Combinatorial Batch Codes

M. B. Paterson*
Department of Mathematics
Royal Holloway, University of London
Egham, Surrey TW20 0EX, U.K.
m.b.paterson@rhul.ac.uk

D. R. Stinson†
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, ON, N2L 3G1, Canada
dstinson@uwaterloo.ca

R. Wei‡
Department of Computer Science
Lakehead University
Thunder Bay ON, P7B 5E1, Canada
rwei@lakeheadu.ca

July 6, 2008

Abstract

In this paper, we study batch codes, which were introduced by Ishai, Kushilevitz, Ostrovsky and Sahai in [3]. A batch code specifies a method to distribute a database of n items among m devices (servers) in such a way that any k items can be retrieved by reading at most t items from each of the servers. It is of interest to devise batch codes that minimize the total storage, denoted by N, over all m servers.

In this paper, we study the special case $t = 1$, under the assumption that every server stores a subset of the items. This is purely a combinatorial problem, so we call this kind of batch code a "combinatorial batch code". For various parameter situations, we are able to present batch codes that are optimal with respect to the storage requirement, N. We also study uniform codes, where every item is stored in precisely c of the m servers (such a code is said to have rate $1/c$). Interesting new results are presented in the cases $c = 2, k - 2$ and $k - 1$. In addition, we obtain improved existence results for arbitrary fixed c using the probabilistic method.

1 Introduction

Ishai, Kushilevitz, Ostrovsky and Sahai [3] have shown that problems connected with reducing the computational overhead of private information retrieval can be related to the question of how to distribute a database of n items among m devices (servers) so that any k items can be retrieved by reading at most t items from each of the servers [3]. This leads naturally to the concept of a *batch code*, which they define as follows.

*Research supported by EPSRC grant EP/D053285/1
†Research supported by NSERC discovery grant 203114-06
‡Research supported by NSERC discovery grant 239135-06

Definition 1.1. *An* (n, N, k, m, t) batch code *over an alphabet* Σ *encodes a string* $x \in \Sigma^n$ *into an* m*-tuple of strings* $y_1, y_2, \ldots, y_m \in \Sigma^*$ *(also referred to as* servers*) of total length* N*, such that for each* k*-tuple (*batch*) of distinct indices* $i_1, i_2, \ldots, i_k \in \{1, \ldots, n\}$*, the entries* $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ *from* x *can be decoded by reading at most* t *symbols from each server.*

In general, we want N to be as small as possible, given n, m, k and t. It is often useful to study the *rate* of the code, which is defined to be the ratio n/N. A large rate is a desirable property of a batch code.

In this paper we consider batch codes for which the decoding is simply reading; these are referred to as *replication-based* batch codes in [3]. In this case, each server can be represented as a subset of the alphabet set, so the problem of constructing such codes falls naturally within a combinatorial framework. We call these codes "combinatorial batch codes" and define them as follows.

Definition 1.2. *An* (n, N, k, m, t) combinatorial batch code (CBC) *is a set system* $(X, \mathcal{B})$*, where* X *is a set of* n *elements (called* items*),* $\mathcal{B}$ *is a collection of* m *subsets of* X *(called* servers*) and* $N = \sum_{B \in \mathcal{B}} |B|$*, such that for each* k*-subset* $\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\} \subset X$ *there exists a subset* $C_i \subseteq B_i$*, where* $|C_i| \leq t, i = 1, \ldots, m$*, such that*

$$\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\} \subset \bigcup_{i=1}^{m} C_i.$$

A feature of a batch code mentioned in [3] is the *load balancing* property, which is captured by the parameter t. Basically, this says that the k desired items can be recovered by reading (at most) t items from each server; thus the load on the servers is balanced (and upper-bounded) by the parameter t.

We will only consider the case $t = 1$ in this paper; such a batch code permits only one item to be retrieved from each server. Since we are fixing $t = 1$, we will omit the parameter t, and denote the batch code as an (n, N, k, m)-CBC.

It is noted in [3] that an (n, N, k, m)-CBC is equivalent to an unbalanced expander with expansion factor 1. However, most existing literature on expanders is concerned with asymptotic behaviour, while in this paper we focus on exact combinatorial constructions and bounds. Furthermore, most research on expanders concerns expansion factors exceeding 1. Thus our work has a considerably different flavour.

Two trivial cases of combinatorial batch codes are:

1. Each server has a copy of X. In this case, k servers are needed and $N = kn$.

2. Each server stores only one item. In this case, n servers are needed and $N = n = m$.

Therefore, we are only interested in (n, N, k, m) combinatorial batch codes with $N < kn$ and $N > n$.

