BINARY AMIABLE WORDS*

ADRIAN ATANASIU
Faculty of Mathematics and Computer Science,
Bucharest University
Str. Academiei 14, Bucharest 010014, Romania
E-mail: aadrian@gmail.com
http://www.galaxyng.com/adrian_atanasiu

Received (31 July 2006)
Revised (20 October 2006)
Accepted (13 November 2006)
Communicated by (Arto Salomaa)

Using the fact that the Parikh matrix mapping is not an injective mapping, the paper investigates some properties of the set of the binary words having the same Parikh matrix; these words are called "amiable" Some results concerning the conditions when the equivalence classes of amiable words have more than one element, a characterization theorem concerning a graph associated to an equivalence class of amiable words, and some basic properties of a rank distance defined on these classes are the main subjects considered here.

Keywords: Parikh matrix mapping; amiable words; scatered subwords; rank distance.

1991 Mathematics Subject Classification: 15A36, 05B20, 05C38, 11D04, 03D40.

1. Introduction

In this paper we investigate some properties of the Parikh matrix mapping defined only for the binary alphabet. The Parikh matrix mapping (introduced in [6]) is an extension of the Parikh mapping ([7]). The extension is based on a special type of matrices, where the classical Parikh vector appears as the second diagonal.

First of all, let us start with some basic notations and definitions. The set of all nonnegative integers is denoted by $\mathcal{N}$. Let Σ be an alphabet. The set of all words over Σ is Σ^*; if λ is the empty word, then the set of nonempty sequences is $\Sigma^+ = \Sigma^* \setminus \{\lambda\}$. For $\alpha \in \Sigma^*$, $|\alpha|$ denotes the length of α. Besides, for any finite set A we denote $|A|$ the number of elements contained by A.

The mirror image of a word $\alpha \in \Sigma^*$, denoted $mi(\alpha)$, is defined as: $mi(\lambda) = \lambda$, $mi(a_1a_2 \dots a_n) = a_n \dots a_2a_1$, where $a_i \in \Sigma$, $1 \leq i \leq n$. A word α is a "palindrome" iff $\alpha = mi(\alpha)$.

*This paper is dedicated to my colleague and friend Alexandru Mateescu, the first author of Parikh matrix mapping.

The alphabet used in this paper is a binary ordered alphabet $\Sigma = \{a, b\}$ where a relation of order ("<") is defined. Without loss of generality, we consider $a < b$.

Let $x \in \Sigma$ be a letter. The number of occurrences of x in a word $\alpha \in \Sigma^*$ is denoted by $|\alpha|_x$. Let u, v be words over Σ. The word u is a scattered subword of v if there exists a word w such that $v \in shuffle(u, w)$. We denote by $|\alpha|_{ab}$ the number of occurrences of $u = ab$ in $v = \alpha$ as a scattered subword. For instance $|abab|_{ab} = 3$.

For the binary ordered alphabet $\Sigma = \{a < b\}$, the Parikh mapping is a mapping

$$\Psi : \Sigma^* \longrightarrow \mathcal{N}^2$$

defined as $\Psi(\alpha) = (|\alpha|_a, |\alpha|_b)$. This couple represents the Parikh vector of α.

Definition 1. *Let $\Sigma = \{a < b\}$ be a binary ordered alphabet and $\mathcal{M}_3$ be the set of 3-dimensional upper-triangular matrices with nonnegative integral entries and unit diagonal. The Parikh matrix mapping, denoted by $\Psi_M : \Sigma^* \longrightarrow \mathcal{M}_3$ is defined as follows: For each $\alpha \in \Sigma^*$*

$$\Psi_M(\alpha) = \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix}$$

where $x = |\alpha|_a,\ y = |\alpha|_b,\ z = |\alpha|_{ab}$.

We denote $\Psi_M(\alpha)$ also by M_α.

A matrix $M \in \mathcal{M}_3$ with the property $M = M_\alpha$ for a particular word $\alpha \in \Sigma^*$ is called *Parikh matrix*.

In [1, 2] some general properties of the Parikh matrices were proved; for example

$$M_{\alpha\beta} = M_\alpha M_\beta, \quad \forall \alpha, \beta \in \Sigma^*.$$

Remark 1. *This equation was the defining one for Parikh matrix mappings in the original reference ([6]).*

Definition 2. *Two words $\alpha, \beta \in \Sigma^*$ are called "amiable" iff $M_\alpha = M_\beta$.*

In [1] the notion of "palindromic amiable" words is defined; the supplementary condition required is that the words α, β are palindromes.

Denote by $\alpha \sim_a \beta$ the property that α and β are amiable words.

The relation $\sim_a$ is obviously an equivalence relation (in [4] is defined a congruence relation $\equiv_2$ very close to $\sim_a$).

In [1] another equivalence relation is defined. Namely:

$x \equiv_{pa} y$ iff $\exists\, \alpha, \beta \in \Sigma^+$ palindromic amiable words so that $x = u\alpha v,\ y = u\beta v$.

$\equiv^*_{pa}$ is the reflexive and transitive closure of $\equiv_{pa}$.

