# Encoding Databases Satisfying a Given Set of Dependencies

Gyula O.H. Katona[⋆1] and Krisztián Tichler[⋆⋆2]

[1] Rényi Institute, Budapest, Hungary,
ohkatona@renyi.hu
[2] Eötvös University, Faculty of Informatics, Budapest, Hungary,
ktichler@inf.elte.hu

**Abstract.** Consider a relation schema with a set of dependency constraints. A fundamental question is what is the minimum space where the possible instances of the schema can be "stored". We study the following model. Encode the instances by giving a function which maps the set of possible instances into the set of words of a given length over the binary alphabet in a decodable way. The problem is to find the minimum length needed. This minimum is called the information content of the database.
We investigate several cases where the set of dependency constraints consist of relatively simple sets of functional or multivalued dependencies. We also consider the following natural extension. Is it possible to encode the instances such a way that small changes in the instance cause a small change in the code.

**Keywords:** relational database, coding, functional dependency, multivalued dependency

## 1 Introduction

Let $(R, \Sigma)$ be a dependency schema in the relational database model ([1]) where $R$ is a relational schema with a single relation and $\Sigma$ is a set of dependencies on the set of attributes $\Omega$ of size $|\Omega| = n$. Suppose that all the domains of the attributes are finite. Then the number of possible tuples is also finite. Hence the number of possible instances $I$ is finite, too. A fundamental question is "what is the minimum space where a database can be stored?". Some of the possible applications are efficient and error-tolerant data transmission or achiving.

Let us repeat the problem for readers not so familiar with the notations above. A database is a table (matrix) of $n$ columns. A row or a record contains the data of one object or person, where the $i$th element of the row is the $i$th

⋆ The work of the first author was supported by the Hungarian National Foundation for Scientific Research grant numbers T037846.
⋆⋆ Research projects presented in this article are supported by the European Union and co-financed by the European Social Fund (grant agreement no. TÁMOP 4.2.1./B-09/1/KMR-2010-0003).

attribute, the $i$th type of data of the object. The number of possible values in one place in a row is finite. There are some constraints, rules, connections among the values in a row, given by $\Sigma$. There are only finitely many possible rows satisfying these constraints.

An actual situation in the database is a collection of these possible tuples. This is called an instance $I$ of the relation. We need to store the instances in such a way that different instances have different "stored forms". On the other hand the "stored forms" should be relatively small. Our goal is to give a model of this situation.

The following model is suggested. Encode the instances by 0,1 sequences of length $\ell$, that is, give a function $c : \mathcal{I}(R, \Sigma) \to \{0,1\}^\ell$ which maps the set of possible instances $\mathcal{I}(R, \Sigma)$ into $\{0,1\}^\ell$. Of course the map should be decodable, $c$ should give different sequences for different instances. $\ell = |c|$ is called the code length. The problem is to find the minimum of $\ell$. This minimum can be called the information content of the database schema: $\mathrm{Inf}(R, \Sigma)$. Of course, this is nothing else but the log of the total number of of possible instances, $\lceil \log_2 |\mathcal{I}(R, \Sigma)| \rceil$.

Although the definition is simple and natural, there are difficulties in its implementation. In most cases it is impossible to give an exact number for the total number of instances. We will show a simple-looking example of a trivial multivalued dependency when it is not easy to determine even the asymptotical number of instances.

Our other toy-example is when there is only one minimal key in the dependency schema. In that case we were able to give an exact formula for the number of possible instances using elementary steps. Of course there is a code with length $\mathrm{Inf}(R, \Sigma)$. But this code is useful only when it can be obtained by a simple algorithm and can be similarly decoded, that is the instance can be obtained from the code by another easy algorithm. We do not know if this can be done in our case of only-one-key. However we can show a very natural code which is only slightly longer than $\mathrm{Inf}(R, \Sigma)$.

The next problem arises when the instance is subject to an elementary modification. There is a very natural requirement on these codes. If two instances are similar then their codes should also be similar. More precisely we should write "close" in the previous sentence rather than "similar". If this condition is not satisfied it might happen that making a little change in the database (instance) the changes in the encoded version are big, we have to work too much to get the changes.

Consider some elementary changes in $I$, like deleting or adding a row, replacing one entry in one of the rows. We would like to have a small change in the code of an instance if it is a subject of one of such elementary changes. The changes in the codes are measured by the Hamming distance that is the number of different digits.

We will show that if this requirement takes place in a fairly strict manner then the code is much longer.

Let us introduce some basic notations that are used in the paper. For an $n$-tuple $t = (t_{A_1}, \ldots, t_{A_n})$ and $X \subseteq \Omega = \{A_1, \ldots, A_n\}$ let $\pi_X(t)$ denote the $|X|$-

tuple $u$ that has the property $u_A = t_A$ $(A \in X)$. Sometimes $\pi_X(t)$ is also called an *$X$-tuple*. If $I$ is an instance let $\pi_X(I) = \{\pi_X(t) \,|\, t \in I\}$.

For integers $r, s$ $[r]$ denotes the set $\{1, \ldots, r\}$ and $[r, s]$ denotes the set $\{r, \ldots, s\}$.

The paper is organized as follows. In section 2 we consider the case where only one key is given in the schema. The concepts of 2-distance-preserving and strongly 2-distance-preserving codes are introduced. We give lower bounds on the size of these codes in section 4. In section 3 the case of joins is analyzed and an other simple but much different set of multivalued dependencies is considered. Investigations lead to a problem on random bipartite graphs. A partial solution is given in section 5. In section 6 we mention some related works and finish the study with several open problems in section 7.

## 2 Only one minimal key

Let the number of attributes of $R$ be $|\Omega| = n = a + b$ where $a$ and $b$ are positive integers. Suppose that all domains are $\{0, 1\}$. Let the attributes be ordered and suppose that the set $K \subseteq \Omega, |K| = a$ is a key. It can be supposed without loss of generality that $K$ is the set of the first $a$ attributes. Let this dependency schema be denoted by $(R, \{K \to \Omega\})$.

If $t$ is an $n$-tuple in an instance $I$ satisfying $K \to \Omega$ then $\pi_K(t)$ uniquely determines $\pi_{\Omega - K}(t)$. For an instance $I \in \mathcal{I}(R, \{K \to \Omega\})$ and $u \in \pi_K(I)$ let $f(I, u)$ denote the function describing this dependency, that is $f(I, \pi_K(t)) = \pi_{\Omega - K}(t)$ for an $n$-tuple $t$. Of course this function depends on $I$.

**Proposition 1.**

$$\mathcal{I}(R, \{K \to \Omega\}) = \left(2^b + 1\right)^{2^a}. \tag{1}$$

*Proof.* The number of possible $K$-tuples (first part of the $n$-tuples) is $2^a$. Let us denote the $i$th possible $a$-tuple by $\widehat{i}$ $(0 \leqslant i < 2^a)$. For any given $\widehat{i} = \pi_K(t)$ there are $2^b$ choices for $f(I, \widehat{i}\,)$, that is each first part has $2^b$ possible "continuations" in the last $b$ attributes. If $s = |\pi_K(I)|$ then the total number of possible choices for $f(I, \widehat{i}\,)$ for all tuples in the instance is $\left(2^b\right)^s$. There are $2^a$ possible values of $\pi_K(t)$ therefore one can choose $s$ pieces of $K$-tuples $\pi_K(t)$ in $\binom{2^a}{s}$ many ways, so the number of instances of size $s$ is

$$\binom{2^a}{s} \left(2^b\right)^s, \tag{2}$$

and the total number of choices is

$$\sum_{s=0}^{2^a} \binom{2^a}{s} \left(2^b\right)^s = \left(2^b + 1\right)^{2^a}. \tag{3}$$

□

Therefore $\mathrm{Inf}(R,\{K\to\Omega\})=2^a\log(2^b+1)$ what is slightly more than $b2^a$. Of course there is always a code $c$ with the length $\ell=\lceil\log(2^b+1)2^a\rceil$, since we can list all the possible instances and the code of the $j$th instance can be the binary form of $j$. However nothing ensures that this code is algorithmic and easy to decode.

