\; \textit{for} \;\; i \geq 0, \qquad \tau_i \xrightarrow{\tau_{i-1}} \sigma_{i+1} \;\; \textit{for} \;\; i \geq 1, \tag{3}$$

$$\sigma_i \xrightarrow{\tau_{i-1}} \tau_{i+1}, \qquad \sigma_i \xrightarrow{\tau'_i} \bar\tau_{i+1} \;\; \textit{for} \;\; i \geq 2. \tag{4}$$

Proof. The edges of the form (2) can simply be verified by hand, we omit the details.

All the remaining edges can be determined using Observation 8, i.e., for each $u \in V(G)$ and $c \in \{0,1\}^*$ we need to find the shortest $v \in V(G)$ such that $uc \in \mathit{Sub}(v)$.

The edges (3) are obtained using recursive definitions of τ_{i+1} and σ_{i+1}:

$$\tau_{i+1} \;=\; \underbrace{\tau_i}_{} \cdot \bar{\tau}_i, \qquad \sigma_{i+1} \;=\; \underbrace{\tau_i}_{} \cdot \tau_{i-1}.$$

In the latter case, σ_{i+1} is the shortest bispecial factor longer than τ_i, however for the former case we need to prove that $\tau_i 1$ is not a factor of any shorter bispecial factor, namely not a factor of σ_{i+1} and $\bar{\sigma}_{i+1}$. This is, however, a consequence of Observation 6. Thus in both cases the decompositions correspond to the shortest bispecial factor of τ containing $\tau_i 0$ and $\tau_i 1$ as a factor.

The analysis of edges (4) is similar. The corresponding decompositions are as follows:

$$\tau_{i+1} \;=\; \mathit{DelQuart}_4(\tau_i) \cdot \underbrace{\sigma_i}_{} \cdot \tau_{i-1}, \qquad \bar{\tau}_{i+1} \;=\; \bar{\tau}_{i-1} \cdot \underbrace{\sigma_i}_{} \cdot \mathit{DelQuart}_4(\tau_i),$$

see also Fig. 3. Here we need to verify that σ_i is not a factor of any of the shorter bispecial factors of τ: τ_i, $\bar{\tau}_i$, σ_{i+1}, $\bar{\sigma}_{i+1}$. As for the first two, it is a consequence of Observation 5 (note that σ_i is a factor of $\bar{\tau}_i$ iff $\bar{\sigma}_i$ is a factor of τ_i). The last two cases are, again, consequences of Observation 6. □

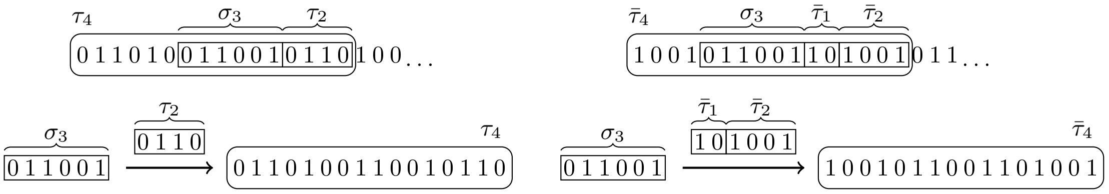

Figure 3: The edges $\sigma_3 \xrightarrow{\tau_2} \tau_4$ and $\sigma_3 \xrightarrow{\bar{\tau}_1\bar{\tau}_2} \bar{\tau}_4$.

4 Structure of Cdawgs of Finite *TM* Words

The description of vertices of G_n is obtained using the vertices of G.

Fact 10. *The vertices of* $\mathsf{cdawg}(\tau_n)$ *are all words* τ_i*,* σ_i *and their bitwise negations for* $i = 0, 1, \ldots, n-2$*, together with the source vertex corresponding to* ε *and the sink vertex corresponding to* τ_n*.*

Proof. The proof follows from Fact 7 and Observation 5. □

The structure of $E(G_n)$ differs from $E(G)$ only by the introduction of edges pointing to the sink.