1.1 Representations of Batch Codes

We have defined combinatorial batch codes to be set systems whose points represent the items in a database, with the servers being represented by subsets of these points. Throughout this paper it will frequently be convenient to consider instead the *dual set system*, in which the servers are represented by points, and each item in the database is represented by a set (referred to as a *block*) containing the points (i.e., the servers) that store that item. It is permitted for the dual set system

to contain "repeated blocks"; this will happen if two (or more) items are assigned to the same set of servers.

Given a set system $(X, \mathcal{B})$ with $X = \{x_1, x_2, \ldots, x_v\}$ and $\mathcal{B} = \{B_1, B_2, \ldots, B_b\}$, the *incidence matrix* of $(X, \mathcal{B})$ is the $b \times v$ matrix $A = (a_{i,j})$, where

$$a_{i,j} = \begin{cases} 1 & \text{if } x_j \in B_i \\ 0 & \text{if } x_j \notin B_i. \end{cases}$$

Conversely, given an incidence matrix, we can define an associated set system in the obvious way.

The blocks of the dual set system of an (n, N, k, m)-CBC are represented by the columns of the corresponding incidence matrix, with the points of the j^{th} block corresponding to rows i for which the matrix entry $a_{i,j} = 1$. The incidence matrix of an (n, N, k, m)-CBC is thus an $m \times n$ matrix with entries in $\{0, 1\}$, such that for any k columns, there are k rows such that the resulting $k \times k$ submatrix has at least one transversal containing only 1s (*i.e.*, a set of k cells in different rows and different columns that all contain the entry 1). So we have the following lemma.

Lemma 1.1. *An $m \times n$ 0-1 matrix is an incidence matrix of an (n, N, k, m)-CBC if and only if, for any k columns, there is a $k \times k$ submatrix which has at least one transversal containing k ones.*

This result leads directly to the following version of Hall's marriage theorem, which is described in [3] using the language of matchings in a bipartite graph.

Lemma 1.2 (Hall's marriage theorem)**.** *Suppose $(X, \mathcal{B})$ is a set system. Then any collection of k blocks $B_1, B_2, \ldots, B_k \in \mathcal{B}$ has a system of distinct representatives if and only if the following condition, denoted $\mathsf{SDR}(i)$, is satisfied for all i, $1 \le i \le k$:*

$$\text{for any collection of } i \text{ blocks } B_{j_1}, B_{j_2} \ldots, B_{j_i}, \left| \bigcup_{l=1}^{i} B_{j_l} \right| \ge i. \tag{1}$$

Example 1.1. *Here is the incidence matrix of a $(7, 15, 5, 5)$-CBC:*

1	0	0	0	0	1	1
0	1	0	0	0	1	1
0	0	1	0	0	1	1
0	0	0	1	0	1	1
0	0	0	0	1	1	1

The set system represented by this incidence matrix is

$$\{\{1, 6, 7\}, \{2, 6, 7\}, \{3, 6, 7\}, \{4, 6, 7\}, \{5, 6, 7\}\},$$

and the dual set system is

$$\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{1, 2, 3, 4, 5\}, \{1, 2, 3, 4, 5\}\}.$$

For $1 \le i \le 5$, it is easy to see that any i blocks of the dual system contain at least i points. Therefore, we have a CBC with $k = 5$.

1.2 Our Contributions

For various parameter situations, we are able to present batch codes that are optimal with respect to the storage requirement, N. In Section 2.1, we study codes with small and large values of m, while in Section 2.2 we determine optimal batch codes when n is sufficiently large. In Section 3, we study codes where every item is stored in precisely c of the m servers (such a code is said to have rate $1/c$). Interesting new results are presented in the cases $c = 2, k-2$ and $k-1$. In addition, we obtain improved existence results for arbitrary fixed c using the probabilistic method.

2 Combinatorial Batch Codes with Minimal Total Storage

The parameter N in an (n, N, k, m)-CBC represents the total amount of information collectively stored by all the servers. Hence, given n, k and m, we would like to find combinatorial batch codes for which N is as small as possible. We say that an (n, N, k, m)-CBC is *optimal* if $N \leq N'$ for all (n, N', k, m)-CBC and we denote the corresponding value of N by $N(n, k, m)$. We would like to determine $N(n, k, m)$ for all $k > 1$ and all m, n with $k \leq m \leq n$.

2.1 Batch Codes with Minimum and Maximum Values of m

In this section, we obtain optimal solutions for the special cases $m = k, n-1$ and n. We also give a construction that applies when m is a bit smaller than n.

The first of these cases is trivial; it was already mentioned in Section 1.

Theorem 2.1. $N(n, k, n) = n$.

The case of $m = k$ is also not difficult. We give a construction that generalizes Example 1.1, and prove that the construction is optimal.

Theorem 2.2. $N(n, k, k) = kn - k(k-1)$.