Example 1. *$aabbabaaa \equiv^*_{pa} babaaaaab$. Indeed,*

$a\underbrace{abba}baaa \equiv_{pa} ab\underbrace{aabbaa}a \equiv_{pa} \underbrace{abba}aaaba \equiv_{pa} ba\underbrace{abaaaba} \equiv_{pa} babaaaaab$.

The next result can be proved using Theorem 3.9 ([1]):

Proposition 1. *For any $\alpha, \beta \in \Sigma^*,\ \alpha \sim_a \beta \iff \alpha \equiv^*_{pa} \beta$.*

Regarding other issues about Parikh matrix mapping (in general form), as well as for language-theoretic considerations not detailed here, the reader is referred to [2, 4, 5, 8, 9] and the references given therein.

2. Classes of binary amiable words

The first part of this section contains some direct computational proofs of the results concerning Parikh matrix mappings presented in the literature.

Let $\Sigma = \{a < b\}$ be a binary ordered alphabet. A general result about binary amiable words is:

Lemma 1.

(1) If $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \Sigma^*$, $\alpha_1 \sim_a \beta_1$, $\alpha_2 \sim_a \beta_2$, *then* $\alpha_1\alpha_2 \sim_a \beta_1\beta_2$.
(2) If $\alpha, \beta, \gamma \in \Sigma^*$, *then* $\alpha ab\beta ba\gamma \sim_a \alpha ba\beta ab\gamma$.

Proof. The first assertion is obvious.

Because a Parikh matrix always has an inverse, for the second assertion, it is enough to prove the equality $M_{ab}M_\beta M_{ba} = M_{ba}M_\beta M_{ab}$. Indeed, if

$$M_{ab} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad M_{ba} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad M_\beta = \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix}$$

then

$$M_{ab}M_\beta M_{ba} = M_{ba}M_\beta M_{ab} = \begin{pmatrix} 1 & x+2 & x+y+z+2 \\ 0 & 1 & y+2 \\ 0 & 0 & 1 \end{pmatrix}$$ □

As a remark, *abba* $\sim_a$ *baab*; moreover, *abba* and *baab* are the only binary amiable words of minimal length.

For a word $\alpha \in \Sigma^*$ we denote by C_α its equivalence class:

$$C_\alpha = \{\beta \in \Sigma^* \mid \beta \sim_a \alpha\}$$

An obvious isomorphism can be established between the multiplicative group of 3×3 Parikh matrices and the group of classes C_α with the rule $C_\alpha \circ C_\beta = C_{\alpha\beta}$.

The main problem we whish to solve in this section is:

Given a Parikh matrix $M_\alpha = \begin{pmatrix} 1 & p & q \\ 0 & 1 & n \\ 0 & 0 & 1 \end{pmatrix}$ **with** $n, p, q \in \mathcal{N}$**, how many words does** C_α **contain ?**

In particular, under which conditions $C_\alpha = \{\alpha\}$ **?**

In [1] a recursive function is defined for computing the value $|C_\alpha|$. We try in this section to simplify the answer and to complete those results.

Let $\alpha \in \Sigma^+$ be a binary sequence (the case $\alpha = \lambda$ is trivial and will be ignored); α can be represented in the following form (by detailing the appearances of letter b):

$$\alpha = a^{x_1}ba^{x_2}b \ldots a^{x_n}ba^{x_{n+1}} \qquad (x_i \geq 0) \tag{1}$$

The Parikh matrix $M_\alpha = \begin{pmatrix} 1 & p & q \\ 0 & 1 & n \\ 0 & 0 & 1 \end{pmatrix}$ corresponds to this word if and only if $(x_1, x_2, \ldots, x_{n+1}) \in \mathcal{N}^{n+1}$ is a solution of the system

$$\begin{cases} x_1 + \quad x_2 + \ldots + \; x_{n-1} + x_n + x_{n+1} = p \\ nx_1 + (n-1)x_2 + \ldots + 2x_{n-1} + x_n \quad = q \end{cases} \tag{2}$$

It is clear that for every solution of this system there is a corresponding sequence in C_α and vice-versa. Therefore the number of solutions of the system (2) equals $|C_\alpha|$.

Remark 2. *The number of solutions of the system* (2) *equals also the number of solutions of the equation* $x_1 + x_2 + \ldots + x_n = q$ *where* $x_i \in \mathcal{N}$ $(1 \leq i \leq n)$ *and* $0 \leq x_1 \leq \ldots \leq x_n \leq p$ *(see* [1]*).*

Example 2. *Let* $M = \begin{pmatrix} 1 & 2 & 4 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{pmatrix}$ *be a Parikh matrix with* $p = 2,\ q = 4,\ n = 4$. *The associated system is*

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 + x_5 &= 2 \\ 4x_1 + 3x_2 + 2x_3 + x_4 \quad &= 4 \end{aligned}$$

This system with 5 *variables has three solutions in* $\mathcal{N}^5$. *Namely:*

(1) $(x_1, x_2, x_3, x_4, x_5) = (1, 0, 0, 0, 1)$ *which corresponds to the word* $\beta_1 = abbbba$;
(2) $(x_1, x_2, x_3, x_4, x_5) = (0, 1, 0, 1, 0)$ *which corresponds to the word* $\beta_2 = babbab$;
(3) $(x_1, x_2, x_3, x_4, x_5) = (0, 0, 2, 0, 0)$ *which corresponds to the word* $\beta_3 = bbaabb$.