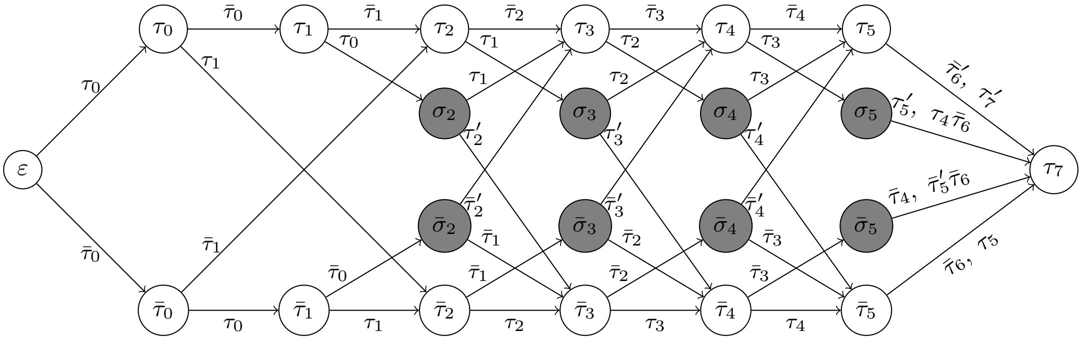

Figure 4: The cdawg for τ_7. Note the power of compaction: $|\tau_7| = 128$ and it contains 6 232 different factors (see Table 1), however $|V(G_7)| = 22$ and $|E(G_7)| = 42$.

Theorem 11. *The edges of G_n (for $n \geq 4$) are of the form (2), (3), (4) for $i \leq n-2$ (and their negations), and additionally the following edges pointing to the sink:*

$$\tau_{n-2} \xrightarrow{\bar\tau'_{n-1},\ \tau'_n} \tau_n, \qquad \bar\tau_{n-2} \xrightarrow{\bar\tau_{n-1},\ \tau_{n-2}} \tau_n, \tag{5}$$

$$\sigma_{n-2} \xrightarrow{\tau'_{n-2},\ \tau_{n-3}\bar\tau_{n-1}} \tau_n, \qquad \bar\sigma_{n-2} \xrightarrow{\bar\tau_{n-3},\ \bar\tau'_{n-2}\bar\tau_{n-1}} \tau_n. \tag{6}$$

Proof. Most of the edges of G_n are also edges of G. The only difference are edges outgoing from vertices τ_{n-2}, σ_{n-2} and their bitwise negations. Indeed, for τ_{n-2} and $\bar\tau_{n-2}$ there are no bispecial factors in $V(G_n)$ that would be longer than them (this is due to Fact 10), and for σ_{n-2} and $\bar\sigma_{n-2}$ the only longer bispecial factors are τ_{n-2} and $\bar\tau_{n-2}$, which, by Observation 5, do not contain them as factors. Hence, the edges outgoing from these four vertices can lead only to the sink. The labels on these edges are uniquely determined by Observations 5 and 8 as suffixes of τ_n starting at positions $a_{n-2}+1, b_{n-2}+1, \ldots, h_{n-2}+1$. More precisely, the labels match the following decompositions of τ_n, obtained by repetively using the recursive definition of τ_n and $\bar\tau_n$:

$$\begin{aligned}
\tau_n &= \underbrace{\tau_{n-2}}_{}\cdot DelQuart_1(\tau_n) = DelQuart_4(\tau_{n-1}) \cdot \underbrace{\tau_{n-2}}_{}\cdot DelQuart_1(\bar\tau_{n-1}),\\
\tau_n &= \tau_{n-2} \cdot \underbrace{\bar\tau_{n-2}}_{}\cdot\bar\tau_{n-1} = \tau_{n-1} \cdot \underbrace{\bar\tau_{n-2}}_{}\cdot\tau_{n-2},\\
\tau_n &= \tau_{n-1}\bar\tau_{n-3} \cdot \underbrace{\sigma_{n-2}}_{}\cdot DelQuart_1(\tau_{n-2}) = DelQuart_4(\tau_{n-2}) \cdot \underbrace{\sigma_{n-2}}_{}\cdot\tau_{n-3}\bar\tau_{n-1},\\
\tau_n &= \tau_{n-3} \cdot \underbrace{\bar\sigma_{n-2}}_{}\cdot DelQuart_1(\bar\tau_{n-2})\bar\tau_{n-1} = \tau_{n-1} DelQuart_4(\bar\tau_{n-2}) \cdot \underbrace{\bar\sigma_{n-2}}_{}\cdot\bar\tau_{n-3}.
\end{aligned}$$

□

The following theorem is a corollary of Fact 10 and Theorem 11.