Proof. Let $B_j = \{x_j\}$ for $j = 1, 2, \ldots, k$, and let $B_j = \{x_1, x_2, \ldots, x_k\}$ for $k+1 \leq j \leq n$. Then it is easy to check that $\{B_1, \ldots, B_n\}$ is the dual set system of an (n, N, k, k)-CBC. On the other hand, if $N < k(n-k+1)$, then at least one server contains $n-k$ or fewer items, so the k items missing from that server cannot be recovered. □

We now describe a construction that can be applied when n is not too much bigger than m. For positive integers k and p, we define a graph that we term a (k, p)-*flying saucer*, which we denote as (k, p)-FS. For simplicity, we first assume that $k \equiv 2 \bmod 3$.

We begin by constructing p paths of length $(k+1)/3$. These paths should all have the same two endpoints, say x and y, but otherwise they are vertex-disjoint. Then attach paths of length $(k-2)/3$ to both x and y. The other endpoint of the path having endpoint x (y, resp.) is denoted u (v, resp.).An example of a flying saucer is given in Figure 1.

Some basic properties of flying saucers are given in the next lemma.

Lemma 2.3. *Suppose $k \equiv 2 \bmod 3$. Then the following properties hold:*

1. *A (k, p)-FS contains*

$$\nu(k, p) = \frac{(p+2)(k-2)}{3} + 2$$

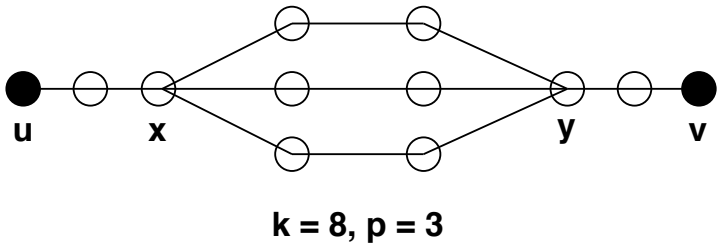

Figure 1: An $(8, 3)$-flying saucer

vertices and

$$\epsilon(k,p) = \nu(k,p) + p - 2 = \frac{p(k+1)}{3} + \frac{2(k-2)}{3}$$

edges.

2. *The distance (i.e., the length of the shortest path) between the two vertices of degree one (namely, u and v) is $k-1$.*

3. *A connected subgraph of a (k,p)-FS that contains a cycle and a vertex of degree one (i.e., one of x or y) has at least k edges (such a subgraph would contain the path from u to x together with two paths from x to y; or the path from v to y together with two paths from y to x).*

4. *A connected subgraph of a (k,p)-FS that contains two cycles has at least $k+1$ edges (such a subgraph would contain three paths from x to y).*

Now we can state our construction and prove that it yields a certain class of CBC.

Theorem 2.4. *Suppose that k and p are positive integers, and $k \equiv 2 \bmod 3$. Define $\nu(k,p)$ and $\epsilon(k,p)$ as in Lemma 2.3. Suppose that $m \geq \nu(k,p)$. Then there exists an $(m+p, m+p+\epsilon(k,p), k, m)$-CBC.*

Proof. First, construct a (k,p)-FS. Then add a sufficient number of isolated vertices so the resulting graph, say G, contains m vertices. Now we construct an $m \times n$ incidence matrix A whose rows are labelled by the vertices of G. For every edge st in G, construct a column of A that has 1s in rows s and t. Then, for every vertex s of G that is either a vertex of degree one (i.e., $s = u$ or v) or an isolated vertex, construct a column of A that has a 1 in row s. The resulting dual set system consists of blocks of size two (corresponding to edges of G) and blocks of size one (corresponding to vertices of G having degree zero or one).

Suppose there is a set of $i \leq k$ blocks of the dual set system that spans fewer than i points. We can ignore blocks corresponding to isolated vertices in G. A bit of thought shows that there are three cases we need to consider:

1. The set of i blocks contains the blocks corresponding to both of the vertices u and v. It is easy to check that we would need to include blocks corresponding to all the edges in a path from u to v, but then we would have at least $2 + k - 1 = k + 1$ blocks, from property 2 of Lemma 2.3. This yields a contradiction.

2. The set of i blocks contains exactly one of the two blocks corresponding to the vertices u and v. Without loss of generality, suppose that we include the block corresponding to u and omit the block corresponding to v. It is easy to check that we would need to include blocks corresponding to all the edges in a path from u to x as well as the edges in two paths from x to y, but then we would have at least $1 + k = k + 1$ blocks, from property 3 of Lemma 2.3. This yields a contradiction.

3. The set of i blocks contains neither of the blocks corresponding to the vertices u or v. It is easy to check that we would need to include blocks corresponding to all the edges in three paths from x to y, but then we would have at least $k + 1$ blocks, from property 4 of Lemma 2.3. This yields a contradiction.

