Some Contributions to the Minimum Representation Problem of Key Systems

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

[1] Alfréd Rényi Institute of Mathematics, 1053 Budapest, Hungary
ohkatona@renyi.hu
[2] Technische Universität Berlin, Strasse des 17. Juni 135,
10623 Berlin, Germany
krisz@renyi.hu

Abstract. Some new and improved results on the minimum representation problem for key systems will be presented. By improving a lemma of the second author we obtain better or new results on badly representable key systems, such as showing the most badly representable key system known, namely of size

$$2^{n(1-c\cdot\log\log n/\log n)},$$

where n is the number of attributes. We also make an observation on a theorem of J. Demetrovics, Z. Füredi and the first author and give some new well representable key systems as well.

Keywords: Labelled directed tree, relational database, minimum matrix representation, extremal problems.

1 Introduction

Consider a relation name R in the *relational database model*. Let $\Omega(R)$ be a finite set, its elements are called the *attributes* of R. If $|\Omega| = n$ we say that the *arity* of R is n. The relation name and the set of attributes together are called the *relation schema* and denoted by $R[\Omega]$. Suppose, that there is given a (countably) infinite set **dom**. An *n-tuple* over the relation schema $R[\Omega]$ is a total mapping u from Ω to **dom**. A *relation instance* over a relation schema $R[\Omega]$ is a (possibly empty) finite (multi)set I of n-tuples over Ω.

The value of the n-tuple u on an attribute A is denoted by $u(A)$. If $X \subseteq \Omega$ then $\pi_X(u)$ is an $|X|$-tuple v over X, such that $v(A) = u(A)$ for all $A \in X$. A relation instance I over $R[\Omega]$ *satisfies* $K \to \Omega$, denoted by $I \models K \to \Omega$, if for each pair u and v of tuples in I, $\pi_K(u) = \pi_K(v)$ implies $\pi_{\Omega\setminus K}(u) = \pi_{\Omega\setminus K}(v)$. $K \to \Omega$ is a *key dependency* where $K \subseteq \Omega$ is called a *key*. (Less formally, $K \subseteq \Omega$ is a key, if the values in K of an n-tuple determine the whole n-tuple.)

* The work was supported by the Hungarian National Foundation for Scientific Research, grant numbers T037846 and T034702.

** The work was supported by the Young Researcher Fellowship of the Alfréd Rényi Institute and the COMBSTRU program (Marie Curie Fellowship of the Europian Union) at Universität Bielefeld and Technische Universität Berlin.

A key is called *minimal key,* if it does not contain any other key as a proper subset. Since the set of keys and minimal keys determine each other (each subset of the attributes that contain a key is also a key), it is natural to investigate the system of minimal keys, which usually contains fewer members, smaller in size. Keys can be widely applied in database management, see [1].

A family $\mathcal{F}$ of subsets of a finite set is called a *Sperner system,* if for $F_1, F_2 \in \mathcal{F}$ the property $F_1 \not\subseteq F_2$ holds. The system of minimal keys is clearly a non-empty Sperner system. For a Sperner system $\mathcal{K}$ let us introduce the notation

$$I(\mathcal{K}) = \{I | I \models K \rightarrow \Omega \text{ if and only if } \exists K', K' \subseteq K, K' \in \mathcal{K}\}.$$

We call an element of $I(\mathcal{K})$ a *representation* of $\mathcal{K}$. The following basic theorem of W.W. Armstrong and J. Demetrovics states, that for every non-empty Sperner system, there is always a representation of it, i.e., there exists a relation, in which the system of minimal keys is exactly the given family of sets.

Theorem 1.1. *[2, 4] If $\mathcal{K}$ is non-empty, then $I(\mathcal{K}) \neq \emptyset$.*

In view of Theorem 1.1 it is natural to ask for the minimum size of a relation, that represents a given system of keys. Formally let $s(\mathcal{K}) = \min\{|I| \mid I \in I(\mathcal{K})\}$ denote this minimum.

Suppose, that little a priori information is known about the structure of a given database instance. If a theorem ensures the validity of an inequality among the parameters of a database and we have information on the actual values of a part of these parameters then a statement may be deduced for the rest of the parameters of the given instance. In our case, we have a theorem for the following three parameters: number of attributes, system of minimal keys (this is not a number!) and the size of the relation. So if the size of the instance is less than the size of this minimal sample database, then the system of minimal keys can not be this one, our hypothesis on the system of keys can be rejected. This argument is trying to justify the investigation of the quantity $s(\mathcal{K})$. The goal of the present paper is to extend our knowledge on the minimum representation problem of key systems. In addition to its importance they usually raise interesting and sometimes challenging mathematical problems.

