Complexity of Dependencies in Bounded Domains, Armstrong Codes, and Generalizations

Yeow Meng Chee, Hui Zhang, and Xiande Zhang
School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
email: {ymchee, huizhang, xiandezhang}@ntu.edu.sg

***Abstract*—The study of Armstrong codes is motivated by the problem of understanding complexities of dependencies in relational database systems, where attributes have bounded domains. A (q, k, n)-Armstrong code is a q-ary code of length n with minimum Hamming distance $n - k + 1$, and for any set of $k - 1$ coordinates there exist two codewords that agree exactly there. Let $f(q, k)$ be the maximum n for which such a code exists. In this paper, $f(q, 3) = 3q - 1$ is determined for all $q \geq 5$ with three possible exceptions. This disproves a conjecture of Sali. Further, we introduce generalized Armstrong codes for branching, or (s, t)-dependencies and construct several classes of optimal Armstrong codes in this more general setting.**

I. Introduction

Let A be a set of n *attributes*. Each attribute $x \in A$ is associated a set Ω_x, called its *domain*. A *relation* is a finite set R of n-tuples (called *data items*), such that $R \subseteq \times_{x \in A} \Omega_x$. A relational database *table* is an $m \times n$ array where each column is indexed by an *attribute* and each row corresponds to a data item in R. We denote this table by $R(A)$. More specifically, if $R = \{(d_{i,x})_{x \in A} : 1 \leq i \leq m\}$, then the cell in $R(A)$ with row index i and column index x has entry $d_{i,x}$. A *relational database* is a set of tables, where different tables may be defined over different attribute sets.

For a given table $R(A)$ and $X \subseteq A$, the *X-value* of a data item $d = (d_x)_{x \in A}$ in $R(A)$ is the $|X|$-tuple $d|_X = (d_x)_{x \in X}$. Let $X \subseteq A$ and $y \in A$ for a given table $R(A)$. We say that y *(functionally) depends* on X, written $X \to y$, if no two rows of $R(A)$ agree in X but differ in y. In other words, if the X-value of a data item is known, then its $\{y\}$-value can be determined with certainty. A *key* for $R(A)$ is a subset $K \subseteq A$, such that $K \to b$ for all $b \in A$. A key K is called *minimal* if no subset of K is a key.

Identifying functional dependencies, especially key dependencies, is important in relational database design [1]–[4]. From the schema design point of view, the question of whether a given collection Σ of functional dependencies has an *Armstrong instance* for Σ, that is, a table that satisfies a functional dependency $X \to y$ if and only if $X \to y$ is in Σ, is well studied. The existence of an Armstrong instance for any given set of functional dependencies was proved by Armstrong [1] and Demetrovics [5]. Further investigations (see for example, [6]) concentrated on the minimum size of an Armstrong instance, since it is a good measure of the complexity of a set of functional dependencies, or a set of minimal keys.

Earlier work on Armstrong instances were mostly studied by assuming that the domain of each attribute is countably infinite. Recently, the study of *higher order data model* in [7], [8] considered the question of Armstrong instances with bounded domains. Another reason for considering bounded domains is that for many attributes, their domains are well defined finite sets. For example, the age of a person can take values from the set $\{0, 1, \ldots, 120\}$.

Thalheim [9] investigated the maximum number of minimal keys in the case of bounded domains and showed that restrictions on the sizes of domains make significant differences. It is natural to ask what one can say about Armstrong instances if all attributes have domains restricted to size q. Let K_n^k denote the collection of all k-subsets of an n-element attribute set A.

Definition 1. Let $q, k > 1$ be integers. Let $f(q, k)$ denote the maximum n such that there exists an Armstrong instance for K_n^k being the system of minimal keys.

The problem of determining $f(q, k)$ was introduced in [10] and investigated in [11], [12]. The only known values of $f(q, k)$ are $f(q, 2) = \binom{q+1}{2}$, which were determined in [11].

The main contribution of this paper is the determination of $f(q, 3)$. We prove that $f(q, 3) = 3q - 1$ for all $q \geq 5$, except possibly for $q \in \{14, 16, 20\}$. This disproves a conjecture of Sali [13]. We also consider the analogous problem of determining $f(q, k)$ when extended to more general dependencies.