There is a nice way to encode the instances in the following way. The code $c(I)$ will be determined in the form $c_0(I)c_1(I)\dots c_{2^a}(I)$ where $c_0(I)$ is a 0,1 sequence of length $2^a$: its $i$th digit is 1 iff $\widehat{i}\in\pi_K(I)$ and

$$c_{i+1}(I)=\begin{cases} f(I,\widehat{i}\,) & \text{if } \widehat{i}\in\pi_K(I)\\ (0,0,\dots,0)(\text{ of length } b) & \text{if } \widehat{i}\notin\pi_K(I)\end{cases} \tag{4}$$

for $0\leqslant i<2^a$. The length of this code is $(1+b)2^a$, that is only a little more than $\mathrm{Inf}(R)$. It is easy to see that this code is uniquely decodable, since $c_0(I)$ determines which first parts are in the instance, the second parts are all determined by the function $f(I,\widehat{i}\,)$.

Let us show by a small example why we have to define the second row of (4) to be 0. Choose $a=2,b=1$ and consider the instance

$$I=\begin{matrix}0\,0\,1\\0\,1\,1\end{matrix}$$

Here $c_0(I)$ is clearly $(1,1,0,0)$ since the first two of the possible four first parts ( (0,0), (0,1), (1,0), (1,1) ) occur in the instance. $f(I,\widehat{0})=1, f(I,\widehat{1})=1$ are also well-defined. But $f(I,\widehat{2})$ and $f(I,\widehat{3})$ are not defined by the instance, so all of the following sequences could be codes of the present $I$: $(1,1,0,0,1,1,0,0)$, $(1,1,0,0,1,1,0,1)$, $(1,1,0,0,1,1,1,0)$, $(1,1,0,0,1,1,1,1)$. To avoid this ambiguity we chose the first one.

We have actually proved the following statement.

**Proposition 2.** *There is a code of length $\ell=(1+b)2^a$, defined by an easy algorithm that can be easily decoded.*

Deleting the parts in the second line of (4) we obtain a variable length code what is shorter in average, but it is inconvenient otherwise. Therefore define $c^*(I)$ skipping the 0 sequences in the second row of (4).

**Proposition 3.** *The average length of the code $c^*(I)$ is*

$$\left(1+b\frac{2^b}{2^b+1}\right)2^a. \tag{5}$$

*Proof.* The length of $c_0^*(I)=c_0(I)$ does not depend on $I$, it is $2^a$, we can consider the average length of the rest of the code. If $s=|\pi_K(I)|$ then the length of the second part of the code is $bs$. By (2) the sum of the lengths is

$$\sum_{s=0}^{2^a}\binom{2^a}{s}\left(2^b\right)^s bs=2^a2^bb\sum_{s=1}^{2^a}\binom{2^a-1}{s-1}\left(2^b\right)^{s-1}=2^a2^bb(2^b+1)^{2^a-1}. \tag{6}$$

We have to divide this by the number of instances (codewords) what was determined in Proposition 1. The ratio is really

$$2^a b \frac{2^b}{2^b+1}. \tag{7}$$

□

Let us now investigate the condition that small changes cause a small change in the code. Of course it is hopeless to achieve this in the case of codes of variable length. Therefore we return to the codes of fixed length.

3 elementary changes of $I$ will be considered.

(i) Delete one tuple.
(ii) Flip one digit in the first part, $\pi_K(t)$.
(iii) Flip one digit in the second part, $\pi_{\Omega-K}(t)$.

Let us see what the changes are in our code described in Proposition 2. In case (i), if the $i$th tuple is deleted then $f(I, \widehat{i}\,)$ becomes a sequence of zeros. The number of changes is $b+1$ since a 1 is also replaced by a 0 in $c_0$. (ii) This change causes 2 changes in $c_0$ and $2b$ in the rest, altogether $2(b+1)$. Finally, in case (iii) there is only one change in $f(I, \widehat{i}\,)$ if the original change was in the $i$th tuple.

But can we expect anything better? In case of (ii) we have to give a lot of new information, because the unchanged second part becomes the continuation of another, new first part, that is unrelated to the present one. So, only the changes (i) and (iii) might induce a small change in the code.

(There is an embarrassing question here. Deleting one tuple from the instance seems to be easy, we do not add too much new information. But adding a new tuple needs that. So, one feels that the first change should induce a small change in the code, while the second one does much more. This really happens in the case of the code of variable length given in Proposition 3. But in the case of codes of fixed length the situation is symmetric, the Hamming distance has no direction. This is why we do not treat "adding a new tuple" as an elementary change.)

Can the Hamming distance of the codes of instances obtained by both elementary changes (i) and (iii) be equal to one? Let

$$I_1 = \begin{array}{l} 0\,0\ldots0\,0 \,|\, 0\,0\ldots0\,0 \\ 0\,0\ldots0\,1 \,|\, 0\,0\ldots0\,0 \end{array},$$

$$I_2 = \begin{array}{l} 0\,0\ldots0\,0 \,|\, 0\,0\ldots0\,0 \\ 0\,0\ldots0\,1 \,|\, 0\,0\ldots0\,1 \end{array}$$

and

$$I_3 = 0\,0\ldots0\,0 \,|\, 0\,0\ldots0\,0.$$

Deleting the second tuple from $I_1$ or from $I_2$, the instance $I_3$ is obtained. Hence $d(c(I_1), c(I_3)) = 1, d(c(I_2), c(I_3)) = 1$ should hold. On the other hand, $I_2$ can be obtained from $I_1$ by flipping one digit in the second part, therefore

$d(c(I_1), c(I_2)) = 1$. This is a contradiction, since there are no three 0,1 sequences with pairwise Hamming distance 1.

The conclusion is that we can only suppose that the Hamming distance is at most 2 in the cases (i) and (iii). We say that a code $c(I)$ is *2-distance-preserving* if the following two conditions are satisfied.

(iv) If $I_2$ is obtained from $I_1$ by deleting one tuple then $d(c(I_1), c(I_2)) \leqslant 2$.
(v) If $I_2$ is obtained from $I_1$ by flipping one digit in $\pi_{\Omega - K}(t)$ (the second part) then $d(c(I_1), c(I_2)) \leqslant 2$.

It is easy to construct a code satisfying this condition. Define the following $2^a \times (1 + 2^b)$ matrix where the rows represent the first part of $I$. The 0th column contains a 1 in the $i$th row if $\widehat{i} \in \pi_K(I)$, otherwise it is 0. The $i$th row contains only 0s if $\widehat{i} \notin \pi_K(I)$. Otherwise the $j$th entry ($1 \leqslant j \leqslant 2^b$) in the $i$th row is 1 iff $f(I, \widehat{i})$ is the $j$th element of the set of all $b$-tuples, $\{0,1\}^b$ in a given order. Arranging the entries of this matrix in any way a 2-distance-preserving code is obtained. Its length is $2^a(1 + 2^b)$ what is much larger than $\mathrm{Inf}(R)$ if $b$ is large. A lot is lost as a tradeoff for having the distance preserving property. Call this code the *trivial code* and denote it by $c_{\mathrm{tr}}$. So we have

$$|c_{\mathrm{tr}}| = 2^a(1 + 2^b). \tag{8}$$

The following bound is exponential in $b$, unlike the construction of Proposition 2.

**Theorem 1.** *If $c$ is a 2-distance-preserving code of $(R, \{K \to \Omega\})$ then $|c| \geqslant \sqrt{2} \cdot 2^{a/2} \cdot 2^{b/2} - 1$.*

Observe that $c_{\mathrm{tr}}$ has an additional property what cannot be naturally supposed. Namely, changing the second part of any tuple in an instance, the code changes at at most two places:

(vi) If $I_2$ is obtained from $I_1$ by changing $\pi_{\Omega - K}(t)$ (the second part) in any arbitrary way then $d(c(I_1), c(I_2)) \leqslant 2$.

Codes satisfying properties (iv) and (vi) are called *strongly 2-distance-preserving*. The following theorem gives an improved bound on the length of such codes. This bound has the right order of magnitude in $b$.

**Theorem 2.** *Let $b \geqslant 2$. If $c$ is a strongly 2-distance-preserving code of $(R, \{K \to \Omega\})$ then $|c| \geqslant 2^b$.*

## 3 A simple multivalued dependency

In this section we consider the information content of multivalued dependency schemas. First, let us remind the reader of the definition of a multivalued dependency (mvd).

Let $R$ be a relation schema with attribute set $\Omega$ and $A, B \subseteq \Omega$. An instance $I = I(R)$ satisfies the *multivalued dependency* $A \twoheadrightarrow B$ if for any tuples $u, v \in I$ and $\pi_A(u) = \pi_A(v)$ implies that there exists a tuple $w \in I$ such that $\pi_A(w) = \pi_A(u) = \pi_A(v)$, $\pi_B(w) = \pi_B(u)$,$\pi_{\Omega-A-B}(w) = \pi_{\Omega-A-B}(v)$ holds.