Let us start with presenting some earlier results on minimum representation.

$A \subseteq \Omega$ is an *antikey* if it is not a key. An antikey is called a *maximal antikey,* if other antikeys do not include it. If $\mathcal{K}$ is the system of minimal keys, denote the system of maximal antikeys by $\mathcal{K}^{-1}$. There is a strong connection between $s(\mathcal{K})$ and $|\mathcal{K}^{-1}|$, namely the magnitude of $s(\mathcal{K})$ is between $|\mathcal{K}^{-1}|$ and its square root.

Theorem 1.2. *[5] If $\mathcal{K} \neq \emptyset$ is a Sperner system, then the following two inequalities hold,*

$$|\mathcal{K}^{-1}| \leq \binom{s(\mathcal{K})}{2} \quad \textit{and} \quad s(\mathcal{K}) \leq 1 + |\mathcal{K}^{-1}|. \tag{1}$$

Informally, we say that a Sperner system $\mathcal{K}$ is well/badly representable if $s(\mathcal{K})$ is close to the lower/upper bound of Theorem 1.2. It is easy to see, that a minimal representation have the following two basic properties.

Proposition 1.1. *Suppose, that $I \in I(\mathcal{K}), |I| = s(\mathcal{K})$ is a minimal representation of the Sperner system $\mathcal{K}$, then the following properties hold,*

(i) *for every $A \in \mathcal{K}^{-1}$ there exist two tuples, u and v in I, such that $\pi_A(u) = \pi_A(v)$,*

(ii) *there are no two different tuples u and v in I, such that $\pi_K(u) = \pi_K(v)$ holds for some $K \in \mathcal{K}$.*

Let us mention two results from the 1980's. $\mathcal{K}_k^n$ denotes the family of all k-element subsets of the n-element Ω.

Theorem 1.3. *[5]* $2 \leq k < n \;\Rightarrow\; \exists c_1 = c_1(k), c_2 = c_2(k)$

$$c_1 n^{(k-1)/2} < s(\mathcal{K}_k^n) < c_2 n^{(k-1)/2}.$$

Theorem 1.4. *[8]* $n > n_0(k), k \geq 1 \;\Rightarrow\; \exists c_3 = c_3(k), c_4 = c_4(k)$

$$c_3 n^{(2k+1)/3} < s(\mathcal{K}_{n-k}^n) < c_4 n^k, \qquad \frac{1}{12} n^2 < s(\mathcal{K}_{n-2}^n) < \frac{1}{2} n^2.$$

It has been proved in [7] that there is a Sperner system $\mathcal{K}$ such that

$$s(\mathcal{K}) > \frac{1}{n^2}\binom{n}{\lfloor \frac{n}{2} \rfloor}.$$

Its proof is, however probabilistic. It does not give a construction. No constructed Sperner system $\mathcal{K}$ exists in the literature with exponential $s(\mathcal{K})$. We will show such a construction in section 3. More precisely, $s(\mathcal{K})$ will have an exponent nearly n. The method of proving a lower estimate on $s(\mathcal{K})$ is the same as that of [11]. The method was based on a lemma on labelled (by subsets) trees. Section 2 improves the statement of this lemma, giving a sharp estimate replacing the estimate of [11]. This improvement makes us able to prove the exponential lower estimate.

In section 4 we return to Theorem 1.3. First (subsection 4.1) the upper estimate is improved (it becomes independent of k). Subsection 4 is a small observation showing that the method of [5] can be used to prove a good upper estimate for other (non-uniform) $\mathcal{K}$s. In section 5 we summarize the related questions to be done.

2 An Extremal Problem on Labelled Directed Trees Revisited: A Tool for Deriving Results on Badly Representable Key Systems

A tree F is called a *directed tree,* if there is a direction on the edges, so that a vertex v_0 *(root)* has only out-neighbours, and an arbitrary vertex $v \neq v_0$ has a uniquely determined in-neighbour $n(v)$. $N(v)$ denotes the out-neighbourhood of v. The set of the leaves of a tree F is denoted by $\ell(F)$. Let U be a (finite) set.

A tree $F = F(U)$ is called *labelled*, if a subset $A(v)$ of U is associated with each vertex v of F.