Thus the set of sequences with the Parikh matrix M *is* $C = \{abbbba, babbab, bbaabb\}$.

The following result is not new (see [5]). We give a proof based on considerations concerning the system (2). A similar remark applies for Theorem 2.

Theorem 1. *Let* $M = \begin{pmatrix} 1 & p & q \\ 0 & 1 & n \\ 0 & 0 & 1 \end{pmatrix}$ *be a matrix with* $p, q, n \in \mathcal{N}$. M *is a Parikh matrix iff* $q \in [0, n \cdot p]$.

Proof. If M is a Parikh matrix, there is at least one binary word α of the type (1) which verifies the system (2). We have to prove that for this word the relations $x_1 + x_2 + \ldots x_n + x_{n+1} = p$ and $nx_1 + (n-1)x_2 + \ldots + x_n \leq np$ are fulfilled. The equality results directly from the construction of (1).

To prove the inequality, let us evaluate

$\sum_{i=1}^{n}(n+1-i)x_i = n\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}(i-1)x_i = n(p-x_{n+1}) - \sum_{i=1}^{n}(i-1)x_i =$

$np - \sum_{i=1}^{n+1}(i-1)x_i \leq n \cdot p$

Because $x_i \in \mathcal{N}$, the values are at least 0.

Now we prove the "if" part of the theorem.

Assume that $n, p, q \in \mathcal{N}$ such that $q \leq n \cdot p$. Then a solution $(x_1, x_2, \ldots, x_{n+1}) \in \mathcal{N}^{n+1}$ of the system (2) can be

$$x_1 = \left\lfloor \frac{q}{n} \right\rfloor,\ x_{i+1} = \left\lfloor \frac{q - \sum_{j=1}^{i}(n+1-j)x_j}{n-i} \right\rfloor\ (1 \leq i \leq n-1),\ x_{n+1} = p - \sum_{i=1}^{n} x_i \quad (3)$$

Checking that (3) is a solution is easy. □

Remark 3. *In the cases $n = 0$ or $p = 0$ we immediately conclude that $q = 0$. So, in the following there will be treated only the nontrivial situations $n \cdot p \neq 0$.*

For the next theorem, a preliminary result is necessary:

Lemma 2. *If* $M_\alpha = \begin{pmatrix} 1 & p & q \\ 0 & 1 & n \\ 0 & 0 & 1 \end{pmatrix}$ *then* $M_{mi(\alpha)} = \begin{pmatrix} 1 & p & n \cdot p - q \\ 0 & 1 & n \\ 0 & 0 & 1 \end{pmatrix}$.

Proof. Let us consider $\alpha = a^{x_1}ba^{x_2}b \ldots a^{x_n}ba^{x_{n+1}}$ $(x_i \geq 0)$, which corresponds to the system (2). Then for $mi(\alpha) = a^{x_{n+1}}ba^{x_n}b \ldots a^{x_2}ba^{x_1}$ will correspond a system in which the first equation is the same, but the second equation is

$$nx_{n+1} + (n-1)x_n + \ldots + x_2 = A$$

where A is a value to be found.

We evaluate $n \cdot p = n(x_1 + x_2 + \ldots + x_{n+1}) = [nx_1 + (n-1)x_2 + \ldots + x_n] + x_2 + 2x_3 + \ldots + nx_{n+1} = q + A$. Thus $A = n \cdot p - q$ and the system built for the word $mi(\alpha)$ corresponds to the Parikh matrix $M_{mi(\alpha)}$ defined above. □

Theorem 2. *Let $\alpha \in \Sigma^+$ be a binary word with the Parikh matrix*

$M_\alpha = \begin{pmatrix} 1 & p & q \\ 0 & 1 & n \\ 0 & 0 & 1 \end{pmatrix}$, $p, q, n \in \mathcal{N}$. *For each of the cases*

(1) $n \leq 1$;
(2) $p \leq 1$;
(3) $q \in \{0, 1, n \cdot p - 1, n \cdot p\}$,

$|C_\alpha| = 1$ *(thus* $C_\alpha = \{\alpha\}$*) holds.*

Proof.

(1) For $n = 0$ (thus $q = 0$) there is only one sequence: $\alpha = a^p$. For $n = 1$ (therefore, accordingly with the Theorem 1, $p \leq q$) there is again only one solution: $\alpha = a^q b a^{p-q}$.
(2) Similarly.
(3) We have to prove that for the values of q given above, the system (2) has only one solution.

Using Lemma 2 we conclude that $|C_\alpha| = |C_{mi(\alpha)}|$; therefore it is enough to prove the assertion only for values $q = 0$ and $q = 1$.

We consider each case.