Note, that due to symmetry reasons, there is also a $w' \in I$, such that $\pi_A(w') = \pi_A(u) = \pi_A(v)$, $\pi_B(w') = \pi_B(v)$,$\pi_{\Omega-A-B}(w') = \pi_{\Omega-A-B}(u)$ holds. Let $\Sigma$ be a set of mvd's. So, according to the previous more general definition, we can say that an instance $I = I(R)$ satisfies the *multivalued dependency schema (mvd schema)* $(R, \Sigma)$ if $I$ satisfies all mvd's of $\Sigma$. We denote the set of such $I$'s by $\mathcal{I}(R, \Sigma)$.

Let us introduce the following notations for a family of sets $\mathcal{F} = \{X_i \,|\, 1 \leqslant i \leqslant r\}$. Let

$$\mathcal{M}_{\mathcal{F}} = \left\{ \varnothing \neq X \subseteq \Omega \,|\, X = \bigcap_{i \in H} X_i \text{ for some } H \subseteq [r] \right\} \tag{9}$$

and

$$a(\mathcal{F}) = \{H \in \mathcal{M}_{\mathcal{F}} \,|\, \forall H' \subset H : H' \notin \mathcal{M}_{\mathcal{F}}\}. \tag{10}$$

First, let us consider a simple case when $\Sigma = \{\varnothing \twoheadrightarrow X_i \,|\, X_i \subseteq \Omega, 1 \leqslant i \leqslant r\}$. If $\varnothing \twoheadrightarrow X_i$ holds, then $\varnothing \twoheadrightarrow \Omega - X_i$ holds, too.

*Example 1.* Let $\Omega$ = {Professor, Course, Book} and consider the following instance $I$:

| Professor | Course | Book |
|---|---|---|
| Taylor | Ocean Studies | Corals |
| Smith | Ocean Studies | Corals |
| Taylor | Ocean Studies | Whales |
| Smith | Ocean Studies | Whales |
| Taylor | Mammals | Whales |
| Smith | Mammals | Whales |
| Taylor | Mammals | Monkeys |
| Smith | Mammals | Monkeys |

We have $\varnothing \twoheadrightarrow$ {Course, Book}, since if we have two tuples of $I$ and we exchange the Course-Book pair, the new tuples are in $I$ as well. This can be also viewed as we changed the professor, so $\varnothing \twoheadrightarrow$ {Professor} holds, too. On the other hand $\varnothing \twoheadrightarrow$ {Course} does not hold. Consider the tuples (Taylor, Ocean Studies, Corals) and (Smith, Mammals, Whales), but Prof. Taylor does not teach Mammals from the books on Corals. Note, that {Course} $\twoheadrightarrow$ {Book} holds as well. □

Let $\mathcal{F}_\Sigma = \{X_i \,|\, 1 \leqslant i \leqslant r\} \cup \{\Omega - X_i \,|\, 1 \leqslant i \leqslant r\}$. Then clearly $a(\mathcal{F}_\Sigma) = \{S_1, \ldots, S_m\}$ is a partiton of $\Omega$, moreover each $F \in \mathcal{F}$ is the union of some elements of $a(\mathcal{F}_\Sigma)$. Let us introduce the notation $\mathcal{D}_Y = \times_{j \in Y} D_j$ for any $Y \subseteq \Omega$.

**Theorem 3.** *Let $D_1, \ldots, D_{|\Omega|}$ be the domains of the attributes of $R$ and $\Sigma = \{\varnothing \twoheadrightarrow X_i \mid X_i \subseteq \Omega, 1 \leqslant i \leqslant r\}$. Then*

$$\mathcal{I}(R, \Sigma) = \{\varnothing\} \cup \{T_1 \times \cdots \times T_m \mid \varnothing \neq T_i \subseteq \mathcal{D}_{S_i}, 1 \leqslant i \leqslant m\}. \tag{11}$$

*Proof.* Consider a nonempty instance of the type of the right hand side of (11) and an mvd $\varnothing \twoheadrightarrow X_i \in \Sigma$. Since $X_i = \bigcup\{S_j \mid j \in H\}$ and $\Omega - X_i = \bigcup\{S_j \mid j \in [m] - H\}$ holds for some $H \subseteq [m]$ then, according to the definition of direct product, for any tuples $u_1$ and $u_2$ of $I$ there exist two tuples $v_1$ and $v_2$ such that $\pi_{X_i}(v_1) = \pi_{X_i}(u_1)$, $\pi_{\Omega - X_i}(v_1) = \pi_{\Omega - X_i}(u_2)$, $\pi_{X_i}(v_2) = \pi_{X_i}(u_2)$ and $\pi_{\Omega - X_i}(v_2) = \pi_{\Omega - X_i}(u_1)$ holds, so $I$ satisfies the dependency $\varnothing \twoheadrightarrow X_i$. This holds for each $i$, so $I \in \mathcal{I}(R, \Sigma)$.

On the other hand consider an instance $I \in \mathcal{I}(R, \Sigma)$ and two tuples $u$ and $v$. We claim that there is a tuple $w$ in $I$ having the property

$$\pi_{S_1}(w) = \pi_{S_1}(v) \text{ and } \pi_{\Omega - S_1}(w) = \pi_{\Omega - S_1}(u). \tag{12}$$

Since $S_1 \in a(\mathcal{F}_\Sigma)$ we know, that $S_1 = \bigcap \mathcal{F}'$ holds for some $\mathcal{F}' \subseteq \mathcal{F}_\Sigma$. Let $G_i = \bigcap\{F_j \mid 1 \leqslant j \leqslant i\}$ $(0 \leqslant i \leqslant |\mathcal{F}'|)$. There exists a tuple $w_i \in I$ with

$$\pi_{G_i}(w_i) = \pi_{G_i}(v) \text{ and } \pi_{\Omega - G_i}(w_i) = \pi_{\Omega - G_i}(u). \tag{13}$$

In fact, $w_0 = u \in I$ holds and suppose that $w_i \in I$ exists with property (13) for some $0 \leqslant i < |\mathcal{F}'|$. Let us define $w_{i+1}$ by

$$\pi_{F_{i+1}}(w_{i+1}) = \pi_{F_{i+1}}(w_i) \text{ and } \pi_{\Omega - F_{i+1}}(w_{i+1}) = \pi_{\Omega - F_{i+1}}(u). \tag{14}$$

Since $G_{i+1} = G_i \cap F_{i+1}$ $w_{i+1}$ satisfies property (13) (for $i+1$). By induction on $i$ we have $w_i \in I$ and we know that $u \in I$ and $\varnothing \twoheadrightarrow F_{i+1}$ holds for $I$. Therefore we have $w_{i+1} \in I$ by (14) and the definition of mvd.

Setting $w$ to be $w_{|\mathcal{F}'|}$ proves the claim since $w$ has property (12).

Analogously, for each $1 \leqslant i \leqslant m$ there exist a tuple $w_{\{S_i\}} \in I$ having the property

$$\pi_{S_i}(w_{\{S_i\}}) = \pi_{S_i}(v) \text{ and } \pi_{\Omega - S_i}(w_{\{S_i\}}) = \pi_{\Omega - S_i}(u). \tag{15}$$

We prove by induction on $|\mathcal{H}|$, that there exists a tuple $w_\mathcal{H} \in I$ for any $\mathcal{H} \subseteq a(\mathcal{F}_\Sigma)$ satisfying

$$\pi_{\bigcup\mathcal{H}}(w_\mathcal{H}) = \pi_{\bigcup\mathcal{H}}(v) \text{ and } \pi_{\Omega - \bigcup\mathcal{H}}(w_\mathcal{H}) = \pi_{\Omega - \bigcup\mathcal{H}}(u). \tag{16}$$

For $|\mathcal{H}| \leqslant 1$ the statement holds, so let $|\mathcal{H}| \geqslant 2$. Suppose, that it is true for any $\mathcal{H}' \subseteq a(\mathcal{F}_\Sigma)$ such that $|\mathcal{H}'| < |\mathcal{H}|$. There exists an $F \in \mathcal{F}_\Sigma$ satisfying $F \cap \bigcup\mathcal{H} \neq \varnothing$ and $F \cap \bigcup\mathcal{H} \neq \bigcup\mathcal{H}$, otherwise $a(\mathcal{F}_\Sigma)$ contains $\bigcup\mathcal{H}$ or a superset of it, but this is not possible for $|\mathcal{H}| \geqslant 2$. $F \cap \bigcup\mathcal{H}$ and $\bigcup\mathcal{H} - F$ are disjoint unions of some $S_i$'s, fewer than $|\mathcal{H}|$. By the induction hypothesis there exists $w_{F \cap \bigcup\mathcal{H}} \in I$ and $w_{\bigcup\mathcal{H} - F} \in I$ satisfying (16) (for $F \cap \bigcup\mathcal{H}$ and $\bigcup\mathcal{H} - F$ respectively). But then since $\varnothing \twoheadrightarrow F$ holds for $I$, $w_{\bigcup\mathcal{H}} \in I$ satisfying (16) exists.