For fixed integers $k \geq 1$, $\ell \geq 2$ and $U = \{1, 2, ..., m\}$ consider the family of labelled directed trees $\mathcal{F}_{k,\ell}^{(m)}$, for which the vertices of each tree $F \in \mathcal{F}_{k,\ell}^{(m)}$ are labelled as follows. The label of the root v_0 of F is $A(v_0) = U$. For an arbitrary vertex v of F there is a disjoint partition $N(v) = \bigcup_{i=1}^{\ell} N_i(v)$ of its out-neighbourhood satisfying the following properties.

$$A(v) \subseteq A(n(v)) \quad (v \neq v_0), \tag{2}$$

$$|A(v)| \geq k+1, \tag{3}$$

$$w_1, w_2 \in N_i(v) \Rightarrow A(w_1) \cap A(w_2) = \emptyset \quad (1 \leq i \leq \ell), \tag{4}$$

$$w_1 \in N_i(v), w_2 \in N_j(v) \Rightarrow |A(w_1) \cap A(w_2)| \leq k \quad (1 \leq i < j \leq \ell). \tag{5}$$

Introduce the notation $T_{k,\ell}(m) = \max\{|\ell(F)| \mid F \in \mathcal{F}_{k,l}^{(m)}\}$. If $k = 1$, we simply write $\mathcal{F}_{\ell}^{(m)}$ for $\mathcal{F}_{k,\ell}^{(m)}$ and $T_\ell(m)$ for $T_{k,\ell}(m)$.

Throughout the rest of the paper we write simply log for $\log_2$. We have for $\mathcal{F}_\ell$ the following:

Theorem 2.1.

$$T_2(m) \leq \frac{1}{2} m \log m. \tag{6}$$

and equality holds if and only if m is a power of 2.

Theorem 2.2.

$$T_\ell(m) = \Theta_\ell(m \log^\alpha m) \tag{7}$$

for $\ell \geq 3$. Where $\alpha = \alpha(\ell) = \log \ell$.

In [11], the magnitude of $T_{k,\ell}$ was determined.

2.1 Case $\ell = 2$

We will use the concept of *entropy* [3] in the proof. Entropy is a measure of a random variable X:

$$H(X) = -\sum_i p_i \log p_i, \tag{8}$$

where $\mathrm{Prob}(X = i) = p_i$. It is known, that

$$H((X, Y)) \leq H(X) + H(Y). \tag{9}$$

The proof is by induction on m. (6) holds for $m = 2$. Suppose that the statement holds for every integer smaller than m.

Let $F \in \mathcal{F}_2^{(m)}$ be a tree with $|\ell(F)| = T(m)$. Furthermore let $N(v_0) = \{v_1, \ldots, v_s, w_1, \ldots, w_t\}$, $N_1(v_0) = \{v_1, \ldots, v_s\}$, $N_2(v_0) = \{w_1, \ldots, w_t\}$. Let us use the short notations $A_i = A(v_i)$, $a_i = |A_i|$, $(1 \leq i \leq s)$, $B_i = A(w_i)$, $b_i = |B_i|$, $(1 \leq i \leq t)$. The subtree of F of root v_i (w_j) is denoted by F_i (F_{s+j}), $1 \leq i \leq s$ $(1 \leq j \leq t)$.

By the induction hypothesis

$$T(a_i) \le \frac{1}{2} a_i \log a_i, \ (1 \le i \le s), \qquad \text{and} \qquad T(b_i) \le \frac{1}{2} b_i \log b_i, \ (1 \le i \le t)$$

holds. So it is enough to prove that

$$\sum_{i=1}^{s} a_i \log a_i + \sum_{i=1}^{t} b_i \log b_i \le m \log m, \tag{10}$$

since then

$$T(m) = |\ell(F)| = \sum_{i=1}^{s+t} |\ell(F_i)| \le \sum_{i=1}^{s} T(a_i) + \sum_{i=1}^{t} T(b_i)$$

$$\le \sum_{i=1}^{s} \frac{1}{2} a_i \log a_i + \sum_{i=1}^{t} \frac{1}{2} b_i \log b_i \le \frac{1}{2} m \log m.$$