II. Preliminaries

Throughout this paper, we view an $m \times n$ Armstrong instance with domains of size q as a q-ary code C of length n and size m, where the codewords are precisely the rows of the instance.

For a positive integer k, $[k]$ denotes the set of integers $\{1, 2, \ldots, k\}$.

A. Armstrong Codes

Katona et al. [11] characterized the code C corresponding to an Armstrong instance with K_n^k as the set of minimal keys, as follows:

(i) C has minimum Hamming distance at least $n - k + 1$;
(ii) for any set of $k-1$ coordinates there exist two codewords agreeing in exactly those coordinates.

A $(k - 1)$-set of coordinates can be considered as a "direction", so in C the minimum distance is *attained in all directions*. Such a code C is called an *Armstrong code*, or

978-1-4799-0446-4/13/$31.00 ©2013 IEEE

more precisely, a (q,k,n)-*Armstrong code*. It is obvious that $f(q,k)$ is the maximum n such that there exists a (q,k,n)-Armstrong code. The following bounds on $f(q,k)$ are known.

Theorem 1 (Katona et al. [11], Sali and Székely [12])**.**

(i) Let $q > 4$. Then $f(q,k) \geq \lceil \frac{k}{2} \log q - 1 \rceil$ for all sufficiently large k.

(ii) There exists a constant $c > 1$ such that $f(2,k) \geq \lfloor ck \rfloor$ for all sufficiently large k.

(iii) Let $q > 1$ and $k > 2$. Then

$$f(q,k) \leq q(k-1)\left(1 + \frac{q-1}{\sqrt{\frac{2(qk-q-k+2)^{k-1}}{(k-1)!}} - q}\right). \quad (1)$$

(iv) If $q \geq 2$ and $k \geq 5$, then the bound (1) can be improved to $f(q,k) \leq q(k-1)$, except when $(k,q) \in \{(5,2),(5,3),(5,4),(5,5),(6,2)\}$.

(v) For fixed $q > 1$, we have

$$\frac{\sqrt{q}}{e}k < f(q,k) < (q - \log q)k$$

for all sufficiently large k.

Proposition 1 (Katona et al. [11])**.** *For $q > 1$, $f(q,3) \leq 3q-1$.*

B. Orthogonal Double Covers

The concept of orthogonal double covers originates in conjectures of Demetrovics et al. [14] concerning database constraints and formalized later by Ganter et al. [15]. Let X be a finite set. A partition of X is said to *cover* $T \subseteq X$ if T is contained in some part of the partition.

Definition 2. Let X be a set of size m. A set of partitions of X is called an *orthogonal double cover* (ODC) of K_m (with its vertices identified with elements of X) if it satisfies the following properties:

(i) for any two partitions, there is exactly one 2-subset of X that is covered by both partitions;

(ii) each 2-subset of X is covered by exactly two different partitions.

We view each part of a partition of X as a complete subgraph of K_m over X. Then each partition can be regarded as a disjoint union of complete subgraphs of K_m. If an ODC consists of n partitions, each of which is isomorphic to a graph G, then we say the ODC is an *ODC by n G's*. Suppose that G is a disjoint union of q complete subgraphs, then an ODC of K_m by n G's gives an $m \times n$ Armstrong instance over $[q]$ as follows. For each partition of the ODC, arbitrarily order the q parts, and construct a column u of length m, with coordinates indexed by elements of X, such that for $i \in X$, $\mathsf{u}_i = \mathsf{j}$ if and only if i is contained in the j-th part of the partition. It is easy to check that the set of rows of this Armstrong instance is a $(q,3,n)$-Armstrong code.

Ganter and Gronau [16] proved that for $q \geq 5$, there exists an ODC of K_{3q-2} by $3q-2$ $(q-1)K_3 \cup K_1$'s, settling a conjecture of Demetrovics et al. [14]. This result also implies the existence of a $(q,3,3q-2)$-Armstrong code. Hence, we have $f(q,3) \geq 3q-2$. Furthermore, it is easy to show that $f(2,3) = 4$. This led Sali [13] to make the following conjecture.

Conjecture 1 (Sali [13])**.** For all $q \geq 2$, $f(q,3) = 3q-2$.

