De Bruijn Cycles for Covering Codes

Fan Chung,[1] Joshua N. Cooper[1,2]
[1]*Department of Mathematics, University of California, San Diego, La Jolla, California 92093*
[2]*Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York, New York 10012; e-mail: cooper@cims.nyu.edu*

Received 9 September 2003; revised 5 May 2004; accepted 18 May 2004
DOI 10.1002/rsa.20033
Published online 16 July 2004 in Wiley InterScience (www.interscience.wiley.com).

ABSTRACT: A de Bruijn covering code is a q-ary string S so that every q-ary string is at most R symbol changes from some n-word appearing consecutively in S. We introduce these codes and prove that they can have size close to the smallest possible covering code. The proof employs tools from field theory, probability, and linear algebra. Included is a table of the best known bounds on the lengths of small binary de Bruijn covering codes, up to $R = 11$ and $n = 13$, followed by several open questions in this area. © 2004 Wiley Periodicals, Inc. Random Struct. Alg., 25: 421–431, 2004

1. INTRODUCTION

A covering code $\mathcal{C}$ of radius R and dimension n on q symbols is a subset of the space $[q]^n$ such that every string in $[q]^n$ differs from some element of $\mathcal{C}$ in at most R coordinates. We do not require R to be as small as possible in this definition.

Question: Given n, R, and q, what is the smallest $M = M(n, R, q)$ so that there exists an q-ary string $S = (s_0, \ldots, s_{M-1})$ with the property that the set of n-strings appearing as $(s_i, \ldots, s_{i+n-1})$, with indices taken modulo M, form a covering code of radius R? Call such a string a (n, R, q)-*de Bruijn covering code.*

Correspondence to: J. N. Cooper

For example, 111000 is a (4, 1, 2)-de Bruijn covering code, because every binary 4-string is at most one bit change from an element of

$$\{1110, 1100, 1000, 0001, 0011, 0111\}.$$

On the alphabet {A, G, T, C}, the string

$$\text{AGATCGCAGATATGGTCTATG}$$

is a (4, 2, 4)-de Bruijn covering code, by Proposition 3 below.

Clearly, $M(n, 0, q) = q^n$, since any de Bruijn covering code of radius 0 is actually a de Bruijn cycle, and de Bruijn cycles of all orders over an arbitrary alphabet exist (see, e.g., [5]). If we fix $R > 0$ and $q \geq 2$, how does $M(n, R, q)$ grow as $n \to \infty$?

It is easy to see that the growth is at least $\Omega(q^n/n^R)$, by the so-called "sphere-covering" bound. The set of strings which differ from any given S in at most R places has the same cardinality, $\sum_{k=0}^{R} \binom{n}{k}(q-1)^k$. Therefore, if we are to cover all q^n strings, we need at least

$$\frac{q^n}{\sum_{k=0}^{R} \binom{n}{k}(q-1)^k}$$

codewords. On the other hand, it is well known that the size of the smallest q-ary covering code of radius R actually achieves this bound, up to a multiplicative constant which depends on R and q (see [4] for the latest results on the size of this constant). We may concatenate all the codewords of such a minimal code to yield a (n, R, q)-de Bruijn covering code of length $O(q^n/n^{R-1})$. This construction is clearly very wasteful, however. Can we do better, i.e., is the true order of magnitude of $M(n, R, q)$ closer to the sphere-covering bound? In particular, can we say something nontrivial in the case of $R = 1$? In fact, in Section 3 we prove the following.

Theorem 1. *For each n and q a prime power, there exists a* (n, R, q)*-de Bruijn covering code of length* $\leq (R + 1 + o(1))q^n \log n/(\binom{n}{R}(q-1)^R)$.

Section 2 states several definitions and preliminary results we will need to prove this. The next section contains the proof itself. In Section 4 we present bounds for special values of n, R, and q, and include a table of bounds on $M(n, R, 2)$ for $2 \leq n \leq 13$ and $1 \leq R \leq 11$. We end with several remarks and questions for further work in Section 5.

2. PRELIMINARIES

We fix a prime power $q \geq 2$ throughout this section and the next, and take our alphabet to be $\mathbb{F}_q$. (If q is not a prime power, we take the alphabet to be $\mathbb{Z}/q\mathbb{Z}$.) Write $b_R(v)$ for v an n-string drawn from $\mathbb{F}_q$, to denote the set of those strings differing from v in at most R coordinates. That is, $b_R(v)$ is v's radius R neighborhood in the Hamming metric. Also, write $\text{wt}(v)$ for the Hamming weight of the vector v, the number of nonzero symbols it contains.