So any $I \in \mathcal{I}(R, \Sigma)$ is of the type of the right hand side of (11). □

**Corollary 1.**

$$|\mathcal{I}(R,\Sigma)| = 1 + \prod_{S_i \in a(\mathcal{F}_\Sigma)} \left(2^{\prod_{j \in S_i} |D_j|} - 1\right) \tag{17}$$

□

We suggest the following coding of the instances. The code consists of $m(= |a(\mathcal{F}_\Sigma)|)$ blocks. The $i$th block has length $\prod_{j \in S_i} |D_j|$, each bit corresponds to an element of $\times_{j \in S_i} D_j$. Note, that only those codewords that have all 0's in a particular block are not used, except for the all 0 codeword.

By corollary 1 the length of this code is the best possible if

$$2\left(1 + \prod_{S_i \in a(\mathcal{F}_\Sigma)} \left(2^{\prod_{j \in S_i} |D_j|} - 1\right)\right) > \prod_{S_i \in a(\mathcal{F}_\Sigma)} \left(2^{\prod_{j \in S_i} |D_j|}\right) \tag{18}$$

holds. Note that there are cases in which (18) does not hold, typically in the case where the cardinality of the domains are small.

Coding and decoding is easily computable.

There is a natural partial ordering on $\mathcal{I}(R,\Sigma)$, namely let $T_1 \times \ldots \times T_m \preccurlyeq T_1' \times \cdots \times T_m'$ if $T_i \subseteq T_i'$ holds for all $1 \leqslant i \leqslant m$ ($\varnothing$ is smaller than anything).

Let us consider various updates on instance $I$. Note, that the modified instance $I'$ should be in $\mathcal{I}(R,\Sigma)$ as well. So restoring the dependency schema maybe neccessary. First, we define quite natural ways to restore the schema, then we investigate the changes in the codeword of $I$. (Of course there is always a natural way to restore the schema: undo. But we exclude this way.)

*Tuple deletion:* To restore the dependencies after deleting $u$ let the resulting instance be such an instance $I'$, that satisfiy $I' \preccurlyeq I$, that does not contain $u$ and that is maximal in the partial ordering with this property. This can be done by modifying a single bit in the codeword, unless $I$ consists of a single tuple.

*Tuple insertion:* To restore the dependencies after inserting $u$ let the resulting instance be such an instance $I'$, that satisfiy $I \preccurlyeq I'$, that does contain $u$ and that is minimal in the partial ordering with this property. This can be done by modifying as many bits in the codeword as "new" $\pi_{T_i}(u)$'s are among the corresponding projections of $u$.

*Entry modification:* To restore the dependencies after a modification in the attribute $j(\in S_i)$ of $u$ to $u'$ (i.e., $\pi_i(u) \neq \pi_i(u') \Leftrightarrow i = j$ $(1 \leqslant i \leqslant |\Omega|)$) we have 2 cases. Either $\pi_{S_i}(u') \in T_i$ (that is the modified tuple is the same as one the other tuples of $I$) or not (the modifed tuple was not a tuple of $I$ yet). In the first case, let $I'$ be such an instance, that satisfies $I' \preccurlyeq I$, that does not contain $u$ and that is maximal in the partial ordering with this property. For the the second case we use $I''$ determined by the first case. Let $I'$ be such an instance that satisfiy $I'' \preccurlyeq I'$, that does contain $u'$ and that is minimal in the partial ordering with this property. In the first case this corresponds to a modification of a single bit in the $i$th block (see tuple deletion), in the second case 2 bits are modified in the same block.

This natural code is quite good in the sense that small changes in the instance result in small changes in the code.

What can we say if the left hand sides of the mvd's are not the empty sets, but say $X$ (the same for all mvd's)? Since $X \twoheadrightarrow Y$ implies $X \twoheadrightarrow Y - X$ and vica versa we can consider $\Sigma'' = \{X \twoheadrightarrow X_i - X \mid X_i \subseteq \Omega, 1 \leqslant i \leqslant r\}$ instead of $\Sigma' = \{X \twoheadrightarrow X_i \mid X_i \subseteq \Omega, 1 \leqslant i \leqslant r\}$. The instances satisfying $\Sigma''$ are

$$\{\varnothing\} \cup \left\{ \bigcup_{w \in \mathcal{H}} \{w\} \times F(w) \,\middle|\, \varnothing \neq \mathcal{H} \subseteq \mathcal{D}_X, F : \mathcal{H} \to \bigtimes_{i=1}^{m} \left(2^{\mathcal{D}_{S_i}} - \{\varnothing\}\right) \right\}. \quad (19)$$

In other words, for a given instance $I$ satisfying $(R, \Sigma'')$ and an $X$-tuple $w \in \pi_X(I)$ the continuations that complete $w$ to an $n$-tuple of $I$ are of the form (11) (for $\Omega - X$ instead of $\Omega$). The proof of (19) is analogous to the proof of Theorem 3.

This similarity motivates to study the case of a simple set of mvd's that looks completely different. Let $\Sigma = \{A \twoheadrightarrow C, B \twoheadrightarrow D\}$, where $B = C - A$ and $A = D - B$. If these 2 conditions hold, $\Sigma$ is equivalent to the set $\{A \twoheadrightarrow B, B \twoheadrightarrow A\}$ (since $A \twoheadrightarrow C$ implies $A \twoheadrightarrow C - A$ and vica versa) or more simply to the situation where $\Omega = \{1, 2, 3\}$ and $\Sigma = \{\{1\} \twoheadrightarrow \{2\}, \{2\} \twoheadrightarrow \{1\}\}$. Let $n = |D_1|, m = |D_2|, k = |D_3|$.

For any $I \subseteq D_1 \times D_2 \times D_3$ let $G = G_I = (D_1, D_2, E)$ be the bipartite graph, that has $\{x, y\} \in E(G) \Leftrightarrow \exists z \in D_3 \ (x, y, z) \in I$. Furthermore for an edge $e = \{x, y\} \in E(G)$ let $S_e = \{z \in D_3 \mid (x, y, z) \in I\}$.

**Lemma 1.** *Let $\Omega = \{1, 2, 3\}$ and $\Sigma = \{\{1\} \twoheadrightarrow \{2\}, \{2\} \twoheadrightarrow \{1\}\}$. $\varnothing \neq I \in \mathcal{I}(R, \Sigma)$ if and only if for any $e, f \in E(G_I)$, that are in the same connected component of $G$ $S_e = S_f$ holds.*

*Proof.* Suppose that for any $e, f \in E(G_I)$, that are in the same component of $G$ $S_e = S_f$ holds. Consider two tuples $(x, y, z)$ and $(x, y', z')$. Since $\{x, y\}$ and $\{x, y'\}$ have a common vertex they are in the same component. So $S_{\{x,y\}} = S_{\{x,y'\}}$, which implies $(x, y, z'), (x, y', z) \in I$, so $\{1\} \twoheadrightarrow \{2\}$ holds. Similarly $\{2\} \twoheadrightarrow \{1\}$ holds as well, so $I \in \mathcal{I}(R, \Sigma)$.

On the other hand if $I \in \mathcal{I}(R, \Sigma)$ and $e, f \in E(G_I)$ are in the same edge-connected component then there exists a path $v_0, e_1, v_1, e_2, \ldots v_{\ell-1}, e_\ell, v_\ell$ in $G$ such that $e_1 = e$ and $e_\ell = f$. We can suppose w.l.o.g., that $v_0 \in D_1$. Let $z$ be an arbitrary element of $S_e$, i.e., $(v_0, v_1, z) \in I$. $\{v_1, v_2\} \in E(G)$ implies that there exist some $z' \in D_3$, such that $(v_2, v_1, z') \in I$. $\{2\} \twoheadrightarrow \{1\}$ implies $(v_2, v_1, z) \in I$. By a similar arguement $\{1\} \twoheadrightarrow \{2\}$ implies $(v_2, v_3, z) \in I$. By easy induction $(v_{\ell-1}, v_\ell, z) \in I$, (or $(v_\ell, v_{\ell-1}, z) \in I$,) which implies $z \in S_f$. Conversely, $z \in S_f$ implies $z \in S_e$, therefore we have $S_e = S_f$. □

Let $c_{n,m,s}$ $(1 \leqslant s \leqslant \min\{n, m\})$ denote the number of bipartite graphs with partition sizes $n$, $m$ and exactly $s$ connected components that contain at least one edge (so isolated vertices do not count as a component).

