

## NOTE ON OCCURRENCES OF FACTORS IN CIRCULAR WORDS

PIERRE ARNOUX<sup>1</sup>

**Abstract.** We give an elementary proof of a property discovered by Xavier Grandsart: let  $W$  be a circular binary word; then the differences in the number of occurrences  $|W|_{0011} - |W|_{1100}$ ,  $|W|_{1101} - |W|_{1011}$ ,  $|W|_{1010} - |W|_{0101}$  and  $|W|_{0100} - |W|_{0010}$  are equal; this property is easily generalized using the De Bruijn graph.

**Mathematics Subject Classification.** 68R15.

Binary words are finite sequences  $W = W_0W_1 \dots W_{n-1}$  taking values in the alphabet  $\{0, 1\}$ ;  $n$  is the *length* of the word  $W$ , denoted by  $|W|$ . We say that the word  $U = U_0 \dots U_{k-1}$  of length  $k \leq n$  *occurs* in  $W$  at position  $i$  if  $U_j = W_{i+j}$  for  $0 \leq j \leq k-1$ , and in that case we say that  $i$  is an *occurrence* of  $U$  in  $W$ , and that  $U$  is a *factor* of  $W$ .

To avoid special effects due to the extremities of the word  $W$ , we will consider *circular words* of length  $n$ , that is, words indexed by the cyclic group of integers mod  $n$ . This allows occurrences at the end of the word; for example, the circular word  $W = 0001$  admits the factor  $010$ , with an occurrence at index 2.

**Remark 1.1.** One can equivalently define a circular word as an infinite word which is periodic, of period  $W$ , and in that case we consider occurrences as defined modulo  $|W|$ . If we identify two circular words which differ only in a shift of index, as  $001$  and  $010$ , we can also define a circular word as a conjugacy class of words, where two words  $W, W'$  are conjugate if there exist two words  $U, V$  such that  $W = UV$  and  $W' = VU$ . This is the usual definition in combinatorics of words.

We consider the number of occurrences of a factor  $U$  in a circular word  $W$ , denoted by  $|W|_U$ . For example, if  $W = 00101$ ,  $|W|_{010} = 2$ , since  $010$  occurs in position 1 and 3 in  $W$ . For a random circular word  $W$  of large length and a fixed factor  $U$ , there is of course a probabilistic aspect to  $|W|_U$ , but there are also combinatorial relations. For example, it is easy to check that  $|W|_{001} = |W|_{100}$ ; these two numbers are also the number of *runs* (sequences of maximal length) of  $0$  of length at least 2, since each such sequence starts with  $100$  and ends with  $001$ .

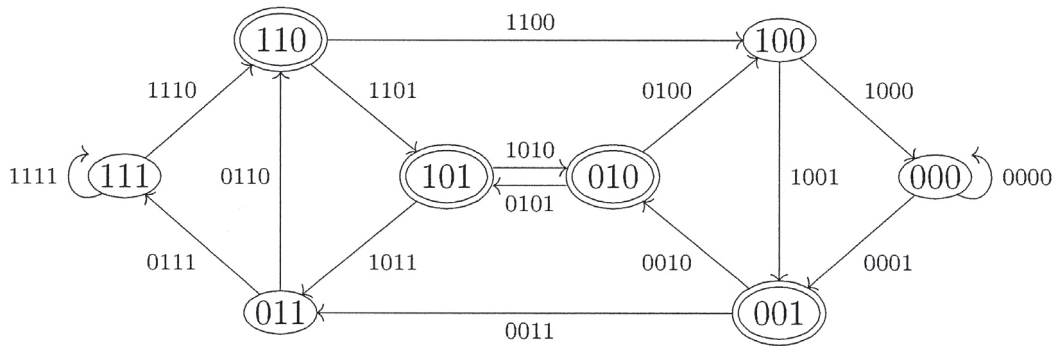
Around 2010, Xavier Grandsart, based on numerical experimentations, generalized this remark; he conjectured and proposed as a challenge the following proposition:

**Proposition 1.2.** *Let  $W$  be a circular binary word. The differences in the number of occurrences  $|W|_{0011} - |W|_{1100}$ ,  $|W|_{1101} - |W|_{1011}$ ,  $|W|_{1010} - |W|_{0101}$  and  $|W|_{0100} - |W|_{0010}$  are equal.*

---

*Keywords and phrases.* Circular words, factors, De Bruijn graphs.

<sup>1</sup> Institute of mathematics, CNRS UPR 9016, Marseille, France. [arnoux@iml.univ-mrs.fr](mailto:arnoux@iml.univ-mrs.fr)

FIGURE 1. The De Bruijn graph  $B(2, 3)$ .