Unfortunately, this conjecture is false. When $m \geq 2$, an ODC of K_{6m+2} by $6m+2$ $2mK_3 \cup K_2$'s has been constructed by Gronau et al. [17]. This gives a $(2m+1,3,6m+2)$-Armstrong code and hence $f(2m+1,3) \geq 6m+2$. Thus, Conjecture 1 is false for all odd $q \geq 5$. One of the primary aims of this paper is to prove that Conjecture 1 is also false for even q. In fact, we determine that $f(q,3) = 3q-1$ for all q, with three possible exceptions.

III. $(q,3,3q-1)$-Armstrong Codes

We prove $f(q,3) = 3q-1$ by showing the existence of $(q,3,3q-1)$-Armstrong codes. Our proof is constructive and uses techniques from combinatorial design theory. We briefly review some required concepts below.

A. Combinatorial Designs

A *set system* is a pair $\mathfrak{S} = (X, \mathcal{A})$, where X is a finite set of *points* and $\mathcal{A} \subseteq 2^X$. Elements of $\mathcal{A}$ are called *blocks*. The *order* of $\mathfrak{S}$ is the number of points in X, and the *size* of $\mathfrak{S}$ is the number of blocks in $\mathcal{A}$. Let K be a set of positive integers. A set system $(X,\mathcal{A})$ is *K-uniform* if $|A| \in K$ for all $A \in \mathcal{A}$. A *parallel class* of a set system $(X,\mathcal{A})$ is a set $\mathcal{P} \subseteq \mathcal{A}$ that partitions X. A *resolvable set system* is a set system whose set of blocks can be partitioned into parallel classes.

Definition 3. A triple system TS(m,λ) is a $\{3\}$-uniform set system $(X,\mathcal{A})$ of order n such that every 2-subset of X is contained in exactly λ blocks of $\mathcal{A}$.

Definition 4. Let $(X,\mathcal{A})$ be a set system and let $\mathcal{G}$ be a partition of X into subsets, called *groups*. The triple $(X,\mathcal{G},\mathcal{A})$ is a *group divisible design* (GDD) when every 2-subset of X not contained in a group is contained in exactly one block, and $|A \cap G| \leq 1$ for all $A \in \mathcal{A}$ and $G \in \mathcal{G}$.

We denote a GDD $(X,\mathcal{G},\mathcal{A})$ by k-GDD if $(X,\mathcal{A})$ is $\{k\}$-uniform. The *type* of a GDD $(X,\mathcal{G},\mathcal{A})$ is the multiset $\langle |G| : G \in \mathcal{G} \rangle$. When more convenient, the exponential notation is used to describe the type of a GDD: a GDD of type $g_1^{t_1} g_2^{t_2} \cdots g_s^{t_s}$ is a GDD where there are exactly t_i groups of size g_i, $i \in [s]$. The following results are known (see, for example, [18], [19]).

Theorem 2.

(i) A resolvable TS$(m,2)$ exists if and only if $m \equiv 0 \pmod 3$ and $m \neq 6$.

(ii) There exists a 4-GDD of type $2^u m^1$ for each $u \geq 6$, $u \equiv 0 \pmod 3$ and $m \equiv 2 \pmod 3$ with $2 \leq m \leq u-1$, except for $(u,m) = (6,5)$ and possibly except for $(u,m) \in \{(21,17),(33,23),(33,29),(39,35),(57,44)\}$.

B. Extorthogonal Double Covers

A *suborthogonal double cover* (subODC) is a collection of partitions of $[m]$ similar to an ODC except that for any two partitions there is *at most* one 2-subset of $[m]$ covered by both partitions. SubODCs were first studied by Hartmann and Schumacher [20], who considered them as generalized ODCs under circumstances when ODCs do not exist. Here, we consider another generalization, called *extorthogonal double covers* (extODC). These are similar to ODCs, except that for any two partitions there is *at least* one 2-subset of $[m]$ covered by both partitions. We construct $(q, 3, 3q-1)$-Armstrong codes from a special class of extODCs of K_{3q} by qK_3's.

