Combinatorial Structures in the de Bruijn Graph

L. J. Cummings
University of Waterloo

Dedicated to Anne Penfold Street.

Abstract

Anne Street wrote an expository article about de Bruijn graphs in the 1970's. We review some subsequent lines of research, at least one of which was inspired by her article.

1 Introduction

At the first Australian Conference on Combinatorial Mathematics held in 1972 at the University of Newcastle, New South Wales, Anne Street gave an an expository talk on graph theory and the application of Eulerian cycles to the design of circular switches, such as those found in automatic washing machines. A brief summary of this talk appeared in the proceedings of the Second Australian Conference with the amusing title "Eulerian Washing Machines", [21].

This survey was subsequently reviewed [8] by N. G. de Bruijn whose name is most frequently associated with these graphs. Indeed, for many years he was credited with the first construction and enumeration of these graphs, [7]. However, de Bruijn acknowledged in his review the priority of C. Flye Sainte-Marie [20] who published in 1894. de Bruijn had rediscovered Sainte-Marie's result that the number of circular arrangments of 2^n 0's and 1's with the property that every binary string of length n appears once and only once as a substring is given by $2^{2^{n-1}-n}$. For details see [7].

2 Definitions and Notation

The binary *de Bruijn graph* of order n, $B_n = B_n(2)$, is the directed graph with vertex set $\{0,1\}^n$ and edges between vertices $x = a_1 \cdots a_n$ and $y =$

$b_1 \cdots b_n$ precisely when $a_2 \cdots a_n = b_1 \cdots b_{n-1}$. The edge of B_n between vertex x and vertex y can be labeled $a_1 \cdots a_n b_n$ or, equivalently, $a_1 b_1 \cdots b_n$. Since B_n has 2^n vertices and every binary string of length 2^{n+1} is the label of some pair of vertices, B_n has 2^{n+1} edges. Similiarly a de Bruijn graph of each order $n > 1$ can be defined for every finite alphabet. A 2-dimensional version of the de Bruijn graph is considered in [14]. The following properties of de Bruijn graphs over arbitrary finite alphabets are easy to establish.

Properties 1 *All vertices of B_n have in-degree and out-degree equal to the cardinality of the alphabet, A. That is, B_n is a regular digraph of degree $2|A|$.*

Here loops at each vertex $(a, a, \ldots, a), a \in A$ are counted.

Properties 2 *If w, x, y, z are any vertices of B_n and $(w, x), (w, y)$ are edges then, for any vertex z such that (z, x) is an edge, (z, y) is also an edge.*

Properties 3 *If $(x, y), (y, z)$ are edges in B_n with labels α, β respectively then (α, β) is an edge of G_{n+1}.*

Definition 1 *An (n, a) de Bruijn sequence is a circular string with the property that all strings of length n over an alphabet of size a occur as substrings exactly once.*

Over the alphabet $\{0, 1, 2\}$, for example,

22221222022112210220122002121202111211021012100202011201020012000111101100101000 0

is a de Bruijn sequence which contains all strings of length 4 exactly once when viewed as a circular string.

An (n, a) de Bruijn sequence corresponds to an Eulerian cycle of B_{n-1}. For example, 00011101 is a $(3, 2)$ de Bruijn sequence which corresponds to the Eulerian cycle

$$000, 001, 011, 111, 110, 101, 010, 100$$

of edges in B_2. Conversely, we can view an Eulerian cycle of B_n as a circular string. There are at least three different ways of defining circular strings or "necklaces" in the literature:

- a sequence of letters written in a circle [1]
- a doubly infinite periodic sequence [5]
- an equivalence class of strings under circular shifts (i.e., an orbit of the action of the cyclic group of order n on the set of n-strings) [4].

