Comma-free Bipartite Subgraphs of the De Bruijn Graph

L.J. Cummings
Faculty of Mathematics
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1

In Memoriam: Andrew Ball 1954-2000

Abstract

By definition the vertices of a de Bruijn graph are all strings of length $n-1, (n>1)$, over a fixed finite alphabet. The edges are all strings of length n over the same alphabet. The directed edge $a_1 \cdots a_n$ joins vertex $a_1 \cdots a_{n-1}$ to vertex $a_2 \cdots a_n$. A block code over an alphabet of σ elements is comma-free if it does not contain any overlap of codewords. Representing the codewords of comma-free codes as directed edges of the de Bruijn graph, we give sufficient conditions that a bipartite subgraph of the de Bruijn graph whose underlying undirected graph is connected is a comma-free code.

1 Introduction

The de Bruijn graph, $B_n(\sigma)$, $(n>1)$ is the directed graph whose vertices are the σ^{n-1} strings $\mathbf{a} = a_1 \ldots a_{n-1}$ of length $n-1$ with entries from an arbitrary finite alphabet Σ of cardinality σ. It is often convienient to take $\Sigma = \{0, \cdots, \sigma-1\}$. We will always assume that $\sigma > 1$. There is a directed edge from vertex $\mathbf{a} = a_1 \cdots a_{n-1}$ to vertex $\mathbf{b} = b_1 \cdots b_{n-1}$ in $B_n(\sigma)$ precisely when $a_2 \cdots a_{n-1} = b_1 \cdots b_{n-2}$. The edge is labelled by $a_1 \cdots a_{n-1} b_{n-1}$ or equivalently by $a_1 b_1 \cdots b_{n-1}$. The de Bruijn graph is clearly regular with indegree and outdegree equal to σ at every vertex. It contains σ^n directed edges including σ loops, one at each vertex $a^{n-1} = a \cdots a$, $a \in \Sigma$. Many authors prefer to define this graph by taking as vertices the strings of length n and the edges as the strings of length $n+1$, but the above definition is more convenient here.

Any code with block length n over the finite alphabet Σ may be viewed as a set of edges in $B_n(\sigma)$ or, equivalently, as a set of vertices in $B_{n+1}(\sigma)$. It will be convenient to use the first representation.

A subgraph G of $B_n(\sigma)$ with vertex set $V(G)$ is *bipartite* if its vertices can be partitioned into two sets U, V in such a way that every edge of G

JCMCC **42** (2002), pp. 161-166

is directed from a vertex of U to a vertex of V and $V(G) = U \cup V$. If this is the case we write $G = B(U, V)$. Equivalently, a bipartite subgraph of $B_n(\sigma)$ is a collection of edges no two of which form a directed path of length 2. Although $B_n(\sigma)$ is a directed graph, we will say that a subgraph of $B_n(\sigma)$ is *connected* if its underlying undirected graph is connected. If $\mathbf{u} \in U$ we denote the set of edges starting at $\mathbf{u}$ by $D(\mathbf{u})$. Note that no bipartite subgraph of $B_n(\sigma)$ can contain a loop at vertices $a^{n-1} = a \cdots a$, $a \in \Sigma$.

We will say that a code with block length n over the alphabet Σ is *comma-free*, or a $CF(n, \sigma)$ code, if, whenever $\mathbf{a} = a_1 \cdots a_n$ and $\mathbf{b} = b_1 \cdots b_n$ are codewords, each of the words

$$a_{i+1} \cdots a_n b_1 \cdots b_i, \quad i = 1, \ldots, n$$

(called *overlaps*) are **not** in the code. No $CF(n, \sigma)$ code can contain both a codeword and one of its $n - 1$ proper cyclic shifts since then the word would be an overlap of two copies of itself. In particular, no $CF(n, \sigma)$ code can contain a periodic word. We will discuss the meaning of periodicity in this context in section 3.