Proposition 2. *If there exists an extODC of K_{3q} by qK_3's, then $f(q, 3) = 3q - 1$.*

Proof: By considering 2-subsets, the number of partitions in an extODC of K_{3q} by qK_3's is easily seen to be $2\binom{3q}{2}/3q = 3q-1$. For each partition, arbitrarily order the q parts. Define a $3q \times (3q-1)$ q-ary array by indexing each column by a partition and each row by a point of the extODC. For each partition, the corresponding column has the symbol i in the rows indexed by the points in the ith part. The set of rows in this array is a $(q, 3, 3q-1)$-Armstrong code, by the definition of an extODC. This, together with Proposition 1, implies that $f(q, 3) = 3q - 1$. ∎

It is easy to see that an extODC of K_{3q} by $3q-1$ qK_3's is a resolvable TS$(3q, 2)$ with the additional property that every two parallel classes covers a common 2-subset. Although $f(q, 3)$ is known for odd q, it is still interesting to know when extODCs of K_m, m odd, can exist. We have the following result for $m = 3q$, q odd.

Proposition 3. *There exists an extODC of K_{3q} by qK_3's, for all odd $q \geq 5$.*

Proof: Let $u = (3q-1)/2$. Starting from a 4-GDD $(X, \mathcal{G}, \mathcal{A})$ of type 2^u, whose existence is guaranteed by Theorem 2, we construct an extODC of K_{2u+1} over $X \cup \{\infty\}$ from $\mathcal{A}$. For each $x \in X$, let $\mathcal{B}_x = \{B \setminus \{x\} : x \in B \in \mathcal{A}\} \cup \{G \cup \{\infty\} : x \in G \in \mathcal{G}\}$. Then $\mathcal{B}_x$ is a partition of $X \cup \{\infty\}$. We claim that $\{\mathcal{B}_x : x \in X\}$ is an extODC of K_{2u+1}.

Indeed, for any two partitions, say $\mathcal{B}_x$ and $\mathcal{B}_y$, both of which cover $\{x, y\}$ if x, y are in the same group, and cover $B \setminus \{x, y\}$ if $x, y \in B$ are in distinct groups. For each pair $\{x, y\} \subset X \cup \{\infty\}$, if $\{x, y\} \subset G \cup \{\infty\}$ for some $G \in \mathcal{G}$, then $\{x, y\}$ is covered by two partitions $\mathcal{B}_g$, $g \in G$; if $x, y \in X$ are in distinct groups, then there exists exactly one block $B \in \mathcal{A}$ such that $\{x, y\} \subset B$, while $\{x, y\}$ is covered by two partitions $\mathcal{B}_g$, $g \in B \setminus \{x, y\}$. Hence, $\{\mathcal{B}_x : x \in X\}$ is an extODC. ∎

We now construct extODCs of K_m, where $m = 3q$ is even. Define a *base partition* of order m, which is a partition P of $\mathbb{Z}_{m-1} \cup \{\infty\}$ into triples with the following two properties:

(i) $\langle \pm(a-b) : \{a, b\} \subset C \in P \text{ and } \infty \notin \{a, b\} \rangle = 2(\mathbb{Z}_{m-1} \setminus \{0\})$.

TABLE I
BASE PARTITIONS FOR SOME SMALL EXTODCS

q	triples
6	$\{0,1,2\}$ $\{3,7,12\}$ $\{4,15,17\}$ $\{5,8,14\}$ $\{6,10,13\}$ $\{9,11,16\}$
8	$\{0,1,2\}$ $\{3,5,8\}$ $\{4,12,18\}$ $\{6,15,19\}$ $\{7,14,23\}$ $\{9,17,21\}$ $\{10,13,20\}$ $\{11,16,22\}$
10	$\{0,1,2\}$ $\{3,5,8\}$ $\{4,10,20\}$ $\{6,23,29\}$ $\{7,11,22\}$ $\{9,17,27\}$ $\{12,16,25\}$ $\{13,21,28\}$ $\{14,19,26\}$ $\{15,18,24\}$
12	$\{0,1,2\}$ $\{3,5,8\}$ $\{4,7,15\}$ $\{6,19,34\}$ $\{9,13,27\}$ $\{10,20,25\}$ $\{11,22,28\}$ $\{12,24,33\}$ $\{14,23,30\}$ $\{16,26,32\}$ $\{17,21,29\}$ $\{18,31,35\}$

(ii) $\langle i : \{a, b\} + i = \{c, d\} \text{ for some } \{a, b\} \subset C, \{c, d\} \subset C' \text{ and } C, C' \in P \rangle \supset (\mathbb{Z}_{m-1} \setminus \{0\})$, where $\infty + i := \infty$.