3 Golomb's conjecture

By a *factor* of B_n we mean a subgraph determined by a set of vertex-disjoint directed circuits (cycles) which contain all vertices of B_n. An *extremal factor* is a factor which contains a maximal number of cycles. In the binary case Solomon Golomb had conjectured [11] that the number of cycles in any extremal factor of B_n is

$$Z(n) = \frac{1}{n} \sum_{d|n} \phi(d) 2^{\frac{n}{d}}$$

Here ϕ is the Euler function and the summation is over all positive divisors d of n.

A proof of Golomb's conjecture was given by Mykkeltveit [19] in 1972. Mykkeltveit's proof was obtained by establishing another conjecture due to Lempel [15] which in turn implied Golomb's conjecture. Lempel's conjecture stated that $Z(n)$ is the mimimal number of vertices which if removed from B_n would leave a directed graph with no cycles.

Quite generally one may ask whether there exists a circular string of length $k \leq m^n$ chosen from the alphabet $\{0, 1, \ldots, m\}$ with the property that its substrings of length k are all distinct. This was resolved in the affirmative by Abraham Lempel [15] in 1971.

4 Embedded de Bruijn Sequences

It is a natural question to ask whether one Eulerian tour of the de Bruijn graph can be found as a subsequence of an Eulerian tour of a higher order de Bruijn graph over the same alphabet or a de Bruijn graph of the same order but over a larger alphabet. As shown in [5] the answer is easy in the latter case because $B_n(p)$ is always an induced subgraph of $B_n(p+1)$. For example, the de Bruijn sequence

$$00110\ 21220\ 3132330\ 414243440\cdots$$

is obtained by iteratively embedding sequences of orders $(n, 2), (n, 3), (n, 4)$, ... The "break points" have been indicated by inserting spaces. In [5] the Euler tours of the de Bruijn graphs and its subgraphs were viewed as periodic infinite sequences, but they can equally be viewed as circular strings.

5 Codes in the de Bruijn graph

It is well-known that the parity check matrix of a Hamming code of length $n = 2^k - 1, k \geq 2$, has as columns all non-zero binary strings of length k appearing exactly once. [17, p 23]. The order of the columns is not important since any two different orderings lead to equivalent codes. Accordingly, deleting $0^n = 0 \cdots 0$ from any Euler tour of B_n leaves a Hamilton path containing all the columns of a generator matrix of the Hamming code. Stated differently, we can claim that the all information required to generate a Hamming code can be stored as a de Bruijn sequence which has been cut to form a linear string with 0^n omitted.

The extended Golay code has as generator matrix a 12×24 matrix which has the form [I,A] bordered by a first column of the form $1^{11}0$ and last row of the form $0^{12}1^{12}$. Here, I is the 11×11 identity matrix and A is a 11×11 matrix obtained from a Hadamard matrix of Paley type. In fact, A is a cyclic matrix so each row is obtained from the preceding one by a cyclic shift. In addition, the first row is a cylic shift of the last one [17, p 65]; thus, the matrix A is equivalent to a 'pure' cycle of length 11 in B_{11}. The matrix A contains the "informational content" of the extended Golay code. Indeed, any cyclic code, such as the BCH codes can be represented in this manner. The importance of this particular representation of these well-known error-correcting codes is not clear, but the de Bruijn graph is valuable in the study of synchronizable codes.

A great deal of research has gone into correcting bit errors in the study of error-correcting codes. Another kind of error is a mis-framing error. To correct mis-framing errors a class of codes known as synchronization codes have been developed. One class of synchronizable codes are the comma-free codes.

As its name implies, a comma-free code is a set of codewords which do not require a "comma" to establish synchronization; that is,

Definition 2 *A set of strings C over an alphabet A is a comma-free code of length n if all strings have length n and*

$$a_1 \cdots a_n, b_1 \cdots b_n \in C$$

implies each of the "overlaps" of length n

$$a_2 \cdots b_1, a_3 \cdots b_1 b_2, \ldots, a_n b_1 \cdots b_{n-1} \notin C$$

In [2] a theorem of Golomb, Gordon, and Welch [13] about comma-free codes is interpreted as a statement about edges in $B_n(|A|)$.