- $\mathbf{q = 0}$: The diophantine equation $nx_1 + (n-1)x_2 + \ldots + x_n = 0$ has only one solution: $x_i = 0$ $(i = 1, \ldots, n)$. By replacing these values in the first equation of the system (2) we obtain $x_{n+1} = p$. Therefore the solution $(0, \ldots, 0, p)$ is unique and corresponds to the binary word $\alpha = b^n a^p$.
- $\mathbf{q = 1}$: The diophantine equation $nx_1 + (n-1)x_2 + \ldots + x_n = 1$ has the unique solution $x_1 = x_2 = \ldots = x_{n-1} = 0,\ x_n = 1$. By replacing it in the first equation of the system (2) we obtain $x_{n+1} = p - 1$.
  The constraint $1 \leq n \cdot p$ (obtained from Theorem 1 with $q = 1$) assures $n \cdot p \neq 0$; in peculiar, $p \geq 1,\ n \geq 1$, thus the solution $(0, \ldots, 0, 1, p-1)$ can be constructed. It is unique and corresponds to the word $\alpha = b^{n-1} a b a^{p-1}$. □

Theorem 3. *Let* $\alpha \in \Sigma^+$ *with the Parikh matrix* $M_\alpha = \begin{pmatrix} 1 & p & q \\ 0 & 1 & n \\ 0 & 0 & 1 \end{pmatrix}$. *If* $n \geq 2,\ p \geq 2$ *and* $q \in [2, n \cdot p - 2]$, *then* $|C_\alpha| \geq 2$.

Proof. Two cases are possible:

(1) $\mathbf{p \geq q}$: Let us denote $p = q + s$ $(s \geq 0)$. Using the hypothesis, $(0, \ldots, 0, t, q - 2t, t + s)$ (with 0 in the first $n - 2$ positions) is a solution of the system (2). It corresponds to the word $\alpha = b^{n-2} a^t b a^{q-2t} b a^{t+s}$. The constraint $q - 2t \geq 0$ assures a distinct solution for each $t \in \left[0, \frac{q}{2}\right]$. Because $q \geq 2$, there are at least two solutions, therefore $|C_\alpha| \geq 2$.
(2) $\mathbf{p < q}$: We know that (3) is a solution of the system (2).

For another (distinct) solution, let us consider $q = s \cdot p + r$ with $1 \leq s < n,\ 0 \leq r < p$. Then

$$x_{n-s} = r,\ x_{n-s+1} = p - r,\ x_i = 0 \quad (i \neq n - s, n - s + 1)$$

is a solution for the system (it corresponds to the sequence $b^{n-s-1} a^r b a^{p-r} b^s$).

Indeed, the first equation is verified with $r + (p - r) = p$, and the second with $(s + 1)r + s(p - r) = s \cdot p + r = q$.

The fact that these two solutions are distinct can be easily proven. □

Theorems 2 and 3 cover all cases. Therefore all the cases when $C_\alpha = \{\alpha\}$ are defined by the Theorem 2.

Corollary 1. *Let $L \subseteq \{a,b\}^*$ and $W = a^*b^* + a^*bab^* + a^*ba^* + b^*ab^*$. The Parikh matrix mapping $\Psi : L \longrightarrow \mathcal{M}_3$ is injective iff $L \subseteq W \cup mi(W)$.*

Unfortunately it is not easy to check an upper limit for the number of words which are amiable with a given binary word α. The recursive mapping

$$\phi(q,p,n) = \sum_{i=0}^{min\{p,q\}} \phi(q-i,i,n-1), \qquad \phi(q,p,1) = \begin{cases} 1 \ , \ q \leq p \\ 0 \ , \ q > p \end{cases}$$

established in [1] (Theorem 4.7) is difficult to be evaluated.
The next Theorem assures only a weak lower bound of this limit:

Theorem 4. *There is at least a word $\alpha \in \Sigma^+$ with $|C_\alpha| \geq \frac{1}{n \cdot p+1} \binom{n+p}{p}$.*

Proof. For two given values $n, p \in \mathcal{N}$ exist $\binom{n+p}{p}$ binary sequences with p $a's$ and n $b's$. For each $q \in [0, n \cdot p]$ exists (Theorem 1) a Parikh matrix, thus there exists at least a word $\alpha \in \Sigma^*$. There are $n \cdot p + 1$ possible classes of amiable words, thus at least one of them will contain a number of binary words greater than or equal to the average. □

Some results are obvious:

Proposition 2.

(1) If $q_1 < q_2 \leq \left\lfloor \frac{n \cdot p}{2} \right\rfloor$ then $|C_{\alpha_1}| \leq |C_{\alpha_2}|$, where $M_{\alpha_i} = \begin{pmatrix} 1 & p & q_i \\ 0 & 1 & n \\ 0 & 0 & 1 \end{pmatrix}$, $i = 1, 2$.

(2) If $q_1 + q_2 = n \cdot p$ then $|C_{\alpha_1}| = |C_{\alpha_2}|$.

Corollary 2. *For a fixed Parikh vector $\Psi = (p, n)$, $\max\limits_{\Psi(\alpha)=\Psi} \{|C_\alpha|\}$ is established for $q = \left\lfloor \frac{n \cdot p}{2} \right\rfloor$.*

Example 3. *Let us consider $p = 19$, $n = 2$. The next table shows $|C_\alpha|$ for $q = 1, 2, \ldots 38$, where $M_\alpha = \begin{pmatrix} 1 & 19 & q \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$:*

The lower bound given by Theorem 4 is $\frac{1}{39} \binom{21}{2} = 5,7$. There are 19 classes C_α which verify this Theorem.