Developing a base partition of order m under the group $\mathbb{Z}_{m-1}$ gives a set of $m-1$ partitions of $\mathbb{Z}_{m-1} \cup \{\infty\}$, which forms an extODC of K_m. The first property ensures that each pair occurs exactly twice, while the second ensures that any two partitions cover at least one common 2-subset.

Proposition 4. *There exists an extODC of K_{3q} by qK_3's, for $q \in \{6, 8, 10, 12\}$.*

Proof: The base partitions for extODC of K_{3q}, for $q \in \{6, 8, 10, 12\}$, are given in Table I. ∎

Proposition 5. *There exists an extODC of K_{3q} by qK_3's, for all even $q \geq 18$, $q \neq 20$.*

Proof: Let $u = (3q-18)/2$. There exists a 4-GDD $(X, \mathcal{G}, \mathcal{A})$ of type $2^u 17^1$ by Theorem 2. We construct an extODC of K_{3q} (on $X' = X \cup \{\infty\}$) from $\mathcal{A}$. Let G_0 be the long group in $\mathcal{G}$ of size 17. By Proposition 4, there exists an extODC of K_{18} (on $G_0 \cup \{\infty\}$) by 17 $6K_3$'s over $G_0 \cup \{\infty\}$. Let the set of partitions be $\{\mathcal{C}_x : x \in G_0\}$. For each $x \in (X \setminus G_0)$, let $\mathcal{B}_x = \{B \setminus \{x\} : x \in B \in \mathcal{A}\} \cup \{G \cup \{\infty\} : x \in G \in \mathcal{G}\}$. For each $x \in G_0$, let $\mathcal{B}_x = \{B \setminus \{x\} : x \in B \in \mathcal{A}\} \cup \mathcal{C}_x$. There are $3q-1$ $\mathcal{B}_x$'s in total and each $\mathcal{B}_x$ is a partition of X'. We claim that the set of all $\mathcal{B}_x$'s is an extDOC.

Indeed, for any two partitions $\mathcal{B}_x$ and $\mathcal{B}_y$, they both cover $\{x, y\}$ if x, y are in the same group of size 2; cover a common 2-subset if $x, y \in G_0$ since $\mathcal{C}_x$ and $\mathcal{C}_y$ have a common 2-subset, and both cover $B \setminus \{x, y\}$ if $x, y \in B$ are in distinct groups. For each pair $\{x, y\} \subset X'$, if $\{x, y\} \subset G \cup \{\infty\}$ for some $G \neq G_0$, then $\{x, y\}$ is covered in two partitions $\mathcal{B}_g$, $g \in G$. If $\{x, y\} \subset G_0 \cup \{\infty\}$, then $\{x, y\}$ is covered by both $\mathcal{B}_u$ and $\mathcal{B}_v$, where $\{x, y\}$ is contained in $\mathcal{C}_u$ and $\mathcal{C}_v$. If x, y are in distinct groups, then there exists exactly one block $B \in \mathcal{A}$ such that $\{x, y\} \subset B$, while $\{x, y\}$ occurs in $\mathcal{B}_g$, $g \in B \setminus \{x, y\}$. Hence, $\{\mathcal{B}_x : x \in X\}$ is an extODC of K_{3q}. ∎

Combining Propositions 2, 3 and 5 gives the main result of this section.

Theorem 3. *For all $q \geq 5$ and $q \neq 14, 16, 20$, there exists an extODC of K_{3q} by qK_3's, and consequently $f(q, 3) = 3q - 1$.*

IV. Generalized Armstrong Codes

The concept of functional dependencies has been generalized by Demetrovics, Katona, and Sali [6].

Definition 5. Let $X \subseteq A$ and $y \in A$ for a given table $R(A)$. Then for positive integers $s \leq t$, we say that y (s,t)-*depends* on X, written $X \xrightarrow{(s,t)} y$, if there do not exist $t+1$ data items (rows) $d_1, d_2, \ldots, d_{t+1}$ of $R(A)$ such that

(i) $|\{d_i|_{\{x\}} \colon 1 \leq i \leq t+1\}| \leq s$ for each $x \in X$, and
(ii) $|\{d_i|_{\{y\}} \colon 1 \leq i \leq t+1\}| = t+1$.