Theorem 12. $|\mathsf{cdawg}(\tau_n)| = O(n) = O(\log|\tau_n|)$.

The following observation provides a classification of accepting nodes in G_n, which is useful in some applications of the cdawg (Theorem 16). These nodes are also highlighted in Fig. 5 below.

Observation 13. *The accepting nodes of G_n are $\tau_n,\ \tau_{n-2},\ \bar{\tau}_{n-3},\ \tau_{n-4},\ \bar{\tau}_{n-5},\ldots$*

Proof. A node of G_n is accepting if and only if the corresponding word is a suffix of τ_n. Note that τ_{n-2} is a suffix of τ_n and that each word on the above list is a suffix of the previous word on the list, hence all these words are accepting nodes. It remains to show that there are no more accepting nodes in the cdawg.

Clearly, none of the nodes $\bar{\tau}_{n-2},\ \tau_{n-3},\ \bar{\tau}_{n-4},\ldots$ is accepting, since it is not possible for both τ_i and $\bar{\tau}_i$ to be a suffix of τ_n.

If any of the words σ_i was a suffix of τ_n, then its suffix $\bar{\tau}_{i-1}$ would also be a suffix of τ_n. Thus $i \in \{n-2, n-4, \ldots\}$. We already proved that for these values of i, the word τ_i is a suffix of τ_n. However, it is not possible for both of the words τ_i, σ_i to be suffixes of τ_n, since $\tau_i = \tau_{i-2}\bar{\tau}_{i-2}\bar{\tau}_{i-1}$ and $\sigma_i = \tau_{i-2}\bar{\tau}_{i-1}$.

The proof for $\bar{\sigma}_i$ is completely analogical.

Finally, none of the implicit nodes could be an accepting node since end-occurrences of any implicit node are the same as end-occurrences of one or two explicit nodes. □

5 Applications of Cdawgs of *TM* Words

In this section we show several benefits of knowing the exact structure of the cdawgs of *TM* words. We consider both algorithmic and combinatorial applications of the cdawgs.

Theorem 14. *The number of different factors of τ_n for $n \ge 4$ equals $\frac{73}{192}|\tau_n|^2 + \frac{8}{3}$.*

Proof. Denote by $\mathsf{mult}(v)$ the multiplicity of vertex $v \in V(G_n)$, i.e. the number of paths from ε to v. Note that

$$\mathsf{mult}(\varepsilon) = \mathsf{mult}(\tau_0) = \mathsf{mult}(\bar{\tau}_0) = \mathsf{mult}(\tau_1) = \mathsf{mult}(\bar{\tau}_1) = 1.$$

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

$$\mathsf{mult}(\tau_i) = \mathsf{mult}(\bar{\tau}_i) = 2^{i-1} \quad \text{and} \quad \mathsf{mult}(\sigma_i) = \mathsf{mult}(\bar{\sigma}_i) = 2^{i-2}.$$

Indeed, the inductive step follows from the equalities:

$$\begin{array}{rclclcl} \mathsf{mult}(\tau_i) &=& \mathsf{mult}(\tau_{i-1}) + \mathsf{mult}(\sigma_{i-1}) + \mathsf{mult}(\bar{\sigma}_{i-1}) &=& 2^{i-2} + 2^{i-3} + 2^{i-3} &=& 2^{i-1} \\ \mathsf{mult}(\sigma_i) &=& \mathsf{mult}(\tau_{i-1}) &=& 2^{i-2} && \end{array}$$

and their symmetric copies for $\mathsf{mult}(\bar{\tau}_i)$ and $\mathsf{mult}(\bar{\sigma}_i)$. Finally, $\mathsf{mult}(\tau_n) = 3 \cdot 2^{n-2}$.

The total number of different factors of τ_n equals

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

n	1	2	3	4	5	6	7	8	9	10
$S(\tau_n)$	3	8	27	100	392	1 560	6 232	24 920	99 672	398 680

Table 1: The number of different factors of τ_n for $n \leq 10$.