These cases cover all the possibilities, so we have proved that we have a CBC with the desired value of k. □

Here is a small example to illustrate the construction.

Example 2.1. *Taking $k = 8$, $p = 3$ and $m = 12$ in Theorem 2.4, we construct a $(15, 28, 8, 12)$-CBC. We start with an $(8, 3)$-FS, as depicted in Figure 1. The incidence matrix of the desired CBC is as follows:*

1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	1	0	0	1	1	0	0	0
0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	0	0

Suppose we fix k and m, and consider how N behaves as n takes on the values $m+1, m+2, \ldots$. From the formulas in Theorem 2.4, we see that we add $(k + 1)/3$ to N every time we add 1 to n.

Theorem 2.4 only considered the case when $k \equiv 2 \bmod 3$. However, similar constructions can be used when $k \equiv 0, 1 \bmod 3$. Basically, what we do is adjust the lengths of the paths in the flying saucer, taking care to ensure that properties 2, 3, and 4 of Lemma 2.3 are satisfied. Suppose that the path from u to x has length a, the path from y to v has length b, and the p paths from x to y have lengths $c_1, \ldots, c_p$. We choose the values $a, b, c_1, \ldots, c_p$ as follows:

- When $k \equiv 0 \bmod 3$, let $a = k/3 - 1$, $b = k/3$, $c_1 = k/3$, $c_2 = \cdots = c_p = k/3 + 1$.
- When $k \equiv 1 \bmod 3$, let $a = b = (k - 1)/3$, $c_1 = (k - 1)/3$, $c_2 = \cdots = c_p = (k + 2)/3$.

The resulting construction has behaviour similar to the case $k \equiv 2 \bmod 3$; we omit the details.

In the case $n = m + 1$, we can show that our construction is optimal, for all values of k.

Theorem 2.5. $N(m+1,k,m) = m+k$.

Proof. First, we observe that, when we let $p = 1$ in Theorem 2.4, we obtain an $(m+1, m+k, k, m)$-CBC when $k \equiv 2 \bmod 3$. In fact, an $(m+1, m+k, k, m)$-CBC exists for all positive integers k (a flying saucer is always a path of length $k-1$ when $p = 1$). Therefore $N(m+1,k,m) \leq m+k$.

To complete the proof, we show that $N(m+1,k,m) \geq m+k$. Let A be the incidence matrix of an $(m+1, m+j, k, m)$-CBC. Suppose the rows of A are ordered in nonincreasing order of weight, and let B consist of the first j rows of A. Denote the remaining $m-j$ rows of A by C. Let w_B (w_C, resp.) denote the number of 1s in B (C, resp.). Then it is clear that $w_B \geq 2j$ and $w_C \leq m-j$.

The maximum number of non-zero columns in C is w_C, so C contains an $(m-j) \times (j+1)$ matrix of 0s, say C'. Let B' denote the $j \times (j+1)$ submatrix of B that has the same columns as C'. The blocks of the dual design corresponding to the $j+1$ columns in B' contain at most j points. If $j \leq k-1$, then we have a contradiction, so we conclude that $j \geq k$. □

2.2 Batch Codes for Large Values of n

We observe that it is never necessary for any of the blocks in the dual set system to contain more than k points. For, if we consider any k columns of the incidence matrix, a $(k-1)$-transversal of any $k-1$ of the columns can always be completed to a k-transversal of all k columns provided the final column contains k ones. Thus, in seeking to determine $N(n,k,m)$ we can restrict our attention to (n, N, k, m) combinatorial batch codes in which each block of the dual set system has at most k points. For sufficiently large n we have the following construction.

Construction 2.6. *For $n \geq (k-1)\binom{m}{k-1}$ we can construct an $(n, kn-(k-1)\binom{m}{k-1}, k, m)$-CBC as follows.*

- *Let the first $(k-1)\binom{m}{k-1}$ blocks of the dual set system consist of $k-1$ copies of each possible set of $k-1$ points.*
- *Each remaining block of the dual set system can be taken to be any set of k points.*

Theorem 2.7. *Construction 2.6 gives rise to an (n, N, k, m)-CBC with $N = kn-(k-1)\binom{m}{k-1}$.*

Proof. According to Construction 2.6 each block contains either k or $k-1$ points, so the value of N is equal to kn minus the number of blocks that contain $k-1$ points, that is, $N = kn-(k-1)\binom{m}{k-1}$.

As each block contains at least $k-1$ points, then, for $i = 1, \ldots, k-1$, the union of any i blocks contains at least i points. Furthermore, the union of any k blocks must contain at least k points, as each set of $k-1$ points gives rise to just $k-1$ blocks that include no further points. Hence, by Lemma 1.2, the set system defined in Construction 2.6 is a (n, N, k, m)-CBC. □

This construction thus yields an upper bound for $N(n,k,m)$.