Table 1. The table of amiable binary words having the Parikh vector $\Psi = (19, 2)$.

q	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
$\lvert C_\alpha\rvert$	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	10	9	9	8	8	7	7	6	6	5	5
q	31	32	33	34	35	36	37	38																						
$\lvert C_\alpha\rvert$	4	4	3	3	2	2	1	1																						

Table 2. The table of amiable binary words having the Parikh vector $\Psi = (10, 10)$.

q	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
$\lvert C_\alpha\rvert$	1	1	2	3	5	7	11	15	22	30	42	54	73	93	121	152	193	237	295	356	433	515	615
q	23	24	25	26	27	28	29	30	31	32	33	34	35	36									
$\lvert C_\alpha\rvert$	720	847	978	1131	1289	1420	1652	1860	2065	2293	2517	2761	2994	3246									
q	37	38	39	40	41	42	43	44	45	46	47	48	49	50									
$\lvert C_\alpha\rvert$	3481	3729	3956	4192	4397	4609	4784	4959	5095	5226	5311	5392	5424	5448									

Example 4. *For the values* $p = 10,\ n = 10$ *the table is (we reprezent only the values* $q = 0, 1 \ldots, 50$*):*

The lower bound given by the Theorem 4 is $\dfrac{1}{101}\binom{20}{10} = 1829,27$.

There are 43 *sets which exceed this limit.*

We made many other tests; it seems that if we set a Parikh vector and construct all the $n \cdot p + 1$ classes C_α corresponding to all values q, then at least 40% of them verify Theorem 4. This percentage equals 100% if and only if the Parikh matrix mapping is injective.

3. Theorem of characterization for amiable words

In this section we will consider an equivalence class C, corresponding to a given Parikh matrix M.

Let us define the unoriented graph $\Gamma_M = (V, E)$ as follows:

- $V = C$;
- $(\alpha, \beta) \in E \iff \exists \gamma_1, \gamma_2, \gamma_3 \in \{a, b\}^*,\ \alpha = \gamma_1 ab\gamma_2 ba\gamma_3,\ \beta = \gamma_1 ba\gamma_2 ab\gamma_3$.

From Lemma 1 it results that the sequences α, β are amiable, thus they belong to the same equivalence class.

The main result of this section is:

Theorem 5. *The graph* Γ_M *is connected.*

Proof. The assertion is trivial for $|C| = 1$.

Let us consider only the situation studied by the Theorem 3. We have to show that for every $\alpha, \beta \in C$ there exists a path between the nodes α and β; hence there is a sequence of transformations of the type $ab - ba$ by which the word β can be obtained from α (and vice-versa).

On the set C we define the next derivation rule:

$\forall \alpha, \beta \in C$
$\{\alpha \implies \beta\} \Longleftrightarrow \{\exists \gamma_1, \gamma_2, \gamma_3 \in \{a,b\}^*,\ \alpha = \gamma_1 ab\gamma_2 ba\gamma_3,\ \beta = \gamma_1 ba\gamma_2 ab\gamma_3\}$

Lemma 1 assures that this derivation is well defined on C. For $\alpha \in C$ we can apply this rule as long as it is possible. Because, sometimes, several variants may appear on each step, we will define a constraint in order to choose only one continuation. Namely

(*i*) *Let* $\alpha = a^{x_1}ba^{x_2}b\ldots a^{x_n}ba^{x_{n+1}} \in C$ *and* x_i, x_j *be the first two positive exponents such that* $j > i+1$. *Then we can write* $\alpha \Longrightarrow \beta$, *where* $\alpha = \gamma_1 a^{x_i} b\gamma_2 ba^{x_j}\gamma_3$ *and* $\beta = \gamma_1 a^{x_i-1}ba\gamma_2 aba^{x_j-1}\gamma_3$.

This rule cannot be applied when there remain at most two positive exponents. In more detail:

- If there remains only one positive exponent, then α has the form $\alpha = b^*a^pb^*$;
- If two consecutive positive exponents remain, then α has the form $\alpha = b^*a^+ba^+b^*$

For each of the two cases discussed in the proof of Theorem 3, the word (denoted α_0) when the derivation rule cannot be applied accordingly to the constraint (*i*) is unique.

Let $q = s \cdot p + r$ where $0 \le s < n,\ 0 \le r < p$ (the case $s = 0$ covers the variant $q < p$). Then

$$\alpha_0 = b^{n-s-1}a^r ba^{p-r}b^s$$

Therefore, for each $\alpha \in C$ we have a (possible empty) sequence $\alpha \Longrightarrow \ldots \Longrightarrow \alpha_0$.

In the graph Γ_M associated to C, this assures (at least) one path from the node α to the node α_0.

The theorem results now from the fact that the graph Γ_M is unoriented: for each $\alpha, \beta \in C,\ (\alpha \ne \beta)$ there is a path between α and β which goes through α_0. □

Example 5. *Let* $n = p = 4,\ q = 8$. *The system* (2) *has eight solutions which correspond to the set of amiable words* $C = \{$*aabbbbaa*, *ababbaba*, *abbaabba*, *baabbaab*, *babaabab*, *bbaaaabb*, *baababba*, *abbabaab*$\}$
Using the derivation rule defined above, we can construct the graph Γ_M *(Figure* 1*).*

The word α_0 *considered by the constraint* (*i*) *is* $\alpha_0 = bbaaaabb$.
It can be considered as the unique representative of the equivalence class.