Let α be a generator of the multiplicative group of the finite field $\mathbb{F}_{q^n}$. Denote by $\mathscr{C}$ the

elementary basis for $\mathbb{F}_q^n$ over $\mathbb{F}_q$. Given a basis $\mathcal{B} = \{b_1, \ldots, b_n\}$ of $\mathbb{F}_{q^n}$ over $\mathbb{F}_q$ and an element $\gamma \in \mathbb{F}_{q^n}$, write $f_{\mathcal{B}}(\gamma)$ for the element of $\mathbb{F}_q^n$ whose jth coordinate is the coefficient of b_j in the $\mathcal{B}$-representation of γ. Then, given a nonzero vector $\mathbf{x} \in \mathbb{F}_q^n$, define $\Lambda(\alpha, \mathcal{B}, \mathbf{x})$ to be the string whose jth coordinate [i.e., $\Lambda_j(\alpha, \mathcal{B}, \mathbf{x})$, $1 \le j \le q^n - 1$] is $\mathbf{x}^{\mathbf{T}} f_{\mathcal{B}}(\alpha^j)$. It is well known that, when $\mathcal{B} = \{\alpha^j : 0 \le j \le n - 1\}$ and $\mathrm{wt}(\mathbf{x}) = 1$, $\Lambda(\alpha, \mathcal{B}, \mathbf{x})$ is a de Bruijn cycle of order n if we insert a 0 at the beginning (see, e.g., [3]). We generalize this result as follows. Define $\Lambda^*(\alpha, \mathcal{B}, \mathbf{x})$ to be the sequence $\Lambda(\alpha, \mathcal{B}, \mathbf{x})$ with a zero inserted at the beginning of each occurrence of the string $0 \ldots 01$. Then we have the following.

Proposition 2. *Fix a basis $\mathcal{B}$ of $\mathbb{F}_{q^n}$ over $\mathbb{F}_q$, a generator $\alpha \in \mathbb{F}_{q^n}^{\times}$, and a vector $\mathbf{x} \in \mathbb{F}_q^n$, and write $\Phi(j)$ for the vector*

$$(\Lambda_j(\alpha, \mathcal{B}, \mathbf{x}), \ldots, \Lambda_{j+n-1}(\alpha, \mathcal{B}, \mathbf{x}))^{\mathbf{T}} \in \mathbb{F}_q^n.$$

The map Ψ which sends 0 to 0 and α^j to $\Phi(j)$ is an isomorphism from the additive group of $\mathbb{F}_{q^n}$ to $\mathbb{F}_q^n$.

Proof. First, we show that Ψ is linear. Write e_j for the elementary n-vector whose coordinates are all zero except for a 1 in the jth coordinate. We denote by $M_{\gamma,\mathcal{B}}$ the matrix representing multiplication by $\gamma \in \mathbb{F}_{q^n}$ in the $\mathcal{B}$ basis. It is easy to see that

$$\Lambda_j(\alpha, \mathcal{B}, \mathbf{x}) = \mathbf{x}^{\mathbf{T}} f_{\mathcal{B}}(\alpha^j)$$

and therefore that

$$\Psi(\gamma) = \sum_{j=0}^{n-1} e_{j+1} \mathbf{x}^{\mathbf{T}} f_{\mathcal{B}}(\alpha^j \gamma) = \sum_{j=0}^{n-1} e_{j+1} \mathbf{x}^{\mathbf{T}} M_{\alpha,\mathcal{B}}^j f_{\mathcal{B}}(\gamma), \tag{1}$$

which is obviously linear.

Now, suppose that $\Psi(\gamma) = 0$. We show that $\gamma = 0$. Indeed, suppose that $\{j_1, \ldots, j_n\}$ are n distinct integers so that $\Lambda_{j_i}(\alpha, \mathcal{B}, \mathbf{x}) = 0$ for each i. If we denote by S the subspace of $\mathbb{F}_q^n$ orthogonal to $\mathbf{x}$, then we have $\alpha^{j_i} \in f_{\mathcal{B}}^{-1}(S)$ for each i. However, $f_{\mathcal{B}}$ is linear and has a trivial kernel, so all the α^{j_i} lie in a subspace of $\mathbb{F}_q^n$ of dimension $n - 1$ and are therefore linearly dependent. If we take $j_i = j + i$ for some j [i.e., $\Psi(\gamma) = 0$ with $\gamma = \alpha^j$], then we have that $\{\alpha^i\}_{i=j+1}^{j+n}$ is a dependent set. Since $M_{\alpha,\mathcal{B}}$ is nonsingular, this implies that $\{\alpha^i\}_{i=0}^{n-1}$ is a dependent set. But then we have