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

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

For the zeroth layer (edges from ε) the sum is $S_0 = 2$, for the first (edges from τ_0 and $\bar{\tau}_0$) it equals $S_1 = 6$, and for the second (edges from τ_1 and $\bar{\tau}_1$) it also equals $S_2 = 6$. For the $(i+1)$-th layer ($2 \leq i \leq n-3$), that is for the edges going from τ_i, $\bar{\tau}_i$, σ_i, $\bar{\sigma}_i$, the corresponding part of the sum equals

$$S_{i+1} = 2 \cdot 2^{i-1} \cdot (2^{i-1} + 2^i) + 2 \cdot 2^{i-2} \cdot (2^{i-1} + 3 \cdot 2^{i-2}) = 3 \cdot 2^{2i-1} + 5 \cdot 2^{2i-3} = 17 \cdot 2^{2i-3}.$$

Finally, for the last, $(n-1)$-th layer, the sum equals

$$S_{n-1} = 2^{n-3} \cdot (3 \cdot 2^{n-3} + 3 \cdot 2^{n-2}) + 2^{n-4} \cdot (3 \cdot 2^{n-4} + 2^{n-3} + 2^{n-1}) +$$

$$+ 2^{n-4} \cdot (2^{n-3} + 3 \cdot 2^{n-4} + 2^{n-1}) + 2^{n-3} \cdot (2^{n-1} + 2^{n-2}) = 43 \cdot 2^{2n-7}.$$

Thus, we obtain the following formula:

$$S(\tau_n) = \sum_{i=0}^{n-1} S_i = 14 + \sum_{i=2}^{n-3} (17 \cdot 2^{2i-3}) + 43 \cdot 2^{2n-7} = 14 + 17 \cdot 2 \cdot \sum_{i=0}^{n-5} 4^i + 43 \cdot 2^{2n-7} =$$

$$= 14 + \frac{34}{3}(4^{n-4} - 1) + 86 \cdot 4^{n-4} = \frac{73 \cdot 4^{n-3} + 8}{3}. \quad \square$$

The following two theorems are related to efficient factor indexing of Thue-Morse words.

Theorem 15. *We can test if a word w is a factor of a given TM word τ_n in $O(|w|)$ time and $O(1)$ space.*

Moreover, if w is a factor of τ_n then we can point out the (implicit or explicit) node of G_n corresponding to w within the same time and space complexity.

Proof. First note that we can test if a specified factor of w is a *TM* word in linear time and constant space (using definition (1) of *TM* words).

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, represented in constant space as its type (τ, $\bar{\tau}$, σ, $\bar{\sigma}$) and its index. Traversing an edge of the cdawg reduces to one or several tests if a given factor of w is a *TM* word, which take $O(|w|)$ time and $O(1)$ space in total. □

The result from Theorem 15 can be further extended, as shown in the following theorem. Its proof utilizes cdawgs and is of graph-theoretic nature.

Theorem 16. *The number of occurrences of a word w in the TM word τ_n can be found in $O(|w| + \log n)$ time and $O(1)$ space.*

Proof. A well known property of cdawgs is that the number of occurrences of w in τ_n equals the number of paths from the (implicit or explicit) node corresponding to w to accepting node in G_n. Clearly, in the case of an implicit node, the number of such paths equals the number of such paths from the closest explicit node to accepting node. By Theorem 15, the aforementioned explicit node can be identified (as its type and index) in $O(|w|)$ time and $O(1)$ space.

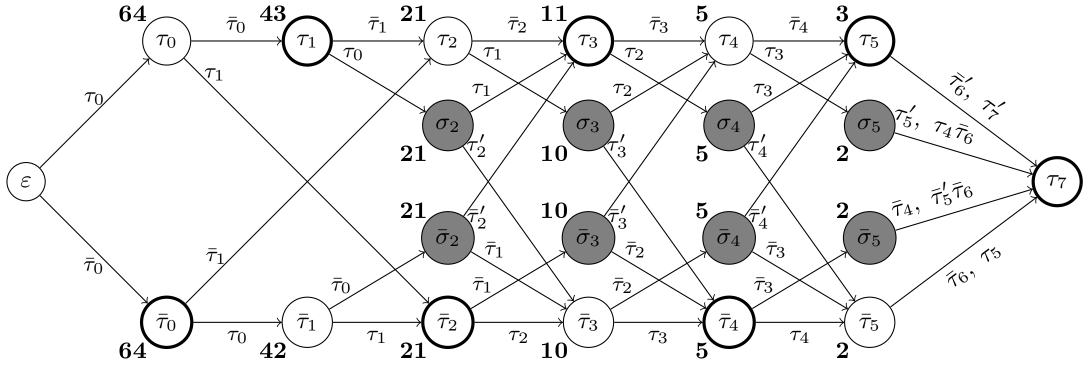

Figure 5: The cdawg G_7 with accepting vertices highlighted (bold circles). The number in bold font next to each node denotes the number of paths from this node to accepting node.