Conjecture: If $C \bigcap b^*a^*b^* = \emptyset$ then the graph Γ_M is hamiltonian. Otherwise, for every $i = 3, 4, \ldots, |C| - 2$ there exists at least a loop with i elements.

Let $\alpha, \beta \in C,\ \alpha = a_1a_2\ldots a_n,\ \beta = b_1b_2\ldots b_n$ be two words with the same Parikh vector; the Rank distance $d_R(\alpha, \beta)$ (defined by L. P. Dinu and A. Sgarro;

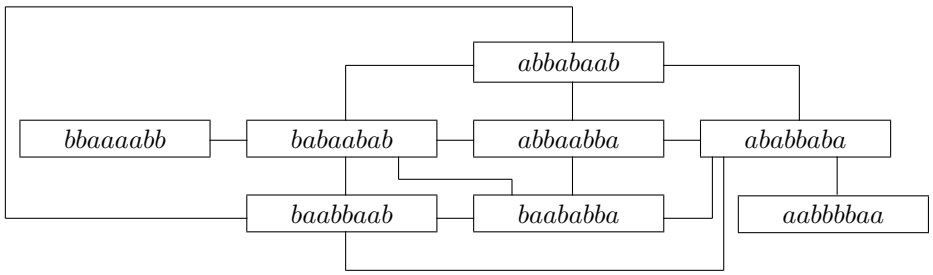

Fig. 1. The graph Γ_M obtained for $n = p = 4,\ q = 8$.

see [3]) counts the number of places between the similar characters from α and β. Formally:

$$d_R(\alpha, \beta) = \sum_{x \in \alpha} |ord_\alpha(x) - ord_\beta(x)|$$

where $ord_u(x)$ represents the position of the character x in the string u, counted from the left to right.

Example 6. $d_R(abba, baab) = 4, \quad d_R(babaabab, aabbbbaa) = 12.$

Theorem 6. *Let $\alpha, \beta \in C$. Then $d_R(\alpha, \beta) = 4k$ where k is the shortest path length in Γ_M between α and β.*

Proof. We will associate to each edge $\alpha - \beta$ from Γ_M an ordered pair on integers (p, q), as following:

$$\alpha = x_1 x_2 \dots x_n,\ \beta = y_1 y_2 \dots y_n \text{ and}$$

$x_i = y_i \qquad \forall i \notin \{p, p+1, q, q+1\}$,
$x_p = x_{q+1} = y_{p+1} = y_q = a$,
$x_{p+1} = x_q = y_p = y_{q+1} = b$.

The order of integers p, q depends on the direction of passing the edge.

So, by example, for $\alpha = abbababa,\ \beta = abbbaaab$, the pair corresponding to the edge $\alpha - \beta$ is $(4, 7)$, and for the edge $\beta - \alpha$ is $(7, 4)$.

The construction of the graph Γ_M assures the constraint $|p - q| \geq 2$.

Let $\alpha = \alpha_0 \longrightarrow \alpha_1 \longrightarrow \dots \longrightarrow \alpha_s = \beta$ be a path in Γ_M of length s, in which all nodes are distinct (otherwise there is a loop which can be avoided). To this path from α to β we will associate a set of pairs $\{(p_1, q_1), (p_2, q_2), \dots (p_s, q_s)\}$ with elements from $\{1, 2, \dots, |\alpha| - 1\}$, where

$$\{p_1, p_2, \dots, p_s\} \cap \{q_1, q_2, \dots, q_s\} = \emptyset.$$

The construction of this set will be inductively obtained as follows:

For $s = 1$ we have the construction defined above ($s = 1$ and the set $\{(p, q)\}$).

Let us consider the construction defined for all paths of lengths at most s and let α, β be two distinct nodes separated by a path of length $s+1$. Let

$$\alpha \longrightarrow \ldots \longrightarrow \alpha_s \longrightarrow \beta$$

be this path, where the first s edges have the set of pairs defined in hypothesis, and the edge $\alpha_s \longrightarrow \beta$ has associated the pair (p, q). Four situations may appear:

(1) $q \notin \{p_1, \ldots, p_s\},\ p \notin \{q_1, \ldots, q_s\}$. Then the set of pairs associated to the path $\alpha_0 \stackrel{*}{\longrightarrow} \beta$ (of length $s+1$) is $\{(p_1, q_1), (p_2, q_2), \ldots (p_s, q_s), (p, q)\}$.