Let $s_1 = m - \sum_{i=1}^{s} a_i$ and $s' = s + s_1$. Add s_1 disjoint sets $A_{s+1}, \ldots, A_{s'}$ of cardinality 1, such that $\{1, 2, \ldots, m\} = \bigcup_{i=1}^{s'} A_i$. We define t_1, t' and the sets $B_{t+1}, \ldots, B_{t'}$ analogously. The sets A_i $(1 \le i \le s')$ and B_j $(1 \le j \le t')$ have the following properties:

$$\{A_i, 1 \le i \le s'\} \ \text{ is a partition of } \{1, 2, \ldots, m\}, \tag{11}$$

$$\{B_j, 1 \le j \le t'\} \ \text{ is a partition of } \{1, 2, \ldots, m\}, \tag{12}$$

$$|A_i \cap B_j| \le 1, 1 \le i \le s', 1 \le j \le t'. \tag{13}$$

Let $\Omega_X = \{1, 2, \ldots, m\}$ be the event space of the random variable X. Furthermore, let $X(\omega) = \omega$, $\omega \in \Omega_X$ and $\text{Prob}(X = \omega) = 1/m$. Let us define another two random variables, $Y(X \in A_i) = i$, $1 \le i \le s'$ and $Z(X \in B_j) = j$, $1 \le j \le t'$. Then

$$\text{Prob}(Y = i) = \frac{a_i}{m} \ (1 \le i \le s') \qquad \text{and} \qquad \text{Prob}(Z = j) = \frac{b_j}{m} \ (1 \le j \le t').$$

The random variables Y and Z are well defined by (11) and (12). Furthermore, by (13) we get

$$\text{Prob}((Y, Z) = (i, j)) = \begin{cases} \text{Prob}(X = k) = 1/m & \text{if} \quad A_i \cap B_j = \{k\}, \\ 0 & \text{if} \quad A_i \cap B_j = \emptyset. \end{cases}$$

So we have for the entropies of Y, Z and (Y, Z):

$$H(Y) = \sum_{i=1}^{s'} \frac{a_i}{m} \log \frac{m}{a_i}, \qquad H(Z) = \sum_{j=1}^{t'} \frac{b_j}{m} \log \frac{m}{b_j},$$

$$H((Y, Z)) = -\sum_{i=1}^{s'} \sum_{j=1}^{t'} \text{Prob}((Y, Z){=}(i, j)) \log \text{Prob}((Y, Z){=}(i, j)) =$$

$$\sum_{i=1}^{m} \frac{1}{m} \log m = \log m.$$

Therefore, by (9) we get

$$\log m \leq \sum_{i=1}^{s'} \frac{a_i}{m} \log \frac{m}{a_i} + \sum_{j=1}^{t'} \frac{b_j}{m} \log \frac{m}{b_j},$$

which is equivalent to (10).

2.2 Case $\ell \geq 3$

To prove Theorem 2.2 we need somewhat more counting. We could not find a straightforward way to generalize the concept of entropy for this case, altough the main idea (see equation (16)) of the proof comes from the proof of the case $\ell = 2$.

In this section we will use the following notations. If $\mathbf{u} = (u_1, \ldots, u_t)$ for some t, then let $P_i(\mathbf{u}) = \sum_{j=1}^{t} u_j^i$ and $\sigma_2(\mathbf{u}) = \sum_{1 \leq i < j \leq t} u_i u_j$. Let $A(\mathbf{u}) = (\sum_{j=1}^{t} u_j)/t$ and $G(\mathbf{u}) = (\prod_{j=1}^{t} u_j)^{1/t}$ denote the arithmetic and geometric mean, respectively. We will use the notations $A_2(\mathbf{u}) = A(\sigma_2(\mathbf{u}))$ and $G_2(\mathbf{u}) = G(\sigma_2(\mathbf{u}))$ as well.

Lower estimation of $T_\ell(m)$. Let $\mathcal{H}_\ell^q$, $(\ell \leq q)$ be a partial affine plane of order q, i.e., a set of ℓq lines on q^2 points, such that the lines form ℓ parallel classes and lines from different parallel classes intersect in 1.

Suppose, that we have for m square

$$T_\ell(\sqrt{m}) \geq C_\ell \cdot \sqrt{m} \log^\alpha \sqrt{m}, \tag{14}$$

and let $F \in \mathcal{F}_\ell^{(\sqrt{m})}$ be a tree with $|\ell(F)| = T_\ell(\sqrt{m})$. Suppose furthermore, that there exists a partial affine plane, $\mathcal{H}_\ell^{\sqrt{m}}$.