Recall the classification of accepting nodes in G_n from Observation 13. Using it we can create simple formulas for the number of accepting paths of explicit vertices of G_n. Denote as the *ith layer* $\mathcal{I}_i$ the nodes τ_i, σ_i, $\bar{\sigma}_i$, $\bar{\tau}_i$ provided that the respective nodes exist. Denote

$$g_i \;=\; \frac{2^i - (-1)^i}{3}.$$

Then for any $v \in \mathcal{I}_i$, $i \geq 1$, we have:

$$\#occ(v, \tau_n) = \begin{cases} g_{n-i+1} & \text{if } 2 \nmid (n-i) \text{ or } v = \tau_i \\ g_{n-i+1} - 1 & \text{otherwise} \end{cases}$$

and additionally $\#occ(v, \tau_n) = 2^{n-1}$ for any $v \in \mathcal{I}_0$, see Fig. 5. The inductive proof of these formulas goes layer by layer by the following recursive formulas, for $1 \leq i \leq n-3$, provided that the respective nodes exist:

$$\begin{aligned}
\#occ(\tau_i, \tau_n) &= \#occ(\tau_{i+1}, \tau_n) + \#occ(\sigma_{i+1}, \tau_n) + ((n-i+1) \bmod 2)\\
\#occ(\sigma_i, \tau_n) &= \#occ(\tau_{i+1}, \tau_n) + \#occ(\bar{\tau}_{i+1}, \tau_n)\\
\#occ(\bar{\sigma}_i, \tau_n) &= \#occ(\tau_{i+1}, \tau_n) + \#occ(\bar{\tau}_{i+1}, \tau_n)\\
\#occ(\bar{\tau}_i, \tau_n) &= \#occ(\bar{\sigma}_{i+1}, \tau_n) + \#occ(\bar{\tau}_{i+1}, \tau_n) + ((n-i) \bmod 2)
\end{aligned}$$

This concludes the proof, since the value g_{n-i+1} can be computed in $O(\log n)$ time and $O(1)$ space. □

Now we investigate the structure of binary representations of occurrences (as natural numbers) of a pattern in the infinite *TM* word τ. Applying some combinatorics of the Thue-Morse word and the properties of its cdawg we obtain a neat characterization of the set of all occurrences of any factor in τ.

Define the predicate $even(\alpha) \equiv$ "α has even number of ones in binary representation". Let X_k be the set of natural numbers with binary representation of the form $\alpha 01^j 0^{k-1}$, where $even(\alpha)$ and j is even (possibly $j = 0$), and let Y_k be the set of numbers with binary representation $\alpha 01^j 0^{k-1}$, where *not* $even(\alpha)$ and j is even (again, possibly $j = 0$). Define also

$$X \oplus c \;=\; \{x + c \;:\; x \in X\}.$$

Lemma 17.

1. *For $k \geq 1$ the pattern τ_k ($\bar{\tau}_k$) starts an occurrence at position i in τ if and only if $i \in X_k$ ($i \in Y_k$).*
2. *For each pattern w of length at least 2 the set of its occurrences in τ is a single set $X_k \oplus c$, $Y_k \oplus c$ or the union of two sets of the form $X_k \oplus c$ or $Y_k \oplus c$ for some constants k, c.*

   Moreover, the constants k, c can be computed in $O(|w|)$ time and $O(1)$ space.

Proof. The word $\tau_1 = 01$ occurs in τ at position i if and only if the representation of i has even number of ones and the representation of $i+1$ has odd number of ones (adding one changes the parity of ones), this can happen exactly when the last block of the same digits is a sequence of ones of even length.

On the other hand, each τ_k occurs as a morphic image of τ_1, we iterate the morphism $(k-1)$ times and this corresponds to adding additional $k-1$ zeros in the end of the binary representation. This proves point (1), the proof for $\bar{\tau}_k$ is analogical.