We denote by $cf(n, \sigma)$ the maximum number of words in any $CF(n, \sigma)$ code. It was first proved in [5] that

$$cf(n, \sigma) \quad \leq \quad \omega(n, \sigma) = \frac{1}{n} \sum_{d|n} \mu(n/d) \sigma^d,$$

where μ is the Möbius function of elementary number theory. This upper bound is known as the Witt function in other contexts [3]. Golomb, Gordon and Welch [5] conjectured that (1) was exact for all odd n. Seven years later, Eastman [4] found a construction which resolved their conjecture affirmatively. An easily implemented algorithm was given by Scholtz in [7]. When $n = 2$ Golomb, Gordon, and Welch [5] showed that $cf(2, \sigma) = \lfloor \frac{\sigma^2}{3} \rfloor$.

2 Bipartite Subgraphs of $B_n(\sigma)$

In this section we explore, quite independently of the alphabet, the constraints imposed on subgraphs of $B_n(\sigma)$ if they are bipartite.

Lemma 1 *If $n > 2$ and $G = B(U, V)$ is any connected bipartite subgraph of $B_n(\sigma)$ then $|U| \leq \sigma$.*

Proof: Every vertex $\mathbf{u} \in U$ has the form $xa_1 \cdots a_{n-2} \in U$ for some $x \in \Sigma$ and there is $\mathbf{v} = a_1 \cdots a_{n-2} y \in V$ such that $xa_1 \ldots a_{n-2} y$ is an edge

of G because G is connected. If

$$\mathbf{u}_i = x_i a_1 \cdots a_{n-2} \neq \mathbf{u}_j = x_j a_1 \ldots a_{n-2}$$

in U then $x_i \neq x_j$ in Σ by the definition of $B_n(\sigma)$. That is, the mapping $\mathbf{u}_i \rightarrow x_i$ is $1-1$.

Theorem 1 *Any connected bipartite subgraph $G = B(U,V)$ of $B_n(\sigma)$ has at most σ^2 edges.*

Proof: Let $U = \{\mathbf{u}_1, \ldots, \mathbf{u}_t\}$. By lemma 1 we know that $t \leq \sigma$. Further, each $D(\mathbf{u}_i) \leq \sigma$ by the definition of $B_n(\sigma)$. Hence the number of edges in $B(U,V)$ is

$$\sum_{i=1}^{t} |D(\mathbf{u}_i)| \leq t\sigma \leq \sigma^2.$$

Corollary 1 *If $\sigma = 2$ and $n > 2$ then $B_n(\sigma)$ has only connected bipartite subgraphs of the form $K_{1,1}, K_{1,2}$, or $K_{2,2}$.*

Theorem 2 *If $n > 2$ and G is a connected bipartite subgraph of $B_n(\sigma)$ then there exist $a_1, \ldots, a_{n-2} \in \Sigma$ such that*

$$G \subseteq \{x a_1 \cdots a_{n-2} y \mid x, y \in \Sigma\}.$$

Proof: Let $G = B(U,V)$. Choose any edge $\mathbf{u}_1 \in C$. Since every edge of $B_n(\sigma)$ is labelled, there exist x_1, y_1 and $a_1, \ldots, a_{n-2} \in \Sigma$ such that $\mathbf{u}_1 = x_1 a_1 \ldots a_{n-2} y_1$.

If $|U| = 1$ then any other edge of G must be of the form $= x_1 a_1 \cdots a_{n-2} y$ for some $y \in \Sigma$, $y \neq y_1$ and we are done.

If $|U| \geq 2$ let $\mathbf{u}_1 = x_1 a_1 \ldots a_{n-2} \in U$. Then,

$$D(\mathbf{u}_1) \subseteq \{x_1 a_1 \cdots a_{n-2} y \mid y \in \Sigma\}$$

by the definition of $B_n(\sigma)$. Since G is connected, there exists an edge from some vertex $\mathbf{u}_2 \in U$, $\mathbf{u}_2 \neq \mathbf{u}_1$ to one of the vertices $a_1 \ldots a_{n-2} y \in V$ such that $x_1 a_1 \ldots a_{n-2} y$ is an edge of $D(\mathbf{u}_1)$. By the definition of edges in $B_n(\sigma)$ the edge $\mathbf{u}_2$ has the form $x_2 a_1 \ldots a_{n-2} y$ for some $x_2, y \in \Sigma, x_2 \neq x_1$. Further every edge in $D(\mathbf{u}_2)$ will have the form $x_2 a_1 \ldots a_{n-2} z$ for some $z \in \Sigma$, i.e.,

$$D(\mathbf{u}_1, \mathbf{u}_2) \subseteq \{x_1 a_1 \cdots a_{n-1} y \mid x, y \in \Sigma\}. \tag{1}$$