Let $T \in \mathcal{F}_\ell^{(m)}$ be the following tree. The root has $\ell\sqrt{m}$ out-neighbours. Each of them are labelled by one of the $\ell\sqrt{m}$ members of the partial affine plane, $\mathcal{H}_\ell^{\sqrt{m}}$. These out-neighbours are roots of one copy of F each. Then

$$T_\ell(m) \geq |\ell(T)| = \ell\sqrt{m}|\ell(F)| \geq \ell\sqrt{m}(C_\ell\sqrt{m} \log^\alpha \sqrt{m}) = C_\ell \cdot m \log^\alpha m.$$

It is known, that there exist a partial affine plane of order q with ℓ parallel classes if and only if there exist ℓ pairwise orthogonal latin squares of order q.

There are only partial results known about the existence of pairwise orthogonal latin squares, [10]. Of course, partial affine planes exist if affine planes exist, i.e., for prime powers.

The statement follows from the fact that prime powers occur densly, see (39).

Upper estimation of $T_\ell(m)$. Let $\ell \geq 3$ arbitrary. We prove by induction on m. We have to prove, that

$$T_\ell(m) \leq cm \log^\alpha m \tag{15}$$

holds for some $c = c(\ell)$ to be chosen later.

For small m (15) is true if c is large enough. Suppose that the statement is true for every integer smaller than m.

Let $F \in \mathcal{F}_\ell^{(m)}$ be a tree with $|\ell(F)| = T(m)$. Let $N(v_0) = \{v_1, \ldots, v_t\}$. The number of the leaves can be maximal only if for the subtrees F_i of F of root v_i, $|\ell(F_i)| = T(m_i)$ holds, where $m_i = |A(v_i)|$. Furthermore let us introduce the short notation $N_j = N_j(v_0)$, $1 \leq j \leq \ell$.

Case A. *$\exists 1 \leq j \leq t$, such that $m_j \geq m/\log^{1/2} m$.*

We need the following observation:

Proposition 2.1. *$T_\ell(m) \leq T_\ell(m-1) + m - 1$.*

Proof. Induction on m. The statement is true for $m = 1$. Suppose that the statement is true for every integer smaller than m. Let $F \in \mathcal{F}_\ell^{(m)}$ be a tree with $|\ell(F)| = T(m)$. Let $N(v_0) = \{v_1, \ldots, v_t\}$. The number of the leaves can be maximal only if for the subtrees F_i of F of root v_i, $|\ell(F_i)| = T(m_i)$ holds, where $m_i = |A(v_i)|$. Let $m_i' = |A(v_i)\backslash\{m\}|$. Consider the following tree $F' \in \mathcal{F}_\ell^{(m-1)}$. Let $N(v_0') \subseteq \{v_1', \ldots, v_t'\}$, $v_i' \in N(v_0') \Leftrightarrow m_i' \geq 2$. $A(v_i') = A(v_i)\backslash\{m\}$. The subtree F_i' of F' of root v_i' is a tree with $|\ell(F_i')| = m_i'$. Then

$$T(m) = \sum_{i=1}^{t} T(m_i) = \sum_{i=1}^{t} T(m_i') + \sum_{i=1}^{t} (T(m_i) - T(m_i')) = |\ell(F')| +$$

$$\sum_{m \in A(v_i)} (T(m_i) - T(m_i - 1)) \leq T(m-1) + \sum_{m \in A(v_i)} (m_i - 1) \leq T(m-1) + m - 1.$$

The last inequality holds by (5), the previous one by the induction hypothesis. □

Let $m_j = \beta m$. It is enough to prove, that $\sum_{i=1}^{t} T(m_i) \leq cm \log^\alpha m$.

$$\sum_{i=1}^{t} T(m_i) = T(m_j) + T(m - m_j) + \sum_{i \neq j : A(v_i) \cap A(v_j) \neq \emptyset} (T(m_i) - T(m_i - 1)) \leq$$

$$c\beta m \log^\alpha \beta m + c(1-\beta) m \log^\alpha (1-\beta) m + \ell(m - m_j) = cm \log^\alpha m +$$

$$c\beta m(\log^\alpha \beta m - \log^\alpha m) + c(1-\beta)m(\log^\alpha (1-\beta)m - \log^\alpha m) + \ell(1-\beta)m \leq$$

$$cm \log^\alpha m - 2\frac{c\alpha\beta(1-\beta)m}{\ln 2} \log^{\alpha-1} m + \ell(1-\beta)m \leq cm \log^\alpha m +$$

$$(1-\beta)m(\ell - \frac{2\alpha c}{\ln 2} \log^{1/20} m) \leq cm \log^\alpha m,$$

if $\log^{1/20} m > (\ell \ln 2)/(2\alpha c)$, which is true for every m if c is large enough. We have used only (4)-(5) (at the first inequality), Proposition 2.1, and the fact that the function $x \mapsto \log^\alpha x$ is concave.