The point (2) follows from our previous results, since each pattern w has the same occurrences (shifted by a constant) as the explicit node following the implicit node corresponding to w in $\mathsf{cdawg}(\tau)$. We obtain a single set or a sum of two sets depending on whether the explicit node is of the form τ_i or $\bar{\tau}_i$ or of the form σ_i or $\bar{\sigma}_i$ — in the latter case we obtain a sum of sets corresponding to τ_{i+1} and $\bar{\tau}_{i+1}$ shifted by some constants, which is due to the structure of the cdawg G.

Finally, the algorithm computing the constants k and c in point (2) follows from Theorem 15. □

Example 18. *The factor* 0011 *occurs in* τ *at positions with binary representation* $\alpha 01^j 101$ *and* $\beta 01^j 111$*, where* $2 \mid j$*,* $even(\alpha)$ *and not* $even(\beta)$*.*

On the other hand, the factor 1011 *occurs at positions of the form* $\alpha 01^j 11$*, where* $2 \mid j$ *and not* $even(\alpha)$*.*

Observe that the binary representation of the constant c from the last lemma has at most $k-1$ bits. We can consider the complexity of the sequence of occurrences of w in τ as an automatic sequence (in the sense of [2]).

Let $BinOcc(w, \tau)$ be the set of binary representations of all starting occurrences of w in τ (the representations are words starting from the least significant binary digit). Due to the very simple structure of representations we have the following fact.

Observation 19. *We can compute in linear time the minimal deterministic automaton accepting* $BinOcc(w, \tau)$*, this automaton has* $O(\log |w|)$ *states.*

The final application of the cdawg which we present is related to periodicity of τ_n. Let us start with recalling several notions.

Let $u = u_0 u_1 \dots u_{m-1}$. A positive integer q is the (shortest) *period* of u (notation: $q = \mathsf{per}(u)$) if q is the smallest number such that $u_i = u_{i+q}$ holds for all $0 \le i \le m-q-1$. We say that a square vv is centered at inter-position p_i of u if both of the following conditions hold, for $x = u[0\,.\,.\,i]$ and $y = u[i+1\,.\,.\,m-1]$:

- v is a suffix of x or x is a suffix of v
- v is a prefix of y or y is a prefix of v.

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

Theorem 20. *The critical factorization point of the TM word* τ_n*, for* $n \ge 4$*, is the inter-position* p_i *for* $i = 2^{n-1}$*.*

Proof. It is a known fact [9] 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 find the suffix (1), we traverse G_n, starting from ε, along the lexicographically largest path, shown by bold straight edges in Fig. 6. The length of this path is $|\tau_n| - 1$.

On the other hand, the suffix (2) corresponds to the maximal path starting from ε that always prefers 0 over 1, shown by bold snaked edges in Fig. 6. Its length equals $|\tau_{n-1}| - 1$.

Thus the suffix (2) is always shorter than (1) and using it we obtain the critical factorization point as specified in the conclusion of the theorem. □

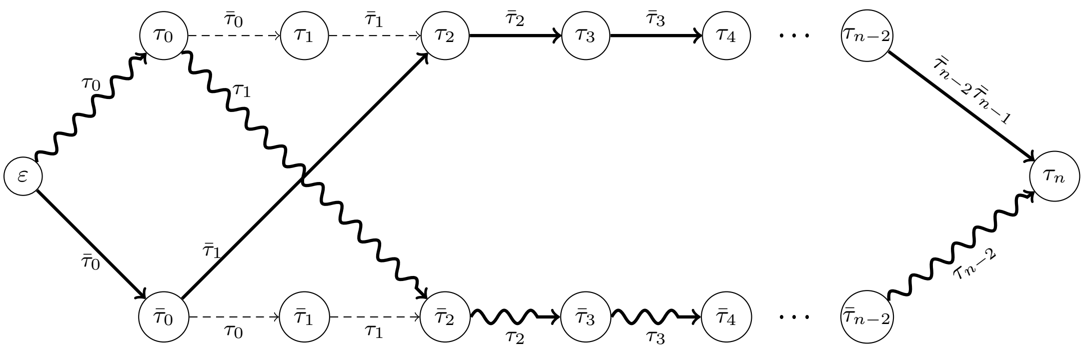

Figure 6: The lexicographically largest path (bold, straight edges) and the lexicographically smallest maximal path (bold, snaked edges) in the cdawg G_n

6 2-Counting Property of Weighted Pseudo-Cdawg of τ

The main result of this section is Theorem 21, in which we show that a slight modification of the cdawg of the infinite *TM* word has 2-counting property. This is related to previous results on counting properties of Sturmian graphs [13].