Corollary 2.8. *For $n \geq (k-1)\binom{m}{k}$, we have $N(n,k,m) \leq kn-(k-1)\binom{m}{k-1}$.*

Theorem 2.9. *If $n \geq (k-1)\binom{m}{k-1}$, then $N(n,k,m) = kn-(k-1)\binom{m}{k-1}$.*

Proof. Let $(\mathcal{X}, \mathcal{B})$ be an (n, N, k, m)-CBC. Let M be an $\binom{m}{k-1} \times n$ matrix whose columns are indexed by the blocks of the dual set system, and whose rows are indexed by all possible subsets of $k-1$ points, with a 1 in position M_{ij} if the j^{th} block is a subset of the i^{th} set, and a 0 otherwise. Counting

the number of nonzero entries in this matrix will allow us to bound the number of blocks of the dual set system that contain fewer than k points.

Each row has at most $k-1$ ones, by Lemma 1.2, so the total number of entries of M that are 1s is at most $(k-1)\binom{m}{k-1}$. If a block B contains $i < k-1$ points, then the corresponding column has $\binom{m-i}{k-1-i}$ entries that are 1s; the column corresponding to a block with $k-1$ points has one 1, and that corresponding to a block with k points has none. For $i = 1, \ldots, k-1$, let A_i denote the number of blocks containing precisely i points. Then the total number of entries of M that are 1s is equal to

$$A_{k-1} + \sum_{i=1}^{k-2} \binom{m-i}{k-1-i} A_i.$$

Hence we have

$$A_{k-1} \leq (k-1)\binom{m}{k-1} - \sum_{i=1}^{k-2} \binom{m-i}{k-1-i} A_i. \tag{2}$$

Now, we have that

$$N = \sum_{i=1}^{k} iA_i = \sum_{i=1}^{k} (k-(k-i))A_i = kn - \sum_{i=1}^{k-1} (k-i)A_i.$$

Combining this with (2), we see that

$$N \geq kn - \sum_{i=1}^{k-2} (k-i)A_i - \left((k-1)\binom{m}{k-1} - \sum_{i=1}^{k-2} \binom{m-i}{k-1-i} A_i \right) \tag{3}$$

$$= kn - (k-1)\binom{m}{k-1} + \sum_{i=1}^{k-2} \left(\binom{m-i}{k-1-i} - (k-i) \right) A_i. \tag{4}$$

As $m \geq k$, the coefficients of the A_i in this expression are all non-negative, hence it is minimised by setting $A_i = 0$ for $i = 1, 2, \ldots, k-2$. Together with Corollary 2.8 this gives the desired result. □

In the case where $m < n < (k-1)\binom{m}{k-1}$, it may be possible to achieve smaller values of $N(n, m, k)$. Construction 2.6 gives an upper bound of $N(m, k) \leq (k-1)n$ for $N(n, m, k)$, but this is not tight for all values of n in this range.

3 Batch Codes with Fixed Rate

In Section 2 we considered the problem of how to construct combinatorial batch codes with small values of N. Another interesting question is that of how to construct batch codes in which the rate n/N is large. For fixed k and m, Theorem 2.9 shows that as $n \to \infty$ the optimal rate of an (n, N, k, m)-CBC approaches k. However, if n/N and k are fixed, we would like to know the largest value of n (as a function of m) for which we can construct a (n, N, k, m)-CBC.

In [3], several constructions of batch codes are given. For example, they constructed batch codes (from *unbalanced expander graphs*) having rate $= 1/d < 1/2$ and $m = O(k \cdot (nk)^{1/(d-1)})$, as well as codes having rate $= \Omega(1/\log n)$ and $m = O(k)$.

In this section, we consider *uniform* batch codes with fixed rate. These are CBCs in which every block of the dual set system contains precisely c points, where $1/c$ is the rate of the CBC. That is, every item is stored in exactly c servers. We denote by $n(m, c, k)$ the maximum value of n for which there exists a uniform (n, cn, k, m)-CBC; we are interested in determining this value for various combinations of c and k. From an application point of view, a large rate is desirable. Hence, we focus in particular on the case $c = 2$; however, we also mention precise results that can be obtained in the cases $c = k - 1$ and $c = k - 2$.

3.1 Batch Codes of Rate $\frac{1}{k-1}$ and $\frac{1}{k-2}$

First, we prove a general lower bound.