Case B. *$\forall 1 \leq j \leq t$, $m_j < m/\log^{1/2} m$.*

Let us introduce the notation $N_j' = N_j \cup \{u_j^{(h)} : 1 \leq h \leq m, \nexists v_i \in N_j : h \in A(v_i)\}$. Let $A(u_j^{(h)}) = \{h\}$, $m_i = A(w)$ for $w \in N_j'$. Then the following notation is well defined. Let $m_j^{(h)} = |A(w)|$, where $w \in N_j'$, $h \in A(w)$.

We will prove that the following inequality holds for every $1 \leq h \leq m$ and $m > M(\ell)$:

$$\log^\alpha m - \sum_{i=1}^{\ell} \log^\alpha m_i^{(h)} \geq \frac{\alpha}{\ln 2} \log^{\alpha-1} m(1 - \frac{\sum_{1\leq j<k\leq \ell} m_j^{(h)} m_k^{(h)}}{\binom{\ell}{2} m}). \tag{16}$$

Using this inequality we get for $m > M$:

$$m \log^\alpha m - \sum_{j=1}^{\ell} \sum_{i:i\in N_j'} m_i \log^\alpha m_i = \sum_{h=1}^{m} (\log^\alpha m - \sum_{j=1}^{\ell} \log^\alpha m_j^{(h)}) \geq$$

$$\sum_{h=1}^{m} \frac{\alpha}{\ln 2} \log^{\alpha-1} m(1 - \frac{\sum_{1\leq j<k\leq \ell} m_j^{(h)} m_k^{(h)}}{\binom{\ell}{2} m}) =$$

$$\frac{\alpha \log^{\alpha-1} m}{(\ln 2)\binom{\ell}{2} m} \sum_{h=1}^{m} \sum_{1\leq j<k\leq \ell} (m - m_j^{(h)} m_k^{(h)}) =$$

$$\frac{\alpha \log^{\alpha-1} m}{(\ln 2)\binom{\ell}{2} m} \sum_{1\leq j<k\leq \ell} \sum_{h=1}^{m} (m - m_j^{(h)} m_k^{(h)}) \geq$$

$$\frac{\alpha \log^{\alpha-1} m}{(\ln 2)\binom{\ell}{2} m} \sum_{1\leq j<k\leq \ell} (m^2 - (\sum_{w\in N_j'} |A(w)|)(\sum_{w\in N_k'} |A(w)|)) = 0.$$

From this, (15) straightforwardly follows in *Case B*, as well. So, we only have to prove, that inequality (16) holds in the case when m is large enough and none of the m_js are too large. Let us fix h, we will omit the upper index of $m_j^{(h)}$ in the rest of the proof, so from now on, we suppose, that $1 \leq m_j \leq m/\log^{1/2} m$ $(1 \leq j \leq \ell)$ holds.

Case B.1. $\forall 1 \leq i \leq \ell : m_i \leq \sqrt{m} \log^5 m$.
Case B.1.1. $\forall 1 \leq j \leq \ell : m_j = \sqrt{m} + \delta_j, |\delta_j| \geq (1/3)\sqrt{m}$.

Proposition 2.2.

$$(\log^\alpha)^{(i)}(x) = \sum_{k=1}^{i} C_k^{(i)} \frac{1}{\ln^k 2} \frac{\alpha!}{(\alpha-k)!} x^{-k} \log^{\alpha-k} x \tag{17}$$

and the constants $C_k^{(i)} \in \mathbb{Z}$ have the following properties:

(i) $C_k^{(i)} = -(i-1)C_k^{(i-1)} + C_{k-1}^{(i-1)}$,
(ii) $C_1^{(i)} = (-1)^{i-1}(i-1)!$,
(iii) $|C_k^{(i)}| \leq \frac{i!}{k!} 2^i$

Proof. Easy induction on i. □

Proposition 2.3.

$$\sum_{k=0}^{\infty}|\binom{\alpha}{k}|<\infty \tag{18}$$

holds for $\alpha > 1$.