$$\sum_{i=0}^{n-1} c_i \alpha^i = 0$$

for some nonzero $(c_1, \ldots, c_n)$, so that α satisfies a polynomial identity of degree less than n. Since α generates $\mathbb{F}_{q^n}^{\times}$, this implies that $\{\alpha^j\}_{j=0}^{d}$ is a basis for $\mathbb{F}_{q^n}$ for some $d < n - 1$, contradicting the fact that the dimension of $\mathbb{F}_{q^n}$ over $\mathbb{F}_q$ is n. We can therefore conclude that $\gamma = 0$. ∎

Note that the map $\gamma \mapsto M_{\gamma,\mathcal{B}}$ is actually an isomorphism of fields. The image is a set of matrices which form a field, i.e., a *matrix field.* These objects have been studied extensively and thoroughly characterized when the matrices take their entries from a finite field [2].

Corollary 3. $\Lambda^*(\alpha, \mathcal{B}, \mathbf{x})$ *is a de Bruijn cycle.*

Proof. By the above argument, $\Lambda(\alpha, \mathcal{B}, \mathbf{x})$ contains all nonzero n-strings. Clearly, the insertion of a 0 causes the occurrence of the all-zeroes string without disrupting the presence of any other string. ∎

Our approach is to find an $\alpha \in \mathbb{F}_{q^n}$, a basis $\mathcal{B}$, and a vector $\mathbf{x}$ so that the first $K \sim q^n \log n/(\binom{n}{R}(q-1)^R)$ length n strings appearing in $\Lambda(\alpha, \mathcal{B}, \mathbf{x})$ are (almost) a covering code of radius R. Specifically, we wish to show that, for only a small fraction of all $v \in \mathbb{F}_q^n$,

$$(v + B_R(0^n)) \cap \Psi(\{\alpha^j\}_{j=1}^K) = \varnothing,$$

where Ψ is the function defined in Proposition 2. Define $\Psi' = f_{\mathcal{B}} \circ \Psi^{-1}$. Setting $w = \Psi^{-1}(v)$, we may bound this quantity from above by asking the number of w so that

$$\Psi'\left(\binom{\mathcal{C}}{R}\right) \cap f_{\mathcal{B}}(w + \{\alpha^j\}_{j=1}^K) = \varnothing,$$

which, by (1), is the same as saying that

$$\left\{\left(\sum_{j=0}^{n-1} e_{j+1}\mathbf{x}^{\mathrm{T}}M_{\alpha,\mathcal{B}}^j\right)^{-1} v : \mathrm{wt}(v) = R\right\} \cap f_{\mathcal{B}}(w + \{\alpha^j\}_{j=1}^K) = \varnothing.$$

We must determine which matrices may appear in the form of the left-hand term. First, a result from linear algebra is needed. A *nonderogatory* matrix is one whose eigenspaces are all one-dimensional, and a matrix in *rational canonical form* is comprised of blocks of the form

$$\begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & a_1 \\ 1 & 0 & 0 & 0 & \cdots & \vdots \\ 0 & 1 & 0 & 0 & \cdots & \vdots \\ 0 & 0 & 1 & 0 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 1 & a_n \end{pmatrix}$$

along the diagonal. The following theorem appears in [7].

Theorem 4. *If $A \in K^{n\times n}$ is nonderogatory and in rational canonical form, then the following are equivalent*:

1. *X commutes with A.*
2. *The successive columns of X are $v, Av, \ldots, A^{n-1}v$ for any $v \in K^n$.*
3. *There exists a polynomial $g \in K[x]$ so that $X = g(A)$.*

Furthermore, $g = \sum_{j=0}^{n-1} v_{j+1}x^j$.

The matrices $M_{\alpha,\mathcal{B}}$ are nonderogatory when α is a generator of $\mathbb{F}_{q^n}$, because their eigenvalues are all distinct, as the next result states.

