On the Maximum Number of Distinct Factors of a Binary String

Jeffrey Shallit

Research Report CS-91-31
July 1991

Department of Computer Science
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1

Faculty of Mathematics

University of Waterloo
Waterloo, Ontario, Canada
N2L 3G1

On the Maximum Number of Distinct Factors of a Binary String

Michel Mendès France
U. E. R. Mathématiques et Informatique
Université de Bordeaux
351, cours de la Libération
33405 Talence Cedex
France

Jeffrey Shallit §
Department of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada
`shallit@graceland.waterloo.edu`

Abstract.
In this note we prove that a binary string of length n can have no more than $2^{k+1}-1+\binom{n-k+1}{2}$ distinct factors, where k is the unique integer such that $2^k+k-1 \leq n < 2^{k+1}+k$. Furthermore, we show that for each n, this bound is actually achieved. The proof uses properties of the de Bruijn graph.

§ Research supported in part by an NSERC operating grant.

I. Introduction.

Let w be a string of 0's and 1's, i.e. $w \in (0+1)^*$. We say that $z \in (0+1)^*$ is a *factor* of w if there exist $x, y \in (0+1)^*$ such that

$$w = xzy.$$

In analogy with the function that counts the number of divisors of a positive integer n, define $d(w)$ to be the *total number of distinct factors* of the string w. For example, $d(10110) = 12$, as its set of factors is given by

$$\{\epsilon, 0, 1, 01, 10, 11, 011, 101, 110, 0110, 1011, 10110\}.$$

Note that we count ϵ, the empty string, as a factor of every string.

In this note we discuss the maximum order of $d(w)$.

II. The Main Results.

Theorem 1. *Let* $|w| = n$. *Then*

$$\begin{aligned} d(w) &\le \sum_{0 \le i \le n} \min(2^i, n-i+1) \\ &= \binom{n-k+1}{2} + 2^{k+1} - 1, \end{aligned}$$

where k is the unique integer such that $2^k + k - 1 \le n < 2^{k+1} + k$.

Proof.

The first inequality is clear, as there are precisely $n - i + 1$ possible factors of length i, of which at most 2^i can be distinct.

To see the second equality, note that if $2^k + k - 1 \le n < 2^{k+1} + k$, then $2^k \le n - k + 1$ and $2^{k+1} > n - k$. Hence

$$\begin{aligned} \sum_{0 \le i \le n} \min(2^i, n-i+1) &= \sum_{0 \le i \le k} 2^i + \sum_{k < i \le n} (n-i+1) \\ &= 2^{k+1} - 1 + \binom{n-k+1}{2}. \end{aligned}$$

This completes the proof. ■

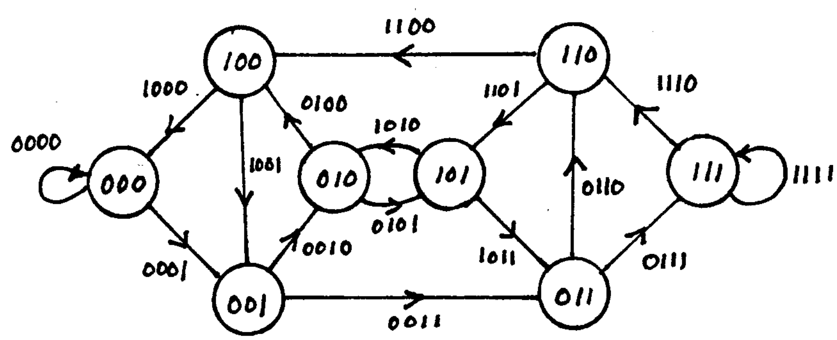
Theorem 2. *The upper bound in Theorem 1 is actually attained for all* n.

To prove Theorem 2, we use the *de Bruijn graph* B_k. This graph was apparently first studied by Flye-Sainte Marie in 1894 [FSM]. Good [G] and de Bruijn [B] independently rediscovered the graph in 1946. A more accessible reference is Bondy and Murty [BM, pp.

181-183] or van Lint [L, pp. 82-92]. For a survey of results on this graph until 1982, see Fredricksen [F].

Recall that B_k is a directed graph with 2^k vertices $\{0,1\}^k$, and 2^{k+1} directed edges with labels $\{0,1\}^{k+1}$. There is a directed edge from the head vertex, labeled $a_1a_2\cdots a_k$, to the tail vertex, labeled $b_1b_2\cdots b_k$, iff $a_2\cdots a_k = b_1\cdots b_{k-1}$. In this case the edge is labeled $a_1a_2\cdots a_kb_k$.

For example, below is the de Bruijn graph B_3:



A *chain* is an alternating sequence of distinct edges and possibly non-distinct vertices, $v_1, e_2, v_2, \ldots, e_j, v_j$, where v_i, $2 \le i \le j$, is the tail of e_i and v_i, $1 \le i \le j-1$, is the head of e_{i+1}. If $v_1 = v_j$, this is a *closed chain.* A closed chain with distinct vertices (other than $v_1 = v_j$) is a *cycle.* The *length* of a chain is the number of edges it contains.

We need the following lemma:

Lemma 3.

For each i with $2^k \le i \le 2^{k+1}$, the graph G_k contains a closed chain of length k that visits every vertex at least once.