Theorem 3.1.

$$n(m, c, k) \leq \frac{(k-1)\binom{m}{c}}{\binom{k-1}{c}}.$$

Proof. As in the proof of Theorem 2.9, we take M to be the matrix whose columns are indexed by the blocks of the dual set system, and whose rows are indexed by all possible subsets of $k-1$ points, with a 1 in position M_{ij} if the j^{th} block is a subset of the i^{th} set, and a 0 otherwise. Each row has at most $k-1$ ones. Hence, the total number of 1s is at most $(k-1)\binom{m}{k-1}$. Each column has precisely $\binom{m-c}{k-1-c}$ ones, so we have that $n\binom{m-c}{k-1-c} \leq (k-1)\binom{m}{k-1}$. This implies that

$$\begin{aligned}\frac{n}{k-1}\binom{k-1}{c} &\leq \frac{\binom{m}{k-1}\binom{k-1}{c}}{\binom{m-c}{k-1-c}}, \\ &= \binom{m}{c}.\end{aligned}$$

□

Theorem 3.2. *For any m and $c < m$ there exists a uniform $\left(c\binom{m}{c}, c^2\binom{m}{c}, c+1, m\right)$-CBC with rate $1/c$.*

Proof. Take the CBC arising from Construction 2.6 with $k = c + 1$ and $n = c\binom{m}{c}$. It is easy to observe that this batch code is uniform. □

Corollary 3.3. $n(m, c, c+1) = c\binom{m}{c}$.

Proof. This follows immediately from Theorem 3.1 (with $k = c + 1$) and Theorem 3.2. □

Theorem 3.4. *For any m and $c < m$, let $\mathcal{B}$ be all the c-subsets of an m-set X. Then $(X, \mathcal{B})$ is the dual set system of a uniform $\left(\binom{m}{c}, c\binom{m}{c}, c+2, m\right)$-CBC with rate $1/c$.*

Proof. It is easy to verify that i blocks span at least $c+1$ points, for $2 \leq i \leq c+1$, and $c+2$ blocks span at least $c+3$ points. Therefore the result follows from from Lemmas 1.1 and 1.2. □

Corollary 3.5. $n(m, c, c+2) = \binom{m}{c}$.

Proof. This follows immediately from Theorem 3.1 (with $k = c + 2$) and Theorem 3.4. □

3.2 Batch Codes of Rate 1/2

For a uniform CBC with rate 1/2, every block of the dual set system has precisely two points. This means that we can view the blocks as the edges of a multigraph whose vertices are the points of the dual set system. It is not difficult to see that a multigraph related to a batch code with parameter k has the property that the graph does not contain any subgraph with i edges and fewer than i vertices, where $i \leq k$.

The case $k = 4$ follows immediately from results proven in the previous section.

Theorem 3.6. *For all positive integers m, there exists a uniform $\left(\frac{m(m-1)}{2}, m(m-1), 4, m\right)$-CBC with rate 1/2. Furthermore, $n(m, 2, 4) = \binom{m}{2}$.*

Proof. Apply Theorem 3.4 and Corollary 3.5 with $c = 2$. □

In the next lemma, we show that a graph of specified girth yields a uniform CBC. (The *girth* of a multigraph is the length of the shortest cycle in the graph. A multigraph containing at least one repeated edge has girth equal to two.)

Lemma 3.7. *If there is a graph G with m vertices, n edges and girth g, then there is a uniform $(n, 2n, k, m)$-CBC with $k = 2g - \lfloor g/2 \rfloor - 1$ and rate $= 1/2$.*

Proof. Suppose there exists a subgraph H of G having i edges that span fewer than i vertices. Then H contains at least two cycles. However, in a graph with girth g, two cycles have at most $\lfloor g/2 \rfloor$ common edges. □

Corollary 3.8. *For all integers $m \geq 2$, there is a uniform $(\lceil (m^2-1)/4 \rceil, 2\lceil (m^2-1)/4 \rceil, 5, m)$-CBC with rate 1/2.*

Proof. A bipartite graph has girth at least 4. We obtain the desired CBC from a complete bipartite graph $K_{\lceil \frac{m}{2} \rceil, \lfloor \frac{m}{2} \rfloor}$. □

We now show that this construction yields CBCs that are very close to optimal.

Theorem 3.9. *If there is a uniform $(n, 2n, 5, m)$-CBC (with rate 1/2), then $n \leq \lceil (m^2 + 2m - 3)/4 \rceil$.*

Proof. Suppose first that the CBC is a simple graph. It is easy to see that a simple graph G is an CBC with $k = 5$ if and only if no subgraph of G is isomorphic to $K_4 - e$, where e is an edge of the K_4 in question. Then an extension of Turán's theorem due to Dirac ([2]) implies that $n \leq \lceil (m^2-1)/4 \rceil$.

Now suppose that G contains one or more multiple edges. G cannot contain any edge of multiplicity three, nor can G contain two adjacent edges of multiplicity two. It follows that the deletion of at most $\lceil (m-1)/2 \rceil$ edges from G yields a simple graph G' that is an CBC with $k = 5$. Therefore $n \leq \lceil (m^2-1)/4 \rceil + \lceil (m-1)/2 \rceil = \lceil (m^2 + 2m - 3)/4 \rceil$. □

Combining Corollary 3.8 and Theorem 3.9, we obtain the following.