**Corollary 2.** *Let* $\Omega = \{1,2,3\}$, $n = |D_1|$, $m = |D_2|$, $k = |D_3|$ *and* $\Sigma = \{\{1\} \twoheadrightarrow \{2\}, \{2\} \twoheadrightarrow \{1\}\}$.

$$|\mathcal{I}(R,\Sigma)| = 1 + \sum_{s=1}^{\min\{n,m\}} c_{n,m,s}\left(2^k - 1\right)^s. \quad \square \tag{20}$$

The following bounds on $\mathrm{Inf}(R,\Sigma)$ follow from Corollary 2.

**Corollary 3.** *Let* $\Omega = \{1,2,3\}$, $n = |D_1| > 0$, $m = |D_2| > 0$, $k = |D_3| > 0$ *and* $\Sigma = \{\{1\} \twoheadrightarrow \{2\}, \{2\} \twoheadrightarrow \{1\}\}$.

$$nm + k - 1 \leqslant \mathrm{Inf}(R,\Sigma) \leqslant nm + \min\{n,m\}k \tag{21}$$

*Proof.* We have by (20)

$$2^{nm+k-1} \leqslant 1 + (2^{nm}-1)(2^k-1) = 1 + \left(\sum_{s=1}^{\min\{n,m\}} c_{n,m,s}\right)(2^k-1) =$$

$$1 + \sum_{s=1}^{\min\{n,m\}} c_{n,m,s}(2^k-1) \leqslant |\mathcal{I}(R,\Sigma)| \leqslant 1 + \sum_{s=1}^{\min\{n,m\}} c_{n,m,s}(2^k-1)^{\min\{n,m\}} =$$

$$1 + \left(\sum_{s=1}^{\min\{n,m\}} c_{n,m,s}\right)(2^k-1)^{\min\{n,m\}} = 1 + (2^{nm}-1)(2^k-1)^{\min\{n,m\}} \leqslant$$

$$2^{nm+\min\{n,m\}k}. \quad \square \tag{22}$$

For small $s$ we prove some asymptotic bounds on $c_{n,m,s}$ in section 5. We also discuss a conjecture on $c_{n,m,s}$ for all $s$ in section 7. Note, that if Conjecture 2 is true both the lower and the upper bounds of (21) can be close for various $n, m, k$. These possibilities are discussed in section 7.

Let us consider again the updates of $I \in \mathcal{I}(R,\Sigma)$ and the poset $\langle \mathcal{I}(R,\Sigma), \subseteq \rangle$. Note, that restoring the mvd schema after insertion of a tuple $t$ may result a big change in $I$ if $\pi_{\{1,2\}}(t)$ is a new edge that connects components $C_1, C_2 \subseteq D_1 \times D_2$. If $S_1, S_2 \in 2^{D_3}$ are the subsets of $D_3$ belonging to $C_1$ and $C_2$ respectively, then the own set of the unified component will be $S_1 \cup S_2 \in 2^{D_3}$. So, maybe there is no positive answer for Problem 3 in section 7.

## 4 Bounds on the size of 2-distance-preserving codes

Define the graph $G_1(a,b) = (V, E_1)$ where $V$ is the set of all instances of $(R, \{K \to \Omega\})$. Therefore, by Proposition 1 we have $|V| = (2^b+1)^{2^a}$. Two vertices are joined by an edge in $G_1(a,b)$ if the corresponding instances can be obtained by a change (i) or (iii). $B_\ell$ is the $\ell$-dimensional cube, more precisely it is the graph with vertex set $\{0,1\}^\ell$ where two vertices are joined by an edge if their Hamming distance is one. $B_\ell^{\leqslant 2}$ is a graph with the same vertex set, but two vertices are adjacent if their Hamming distance is 1 or 2.

Let $H_1 = (U_1, F_1)$ and $H_2 = (U_2, F_2)$ be two graphs. An injective map $m : U_1 \to U_2$ is called an *embedding* if the edges in $F_1$ are mapped into edges in $F_2$. In notation: $m : H_1 \hookrightarrow H_2$. The following is true by the definitions,

$$c \text{ is a 2-distance-preserving code of length } \ell \text{ iff } c : G_1(a, b) \hookrightarrow B_\ell^{\leqslant 2}.$$

The smallest $\ell$ for which an embedding $H = (U, F) \hookrightarrow B_\ell^{\leqslant 2}$ exists is called the $(\leqslant 2)$-*dimension* of $H$. In notation: $\dim^{\leqslant 2}(H)$.

**Lemma 2.** *Let $m : H = (U, F) \hookrightarrow B_\ell^{\leqslant 2}$ be an embedding and $u \in U$ an arbitrary vertex. Then there is another embedding $m' : H \hookrightarrow B_\ell^{\leqslant 2}$ which maps $u$ to the all-zero sequence.*

*Proof.* Add $m(u)$ to all the vectors mod 2, that is $m'(v) = m(v) + m(u)$. This operation does not change the Hamming distance, the modified map is also injective, maps edge to edge and $m'(u) = 2m(u)$ which is zero mod 2. □

**Lemma 3.** *If $H_2$ is a subgraph of $H_1$ then $dim^{\leqslant 2}(H_2) \leqslant dim^{\leqslant 2}(H_1)$.*

*Proof.* Let $m : H_1 \hookrightarrow B_\ell^{\leqslant 2}$ be an embedding where $\ell = \dim^{\leqslant 2}(H_1)$. Then $m|_{H_2}$ is also an embedding of $H_2$ into the same $B_\ell^{\leqslant 2}$. So by the definition of $(\leqslant 2)$-dimension $\dim^{\leqslant 2}(H_2) \leqslant \ell = \dim^{\leqslant 2}(H_1)$ holds. □

Let $K_r$ be the complete graph on $r$ vertices.

**Lemma 4.** *$dim^{\leqslant 2}(K_r) = r - 1$ $(r \neq 4)$, $dim^{\leqslant 2}(K_4) = 2$.*

*Proof.* First we give an embedding. The image of the map are those sequences of length $r - 1$ having at most one 1. The distance of every pair is $\leqslant 2$. Since $B_2^{\leqslant 2}$ is isomorphic to $K_4$ the set of all 0,1 sequences of length 2 gives a better construction for $r = 4$.

Conversely suppose that $m : K_r \hookrightarrow B_\ell^{\leqslant 2}$ is an embedding. The trivial inequality $\ell \geqslant \lceil \log r \rceil$ proves the correct lower bound for $r = 1, 2, 3, 4$. Suppose $r > 4$.

By Lemma 2 it can be supposed that one of the vertices, say $u$ is mapped to the all-zero sequence of length $\ell$. The maps of all other vertices must contain one or two 1s. Let $n(1)$ and $n(2)$ denote the number of sequences containing one and two 1s, respectively. Several cases will be distinguished.

$n(1) \geqslant 3$. The vertices $u_1, u_2, u_3$ have a map containing exactly one 1. Suppose that the map of a vertex $v$ contains two 1s. Then the Hamming distance of $v$ and at least one of $u_1, u_2, u_3$ is $\geqslant 3$ what is a contradiction. Therefore $n(1) = r - 1$, and $\ell \geqslant r - 1$, this case is settled.

$n(1) = 2$. Let $m(u_1)$ and $m(u_2)$ contain one 1. If the map $m(v)$ contains two 1s then the place of the 1 in $m(u_1)$ must be one of the two places of 1s in $m(v)$, otherwise their Hamming distance is 3. The same is true for $m(u_2)$ therefore the two places are uniquely determined in $m(v)$. Since $r > 4$ there must be a map whose 1s are not at these two places, contradicting the conditions. This case is impossible.

$n(1) = 1$. Suppose that the map of $u_1$ has only one 1, say in the first place in the sequence. The maps with two 1s must have one of them in the first place, again. Their second 1s must occupy $r-2$ different places. Hence we have $\ell \geqslant r-1$.