Proof.

$$\sum_{k=0}^{\infty}|\binom{\alpha}{k}|=\sum_{k=0}^{\lceil\alpha\rceil-1}|\binom{\alpha}{k}|+\sum_{i=1}^{\infty}\sum_{i\lceil\alpha\rceil}^{(i+1)\lceil\alpha\rceil-1}|\binom{\alpha}{k}|\leq$$

$$\sum_{k=0}^{\lceil\alpha\rceil-1}|\binom{\alpha}{k}|+\sum_{i=1}^{\infty}\lceil\alpha\rceil\frac{1}{\binom{i\lceil\alpha\rceil}{\lceil\alpha\rceil}}\leq\sum_{k=0}^{\lceil\alpha\rceil-1}|\binom{\alpha}{k}|+\lceil\alpha\rceil\sum_{i=1}^{\infty}\frac{1}{i^{\alpha}}<\infty. \qquad \square$$

Let $a_2(\alpha)=\sum_{k=2}^{\infty}|\binom{\alpha}{k}|$, and $a_{\max}(\alpha)=\max_{k=0}^{\infty}|\binom{\alpha}{k}|$. Note, that $a_{\max}(\alpha)\leq\ell$ and $a_2(\alpha)\leq\ell+2\alpha+2$.

Let $\boldsymbol{\delta}=(\delta_1,\delta_2,\dots,\delta_\ell)$. Then we have to prove the following inequality:

$$\log^{\alpha}m-\sum_{i=1}^{\ell}\log^{\alpha}(\sqrt{m}+\delta_i)\geq-\frac{\alpha}{\ln 2}(\log^{\alpha-1}m)(\frac{2P_1(\boldsymbol{\delta})}{\ell\sqrt{m}}+\frac{\sigma_2(\boldsymbol{\delta})}{\binom{\ell}{2}m}). \tag{19}$$

If δ is small enough we know, that

$$\log^{\alpha}(x+\delta)=\sum_{i=0}^{\infty}\frac{(\log^{\alpha})^{(i)}(x)\delta^{i}}{i!} \tag{20}$$

holds, so using (17) we get for every $1\leq j\leq\ell$

$$\log^{\alpha}(\sqrt{m}+\delta_j)=\log^{\alpha}\sqrt{m}+\frac{\alpha}{\ln 2}\frac{\log^{\alpha-1}\sqrt{m}}{\sqrt{m}}\delta_j-\frac{\alpha}{\ln 2}\frac{\log^{\alpha-1}\sqrt{m}}{(\sqrt{m})^2}\delta_j^2+$$

$$\frac{\alpha(\alpha-1)}{\ln^2 2}\frac{\log^{\alpha-2}\sqrt{m}}{(\sqrt{m})^2}\delta_j^2+\sum_{i=3}^{\infty}\frac{\delta_j^i}{i!(\sqrt{m})^i}\sum_{k=1}^{i}\frac{\alpha!}{(\alpha-k)!}\frac{C_k^{(i)}}{\ln^k 2}\log^{\alpha-k}\sqrt{m}.$$

Using Proposition 2.2 we get

$$|\sum_{i=3}^{\infty}\frac{\delta_j^i}{i!(\sqrt{m})^i}\sum_{k=2}^{i}\frac{\alpha!}{(\alpha-k)!}\frac{C_k^{(i)}}{\ln^k 2}\log^{\alpha-k}\sqrt{m}|\leq$$

$$\sum_{i=3}^{\infty}\frac{|\delta_j|^i}{i!(\sqrt{m})^i\ln^i 2}\sum_{k=2}^{i}|\binom{\alpha}{k}|k!\ln^{i-k}2|C_k^{(i)}|\log^{\alpha-k}\sqrt{m}\leq$$

$$\sum_{i=3}^{\infty}\frac{|\delta_j|^i}{i!(\sqrt{m})^i\ln^i 2}a_{\max}(\alpha)i!2^i\frac{1}{1-\ln 2}\log^{\alpha-2}\sqrt{m} =$$
$$\frac{1}{1-\ln 2}a_{\max}(\alpha)\log^{\alpha-2}\sqrt{m}\sum_{i=3}^{\infty}(\frac{2|\delta_j|}{\sqrt{m}\ln 2})^i <$$
$$2100a_{\max}(\alpha)(\frac{|\delta_j|}{\sqrt{m}})^3\log^{\alpha-2}\sqrt{m}.$$