Proposition 5. *A matrix $M \in \mathbb{F}_q^{n\times n}$ is of the form $M_{\alpha,\mathcal{B}}$ for some generator $\alpha \in \mathbb{F}_{q^n}^{\times}$ and basis $\mathcal{B} \subset \mathbb{F}_{q^n}$ over $\mathbb{F}_q$ if and only if its eigenvalues (over the algebraic closure of $\mathbb{F}_q$) are $\{\alpha^{q^j}\}_{j=0}^{n-1}$.*

Proof. For a given α, fix the basis $\mathcal{A} = \{\alpha^j\}_{j=0}^{n-1}$. Clearly, if we write B for the matrix whose columns are $\mathcal{B}$ written in the basis $\mathcal{A}$, then $M_{\alpha,\mathcal{B}} = B^{-1}M_{\alpha,\mathcal{A}}B$. Therefore, a matrix M is one of the desired ones if and only if it has the same eigenvalues as the matrix $M_{\alpha,\mathcal{A}}$. Let $p_\alpha(\lambda)$ denote the characteristic polynomial of this matrix. By the Cayley-Hamilton Theorem (which applies to all commutative rings), $p_\alpha(M_{\alpha,\mathcal{A}}) = 0$. However, the map $\alpha \mapsto M_{\alpha,\mathcal{B}}$ is an isomorphism of fields for any basis $\mathcal{B}$. Therefore, $p_\alpha(\alpha) = 0$. Since the Galois group of $\mathbb{F}_{q^n}$ over $\mathbb{F}_q$ is cyclic and generated by the Frobenius map $x \mapsto x^q$, and the rest of the roots of p_α are the Galois conjugates of α, the result follows. ∎

Furthermore, if we let Θ_α denote the basis $\{\alpha^j\}_{j=0}^{n-1}$, then M_{α,Θ_α} is in rational canonical form. Its jth column is e_{j+1} for $1 \le j \le n-1$ and its nth column is the vector of coefficients of the minimal polynomial of α (without the leading term). Using this fact, we can prove the following from Theorem 4.

Lemma 6. *Fix a generator α of $\mathbb{F}_{q^n}$. Choose $\mathbf{x} \in \mathbb{F}_q^n\backslash\{0^n\}$ randomly and uniformly, and choose a basis $\mathcal{B}$ randomly and uniformly. Then*

$$\left(\sum_{j=0}^{n-1} e_{j+1}\mathbf{x}^{\mathbf{T}}M_{\alpha,\mathcal{B}}^j\right)^{-1}$$

is distributed uniformly over all invertible matrices.

Proof. Evidently, it suffices to show that $D(\mathcal{B}, \mathbf{x}) = \sum_{j=0}^{n-1} e_{j+1}\mathbf{x}^{\mathbf{T}}M_{\alpha,\mathcal{B}}^j$ is distributed uniformly. This matrix is one whose rows are $\mathbf{x}^{\mathbf{T}}$, $\mathbf{x}^{\mathbf{T}}M_{\alpha,\mathcal{B}}$, ..., $\mathbf{x}^{\mathbf{T}}M_{\alpha,\mathcal{B}}^{n-1}$. Write A for the matrix M_{α,Θ_α} and P for the matrix whose successive columns are the elements of Θ_α written in the $\mathcal{B}$ basis, and write $\mathbf{y}$ for $P^{\mathbf{T}}\mathbf{x}$. Then we may also say that $D(\mathcal{B}, \mathbf{x})$ is the matrix whose rows are $\mathbf{x}^{\mathbf{T}}$, $\mathbf{x}^{\mathbf{T}}PAP^{-1}$, ..., $\mathbf{x}^{\mathbf{T}}PA^{n-1}P^{-1}$, which we may rewrite as $D(A, P^Tx)P^{-1}$. Therefore, by Theorem 4 and the fact that A is nonderogatory and in rational canonical form, $D(\mathcal{B}, \mathbf{x}) = g_{\mathbf{y}}(A)^{\mathbf{T}}P^{-1}$ with $g_{\mathbf{y}}$ denoting the polynomial whose coefficients are the entries of $\mathbf{y}$. Choosing $\mathbf{x}$ uniformly and randomly from the nonzero vectors yields the same distribution on $\mathbf{y}$, independent of the choice of $\mathcal{B}$. Since A is the image of α under the map $\alpha \mapsto M_{\alpha,\Theta_\alpha}$, and $g_{\mathbf{y}}(\alpha)$ is uniformly distributed over $\mathbb{F}_{q^n}\backslash\{0\}$ as $\mathbf{y}$ varies, we have $g_{\mathbf{y}}(A)$ uniformly distributed over all matrices of the form M_{γ,Θ_α} for $\gamma \in \mathbb{F}_{q^n}\backslash\{0\}$. Choosing $\mathcal{B}$ uniformly is the same as choosing P^{-1} uniformly, so we may conclude that $D(\mathcal{B}, \mathbf{x}) = g_{\mathbf{y}}(A)^{\mathbf{T}}P^{-1}$ is uniformly distributed over all invertible matrices. ∎