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, see Fig. 7. We call G' the *pseudo-cdawg* of τ. Let H be a directed weighted graph obtained from G' by replacing labels of edges with their lengths, see Fig. 8. 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 (two series $2, 4, 8, 16, \ldots$).

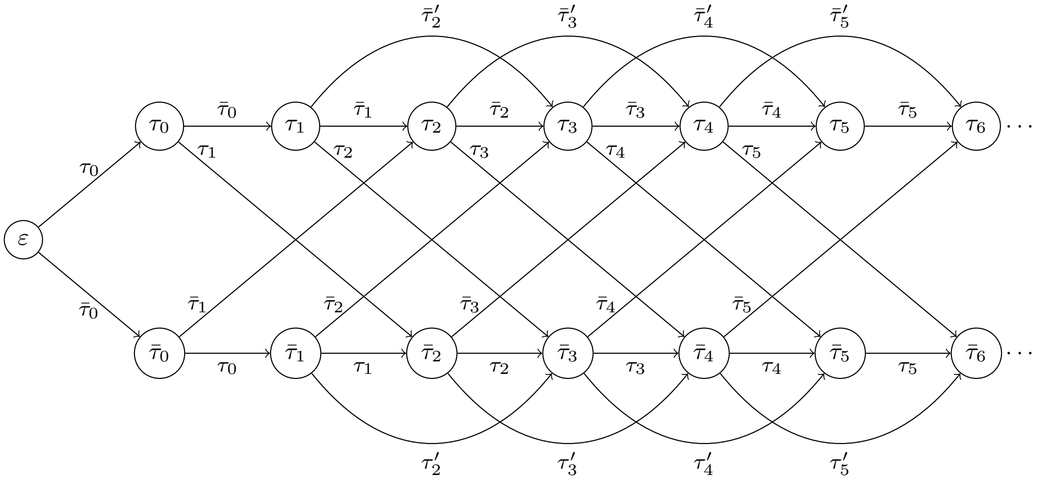

Figure 7: The pseudo-cdawg for τ — an initial part.

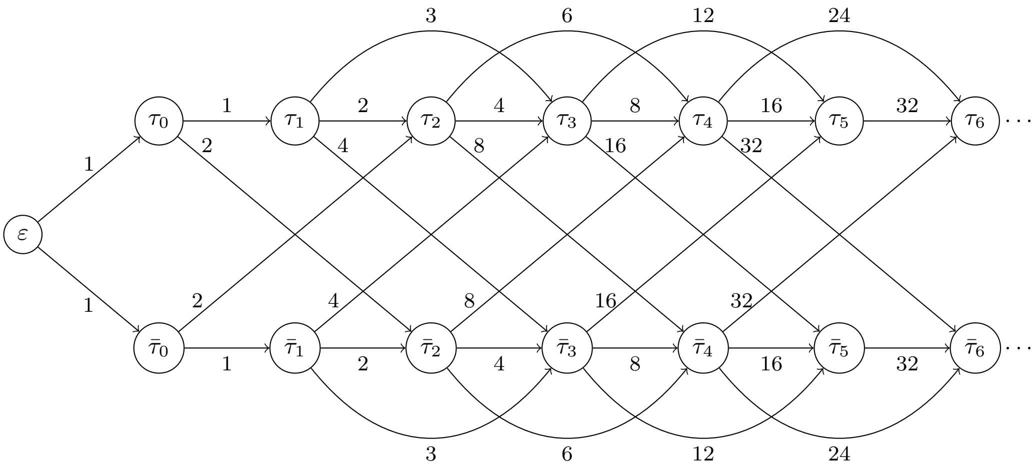

Figure 8: The initial part of the weighted graph H obtained from G' by replacing labels with their lengths. This graph has 2-counting property.

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.

Theorem 21. *Graph H with source vertex ε has 2-counting property.*

Proof. Clearly, in H there are exactly two paths from ε of length 1, ending in τ_0 and $\bar\tau_0$ respectively. To prove the theorem, it suffices to show that for $i \geq 1$, for every $\ell \in [2^{i-1}+1,\ 2^i]$ there exists exactly one path from ε to τ_i of length ℓ (the same claim can be proved for $\bar\tau_i$, since the graph is symmetric). We show this by induction on i.