$n(1) = 0$. All maps have two 1s. Their places in $m(u_1)$ and $m(u_2)$ cannot be 4 distinct ones: this would imply Hamming distance 4. Without loss of generality one can suppose that $m(u_1)$ has 1s in the first and second places, while $m(u_2)$ in the first and third places. If $m(v)$ has 1 at the second and third places, then we cannot choose a good map for the fourth vertex of $K_r$. Therefore the map of all vertices has a 1 in the first place. The $r-1$ maps (different from the all-zero) occupy $r$ places. In this case $\ell \geqslant r$ is obtained, better than our need. □

**Lemma 5.** *If $H = (U, F)$ is a graph with one vertex of maximum degree $|U|-1$ then the smallest integer satisfying*

$$1 + x + \binom{x}{2} \geqslant |U| \tag{23}$$

*is a lower bound on $dim^{\leqslant 2}(H)$.*

*Proof.* Let $m : H \hookrightarrow B_\ell^{\leqslant 2}$ be an embedding and let $u$ be the vertex with degree $|U| - 1$. By Lemma 2 one can suppose that $m(u)$ is the all-zero sequence. The maps of all other vertices have one or two 1s in their maps. Of course $|U| = 1 + n(1) + n(2)$ holds. Now $n(1) \leqslant \ell$ and $n(2) \leqslant \binom{\ell}{2}$ imply

$$1 + \ell + \binom{\ell}{2} \geqslant |U|. \tag{24}$$

□

**Lemma 6.** *$G_1(a, b)$ contains a vertex of degree $2^a 2^b$.*

*Proof.* Consider the empty instance (no tuple) as a vertex of $G_1$. All the instances with one tuple are neighbors in $G_1$ since the deletion of the only tuple leads to the empty instance. The number of such tuples is $2^{a+b}$. □

*Proof (of Theorem 1).* Take the subgraph $H$ of $G_1$ spanned by the empty instance and its neighbors. The number of vertices of $H$ is $1 + 2^a 2^b$. Applying Lemma 5 for $H$ condition (23) becomes $\frac{x(x+1)}{2} \geqslant 2^a 2^b$. The inequality $\frac{(x+1)^2}{2} \geqslant 2^a 2^b$ has smaller or equal solutions. Hence we have $\dim^{\leqslant 2}(H) \geqslant \sqrt{2} 2^{\frac{a}{2}} 2^{\frac{b}{2}} - 1$. By Lemma 3 this is true for $G_1$, too. □

Define now the graph $G_2(a, b) = (V, E_1)$ that has the same vertex set as $G_1(a, b)$ that is $\mathcal{I}(R, \{K \to \Omega\})$. Two vertices are joined by an edge in $G_2(a, b)$ if the corresponding instances can be obtained by a change (i) or (vi). Now we want to give a lower bound on $\dim^{\leqslant 2}(G_2)$.

**Lemma 7.** *$G_2(a, b)$ contains a complete subgraph of size $1 + 2^b$.*

*Proof.* Take the empty instance and all the instances having one tuple with a fixed first part (say $\pi_K(t) = (1, 0, \ldots, 0, 0)$ ) and all possible second parts. The number of these instances is really $1 + 2^b$, and any two of them are joined by an edge in $G_2$. □

*Proof (of Theorem 2).* Use Lemma 4 with the complete subgraph $K_r$ obtained in Lemma 7 where $r = 1 + 2^b$. Lemma 4 gives $\dim^{\leqslant 2}(K_r) \geqslant 2^b$. Lemma 3 completes the proof. □

## 5 Asymptotic bounds on $c_{n,m,s}$

In this section we focus on how to calculate $c_{n,m,s}$. The question is equivalent to counting the probability $p_{n,m,s}$ of a random bipartite graph with partition sizes $n$, $m$ having exactly $s$ connected components that contain at least one edge.

$$p_{n,m,s} = \frac{c_{n,m,s}}{2^{nm}}. \tag{25}$$

Instead of considering the model of taking graphs with probability $1/2^{nm}$ we can consider the random graph model $G(n, m, 1/2)$, where each edge of the complete bipartite graph $K_{n,m}$ with partition sizes $n$ and $m$ has probability $1/2$ to be included in a random bipartite graph with partition sizes $n$ and $m$. This way, each random graph will be equally probable, too. In most of the following calculations, we consider the more general $G(n, m, p)$ model, where each edge of the complete bipartite graph $K_{n,m}$ has probability $p$ to be included, and probability $q = 1 - p$ to be not included in the random graph.

We know, that most of the graphs are connected, and that the probability of being not connected is exponentially small. But in (20) the less probable an event is the bigger is its weight, and the weight is exponential, so we need precise counting.

First of all, we need an estimate with error terms for the probability of a random graph being connected. Let us denote this probability by $p'_{n,m}$. Note, that $p'_{n,m} \neq p_{n,m,1}$, since in the latter case isolated vertices are allowed. E. N. Gilbert ([6]) has determined asymptotically the number of disconnected graphs on $n$ vertices. We adapt the ideas for bipartite graphs, but we need more precise estimates. Throughout the counting we suppose that $m, n$ is large enough if the bounds do not hold for some small values.

For the lower bound on $1 - p'_{n,m}$ we can say, that those graphs that have an isolated vertex are surely disconnected. So let $E_i$ be the event that the $i$th vertex is isolated. Then by Bonferroni's inequality ([5])

$$1 - p'_{n,m} \geqslant \sum_{i=1}^{n+m} \mathrm{P}(E_i) - \sum_{i<j} \mathrm{P}(E_i E_j) = \\ nq^m + mq^n - \binom{n}{2} q^{2m} - \binom{m}{2} q^{2n} - nmq^{m+n-1}. \tag{26}$$

For the upper bound on $1 - p'_{n,m}$ we use the following recursion:

$$1 - p'_{n,m} = \sum_{(i,j)\in\Gamma} \binom{n}{i}\binom{m-1}{j-1} p'_{i,j} q^{i(m-j)+j(n-i)}, \tag{27}$$

where $\Gamma = [0,n] \times [0,m] - \{(n,m)\} - \{(i,0) \,|\, i \in [n]\} - \{(0,j) \,|\, 2 \leqslant j \leqslant m\}$.

Let us separate the main terms of the right hand side of (27),

$$p'_{0,1}q^n + p'_{1,1}nq^{n+m-2} + p'_{n,m-1}(m-1)q^n + p'_{n-1,m}nq^m + p'_{n-2,m}\frac{n(n-1)}{2}q^{2m} + p'_{n-1,m-1}n(m-1)q^{n+m-2} + p'_{n,m-2}\frac{(m-1)(m-2)}{2}q^{2n} + \mathrm{err}_{n,m}\,. \tag{28}$$

So $\mathrm{err}_{n,m}$ is defined by (28) equals to the right hand side of (27). Note, that $p'_{0,1} = 1$ and $p'_{1,1} = p$.

An upper bound on $\mathrm{err}_{n,m}$ is the following:

$$\mathrm{err}_{n,m} \leqslant \sum_{(i,j)\in\Gamma_1} \binom{n}{i}\binom{m-1}{j-1} q^{i(m-j)+j(n-i)}, \tag{29}$$

where $\Gamma_1 = \{(i,j) \in [1,n] \times [1,m] \,|\, 3 \leqslant i + j \leqslant m + n - 3\}$.