Our usual concept of functional dependency is equivalent to the special case of $(1,1)$-dependency. When functional dependencies are not known, (s,t)-dependencies identified in a relational database can still be exploited for improving storage efficiency [6], [21]–[23].

Given $1 \leq s \leq t$, an (s,t)-*dependent key* K is a subset of the attribute set A, such that $R(A)$ satisfies (s,t)-dependencies $K \xrightarrow{(s,t)} y$ for all $y \in A$. A key K is called *minimal* if no subset of K is an (s,t)-dependent key. Here, we generalize Armstrong codes from functional dependencies into (s,t)-dependencies.

A q-ary code C is called a $(q,k,n)_{s,t}$-*Armstrong code* if

(i) for any $t+1$ rows of C, there exist at most $k-1$ columns such that each column has at most s distinct elements in the $t+1$ rows, and
(ii) for any $k-1$ columns of C, there exist $t+1$ rows such that each of the $k-1$ columns has at most s distinct elements in the $t+1$ rows.

It is clear that we need $q > s$ and $k > 1$ for a $(q,k,n)_{s,t}$-Armstrong code to be meaningful. Note that a $(q,k,n)_{1,1}$-Armstrong code is just a (q,k,n)-Armstrong code.

Definition 6. Let $q > s \geq 1$, $t \geq s$ and $k > 1$. Then $f_{s,t}(q,k)$ denotes the maximum n such that there exists a $(q,k,n)_{s,t}$-Armstrong code.

As with the Armstrong codes for functional dependencies [11], we have the following restrictions on $(q,k,n)_{s,t}$-Armstrong codes. Let ϕ be the least number of submultisets $S \subset M$ of size $t+1$ with at most s distinct elements, where M ranges over all multisets of size m over $[q]$.

Proposition 6. *Let C be a $(q,k,n)_{s,t}$-Armstrong code and let $m = |C|$. Then $\binom{m}{t+1} \geq \binom{n}{k-1}$ and $n \cdot \phi \leq (k-1)\binom{m}{t+1}$.*

Proof: Let T be a set of $k-1$ columns of C. By condition (ii), there exists a set R_T of $t+1$ rows such that each column of T has at most s distinct elements in R_T. By the first defining condition (i) of a $(q,k,n)_{s,t}$-Armstrong code, R_T is distinct for distinct T. The first inequality then follows. The second inequality holds by the definition of ϕ and the defining condition (i). ∎

The two inequalities in Proposition 6 combine to give an upper bound on $f_{s,t}(q,k)$ as in [11]. However, it is not explicit and not tight in most cases. We only explore the values of $f_{s,t}(q,k)$ for special cases.

A. The Case $s = 1$ and $k = 2$

Proposition 7. *When $s = 1$ and $q < m$, we have*

$$\begin{aligned}\phi &\geq r\binom{h+1}{t+1} + (q-r)\binom{h}{t+1} \\ &= q\binom{h}{t+1} + r\binom{h}{t},\end{aligned}$$

where $m = qh + r$, with $0 \leq r < q$.

Proof: Similar to the proof in [11, Lemma 3.2], let m_1 and m_2 be the number of two distinct symbols in M. The inequality follows by the fact that $\binom{m_1}{t+1} + \binom{m_2}{t+1} \geq \binom{m_1+1}{t+1} + \binom{m_2-1}{t+1}$ for all m_1 and m_2 satisfying $m_2 - m_1 \geq 2$. ∎

Proposition 8. *The function $g(m) = \frac{(k-1)\binom{m}{t+1}}{q\binom{h}{t+1}+r\binom{h}{t}}$ is decreasing in m, for $r+1 \leq q < m$.*

Proof: The proof is straightforward and is omitted here. ∎

Proposition 9. $f_{1,t}(q,2) = \binom{qt+1}{t+1}$.