The challenge was quickly solved by 3 persons, Maher Younan [5] who proposed a proof similar to the proof given here, Alberto Costa [3] and Pierre Deligne [4], who proposed proofs by induction. We give here a proof based on De Bruijn graphs. While the argument is elementary, it seems that, curiously, nobody had noticed this property before, and up to my knowledge no proof is publicly available, hence it can be useful to publish a self-contained proof of this nice exercise and to show how it can be generalized to slightly less obvious results.

We start with some remarks. Recall that a palindrome is a word  $W$  which is equal to its mirror image, that is, if  $n$  is the length of  $W$ ,  $W_i = W_{n-1-i}$  for all  $i < n$ . A palindromic pair is a pair  $(U, V)$  of words of same length  $n$  which are not palindromes and which are mirror image of each other.

There are 16 binary words of length 4; 4 of them, 1111, 1001, 0110 and 0000 are palindromes. 4 of them contain a run of length 3, and form two palindromic pairs whose two elements have same number of occurrences, by the same proof as above: (1000, 0001) and (1110, 0111). The remaining 8 words form the 4 palindromic pairs considered by Xavier Grandsart: (1010, 0101), (0010, 0100), (1011, 1101), (0011, 1100). These 8 words have the common property that their prefix of length 3 ends with two different letters: 001, 010, 101, 110.

We will use the DeBruijn graph on binary words, and we recall the definition:

**Definition 1.3.** The De Bruijn graph  $B(2, n)$  is the oriented graph whose vertices are the binary words of length  $n$ , and whose edges are the binary words of length  $n + 1$ ; the initial vertex of the edge  $U$  is the prefix of length  $n$  of  $U$ , and the final vertex is the suffix of length  $n$  of  $U$ .

Figure 1 shows the graph  $B(2, 3)$ , were the initial vertices of the edges labeled by the words considered by Grandsart are surrounded by double circles. Each circular word of length  $n$  defines a closed path in the De Bruijn graph  $B(2, 3)$ ; the sequence of consecutive vertices of this path is the sequence of factors of length 3 in their order of occurrence, and the sequence of consecutive edges is the sequence of factors of length 4.

*First proof of the proposition.* Each vertex defines a relation in the number of occurrences: If  $W$  is a circular binary word, and  $U$  is a factor of  $W$  of length 3, then an occurrence of  $U$  is also an occurrence of some  $aU$  and some  $Ub$ , with  $a, b \in \{0, 1\}$ , hence  $|W|_U = |W|_{U0} + |W|_{U1} = |W|_{0U} + |W|_{1U}$ : the number of occurrences of factors of length 4 satisfy Kirchhoff's law on the De Bruijn graph.

With  $U = 101$ , we get  $|W|_{101} = |W|_{1011} + |W|_{1010} = |W|_{0101} + |W|_{1101}$ , and with  $U = 010$ , we have  $|W|_{1010} + |W|_{0010} = |W|_{0100} + |W|_{0101}$ ; these are equivalent forms for the relations  $|W|_{1101} - |W|_{1011} = |W|_{1010} - |W|_{0101}$  and  $|W|_{1010} - |W|_{0101} = |W|_{0100} - |W|_{0010}$ .

Taking  $U = 000$ , we obtain  $|W|_{0000} + |W|_{0001} = |W|_{0000} + |W|_{1000}$ , hence  $|W|_{0001} = |W|_{1000}$  as we saw above. Using vertex 100 (resp. 001), we obtain  $|W|_{0100} + |W|_{1100} = |W|_{1000} + |W|_{1001}$  (resp.  $|W|_{0010} + |W|_{0011} = |W|_{1001} + |W|_{0001}$ ); hence  $|W|_{0100} + |W|_{1100} = |W|_{0010} + |W|_{0011}$  (This quantity is the number of runs of 0 of length at least 2). Hence we have  $|W|_{0100} - |W|_{0010} = |W|_{0011} - |W|_{1100}$ .  $\square$

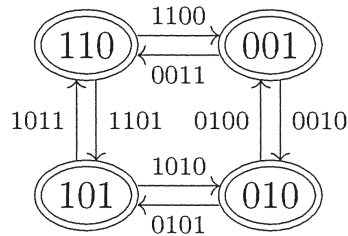


FIGURE 2. The shortened graph.

We sketch another proof which gives an interpretation as a rotation number for the differences in number occurrences that appear in the proposition.

*Second proof of the proposition.* Starting from vertex 110, we can either go directly to 101 through edge 1101, or to 100 through 1100, then  $n \geq 0$  times in 000, then to 001; a similar remark can be made for the other three vertices 101, 010, 001, which are surrounded by double circles in the picture; each of them can only lead to two other vertices.