For an upper estimate on $\mathrm{err}_{n,m}$ we need a lower bound on the exponent of $q$. Let

$$f_1(x) = (n-2)x + m, \quad f_2(x) = \frac{n}{2}\left(x + \frac{m-n}{2}\right), \quad f_3(x) = \frac{n}{2}\left(-x + m + \frac{m+n}{2}\right), \quad f_4(x) = n(-x+m+n). \tag{30}$$

We suppose for the rest of the proof, that $n \leqslant m$ holds. The proofs of the following elementary inequalities are left to the reader.

$$i(m-j) + j(n-i) \geqslant \begin{cases} f_1(i+j) & 3 \leqslant i+j \leqslant \frac{m-n}{2},\ i \neq 0 \\ f_2(i+j) & \frac{m-n}{2} \leqslant i+j \leqslant \frac{m+n}{2} \\ f_3(i+j) & \frac{m+n}{2} \leqslant i+j \leqslant m+n-\frac{m-n}{2} \\ f_4(i+j) & m+n-\frac{m-n}{2} \leqslant i+j \leqslant m+n-3 \end{cases}. \tag{31}$$

So an upper bound on $\mathrm{err}_{n,m}$ is the following:

$$\sum_{t=1}^{4} \sum_{(i,j)\in\Gamma_1} \binom{n}{i}\binom{m-1}{j-1} q^{f_t(i+j)} = \sum_{t=1}^{4} \left( \sum_{(i,j)\in\Gamma_2} \binom{n}{i}\binom{m-1}{j-1} q^{f_t(i+j)} - \sum_{j=1}^{n} \binom{m-1}{j-1} q^{f_t(j)} \right), \tag{32}$$

where $\Gamma_2 = [n] \times [1,m] - \{(1,1)\} - \{(i,j) \mid m-i-j \leqslant 2\}$. So for each $1 \leqslant t \leqslant 4$:

$$\sum_{(i,j)\in\Gamma_2} \binom{n}{i}\binom{m-1}{j-1} q^{f_t(i+j)} = \sum_{(i,j)\in[n]\times[1,m]} \binom{n}{i}\binom{m-1}{j-1} q^{f_t(i+j)} - nq^{f_t(2)} - q^{f_t(n+m)} - (n+m-1)q^{f_t(n+m-1)} - \frac{(n+m-1)(n+m-2)}{2} q^{f_t(n+m-2)}. \tag{33}$$

Let $f_t(x) = a_t(x) + c_t$ $(1 \leqslant t \leqslant 4)$, then

$$\sum_{(i,j)\in[n]\times[1,m]} \binom{n}{i}\binom{m-1}{j-1} q^{f_t(i+j)} - \sum_{j=1}^{m} \binom{m-1}{j-1} q^{f_t(j)} = q^{f_t(1)} \left( \sum_{h=0}^{m+n-1} \binom{m+n-1}{h} q^{a_t h} - \sum_{j'=0}^{m-1} q^{a_t j'} \right). \tag{34}$$

Then $\mathrm{err}_{n,m} \leqslant \mathrm{err}_{n,m,1} + \mathrm{err}_{n,m,2} + \mathrm{err}_{n,m,3} + \mathrm{err}_{n,m,4}$, where

$$\mathrm{err}_{n,m,1} = q^{m+n-2}\left((1+q^{n-2})^{n+m-1} - (1+q^{n-2})^{m-1}\right) - nq^{m+2(n-2)} - q^{(n-2)(m+n-2)+m}\left(\frac{(m+n-1)(m+n-2)}{2} + (m+n-1)q^{n-2} + q^{2(n-2)}\right), \tag{35}$$

$$\mathrm{err}_{n,m,2} = q^{\frac{n}{2}(1+\frac{m-n}{2})}\left((1+q^{\frac{n}{2}})^{n+m-1} - (1+q^{\frac{n}{2}})^{m-1}\right) - nq^{\frac{n}{2}(2+\frac{m-n}{2})} - q^{\frac{n}{2}(m+n-2+\frac{m-n}{2})}\left(\frac{(m+n-1)(m+n-2)}{2} + (m+n-1)q^{\frac{n}{2}} + q^{n}\right), \tag{36}$$

$$\mathrm{err}_{n,m,3} = q^{\frac{n}{2}(m-1+\frac{m+n}{2})}\left((1+q^{-\frac{n}{2}})^{n+m-1} - (1+q^{-\frac{n}{2}})^{m-1}\right) - nq^{\frac{n}{2}(m-2+\frac{m+n}{2})} - q^{\frac{n}{2}(2+\frac{m-n}{2})}\left(\frac{(m+n-1)(m+n-2)}{2} + (m+n-1)q^{-\frac{n}{2}} + q^{-n}\right), \tag{37}$$

$$\mathrm{err}_{n,m,4} = q^{n(m+n-1)}\left((1+q^{-n})^{n+m-1} - (1+q^{-n})^{m-1}\right) - nq^{n(m+n-2)} - 1 - (m+n-1)q^{n} - \frac{(m+n-1)(m+n-2)}{2} q^{2n}. \tag{38}$$

This gives an upper bound of

$$\mathrm{err}_{n,m} \leqslant (n+m)^4 \left(q^{3n} + q^{\frac{n}{2}(3+\frac{m-n}{2})} + q^{m+2(n-2)}\right) \tag{39}$$

for $q < (1/(m+n))^{2/n}$ (which holds for large $m, n$). This gives

$$\mathrm{err}_{n,m} \leqslant \left(\frac{n+m}{q}\right)^4 q^{3n} \tag{40}$$

for $m - n \geqslant 6$. Note, that by more careful, but similar counting one can show, that (40) is valid in the remaining cases, too. We skip this counting, here. Note, that this is a good bound only if we assume $m < 1.999n$ to ensure that a term of $q^{n+m}$ has smaller exponent than $q^{3n}$. Having too unbalanced partition sizes would complicate counting, since there is an existing situation in this case, when having 3 (or more) isolated vertices in the larger partition is more probable than having a component with a single edge.

So we proved the following lemma.

**Lemma 8.** *Let $n \leqslant m$. Then the following holds.*

$$nq^m + mq^n - \binom{n}{2}q^{2m} - \binom{m}{2}q^{2n} - nmq^{m+n-1} \leqslant 1 - p'_{n,m} \leqslant nq^m + mq^n + npq^{n+m-2} + \frac{n(n-1)}{2}q^{2m} + n(m-1)q^{n+m-2} + \frac{(m-1)(m-2)}{2}q^{2n} + \left(\frac{n+m}{q}\right)^4 q^{3n}. \quad (41)$$

Now let us turn our attention to $p_{n,m,1}$. We can have an upper bound on $p_{n,m,1}$ similarly to $p'_{n,m}$. A component with a single edge guarantees that we don't have an edge-connected bipartite graph. We use Bonferroni's inequality ([5]) again.

$$p_{n,m,1} \leqslant 1 - nmpq^{m+n-2} + n(n-1)m(m-1)p^2q^{2n+2m-6} \quad (42)$$

For the upper bound observe, that an edge-connected bipartite graph is such a graph that have $i$ isolated vertices in partition $D_1$ and $j$ isolated vertices in partition $D_2$ and the rest of the graph is connected. So

$$p_{n,m,1} = \sum_{i,j} \binom{n}{i}\binom{m}{j} q^{im+jn-ij} p'_{n-i,m-j}. \quad (43)$$

A lower bound is just a few terms of the left hand side of (43), so we have by (28) and (40)

$$p_{n,m,1} \geqslant 1 - (1 + (m-1)p'_{n,m-1})q^n - npq^{n+m-2} - p'_{m,n-1}nq^m - p'_{n-2,m}\frac{n(n-1)}{2}q^{2m} - p'_{n-1,m-1}n(m-1)q^{n+m-2} - p'_{n,m-2}\frac{(m-1)(m-2)}{2}q^{2n} - \mathrm{err}_{n,m} + mq^n p'_{n,m-1} + nq^m p'_{n-1,m} + \binom{m}{2}q^{2n}p'_{n,m-2} + nmq^{m+n-1}p'_{n-1,m-1} + \binom{n}{2}q^{2m}p_{n-2,m} \geqslant 1 - nmpq^{n+m-2} + q^n(p'_{n,m-1} - 1 + (m-1)q^n p'_{n,m-2} + np'_{n-1,m-1}q^{m-2}) \geqslant 1 - nmpq^{n+m-2} - 2\left(\frac{n+m}{q}\right)^4 q^{3n}. \quad (44)$$

We obtain the following by (42) and (44).

**Lemma 9.** *Let $n \leqslant m$ Then*

$$nmpq^{m+n-2} - n(n-1)m(m-1)p^2q^{2n+2m-6} \leqslant 1 - p_{n,m,1} \leqslant nmpq^{n+m-2} + 2\left(\frac{n+m}{q}\right)^4 q^{3n}. \quad (45)$$

So we have

$$p_{n,m,2} \leqslant nmpq^{n+m-2} + 2\left(\frac{n+m}{q}\right)^4 q^{3n}. \quad (46)$$

For larger $s$ the situation becomes more complicated. We can have the following lower bound for $p_{n,m,\geqslant s+1}$ by Bonferroni's inequality ([5])

$$p_{n,m,\geqslant s+1} \geqslant \beta_{n,m,s} - \sum_{i=s+1}^{\min\{2s,n,m\}} \beta_{n,m,i}, \quad (47)$$

A lower bound on $p_{n,m,s}$ can be obtained by the equality $p_{n,m,s} = p_{n,m,\geqslant s} - p_{n,m,\geqslant s+1}$.