Theorem 3.10. $\lceil (m^2-1)/4 \rceil \leq n(m, 2, 5) \leq \lceil (m^2 + 2m - 3)/4 \rceil$.

Margulis [5] and Lubotzky et al. [4] have constructed d-regular graphs G with the following parameters:

$$g \geq \frac{4}{3}\frac{\log m}{\log(d-1)} - \frac{\log 4}{\log(d-1)},$$

where g is the girth, $d-1$ is any prime $p \equiv 1 \pmod 4$ and m is the number of vertices. So we have the following construction.

Theorem 3.11. *There exists a uniform $(dm/2, dm, 2\log m/\log(d-1), m)$-CBC, where $d-1 \equiv 1 \pmod 4$ is a prime.*

3.3 Batch codes of rate $1/c$ for arbitrary specified values of c

In this section, we give an existence result using the probabilistic method. It is similar to a theorem proved in [1, pp. 59–61].

Theorem 3.12. *For all integers $c \geq 2$ and all integers $k \geq 2$, there is a positive constant $a_{c,k}$ depending on c and k, such that there exists a uniform (n, cn, k, m)-CBC with $n \geq a_{c,k} m^{ck/(k-1)-1}$, having rate $1/c$.*

As a warm-up, we prove a special case of Theorem 3.12 to illustrate the main ideas. Suppose that $c = 2$ and $k = 5$. We will construct a graph that satisfies $\mathsf{SDR}(i)$ for all $i \leq 5$. It is easy to see that this is equivalent to saying that the graph contains no subgraph isomorphic to $K_4 - e$, where e is an edge of the K_4 in question.

Construct a random graph G on m vertices and t edges. Suppose the set of edges in G are denoted by $E = \{e_1, \ldots, e_t\}$. For a subset of edges $F \subseteq E$, $|F| = 5$, a random variable X_F is defined as

$$X_F = \begin{cases} 1 & \text{if } F \text{ is isomorphic to } K_4 - e \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that

$$E[X_F] = Pr[X_F = 1] = \frac{6\binom{m}{4}}{\binom{\binom{m}{2}}{5}}.$$

Defining

$$X = \sum_{F \subseteq E, |F|=5} X_F,$$

we have

$$E[X] = \frac{6\binom{m}{4}\binom{t}{5}}{\binom{\binom{m}{2}}{5}}.$$

Suppose that

$$\frac{6\binom{m}{4}\binom{t}{5}}{\binom{\binom{m}{2}}{5}} \leq \frac{t}{2}. \tag{5}$$

Then we can delete at most $t/2$ edges form G and obtain a graph G' that contains no subgraph isomorphic to $K_4 - e$. Hence, there will exist a $(t/2, t, 5, m)$-CBC having rate $1/2$.

Now, it is easy to see that $\binom{m}{4} \leq m^4/24$ and

$$\frac{\binom{t}{5}}{\binom{\binom{m}{2}}{5}} \leq \left(\frac{t}{\binom{m}{2}}\right)^5.$$

Hence, (5) will hold provided that

$$6 \times \frac{m^4}{24} \times \left(\frac{t}{\binom{m}{2}}\right)^5 \leq \frac{t}{2}. \tag{6}$$

The inequality (6) is equivalent to

$$t^4 \leq \frac{m(m-1)^5}{16}. \tag{7}$$

Finally, (7) holds provided that

$$t \leq \frac{(m-1)^{3/2}}{2}. \tag{8}$$

Therefore, there is a $(n, 2n, 5, m)$-CBC having rate $1/2$. where

$$n \geq \left\lfloor \frac{(m-1)^{3/2}}{4} \right\rfloor.$$

Note that this result is considerably weaker than the result we already proved in Corollary 3.8.

Now let's turn to the general case. It is not hard to see that any collection of distinct c-subsets automatically satisfies $\mathsf{SDR}(i)$ for $i = 1, \ldots, c+2$. Therefore, we need to ensure that $\mathsf{SDR}(i)$ holds for $i = c+3, \ldots, k$. As above, we will construct a random c-hypergraph on a set of m points having t edges.

Suppose the set of edges (blocks) in G are denoted by $E = \{e_1, \ldots, e_t\}$. For a subset of edges $F \subseteq E$, $c+3 \leq |F| \leq k$, define a random variable X_F as follows:

$$X_F = \begin{cases} 1 & \text{if } F \text{ spans fewer than } |F| \text{ points} \\ 0 & \text{otherwise.} \end{cases}$$

Suppose $|F| = i$. Then it is easy to see that

$$E[X_F] = Pr[X_F = 1] \leq \frac{a_i \binom{m}{i-1}}{\binom{\binom{m}{c}}{i}},$$

where the constant a_i denotes the number of ways to construct a c-hypergraph consisting of i blocks on a fixed set of $i-1$ points.