In a finite number of steps we have

$$V = D(U) \subseteq \{x a_1 \cdots a_{n-2} y \mid x, y \in \Sigma\}. \tag{2}$$

Corollary 2 *Every connected bipartite subgraph G of $B_n(\sigma)$ $(n > 2)$ defines an undirected graph $G(\Sigma)$ with vertices $x \in \Sigma$ and edges (x, y) whenever $xa_1 \cdots a_{n-2}y \in G$.*

We will utilize $G(\Sigma)$ in the next section to express sufficient conditions for a connected bipartite subgraph of $B_n(\sigma)$ to be comma-free for all $n > 2$.

3 Comma-Free Bipartite Subgraphs of $B_n(\sigma)$

We remarked in the introduction that we would consider the definition of periodicity in this section. It plays a curious role in this context.

There are two distinct but frequently used definitions of periodicity in strings.

> **P1:** A string $\mathbf{s}$ is *periodic* if there exists a string $\mathbf{u}$ such that $\mathbf{s} = \mathbf{u}^k = \mathbf{u} \cdots \mathbf{u}$, k times.
>
> **P2:** A string $\mathbf{a} = a_1 \cdots a_n$ is *periodic* if there exists $0 < i < n$ such that $a_{k+i} = a_k$ for $k = 1, \ldots n - i$.

It is easy to see that **P1** is a special case of **P2**.

Definition 1 *We say that a string $\mathbf{s} = a_1 \cdots a_n$ is pre-periodic if it is periodic with respect to* **P2** *above and either $x\mathbf{s}$ or $\mathbf{s}x$ is periodic with respect to* **P1** *for some $x \in \Sigma$.*

Theorem 3 *Let C be the edges of a maximal star of $B_n(\sigma)$ $(n > 2)$ of the form*

$$C = \{xa_1 \cdots a_{n-1} | x \in \Sigma\},$$

where $a_1 \cdots a_{n-1}$ is not pre-periodic. Then C is a $CF(n, \sigma)$ code.

Proof: Suppose C contains the ith overlap

$$\mathbf{o}_i = a_{i+1} \cdots a_{n-1} y a_1 \cdots a_i$$

of $xa_1 \cdots a_{n-1}$ and $ya_1 \cdots a_{n-1}$ for some $i = 1, \ldots, n-1$. From the definition of C, there is $z \in \Sigma$ such that $\mathbf{o}_i = za_1 \cdots a_{n-1}$. But if two strings are equal then, in particular, they have the same frequency vectors; i.e., the same number of occurrences of each letter. Therefore, $y = z$. This means $\mathbf{o}_i$ is also a proper cyclic permutation of $za_1 \cdots a_{n-1}$. But a string which is a proper cyclic permutation of itself is periodic in the sense of the definition **P1**. This contradicts the assumption that $a_1 \cdots a_{n-1}$ was not pre-periodic.

We remark that $a_1 \cdots a_{n-1}$ cannot be a constant string c^{n-1} because it would then be pre-periodic.

It is natural to ask whether a set of edges $\{xa_1 \cdots a_{n-2}y \mid x, y \in \Sigma\}$ could represent a $CF(n, \sigma)$ code. Over $\Sigma = \{0, 1\}$ for example, the set $\{x010y \mid x, y \in \Sigma\}$ of edges in $B_5(2)$ is not comma-free since, say, 10100 is an overlap of both 10101 and 00101. Of the edges in this set, both 00100 and 10101 are periodic with respect to **P2**. Even if these edges are removed, the remaining set of edges is still not a CF(5,2) code because 00101 and 10100 are cyclic permutations of one another and hence cannot both be in the same comma-free code.

In the proof of Theorem 3 we used the notion of the frequency vector of a string. We now give a more formal definition.

Definition 2 *If $a \in \Sigma = \{a_1, \ldots, a_n\}$ and $\mathbf{s}$ is a string with entries in Σ, define $|\mathbf{s}|_a$ to be the number of occurrences of a in $\mathbf{s}$. Further define*

$$\phi(\mathbf{s}) = (|\mathbf{s}|_{a_1}, \ldots, |\mathbf{s}|_{a_n}).$$

The non-negative integer vector $\phi(\mathbf{s})$ is called the frequency vector of $\mathbf{s}$.