Proof: The upper bound is obtained by combining Propositions 6–8, and setting $h = t$ and $r = 1$. The lower bound is given by construction. Construct a $(qt+1) \times \binom{qt+1}{t+1}$ array as follows. For each column, we have exactly one subset of $t+1$ rows with equal symbols and all other $q-1$ symbols occurring exactly t times. We do so such that each column has a distinct subset of $t+1$ rows with equal symbols. It is clear that this gives a $(q, 2, \binom{qt+1}{t+1})_{1,t}$-code. ∎

B. The Case $s = t = 2$ and $k = 4$

Proposition 10. *When $s = t = 2$ and $q < m$, we have $\phi \geq h(m)$, where*

$$\begin{aligned}h(m) =& r\binom{h+1}{3} + (q-r)\binom{h}{3} + \\ & r\binom{h+1}{2}(m-h-1) + (q-r)\binom{h}{2}(m-h),\end{aligned}$$

where $m = qh + r$, with $0 \leq r < q$.

Proof: Let m_1 and m_2 be the number of two distinct symbols in M. The inequality follows by the fact that $\binom{m_1}{3} + \binom{m_2}{3} + \binom{m_1}{2}(m-m_1) + \binom{m_2}{2}(m-m_2) \geq \binom{m_1+1}{3} + \binom{m_2-1}{3} + \binom{m_1+1}{2}(m-m_1-1) + \binom{m_2-1}{2}(m-m_2+1)$ for all m_1 and m_2 satisfying $m_2 - m_1 \geq 2$. ∎

Proposition 11. *The function $k(m) = \frac{(k-1)\binom{m}{3}}{h(m)}$ is decreasing in m, for $r+1 \leq q < m$.*

Proof: The proof is straightforward and is omitted here. ∎

When $k = 4$, we have $n \leq m$ by Proposition 6. Combining Propositions 10 and 11, we know that the universal upper bound is $n \leq m$ when $m = k(m)$. The solution is $m = 2q-1$, which is achieved when $h = 1$ and $r = q-1$. Hence, $f_{2,2}(q,4) \leq 2q-1$. In fact, $f_{2,2}(q,4) = 2q-1$, since there

exists an Armstrong instance over q symbols for K^4_{2q-1} being the system of minimal $(2,2)$-dependent keys [24].

V. Conclusion

We investigated the maximum number of minimal keys in relational database systems with attributes having bounded domains via the study of Armstrong codes. We showed that the maximum length n for which a $(q,3,n)$-Armstrong code can exist is $f(q,3) = 3q-1$ for all $q \geq 5$ with three possible exceptions, disproving a conjecture of Sali.

Our determination of $f(q,3)$ involves introducing the new concept of extorthogonal double covers (extODC), a generalization of orthogonal double covers with property that any two partitions cover at least one common 2-subset. This new combinatorial design is interesting not only in database theory, but also in design theory. Similar to ODCs, there are several directions for the study of extODCs. For example, each partition could be extended to any spanning subgraph, or consider similar properties for hypergraphs.

Further, we generalized Armstrong codes to the case of (s,t)-dependencies. Classes of optimal Armstrong codes of this type are constructed.

VI. Acknowledgement

Research of the authors is supported in part by the Singapore National Research Foundation under Research Grant NRF-CRP2-2007-03. The authors are grateful to the anonymous reviewers and the TPC member whose comments greatly improved the presentation of the paper.

References

[1] W. W. Armstrong, "Dependency structures of data base relationships," in *IFIP Conference Proceedings*. North-Holland, 1974, pp. 580–583.

[2] P. A. Bernstein, "Synthesizing third normal form relations from functional dependencies," *ACM Trans. Database Syst.*, vol. 1, no. 4, pp. 277–298, 1976.

[3] C. Beeri, R. Fagin, and J. H. Howard, "A complete axiomatization for functional and multivalued dependencies in database relations," in *SIGMOD '77 – Proceedings of the 1977 ACM SIGMOD International Conference on Management of Data*, 1977, pp. 47–61.

[4] J. Rissanen, "Independent components of relations," *ACM Trans. Database Syst.*, vol. 2, no. 4, pp. 317–325, 1977.

[5] J. Demetrovics, "On the equivalence of candidate keys with Sperner systems," *Acta Cybernet.*, vol. 4, pp. 247–252, 1979.