Defining

$$X_i = \sum_{F \subseteq E, |F| = i} X_F,$$

we have

$$E[X_i] \leq \frac{a_i \binom{m}{i-1}\binom{t}{i}}{\binom{\binom{m}{c}}{i}}.$$

Finally, defining

$$X = \sum_{i=c+3}^{k} X_i,$$

we have

$$\begin{aligned} E[X] &\leq \sum_{i=c+3}^{k} \frac{a_i \binom{m}{i-1}\binom{t}{i}}{\binom{\binom{m}{c}}{i}} \\ &\leq \sum_{i=c+3}^{k} \left(a_i \times \frac{m^{i-1}}{(i-1)!} \times \left(\frac{t}{\binom{m}{c}} \right)^i \right) \\ &\leq \sum_{i=c+3}^{k} \frac{a_i\, m^{i-1}\, t^i\, (c!)^i}{(i-1)!\,(m-c+1)^{ci}}. \end{aligned}$$

Define

$$A = \max \left\{ \frac{a_i (c!)^i}{(i-1)!} : c+3 \leq i \leq k \right\}.$$

Then

$$E[X] \leq A \sum_{i=c+3}^{k} \frac{m^{i-1}\, t^i}{(m-c+1)^{ci}}$$

Observe that the above sum is a geometric sequence with ratio $r = mt/(m-c+1)^c$. Suppose that $r \geq 1$, i.e., $t \geq (m-c+1)^c/m$. Then

$$E[X] \leq A(k-c-2) \frac{m^{k-1}\, t^k}{(m-c+1)^{ck}}. \tag{9}$$

Let $A' = A(k-c-2)$ and suppose that

$$\frac{A' m^{k-1}\, t^k}{(m-c+1)^{ck}} \leq \frac{t}{2}. \tag{10}$$

Then (9) implies that $E[X] \leq t/2$, so there will exist a $(t/2, ct/2, k, m)$-CBC.

It remains to compute a bound on t from (10):

$$2A'\, t^{k-1} \leq (m-c+1)^{ck}\, m^{-k+1},$$

which simplifies to

$$t \leq A''\, m^{ck/(k-1)-1}$$

for some constant A''.

3.4 Comparison

Ignoring constants, we have shown in Theorem 3.12 the existence of CBC with rate $1/c$ in which n is $\Omega(m^{ck/(k-1)-1})$. This compares favourably with the result in [3] where n is $\Omega(m^{c-1})$.

In the case $c = 2$, we showed in Theorem 3.12 that n is $\Omega(m^{(k+1)/(k-1)})$ whereas [3] proved the weaker result that n is $\Omega(m)$. If we set $k \approx 2 \log m / \log d$ in Theorem 3.11, we obtain an CBC in which n is $\Omega(m^{(k+2)/k})$. This is better than the result in [3] but not as good as our Theorem 3.12. However, it should be noted that the graphs in Theorem 3.11 can be constructed explicitly, whereas our results are nonconstructive.

4 Summary

We have initiated a combinatorial study of batch codes. Many interesting problems remain to be settled. Here are three particularly interesting questions:

1. How close to being optimal are the constructions using flying saucers that given in Section 2.1? In particular, is it true that

$$N(m + p, k, m) - N(m + p - 1, k, m) \approx \frac{k}{3}$$

   when $p > 1$ and m is sufficiently large as a function of p and k?

2. Are there explicit constructions for "good" uniform batch codes with fixed rate $1/c$, where $c > 2$ is an integer?

3. Can $N(n, k, m)$ be computed for a range of values of n, where $n < (k - 1)\binom{m}{k-1}$?

References

[1] W. G. Brown, P. Erdös and V. T. Sós. Some extremal problems on r-graphs. In: *New directions in the theory of graphs (Proc. Third Ann Arbor Conf.)*, Academic Press, New York, 1973, pp. 53–63.

[2] G. Dirac. Extension of Turán's theorem on graphs. *Acta Math. Acad. Sci. Hungar* **14** 1963 417–422.

[3] Y. Ishai, E. Kushilevitz, R. Ostrovsky and A. Sahai. Batch codes and their applications. *Proceedings of STOC 2004*, ACM Press, pp. 262–271.

[4] A. Lubotzky, R. Philips and P. Sarnak. Ramanujan graphs. *Combinatorica* **8** (1988), 261–277.

[5] G. A. Margulis. Arithmetic groups and graphs without short cycles. In: *Proc. 6th. Int. Symp. on Information Theory, Tashkent 1984, Vol. 1*, pp. 123–125.