Theorem 1 *If $n \geq 4$ is odd then a collection, C of edges in $B_n(|A|)$ corresponding to the words of a maximum comma-free code is a bipartite subgraph of $B_n(|A|)$.*

A proof of theorem 1 is found in [3].

Another synchronizable code is Λ_n, the set of Lyndon strings of fixed length n. These are the aperiodic strings of length n which are lexicographically least in their equivalence classes under the relation of cyclic permutation. Golomb and Gordon [12] showed that Λ_n, while not comma-free is a bounded synchronization delay code over an arbitrary finite alphabet A. Thus, there is a fixed integer M such that after M letters have been read from a message encoded with Λ_n synchronization can be established. Although bounds exist, the exact value of M given n and A has never been determined. It has been shown in [3] that the Lyndon strings for any fixed n viewed as edges in $B_{n-1}(2)$ are a collection of disjoint paths. A counterexample was given in [3] to show that this is not true when $|A| > 2$. The structure of Λ_n for higher order alphabets remains an open question.

References

[1] J. Berstel and D. Perrin, Codes circulaires, *Combinatorics on Words, Progress and Perspectives*, ed. L. Cummings, Academic Press, 1983.

[2] L. Cummings, Comma-free codes in the de Bruijn Graph, *Caribb. J. Math.* **3**(1983), 65- 68.

[3] L. Cummings, Synchronizable codes in the de Bruijn Graph, *Ars Combinatoria* **19**(1985), 73-80.

[4] L. Cummings, Aspects of synchronizable coding, *JCMCC*, **1**(1987), 67-84.

[5] L. Cummings and D. Wiedemann, Embedded de Bruijn sequences, *Congressus Numeratium*, **53**(1986), 155-160.

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

[7] N. G. de Bruijn, Acknowledgement of priority to C. Flye Sainte-Marie on the counting of circular arrangements of 2^n zeros and ones that show each n-letter word exactly once, *T.H.-Report 75-WSK-06* (Technological University Eindhoven, 1975).

[8] N. G. de Bruijn, Review of Eulerian washing machines by Anne Penfold Street. *Mathematical Reviews* #12538, Vol. 51.

[9] H. Fredricksen, Survey of full length nonlinear shift register cycle algorithms, *SIAM Review*, **24**(1982), 195-221.

[10] H. Fredricksen, A new look at the de Bruijn graph, *Discrete Applied Mathematics*, **37/38**(1992), 193–203.

[11] S. Golomb, *Shift Register Sequences* (Holden-Day, 1967).

[12] S. Golomb and B. Gordon, Codes with bounded synchronization delay, *Information and Control*, **8**(1965), 355–372.

[13] S. Golomb, B. Gordon, and L. Welch, Comma-free codes, *Canadian J. Math.*, **10**(1958), 202–209.

[14] G. Hurlbert and G. Isaak, On the de Bruijn torus problem, *SIAM Review*, **24**(1982), 195–221.

[15] A. Lempel, m-ary closed sequences, *J. Combinatorial Theory* **11**(1971), 17–27.

[16] A. Lempel, On extremal factors of the de Bruijn Graph, *J. Combinatorial Theory* **11**(1971), 17–27.

[17] F.J. MacWilliams and N.J.A. Sloane, *The Theory of Error-Correcting Codes* (North Holland, 1978).

[18] M.H. Martin, A problem in arrangements, *Bull. Amer. Math. Soc.* **40**(1934), 859–864.

[19] J. Mykkeltveit, A proof of Golomb's conjecture for the de Bruijn graph, *J. Combinatorial Theory* **13B**(1972), 40–45.

[20] C. Sainte-Marie, Question 48, *Intermédiaire des Mathématiciens, Rev. Semestrielle Publ. Math.* **1**(1894), 107–110.

[21] A. P. Street, Eulerian washing machines, *Lecture Notes in Mathematics, Vol. 403* (Springer-Verlag, 1973), 105–108.