[6] J. Demetrovics, G. O. H. Katona, and A. Sali, "The characterization of branching dependencies," *Discrete Appl. Math.*, vol. 40, no. 2, pp. 139–153, 1992.

[7] S. Hartmann, S. Link, and K.-D. Schewe, "Weak functional dependencies in higher-order datamodels," in *FoIKS '04 – Proceedings of the Third International Symposium on Foundations of Information and Knowledge Systems 2004*, ser. Lecture Notes in Comput. Sci., vol. 2942. Springer, 2004, pp. 116–133.

[8] A. Sali, "Minimal keys in higher-order datamodels," in *FoIKS '04 – Proceedings of the Third International Symposium on Foundations of Information and Knowledge Systems 2004*, ser. Lecture Notes in Comput. Sci., vol. 2942. Springer, 2004, pp. 242–251.

[9] B. Thalheim, "The number of keys in relational and nested relational databases," *Discrete Appl. Math.*, vol. 40, no. 2, pp. 265–282, 1992.

[10] A. Sali and K.-D. Schewe, "Keys and Armstrong databases in trees with restructuring," *Acta Cybernet.*, vol. 18, no. 3, pp. 529–556, 2008.

[11] G. O. H. Katona, A. Sali, and K.-D. Schewe, "Codes that attain minimum distance in all possible directions," *Cent. Eur. J. Math.*, vol. 6, pp. 1–11, 2008.

[12] A. Sali and L. A. Székely, "On the existence of Armstrong instances with bounded domains," in *FoIKS '08 – Proceedings of the 7th International Symposium on Foundations of Information and Knowledge Systems 2004*, ser. Lecture Notes in Comput. Sci., vol. 4932, 2008, pp. 151–157.

[13] A. Sali, "Coding theory motivated by relational databases," in *SDKB '10 – Proceedings of the 4th International Conference on Semantics in Data and Knowledge Bases*, ser. Lecture Notes in Comput. Sci., vol. 6834. Springer-Verlag, 2011, pp. 96–113.

[14] J. Demetrovics, Z. Füredi, and G. O. H. Katona, "Minimum matrix representations of closure operations," *Discrete Appl. Math.*, vol. 11, no. 2, pp. 115–128, 1985.

[15] B. Ganter, H.-D. O. F. Gronau, and R. C. Mullin, "On orthogonal double covers of K_n," *Ars Combin.*, vol. 37, pp. 209–221, 1994.

[16] B. Ganter and H.-D. O. F. Gronau, "Two conjectures of Demetrovics, Füredi, and Katona, concerning partitions," *Discrete Math.*, vol. 88, no. 2–3, pp. 149–155, 1991.

[17] H.-D. O. F. Gronau, R. C. Mullin, and P. J. Schellenberg, "On orthogonal double covers of K_n and a conjecture of Chung and West," *J. Combin. Des.*, vol. 3, no. 3, pp. 213–231, 1995.

[18] R. J. R. Abel, G. Ge, and J. Yin, "Resolvable and near-resolvable designs," in *The CRC Handbook of Combinatorial Designs*, 2nd ed., C. J. Colbourn and J. H. Dinitz, Eds. Boca Raton: CRC Press, 2007, pp. 124–134.

[19] G. Ge, "Group divisible designs," in *The CRC Handbook of Combinatorial Designs*, 2nd ed., C. J. Colbourn and J. H. Dinitz, Eds. Boca Raton: CRC Press, 2007, pp. 255–260.

[20] S. Hartmann and U. Schumacher, "Suborthogonal double covers of complete graphs," *Congr. Numer.*, vol. 147, pp. 33–40, 2000.

[21] J. Demetrovics, G. O. H. Katona, and A. Sali, "Minimal representations of branching dependencies," *Acta Sci. Math. (Szeged)*, vol. 60, no. 1–2, pp. 213–223, 1995.

[22] ——, "Design type problems motivated by database theory," *J. Statist. Plann. Inference*, vol. 72, no. 1–2, pp. 149–164, 1998.

[23] G. O. H. Katona and A. Sali, "New type of coding problem motivated by database theory," *Discrete Appl. Math.*, vol. 144, no. 1–2, pp. 140–148, 2004.

[24] Y. M. Chee and H. Zhang, "Optimal Armstrong codes for branching dependencies," *preprint*, 2013.