(2) $q \notin \{p_1, \ldots, p_s\},\ p \in \{q_1, \ldots, q_s\}$. Therefore there exists i $(1 \leq i \leq s)$ having $p = q_i$. It results that between the nodes α and β there exists also a path of length s, and the set of pairs associated to it is
$$\{(p_1, q_1), \ldots (p_{i-1}, q_{i-1}), (p_i, q), (p_{i+1}, q_{i+1}), \ldots (p_s, q_s)\}.$$

(3) $q \in \{p_1, \ldots, p_s\},\ p \notin \{q_1, \ldots, q_s\}$. Thus there exists i $(1 \leq i \leq s)$ with $q = p_i$. Like in the previous situation, it results that between the nodes α and β there exists also a path of length s, and the set of pairs associated to it is
$$\{(p_1, q_1), \ldots (p_{i-1}, q_{i-1}), (p, q_i), (p_{i+1}, q_{i+1}), \ldots (p_s, q_s)\}.$$

(4) $q \in \{p_1, \ldots, p_s\},\ p \in \{q_1, \ldots, q_s\}$. Then there exist $i, j \in \{1, \ldots, s\}$ with $q = p_i,\ p = q_j$. We can suppose $i < j$ (the case $i > j$ is similar). Then in the graph Γ_M, will exist also a path of length $s-1$ between α and β, and the set of pairs associated to it is

$$\{(p_1, q_1), \ldots (p_{i-1}, q_{i-1}), (p_{i+1}, q_{i+1}), \ldots (p_{j-1}, q_{j-1}), (p_j, q_i), (p_{j+1}, q_{j+1}), \ldots\}$$

A singular variant is $i = j$. Then the pair (p_i, q_i) is simply avoided from the set and the length of the path having this new set of pairs is $s-1$.

Example 7. *Let us consider the binary sequences*

$$u_1 = abbbababa, \quad u_2 = babbaabba, \quad u_3 = bbabaabab, \quad u_4 = bbaabbaab$$

The set asociated to the path $u_1 \longrightarrow u_2$ *is* $\{(1,6)\}$ *and to* $u_1 \stackrel{*}{\longrightarrow} u_3$ *is* $\{(1,6),(2,8)\}$.

Between u_1 *and* u_4 *there exists the path* $u_1 \longrightarrow u_2 \longrightarrow u_3 \longrightarrow u_4$. *The set associated is* $\{(1,6),(2,8),(6,4)\}$. *Because the integer* 6 *appears once on the first position and once on the second position, this set can be reduced to* $\{(1,4),(2,8)\}$*; that means there exists a shorter path between* u_1 *and* u_4 *(via* $u_5 = bababbaba$*): see Figure* 2.

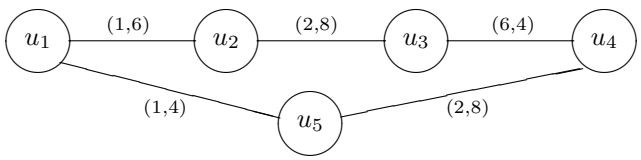

Fig. 2. An example concerning the case (2).

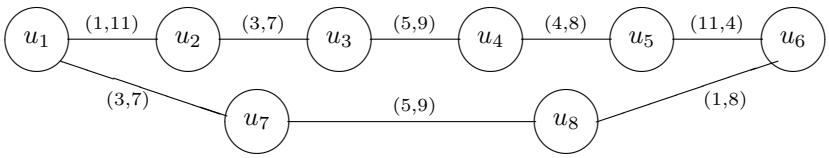

Fig. 3. An example concerning the case (4).

Example 8. *Let us consider the words*

$$u_1 = abababbababa,\ u_2 = baababbabaab,\ u_3 = bababaababab,$$
$$u_4 = babbaaaabbab,\ u_5 = babbaaaabbab,\ u_6 = bababaaabbba$$

Between u_1 *and* u_5 *there exists the path* $u_1 \longrightarrow u_2 \longrightarrow u_3 \longrightarrow u_4 \longrightarrow u_5$ *with the set* $\{(1,11),(3,7),(5,9),(4,8)\}$. *Let us add the edge* $u_5 \longrightarrow u_6$ *having the set* $\{(11,4)\}$. *Because the components of the pairs from the set* $\{(1,11),(3,7),(5,9),(4,8),(11,4)\}$ *are not distinct, there is a shorter path between* u_1 *and* u_6. *Applying the rule* 4, *the new set of pairs is* $\{(3,7),(5,9),(1,8)\}$; *it corresponds to the path* $u_1 \longrightarrow u_7 \longrightarrow u_8 \longrightarrow u_6$ *(Figure* 3*), where* $u_7 = abbaababbaba$ *and* $u_8 = abbabaababba$ *(obviously,* u_7, u_8 *are nodes in* Γ_M*).*

Let us consider now $\alpha, \beta \in C$ ($\alpha \neq \beta$). Because the graph Γ_M is connected, at least a path between α and β will exist. Let $\alpha = \alpha_0 \longrightarrow \alpha_1 \longrightarrow \ldots \longrightarrow \alpha_s = \beta$ be a path obtained according to the rules $1. - 4.$, and $(p_1, q_1)(p_2, q_2)\ldots(p_s, q_s)$ be the set of pairs associated to this path.

The length of this path is minimal. Indeed, if a character $x \in \{a, b\}$ is situated on the position i in α and on the position j ($j \neq i$) in β, then the set of pairs associated using the algorithm above will assure the shift of x always in the same direction (on right if $i < j$, respectively on left if $i > j$). The crossing of one edge in Γ_M assures the shift of x with at most one position (left or right).

We shall prove by induction on s that $d_R(\alpha, \beta) = 4s$.

$s = 1$: then $\alpha = \gamma_1 ab\gamma_2 ba\gamma_3,\ \beta = \gamma_1 ba\gamma_2 ab\gamma_3$ where $\gamma_1, \gamma_2, \gamma_3 \in \{a, b\}^*$; hence, obviously, $d_R(\alpha, \beta) = 4 = 4 \cdot 1$.