3. THE MAIN RESULT

It remains to show that the set of all sums of k columns of a randomly, uniformly chosen invertible matrix are distributed more or less uniformly. Before proceeding, we need to state Suen's Inequality. We follow [1]. Let $\{A_i\}_{i\in I}$ be a set of events, and define a symmetric relation (i.e., a graph) $\sim$ on I. We say that $\sim$ is a *superdependency* graph if, whenever $J_1, J_2 \subset I$ have no edges between them, any Boolean combination of $\{A_i\}_{i\in J_1}$ is independent of any Boolean combination of $\{A_i\}_{i\in J_2}$. Write $M = \prod_{i\in I} \Pr[\overline{A_i}]$.

Theorem 7 (Suen's Inequality). *Define*

$$y(i,j) = (\Pr[A_i \wedge A_j] + \Pr[A_i]\Pr[A_j]) \prod_{l\sim i \text{ or } l\sim j} (1 - \Pr[\overline{A_l}])^{-1}.$$

Then

$$\Pr\left[\bigwedge_{i\in I} \overline{A_i}\right] \le M e^{\sum_{i\sim j} y(i,j)}.$$

The following is a routine application of this result.

Proposition 8. *For $R \in \mathbb{Z}^+$, if M is chosen randomly and uniformly from $GL_n(\mathbb{F}_q)$, then, for any set $S \subset \mathbb{F}_q^n$ with $|S| = q^n K/(\binom{n}{R}(q-1)^R)$,*

$$\Pr[\{Mv : \mathrm{wt}(v) = R\} \cap S = \varnothing] \le e^{-K}(c_q^{-1} + o(1)).$$

where $c_q = \prod_{j=1}^{\infty} (1 - q^{-j})$ and $K = o(\sqrt{n})$.

Proof. The probability that a randomly, uniformly chosen invertible matrix has all sums of k columns lying outside of a set S is given by

$$\begin{aligned}\rho &= \Pr[Mv \in \bar{S} \text{ when } \mathrm{wt}(v) = R \mid M \in GL_n(\mathbb{F}_q)] \\ &= \frac{\Pr[(Mv \in \bar{S} \text{ when } \mathrm{wt}(v) = R) \wedge (M \in GL_n(\mathbb{F}_q))]}{\Pr[M \in GL_n(\mathbb{F}_q)]} \\ &\le \frac{\Pr[Mv \in \bar{S} \text{ when } \mathrm{wt}(v) = R]}{\Pr[M \in GL_n(\mathbb{F}_q)]},\end{aligned}$$

where we are choosing M randomly and uniformly from *all* matrices. It is well known that $|GL_n(\mathbb{F}_q)| = q^{n^2}(c_q + o(1))$ with $c_q = \prod_{j=1}^{\infty}(1 - q^{-j})$. Therefore,

$$\rho \le \Pr[Mv \in \bar{S} \text{ when } \mathrm{wt}(v) = R](c_q^{-1} + o(1)).$$

Now, for a vector v of weight R, define A_v to be the event that $Mv \in S$, and let $I(v)$ denote the set of indices at which v is nonzero. Then $\Pr[Mv \in \bar{S} \text{ when } \mathrm{wt}(v) = R] = \Pr[\wedge_v \overline{A_v}]$.

The relation $v \sim w$ iff $I(v) \cap I(w) \neq \varnothing$ clearly defines a superdependency graph on these events. Furthermore, any pair A_v and A_w, $v \neq w$, are independent, since, if we fix the ith columns of M for $i \in I(v) \cap I(w)$, then $\sum_{i \in I(v) \setminus I(w)} Me_i$ and $\sum_{i \in I(v) \setminus I(w)} Me_i$ are independent and uniformly distributed over $\mathbb{F}_q^n$. Therefore,