Hence if we erase, in the sequence of edges labelling the path of a circular word, the factors which are not in the 4 palindromic pairs considered by Grandsart, the remaining sequence labels a path in the smaller graph shown in Figure 2.

Between each pair of adjacent vertices on this graph, there are two opposite edges labelled by a palindromic pair. The graph is a cyclic graph; let us give this graph an orientation, so that edges labelled 0011, 1101, 1010 and 0100 correspond to a quarter turn in the positive direction, and the four other edges go in the negative direction. A closed path on this graph makes a number  $k$  of turns, counted with orientation, and for any pair of edges between adjacent vertices, the number of occurrences of the positive edge minus the number of occurrences of the negative edge is equal to  $k$ , hence it does not depend on the edge. This proves the proposition.  $\square$

There is a simple way to compute the number  $k$ . Any letter  $a$  occurs in a circular word either as an isolated letter  $\dots bab \dots$  or as a run of length at least 2  $\dots aa \dots$ . We can decompose a circular word in maximal sequences of isolated letters  $\dots ababab \dots$  and sequences of runs of length at least 2; the number  $k$  is equal to the difference between the number of maximal sequences of even length of isolated letters starting with 0, and the number of maximal sequences of even length of isolated letters starting with 1. In particular, for  $W = 010011$ , we have  $k = 1$ , and for  $W = 101100$ , we have  $k = -1$ , as we can check on the graph; this allows us to build shortest circular words for which the difference is maximal, by taking powers.

Julien Cassaigne [1] has pointed to me that this result can be widely generalized. Consider an alphabet with  $d$  letters, and the functions  $W \mapsto |W|_U$ , where  $U$  is any word of length at most  $l$  on this alphabet. There are  $\sum_{k=0}^l d^k$  such functions; they are obviously not independent, since  $|W|_U = |W|_{U0} + |W|_{U1}$ , hence they are all generated by the  $d^l$  functions defined by words of length  $l$ ; one proves that the space generated by these functions has in fact dimension  $(d-1)d^{l-1} + 1$ . More precisely, it can be deduced from [2] that these functions can also be computed from the  $W \mapsto |W|_U$ , where  $U$  are the words of length at most  $l$  whose first and last letter is not 0.

Remark that the function  $W \mapsto |W|_U$ , for  $|U| < l$ , is easily computed from the functions  $W \mapsto |W|_V$ , for all words  $V$  of length  $l$ , since it is the sum of all the occurrences of the words of length  $l$  which admit  $U$  as prefix. Hence we can state the previous result in this way:

**Proposition 1.4.** *Let  $E$  be the set of all circular words on an alphabet of cardinal  $d$ . For any word  $U$  of length  $l$  on this alphabet, consider the function  $E \rightarrow \mathbb{N}$ ,  $W \mapsto |W|_U$ ; these functions satisfy exactly  $d^{l-1} - 1$  independent linear relations.*

*Proof.* These functions must satisfy the Kirchhoff equations for the graph  $B(d, l-1)$ , which has  $d^{l-1}$  vertices and  $d^l$  edges. The numbers of occurrences of edges in a given circular word define an admissible flow on the graph with integral values, and conversely any admissible flow with integral values and connected support defines a circular word. The dimension of the space of admissible flows is the cyclomatic number of the graph  $d^l - d^{l-1} + 1$ : there are  $d^{l-1}$  vertices, each of which gives a relation between incoming and outgoing edges; the sum of all these relations is trivial, and since the graph is connected, this is the only equation between these relations.  $\square$

In our case, the Kirchhoff relations on the De Bruijn graph imply that the 16 functions  $W \mapsto |W|_U$ , where  $U$  is a binary word of length 4, satisfy 7 linear relations. Indeed, the 9 functions  $W \mapsto |W|_U$ , with  $U = 0000$  or  $U = 1111$  are independent (the corresponding words, if we remove the loops 0000 and 1111, generate a spanning tree for the graph), and the seven remaining functions, for  $U$  starting with 0 and different from 0000, can be generated from the other 9.

## REFERENCES

- [1] J. Cassaigne, *Private communication*.
- [2] J. Cassaigne, J. Karhumäki and A. Saarela, On growth and fluctuation of  $k$ -abelian complexity, *Proc. of the 10th CSR*. In Vol. 9139 of *Lect. Notes Comput. Sci.* Springer (2015) 109–122.
- [3] A. Costa, Solution for the problem of the game heads or tails. Available at <http://arxiv.org/abs/1012.0508> (2010).
- [4] P. Deligne, *letter to Xavier Grandsart*.
- [5] M. Younan, *Letter to Xavier Grandsart*.

Communicated by Ch. Choffrut.

Received June 12, 2016. Accepted September 5, 2016.