It is easy to see that if $\mathbf{xy}$ is the concatenation of two strings $\mathbf{x}$ and $\mathbf{y}$ then

$$\phi(\mathbf{xy}) = \phi(\mathbf{x}) + \phi(\mathbf{y}). \tag{3}$$

Theorem 4 *Let C be a connected bipartite subgraph of $B_n(\sigma)$ $(n > 2)$ without pre-periodic vertices. If $C(\Sigma)$ has no path of length 3 then the edges of C form a $CF(n, \sigma)$ code.*

Proof: Since $n > 2$ and C is a connected bipartite subgraph of $B_n(\sigma)$, Theorem 2 shows there exist $a_1, \ldots, a_{n-2} \in \Sigma$ such that

$$C \subseteq \{xa_1 \cdots a_{n-2}y \mid x, y \in \Sigma\}.$$

Suppose for some $i = 1, \ldots, n-1$ that C contains an overlap

$$\mathbf{o}_i = a_i \cdots a_{n-2}y_1x_2a_1 \cdots a_{i-1}$$

of two codewords $x_1a_1 \cdots a_{n-1}y_1$ and $x_2a_1 \cdots a_{n-2}y_2$. By Theorem 2 there exist $x^*, y^* \in \Sigma$ such that

$$\mathbf{o}_i = a_i \cdots a_{n-2}y_1x_2a_1 \cdots a_{i-1} = x^*a_1 \cdots a_{n-2}y^*. \tag{4}$$

Hence $\phi(\mathbf{o}_i) = \phi(x^*a_1 \cdots a_{n-2}y^*)$. It follows from equation (3) that

$$\phi(x^*y^*) = \phi(y_1x_2).$$

Since there are only two entries in each frequency vector there are precisely two cases.

Case 1: $x^* = x_2$ and $y^* = y_1$. Here $\mathbf{o}_i = a_i \cdots a_{n-2} y_1 x_2 a_1 \cdots a_{i-1}$ is a proper cyclic permutation of the edge $x_2 a_1 \cdots a_{n-2} y_1$ and yet equal to it. Therefore, $x_2 a_1 \cdots a_{n-2} y_1$ is periodic (**P1**), contradicting the hypothesis that C contains no pre-periodic vertices, or equivalently no periodic edges.

Case 2: $x^* = y_1$ and $y^* = x_2$. Here $\mathbf{o}_i = a_i \cdots a_{n-2} y_1 x_2 a_1 \cdots a_{i-1}$ is not necessarily a proper cyclic permutation of $y_1 a_1 \cdots a_{n-2} x_2$ unless $x_2 = y_1$ in which case the argument of Case 1 applies. Otherwise,

$$\mathbf{o}_i = a_i \cdots a_{n-2} y_1 x_2 a_1 \cdots a_{i-1} \quad = \quad y_1 a_1 \cdots a_{n-2} x_2.$$

But then the undirected graph $C(\Sigma)$ necessarily contains the edges $(x_1, y_1), (y_1, x_2)$, and (x_2, y_2) which contradicts the hypothesis that $C(\Sigma)$ contained no path of length 3.

References

[1] A. H. Ball, *The construction of comma-free codes with odd word length*, Ph.D thesis, Department of Combinatorics and Optimization, University of Waterloo, Canada, 1980.

[2] L.J. Cummings, *Comma-free codes in the deBruijn graph*, Caribbean J. Math. **2**(1983), 65–68.

[3] L.J. Cummings and M.E. Mays, On the parity of the Witt formula, *Congressus Numerantium* **80**(1991), 49–56.

[4] W.L. Eastman, *On the construction of comma-free codes*, IEEE Trans. Information Theory **11**(1965), 263–267.

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

[6] B.H. Jiggs, *Recent results in comma-free codes*, Canadian J. Math. **15**(1963), 178–187.

[7] R. A. Scholtz, *Maximal and variable word- length comma-free codes*, IEEE Trans. Inform. Theory **IT15**(1969), 300–306.

[8] B. Tang, S.W. Golomb, and R.L. Graham, *A New Result on Comma-Free Codes of Even Word Length*, Canadian J. Math. **39**(1987), 513–526.