$$\begin{aligned} y(v, w) &= 2 \Pr[A_v]\Pr[A_w] \prod_{z \sim v \text{ or } z \sim w} (1 - \Pr[\overline{A_z}])^{-1} \\ &\leq 2\left(\frac{K}{\binom{n}{R}(q-1)^R}\right)^2 \left(1 - \frac{K}{\binom{n}{R}(q-1)^R}\right)^{-2(\binom{n}{R}(q-1)^R - \binom{n-R}{R}(q-1)^R)} \\ &= 2\left(\frac{K}{\binom{n}{R}(q-1)^R}\right)^2 \left(1 - \frac{K}{\binom{n}{R}(q-1)^R}\right)^{\binom{n}{R-1}(-2R^2+o(1))(q-1)^R} \\ &\leq 2\left(\frac{K}{\binom{n}{R}(q-1)^R}\right)^2 e^{-K\binom{n}{R-1}(-2R^2+o(1))/\binom{n}{R}} \\ &= 2\left(\frac{K}{\binom{n}{R}(q-1)^R}\right)^2 e^{-K(-2R^3+o(1))/n}. \end{aligned}$$

Since there are $\binom{n}{R}(\binom{n}{R} - \binom{n-R}{R})(q-1)^{2R}/2 = O(n^{2R-1})$ relations $v \sim w$, the quantity $\sum_{v \sim w} y(v, w)$ tends to 0 as $n \to \infty$ so long as $K = o(\sqrt{n})$. Therefore, Suen's Inequality implies that

$$\begin{aligned} \Pr\left[\bigwedge_{\mathrm{wt}(v)=R} \overline{A_v}\right] &\leq (c_q^{-1} + o(1)) \prod_{\mathrm{wt}(v)=R} \Pr[\overline{A_v}] \\ &= (c_q^{-1} + o(1))\left(1 - \frac{K}{\binom{n}{R}(q-1)^R}\right)^{\binom{n}{R}(q-1)^R} \\ &\leq (c_q^{-1} + o(1))e^{-K}. \end{aligned}$$

■

Taking an initial segment of a random $\Lambda(\alpha, \mathcal{B}, \mathbf{x})$ and adding in all the "uncovered" codewords yields an (n, R, q)-de Bruijn covering code.

Theorem 2. *For each n, there exists an (n, R, q)-de Bruijn covering code of length $\leq (R + 1 + o(1))q^n \log n/(\binom{n}{R}(q-1)^R)$.*

Proof. Fix any generator $\alpha \in \mathbb{F}_{q^n}^{\times}$. Choose the basis $\mathcal{B} = \{b_i\}_{i=1}^n$ and the vector $\mathbf{x} \in \mathbb{F}_q^n \setminus \{0^n\}$ randomly and uniformly. Then define $\bar{\Lambda}(K)$ to be the string of the first $q^n K/(\binom{n}{R}(q-1)^R) + n$ symbols of $\Lambda(\alpha, \mathcal{B}, \mathbf{x})$ [which we will call $\Lambda_1(K)$], followed by a concatenated list [which we will call $\Lambda_2(K)$] of all strings in

$$\mathbb{F}_q^n \Big| \bigcup_{c \in \mathcal{C}} b_R(c),$$

where $\mathcal{C}$ is the set of codewords appearing as n consecutive symbols (without wrap-around) in $\Lambda_1(K)$. Then the resulting expected length of the string is given by

$$\mathrm{E}(|\Lambda_1(K)| + |\Lambda_2(K)|) = \frac{q^n K}{\binom{n}{R}(q-1)^R} + n + nq^n \sum_{v \in \mathbb{F}_q^n} \Pr[b_R(v) \cap \mathcal{C} = \varnothing]. \tag{2}$$

Furthermore, the constructed string is an (n, R, q)-de Bruijn covering code. By the discussion preceding Theorem 4, $\Pr[b_R(v) \cap \mathcal{C} = \varnothing]$ is bounded above by

$$\Pr\left[\left\{\left(\sum_{j=0}^{n-1} e_{j+1}\mathbf{x}^{\mathrm{T}} M_{\alpha,\mathcal{B}}^j\right)^{-1} w : \mathrm{wt}(w) = R\right\} \cap f_{\mathcal{B}}(v + \{\alpha^j\}_{j=1}^K) = \varnothing\right].$$

The matrix in the left-hand term is uniformly distributed over all invertible matrices, by Lemma 6. Therefore, by Proposition 8,

$$\Pr[b_R(v) \cap \mathcal{C} = \varnothing] \le e^{-K}(c_q^{-1} + o(1)).$$

Plugging this and $K = (R + 1)\log n$ into (2) yields

$$\mathrm{E}(|\Lambda_1(K)| + |\Lambda_2(K)|) \le \frac{q^n \log n}{\binom{n}{R}(q-1)^R}(R + 1 + o(1)),$$

so a (n, R, q)-de Bruijn covering code of the desired length exists. ∎

4. NUMERICAL BOUNDS

It is of interest to know $M(n, R, q)$ for small values of its parameters—in particular, for $q = 2$, i.e., the binary case. First, we collect a few simple observations.