Let us consider the equality verified for any pair of binary words connected by a path of minimal length s and let $\alpha, \beta \in C$ be two words connected by a path of minimal length $s + 1$: $\alpha \stackrel{*}{\longrightarrow} \alpha_s \longrightarrow \beta$. Then

$$d_R(\alpha, \beta) = d_R(\alpha, \alpha_s) + d_R(\alpha_s, \beta) = 4k + 4 = 4(k + 1)$$

This is seen as follows. Every character a situated in one of locations $p_1, \ldots p_s, p$ is moved one position to the right; therefore the distance between it and the original a increases by 1; overall, there are $s+1$ moves. The same thing is happening with the characters a situated in positions $q_1 + 1, q_2 + 1, \ldots, q_s + 1, q + 1$ (shifted to the left) and with the characters b situated in positions $p_1 + 1, \ldots, p_s + 1, p + 1$ (shifted to left) and respectively $q_1, q_2, \ldots, q_s, q$ (shifted to right). Every character (of the words situated on this path) stays on its place or is moved in only one direction. Because

$\{p_1, \ldots, p_s\} \cap \{q_1, \ldots, q_s\} = \emptyset$, a character moved from an arbitrary position will not be moved back later on in this position. Therefore the distances from the initial positions are increasing mappings. □

Example 9. *Let us construct a table containing the rank distance between all the words of the class C defined in Example 5.*

Table 3. The rank distance between the words defined in Example 5.

	aabbbbaa	*ababbaba*	*abbaabba*	*baababba*	*baabbaab*	*babaabab*	*abbabaab*	*bbaaaabb*
aabbbbaa	0	4	8	8	8	12	12	16
ababbaba	4	0	4	4	4	8	4	12
abbaabba	8	4	0	4	8	4	4	8
baababba	8	4	4	0	4	4	8	8
baabbaab	8	4	8	4	0	4	4	8
babaabab	12	8	4	4	4	0	4	4
abbabaab	12	4	4	8	4	4	0	8
bbaaaabb	16	12	8	8	8	4	8	0

All these distances verify the Theorem 6.

We close this section with some considerations regarding palindromes.

Lemma 3. *If $p = n$ then $C_{mi(\alpha)} = C_{\overline{\alpha}}$, where $\overline{\alpha}$ is obtained from the word α by replacing the letter a with b and vice-versa.*

Lemma 4. *Let $\alpha, \beta \in \Sigma^*$ be two binary palindromes with the same Parikh vector. Then $\alpha \sim_a \beta$.*

As a result of this lemma, any binary palindrome α with the Parikh vector $(|\alpha|_a, |\alpha|_b)$ fixed, will have the same value for $|\alpha|_{ab}$, therefore the same Parikh matrix.

Remark 4. *From Lemma 4 it results that all palindromes with the same Parikh vector are amiable and thus they are in the same equivalence class C. But the reverse is not true: not all words from C are palindromes. For example, consider $n = p = 4$, $q = 8$. However baababba $\sim_a$ aabbbbaa, although the second word is a palindrome and the first word is not a palindrom. See Example 5. This remark corrects Corollary* 3.4 *(*[1]*).*

4. Conclusion

The paper considers some issues regarding the binary case of the Parikh matrix mapping. In Section 2 we want to give a direct computational proof to some results presented in the literature; also a lower bound for the maximal number of binary amiable words is presented and discussed. Section 3 introduces a graph associated to an equivalence class of amiable words and some properties are proved.

There remain some open problems concerning the extension of the results to alphabets with more than two letters. The characterisation theorem offers more information than mentioned here. A separate investigation of properties of amiable words using the graph will be necessary.

Acknowledgment

The author wishes to thank Dr. Arto Salomaa for his helpful comments an earlier version of this paper.

References

[1] A. Atanasiu, C. Martin - Vide, Al. Mateescu - *On the injectivity of Parikh matrix mapping*, Fundamenta Informaticae 49 (2001), 166-180.
[2] A. Atanasiu, C. Martin - Vide, Al. Mateescu - *Codifiable languages and Parikh matrix mapping*, Journal of Universal Computer Science, vol. 7, nr. 9 (2001), 783-793.
[3] L.P. Dinu, A. Sgarro - *A low complexity distance for DNA strings*, manuscript.
[4] S. Fosse, G. Richmomme - *Some characterisations of Parikh matrix equivalent binary words*, Inf. Processing Letters Vol. 92(2), 77-82 (2004).
[5] Al. Mateescu, A. Salomaa - *Matrix indicators for subword occurences and ambiguity*, Int. J. Found. Comput. Sci. 15 (2004), 277-292.
[6] Al. Mateescu, A. Salomaa, K. Salomaa, S. Yu - *On the extension of the Parikh mapping*, Theoret. Informatics Appl. 35 (2001), 551-564.
[7] R.J. Parikh - *On context-free languages*, J. Assoc. Comput. Mach., 13 (1966), 570-581.
[8] G. Rozenberg, A. Salomaa (eds) - *Handbook of formal languages*, Springer, Berlin, 1997.
[9] A. Salomaa - *Independence of certain quantities indicating subword occurences*, Theoretical Computer Science (2006), to appear.