All bipartite graphs with exactly/at least $s$ component of size at least 2 have a cut to an edge-connected graph with at least an edge and a bipartite graph with exactly/at least $s-1$ component of size at least 2, so we have

$$p_{n,m,s} \leqslant \sum_{(i,j)\in[1,n-s+1]\times[1,m-s+1]} \binom{n}{i}\binom{m}{j} q^{i(m-j)+j(n-i)} p_{i,j,1} p_{n-i,m-j,s-1}. \quad (48)$$

We use induction to upper estimate $p_{n,m,s}$. For $s \geqslant 3$ the main term is when $i = j = 1$ ($p_{1,1,1} = p$), so if $\Gamma_3 = [1, n-s+1] \times [1, m-s+1] - \{(1,1)\}$, then

$$p_{n,m,s} = nmpq^{n+m-2} p_{n-1,m-1,s-1} + \mathrm{err}_{n,m,s}, \quad (49)$$

where

$$\mathrm{err}_{n,m,s} \leqslant \sum_{(i,j)\in\Gamma_3} \binom{n}{i}\binom{m}{j} q^{i(m-j)+j(n-i)} p_{n-i,m-j,s-1}. \quad (50)$$

Note, that for $\beta_{n,m,s}$ we have

$$\beta_{n,m,s} = nmpq^{n+m-2}\beta_{n-1,m-1,s-1}, \quad (51)$$

so (49) and (51) are similar recursions, but unfortunately we could not find good bounds on the error term $\mathrm{err}_{n,m,s}$, especially for large $s$, to get an asymptotic upper bound on $p_{n,m,s}$.

## 6 Related work

The problem of space-efficient encoding of relational databases is related to some other recent research efforts.

The information content of a relational database schema was considered in a the papers of A. Benczúr [3], [4] but his model was entirely different, based on Kolgomorov complexity. Grumbach and Vianu [8] used standard encodings of complex object database instances on Turing tapes for efficient query answering and schema recovery. The size of their standard code of a database domain divided by the cardinality of the database domain has an upper bound of a polylogarithmic function of the cardinality of the database domain. Grumbach and Mecca [7] considered the problem of rediscovering the schema of nested relations that have been encoded as strings for storage purposes. Arenas and Libkin [2] has introduced a new information theoretical concept of relative information content of a position in the database and used it to justify Boyce-Codd normal forms. Kolahi and Libkin has successfully applied the concept for an information theoretic study on 3NF [9], XML design [10], and worst-case redundancy analysis [11]. Köhler[12] proposed and analyzed a new normal form for relational databases based on the idea of minimizing overall storage space and update costs.

## 7 Open problems and future work

We initialized a study on the information content (smallest storage space) of databases of a given dependency schema. Partial results were proved for some simple sets of functional or multivalued dependencies.

*Problem 1.* Determine $\mathrm{Inf}(R, \Sigma)$ for other small sets of dependencies (fd's, mvd's, etc.) and analyze the coding problem for elementary modifications of the instance.

In section 2, we considered the case where the dependency schema consists of a single key. This gave rise to the concept of 2-distance preserving and strongly 2-distance preserving codes which were discussed in section 2 and 4. We proved a considerably weaker bound (Theorem 1) than (8), but it is at least exponential in $b$. We believe that the trivial bound is the best possible.

*Conjecture 1.* Let $2 \leqslant a, b$. If $c$ is a 2-distance-preserving code of $(R, \{K \to \Omega\})$ then $|c| \geqslant 2^a(1 + 2^b)$.

(Strongly) $d$-distance preserving codes can be defined analogously by changing Hamming distance 2 to $d$ in the definitions.

*Problem 2.* Give good bounds on the length of (strongly) $d$-distance preserving codes.

We also discussed the problem of determining $\mathrm{Inf}(R, \Sigma)$ for multivalued dependencies. We considered the example $\Omega = \{1, 2, 3\}$ and $\Sigma = \{\{1\} \twoheadrightarrow \{2\}, \{2\} \twoheadrightarrow \{1\}\}$ in section 3 and 5. This led to the problem of counting $c_{n,m,s}$, the number of bipartite graphs with partition sizes $n$, $m$ and exactly $s$ connected components that contain at least one edge.

The asymptotic bounds for small $s$'s in section 5 support that the following may be true for all $s$.

*Conjecture 2.*

$$c_{n,m,s} \approx \begin{cases} 1 - \beta_{n,m,s} & \text{if } s = 1 \\ \beta_{n,m,s} & \text{if } s > 1 \end{cases}, \tag{52}$$

where

$$\beta_{n,m,s} = \frac{n!}{(n-s)!} \frac{m!}{(m-s)!} p^{s-1} q^{(s-1)(m+n)-s(s-1)} \tag{53}$$

and $p = q = 1/2$.

If Conjecture 2 holds, then we have by substituting (52) into (20)

$$|\mathcal{I}(R, \Sigma)| \approx 1 + 2^{nm}(h_1 + h_2), \tag{54}$$

where

$$h_1 = (1 - \frac{nm}{2^{n+m-1}})2^k \tag{55}$$

and

$$h_2 = \sum_{s=2}^{\min\{n,m\}} \frac{n!}{(n-s)!} \frac{m!}{(m-s)!} \frac{1}{2^{s-1}} \frac{1}{2^{(s-1)(m+n)-s(s-1)}} (2^k - 1)^s. \tag{56}$$

So we have

$$\log_2(\beta_{n,m,s}(2^k - 1)^s) \approx s(k - n - m + 1 + \log_2(n - s) + \log_2(m - s) + s). \tag{57}$$

As an example, if $n \leqslant m$ and $k < m - 3 - \log_2(nm)$ then $h_2$ can be upper estimated by 1, so the information content is close to the lower bound of (21). On the other hand if $k > n+m$, $h_2$ becomes the significant term and the information content is far from the lower bound. As an extreme case, if $n = m$ and $k \gg n$, only the last term itself gives $|\mathcal{I}(R, \Sigma)| > (n!)^2 2^{n(k-n)}$, which implies a less than $2n^2$ difference in $\mathrm{Inf}(R, \Sigma)$ to the upper bound of (21), so it can be close in magnitude.

We also discussed the problem of updates in section 3. The problem behaves badly for tuple insertion. An example was discussed where insertion of a single tuple implies plenty of new tuples and breaks the structure of the instance. So the answer for the following problem might be negative.

*Problem 3.* Give a coding for $(R, \Sigma)$, that has the property that small changes in the instances imply a small change in the code.

## References


1. Abiteboul, S., Hull, R., Vianu, V., Foundations of Databases. Addison-Wesley, 1994.
2. Arenas, M. and Libkin, L., An information-theoretic approach to normal forms for relational and XML data. *Journal of the ACM,* 52:246283 (2005).

3. Benczúr, A., Information measurement in relational data bases. In: Mathematical Fundamentals of Database Systems, Springer Verlag, Lecture Notes in Computer Science, 305:1-9, (1987).
4. Benczúr, A., The Evolution of Human Communication and the Information Revolution- A Mathematical Perspective. *Mathematical and Computer Modelling*, 38(7-9):691-708 (2003).
5. Bonferroni C. E., Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze,* 8:3-62 (1936).
6. Gilbert, E. N., Random graphs. *Annals of Mathematical Statistics,* 30(4):1141-1144 (1959).
7. Grumbach, S. and Mecca, G., In Search of the Lost Schema, in: *Proceedings of the 7th International Conference on Database Theory, ICDT '99, Jerusalem, Israel* Springer-Verlag, 314-331 (1999).
8. Grumbach, S. and Vianu, V., Tractable query languages for complex object databases, *Journal of Computer and System Sciences,* 51(2):149-167 (1995).
9. Kolahi, S. and Libkin, L., On redundancy vs dependency preservation in normalization: an information-theoretic study of 3NF, in: *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, PODS '06, Chicago, IL, USA*, 114-123 (2006).
10. Kolahi, S. and Libkin, L., Dependency-preserving normalization of relational and XML data, *Journal of Computer and System Sciences,* 73(4):636-647 (2007).
11. Kolahi, S. and Libkin, L., An information-theoretic analysis of worst-case redundancy in database design, *ACM Transactions on Database Systems,* 35(1):1-32 (2010).
12. Köhler, H., Global Database Design based on Storage Space and Update Time Minimization, *Journal of Universal Computer Science*, 15(1):195-240 (2009).