1. $M(n, R, q) \le M(n + k, R - l, q + m)$ for any $k, l, m \ge 0$. If a de Bruijn covering code $\mathcal{C}$ exists for parameters $(n + k, R - l, q + m)$, then certainly decreasing the dimension, increasing the radius, or decreasing the number of symbols will leave $\mathcal{C}$ covering everything. (In the case of decreasing the number of symbols, we can replace all occurrences of the excluded symbols to "0." It is easy to check that this operation can only decrease distances from n-strings to the code.)
2. $M(n, 0, q) = q^n$, as noted in the introduction.
3. $M(n, R, q) = 1$ if $R \ge n$, by taking the string "0."
4. $M(n, R, 2) = 2$ if $\lfloor n/2 \rfloor \le R < n$, by taking the string "01." The two resulting codewords are complements in the n-cube, and therefore every string is within $\lfloor n/2 \rfloor$ of one of them. Furthermore, it is clear that at least 2 codewords are necessary.
5. $M(n, R, q) \ge K_q(n, R)$, the smallest number of codewords in a q-ary covering code of dimension n and radius R.

TABLE 1. Best Known Bounds for $M(n, R, 2)$.

$R\backslash n$	2	3	4	5	6	7
1	2	2	6	8	12	22
2	1	2	2	2	8	10
3	1	1	2	2	2	2
4	1	1	1	2	2	2
5	1	1	1	1	2	2
6	1	1	1	1	1	2

$R\backslash n$	8	9	10	11	12	13
1	32	57–130	105–322	180–694	342–1454	598–2937
2	14	20	38	38–117	62–244	97–529
3	6	12	16	20	34–40	34–119
4	2	2	4	8	16	24
5	2	2	2	2	8	8
6	2	2	2	2	2	2
7	2	2	2	2	2	2
8	1	2	2	2	2	2
9	1	1	2	2	2	2
10	1	1	1	2	2	2
11	1	1	1	1	2	2

6. If there exists an (n, R, q)-de Bruijn covering code of length M, then there exists one of length $M + n + k - 1$ for all $k \geq 0$. If S is the shorter string, append a copy of the first $(n - 1)$ symbols and k arbitrary q-ary symbols to the end.

Below, we include a table of the best known bounds on the sizes of binary de Bruijn covering codes with various parameters. A single number in an entry indicates that the exact value of $M(n, R, 2)$ is known; two numbers indicate an upper and lower bound. Bounds were achieved using the observations above, the table in [6], as well as software that searched the string space randomly (for upper bounds), and one which searched it exhaustively (for lower bounds).

5. REMARKS AND FURTHER QUESTIONS

Statement 6 in the previous section highlights a frustrating property of de Bruijn covering codes that stands in stark contrast to ordinary covering codes: it is possible for one to exist of length M but for none to exist of length $M + 1$. For example, a (10, 4, 2) code exists of lengths 4 ("1100"), 6 ("011100"), 8 ("00111100"), and 12 ("000011111100"), but none of lengths 5, 7, 9, 10, or 11 exist. However, by the above, a (10, 4, 2) code of all lengths at least 13 must exist. Therefore, in addition to finding the smallest possible de Bruijn covering code, we would like to know when de Bruijn covering codes with lengths *between* $M(n, R, q)$ and $M(n, R, q) + n - 1$ exist. We also would like to explain why all of the super-diagonal entries in Table 1 are even, at least where we know the precise value.

Another difference between de Bruijn covering codes and ordinary ones is that there is

no easy way to use known efficient codes to build efficient codes for larger n, smaller R, or larger q. It would be desirable to define "product" analogous to direct sums for ordinary covering codes. Unfortunately, interlacing, the obvious candidate for such a product, appears to be very inefficient. We offer a different, though related construction which allows us to increase q when the desired number of symbols is a perfect power of the number of symbols in the original code.