Note that for $i = 2^k$, this is a Hamiltonian cycle, and for $i = 2^{k+1}$, this is an Eulerian tour.

Proof.

This theorem can be derived from results in a paper of Yoeli [Y], although it is not explicitly stated there.

Yoeli proved the following theorems:

Theorem A.

If G_k has a cycle of length i, then it has a closed chain of length $i + 2^k$.

Theorem B.

G_k contains a cycle of length i for any i, $0 < i \le 2^k$.

Combining these two theorems, we see that G_k has a closed chain of any length between 2^k and 2^{k+1}. However, it remains to see there exists such a chain that visits every vertex of G_k. Yoeli's proof of Theorem A does in fact construct a closed chain that visits every vertex of G_k. Since this is nowhere stated in his paper, we briefly go through the argument.

Yoeli proves the following three lemmas:

Lemma 4. *G_k is strongly connected.*

Define a *P-set* of cycles of G_k to be a set of vertex-disjoint cycles covering all the vertices. (Each cycle must have at least one edge; thus a P-set of G_k has 2^k edges.)

Lemma 5.

Let C be a cycle of G_k. Then there exists a P-set of cycles of G_k including no edge of C.

Lemma 6.

Let C' and C'' be vertex-disjoint cycles of G_k and let $e = (u, v)$ be an edge with u in C' and v in C''. Then there is an edge e' from v's predecessor in C'' to u's predecessor in C', and a cycle on the vertex set of $C' \cup C''$ can be formed using edges of $C' \cup C''$ together with e and e'.

Now we can complete the proof of Lemma 3, following the proof Yoeli gave for his Theorem A.

Let C be a cycle in G_k of length i. By Lemma 5 there exists a P-set of cycles P_1 of G_k including no edge of C. Let H_1 be the subgraph of G_k formed by the edges of P_1 and C. If the underlying undirected graph of H_1 consists of more than one connected component, then by Lemma 4 there must be an edge e in G_k joining two components of H_1. Edge e must join two vertex disjoint cycles D' and D'' in P_1, where no edge of H_1 goes between D' and D''. Applying Lemma 6 to combine D' and D'', we obtain a P-set of cycles P_2 including no edge of C, and such that $H_2 = C \cup P_2$ has one fewer connected component. Continuing in this fashion leads to a connected subgraph H_r, consisting of $C \cup P_r$, where P_r is a P-set. Since H_r is connected, with each vertex's in-degree equal to its out-degree, H_r has an Eulerian tour. This provides a closed chain of length $2^k + i$ visiting all vertices. ■

Using Yoeli's result we can construct a string that achieves the upper bound:

Proof of Theorem 2.

Let n be given, and let k be the unique integer such that $2^k + k - 1 \le n < 2^{k+1} + k$. Consider the de Bruijn graph B_k. By Lemma 3 there exists a closed chain C of length $n - (k - 1)$ traversing each vertex in B_k and repeating no edges. Take the string formed by the k letters of the vertex label of the first vertex in C, followed by the last letter in the labels of all subsequent edges in C. The result is a string of length n, and we claim it is the desired one.

Now this closed chain visits every vertex of B_k; hence w contains all factors of length k, and hence all factors of lengths $0, 1, 2, \ldots k - 1$.

On the other hand, the chain C does not repeat any edge, so all the factors of length $k+1$ are distinct. Hence so are all the factors of lengths $k+2, k+3, \ldots, n$, since any two factors of the same length must differ in the first $k+1$ positions.

Thus we see

$$d(w) = \sum_{0 \le i \le k} 2^i + \sum_{k < i \le n} n - (k+1),$$

and so the upper bound is achieved. ■

An Example.

Let $n = 14$. Then $k = 3$ and $n - (k-1) = 12$. Looking at B_3, we see there is a closed chain of length 12, as follows (listing only the vertices):

$$000 \to 001 \to 010 \to 100 \to 001 \to 011 \to 110 \to$$

$$101 \to 011 \to 111 \to 110 \to 100 \to 000.$$

This corresponds to the string 000100110111000 of length 14. It has $15 + 66 = 81$ distinct factors, which is the maximum possible for any binary string of length 14.

III. Acknowledgments.

We are grateful to A. Rosenberg for suggesting the article of Yoeli, and to T. Leighton for suggesting we speak to A. Rosenberg.

We would like to express our thanks to A. Lubiw, who provided the proof of Lemma 3.

References

[B] N. G. de Bruijn, A combinatorial problem, *Nederl. Akad. Wetensch. Proc.* **49** (1946), 758–764.

[BM] J. A. Bondy and U. S. R. Murty, Graph Theory with Applications, Macmillan, 1976.

[F] H. Fredricksen, A survey of full length nonlinear shift register cycle algorithms, *SIAM Review* **24** (1982), 195–221.

[FSM] C. Flye-Sainte Marie, Solution to problem number 58, *L'Intermédiaire des Mathématiciens* **1** (1894), 107–110.

[G] I. J. Good, Normally recurring decimals, *J. London Math. Soc.* **21** (1946), 167–169.

[L] J. H. van Lint, Combinatorial Theory Seminar, Eindhoven University of Technology, *Lecture Notes in Mathematics # 382*, Springer-Verlag, 1974.

[Y] M. Yoeli, Binary ring sequences, *Amer. Math. Monthly* **69** (1962), 852–855.