The base $i \leq 2$ is trivial. Let $i > 2$ and assume the inductive hypothesis for all $i' < i$. There are exactly three edges ending in τ_i in H; for each of these edges we determine the set of lengths of paths from ε that end with that edge e. Note that this set is exactly the interval of lengths of paths ending in the starting vertex of e shifted by the weight of e. For the edge of type $\tau_{i-2} \xrightarrow{3\cdot 2^{i-3}} \tau_i$ we obtain

$$3\cdot 2^{i-3} + [2^{i-3}+1,\ 2^{i-2}] \ = \ [2^{i-1}+1,\ 2^{i-1}+2^{i-3}], \tag{7}$$

for the edge $\bar\tau_{i-2} \xrightarrow{2^{i-1}} \tau_i$:

$$2^{i-1} + [2^{i-3}+1,\ 2^{i-2}] \ = \ [2^{i-1}+2^{i-3}+1,\ 2^{i-1}+2^{i-2}], \tag{8}$$

finally for the edge $\tau_{i-1} \xrightarrow{2^{i-1}} \tau_i$ the set of lengths of paths equals

$$2^{i-1} + [2^{i-2}+1,\ 2^{i-1}] \ = \ [2^{i-1}+2^{i-2}+1,\ 2^i]. \tag{9}$$

The intervals (7)-(9) are pairwise disjoint and the set of integers contained in any of them is $[2^{i-1}+1,\ 2^i]$. This concludes the inductive proof. □

References

[1] J.-P. Allouche and J. Shallit. The ubiquitous Prouhet-Thue-Morse sequence. *Springer Ser. Discrete Math. Theor. Comput. Sci.*, pages 1–16, 1999.

[2] J.-P. Allouche and J. Shallit. *Automatic Sequences.* Cambridge University Press, 2003.

[3] L. Balkova, E. Pelantova, and W. Steiner. Return words in the Thue-Morse and other sequences. *arxiv:math/0608603v2*, 2006.

[4] J. Berstel, A. Lauve, C. Reutenauer, and F. V. Saliola. *Combinatorics on Words: Christoffel Words and Repetitions in Words.* Amer. Mathematical Society, 2009.

[5] A. Blumer, J. Blumer, D. Haussler, A. Ehrenfeucht, M. T. Chen, and J. I. Seiferas. The smallest automaton recognizing the subwords of a text. *Theor. Comput. Sci.*, 40:31–55, 1985.

[6] S. Brlek. Enumeration of factors in the Thue-Morse word. *Discrete Applied Mathematics*, 24(1-3):83–96, 1989.

[7] M. Crochemore. Reducing space for index implementation. *Theor. Comput. Sci.*, 292(1):185–197, 2003.

[8] M. Crochemore and W. Rytter. *Text Algorithms.* Oxford University Press, 1994.

[9] M. Crochemore and W. Rytter. *Jewels of Stringology.* World Scientific, 2003.

[10] M. Crochemore and R. Vérin. Direct construction of compact directed acyclic word graphs. In *CPM*, pages 116–129, 1997.

[11] A. de Luca and L. Mione. On bispecial factors of the Thue-Morse word. *Inf. Process. Lett.*, 49(4):179–183, 1994.

[12] A. de Luca and S. Varricchio. Some combinatorial properties of the Thue-Morse sequence and a problem in semigroups. *Theor. Comput. Sci.*, 63(3):333–348, 1989.

[13] C. Epifanio, F. Mignosi, J. Shallit, and I. Venturini. On Sturmian graphs. *Discrete Applied Mathematics*, 155(8):1014–1030, 2007.

[14] S. Inenaga, H. Hoshino, A. Shinohara, M. Takeda, S. Arikawa, G. Mauri, and G. Pavesi. On-line construction of compact directed acyclic word graphs. *Discrete Applied Mathematics*, 146(2):156–179, 2005.

[15] W. Rytter. The structure of subword graphs and suffix trees of Fibonacci words. *Theor. Comput. Sci.*, 363(2):211–223, 2006.

[16] J. Tromp and J. Shallit. Subword complexity of a generalized Thue-Morse word. *Inf. Process. Lett.*, 54(6):313–316, 1995.