Proposition 3. *If $a^s = b$ for any positive integers a, b, and s, then, for all n, $R > 0$,*

$$M(n, R, b) \leq s^2 \left\lceil \frac{M(sn, R, a) + sn}{s} \right\rceil - s.$$

Proof. Let $t = M(sn, R, a)$ and $m = s^2\lceil (t + sn)/s \rceil - s$, and let $C = (c_0, \ldots, c_{t-1})$ be a minimum-length (sn, R, a)-de Bruijn covering code. We construct an (n, R, a^s)-de Bruijn covering code $C' = (c'_0, \ldots, c'_{m-1})$ of length m. Choose some bijection σ between $(\mathbb{Z}/a\mathbb{Z})^s$ and $\mathbb{Z}/a^s\mathbb{Z}$, and define

$$c'_j = \sigma(c_{|sj \bmod (m/s)|}, \ldots, c_{|s(j+1)-1 \bmod (m/s)|})$$

with indices on the left-hand side taken modulo m and indices on the right hand side taken modulo t. Evidently, C' is well defined, since $s|m$. Now, suppose $X = (x_0, \ldots, x_{n-1})$ is an n-string over a^s symbols. We claim that there is some codeword in the set of consecutive n-strings of C' which is within R symbols of x.

Indeed, let $x'_j = \sigma^{-1}(x_j)$ for $0 \leq j < n$ and define $X' = x'_0 \cdots x'_{sn-1}$, a string of length sn. Then some string X'' which differs from X' in at most R symbols occurs somewhere in C, say, beginning at coordinate k. X'' must occur at least s times in C', at coordinates $k + jm/s$ for $0 \leq j < s$. (If X'' "wraps around" in C, the extra $\geq sn - 1$ symbols at the end of each block of length m/s guarantee X'' appears in C'.) Furthermore, since $(m/s, s) = 1$, the numbers $k + jm/s$, $0 \leq j < s$, represent all residue classes modulo s, so there is some r so that $k + rm/s \equiv 0 \bmod s$. Then the string

$$\sigma^{-1}(c'_{k+rm/s}, \ldots, c'_{k+rm/s+s-1}) \ldots \sigma^{-1}(c'_{k+rm/s+(n-1)s}, \ldots, c'_{k+rm/s+ns-1})$$

appears in C and at most R of its coordinates differ from those of X. ∎

The most obvious question arising from the subject of the present work is the issue of whether the bound stated in Theorem 1 is best possible, i.e., whether the log factor can be dropped or the result can be extended to q's which are not prime powers. Another way to state this is to ask whether

$$\limsup_{n \to \infty} \frac{M(n, R, q) n^R}{q^n} < \infty.$$

On the other hand, it is an interesting question to ask what happens when the lim sup is replaced with a lim inf. In fact, we *do* know that

$$\liminf_{n\to\infty} \frac{M(n, R, q)n^R}{q^n} < \infty,$$

by the following construction. Suppose $q = 2$, $R = 1$, $n = 2^m - 1$, and n is prime. Choose a Hamming code of length n in cyclic form, and partition the codewords into orbits under cyclic shift. Every such orbit of size n can be written as a string of length $2n - 1$ with the words from the orbit appearing as consecutive substrings. Concatenating all these strings gives the desired de Bruijn covering code. To generalize it to arbitrary R, we may use an R-fold interleaving of R Hamming codes. To make the construction work when n is not necessarily prime, it is possible to prove that most of the orbits are still of size proportional to n. For arbitrary q, we simply use q-ary cyclic Hamming codes. Of course, this provides a construction for only very rare examples of n's, and so does not help with the general question of the upper bound.

ACKNOWLEDGMENTS

We thank the two anonymous referees for their helpful corrections and suggestions, particularly the construction at the end of Section 5.

REFERENCES

[1] N. Alon and J. Spencer, The probabilistic method, Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley-Interscience, New York, 2000.

[2] T. B. Beard, Jr., Matrix fields, regular and irregular: a complete fundamental characterization, Linear Algebra Appl 81 (1986), 137–152.

[3] H. Fredricksen, A survey of full length nonlinear shift register cycle algorithms, SIAM Rev 24(2) (1982), 195–221.

[4] M. Krivelevich, B. Sudakov, and V. H. Vu, Covering codes with improved density, IEEE Trans Inform Theory 49(7) (2003), 1812–1815.

[5] M. Landsberg, Feedback functions for generating cycles over a finite alphabet, Discrete Math 219(1–3) (2000), 187–194.

[6] S. Litsyn, Table of the best currently known lower and upper bounds on the smallest size of a covering code, manuscript, http://www.eng.tau.ac.il/~litsyn/tablecr/index.html.

[7] W. V. Parker, The matrix equation $AX = XB$, Duke Math J 17 (1950), 43–51.