Using Proposition 2.2 and substituting into (19) we get that it is enough to prove the following inequality:

$$\frac{\alpha}{\ln 2}\log^{\alpha-1}\sqrt{m}(\frac{P_1^2(\boldsymbol{\delta})+(2\ell-3)P_2(\boldsymbol{\delta})}{(2\ell-2)m}-\sum_{i=3}^{\infty}\frac{(-1)^{i-1}}{i}\frac{P_i(\boldsymbol{\delta})}{(\sqrt{m})^i}-$$
$$\frac{1}{\log\sqrt{m}}(\frac{P_2(\boldsymbol{\delta})}{m}\frac{\alpha-1}{\ln 2}+2100a_{\max}(\alpha)\frac{P_3(\boldsymbol{\delta}^+)}{(\sqrt{m})^3}))\geq 0, \quad (21)$$

where $\boldsymbol{\delta}^+ = (|\delta_1|, |\delta_2|, \ldots, |\delta_\ell|)$. The LHS of the previous inequality can be underestimated by

$$\frac{\alpha}{\ln 2}\log^{\alpha-1}\sqrt{m}(\frac{P_1^2(\boldsymbol{\delta})+(2\ell-3)P_2(\boldsymbol{\delta}^+)}{(2\ell-2)m}-\frac{1}{3}\frac{P_3(\boldsymbol{\delta}^+)}{(\sqrt{m})^3}-$$
$$\frac{1}{\log\sqrt{m}}(\frac{P_2(\boldsymbol{\delta}^+)}{m}\frac{\alpha-1}{\ln 2}+2100a_{\max}(\alpha)\frac{P_3(\boldsymbol{\delta}^+)}{(\sqrt{m})^3}))\geq\frac{\alpha}{\ln 2}\log^{\alpha-1}\sqrt{m}(\frac{P_1^2(\boldsymbol{\delta})}{(2\ell-2)m}+$$
$$\frac{P_2(\boldsymbol{\delta}^+)}{m}(\frac{2\ell-3}{2\ell-2}-\frac{1}{9}-\frac{1}{\log\sqrt{m}}(\frac{\alpha-1}{\ln 2}+700a_{\max}(\alpha)))), \quad (22)$$

which is nonnegative if $m > \ell^5 2^{2200\ell-5}$. (We used in the last estimation that $|\delta_i| < (1/3)\sqrt{m}$, $1 \leq i \leq \ell$.)

Case B.1.2. $\exists 1 \leq j \leq \ell : |m_j - \sqrt{m}| > (1/3)\sqrt{m}$.

Proposition 2.4. *Let* $\mathbf{u} = (u_1, u_2, \ldots, u_\ell)$, $\ell \geq 3$, $u_i > 0, 1 \leq i \leq \ell$, $u_\ell \notin [2/3, 4/3]$. *Then there exist a constant* $C(\ell) > 0$, *such that the following inequality holds.*

$$A_2(\mathbf{u}) - 1 - \ln G_2(\mathbf{u}) \geq C(\ell).$$

Proof. The LHS of (24) is nonnegative, since $A_2(\mathbf{u}) \geq G_2(\mathbf{u})$ holds by the arithmetic-geometric inequality and $G_2(\mathbf{u}) - 1 \geq \ln G_2(\mathbf{u})$ is true as well due to the fact $x-1 \geq \ln x$, $x > 0$. In both inequalities equality can hold for infinitely many $\mathbf{u}$'s under the conditions of the lemma, but the statement says that equality cannot hold simultaneously.

First, suppose that $|G_2(\mathbf{u}) - 1| \geq 1/10$. In this case, $G_2(\mathbf{u}) - 1 - \ln G_2(\mathbf{u}) \geq \min\{0.1 - \ln 1.1, -0.1 - \ln 0.9\} > 1/214$.

So we can assume, that $|G_2(\mathbf{u}) - 1| < 1/10$. Let $\mathbf{u}^* = (u_1, u_2, \dots, u_{\ell-1})$ and $\mathbf{u}' = (u_1, u_2, \dots, u_{\ell-1}, v)$, $v(\neq u_\ell) > 0$ to be chosen later. So we have

$$A_2(\mathbf{u}) - G_2(\mathbf{u}) = (u_\ell - v)\frac{2}{\ell}A(\mathbf{u}^*) + A_2(\mathbf{u}') - G_2(\mathbf{u}) \geq (u_\ell - v)\frac{2}{\ell}G(\mathbf{u}^*) +$$

$$G_2(\mathbf{u}') - G_2(\mathbf{u}) = (u_\ell - v)(\frac{2}{\ell}G(\mathbf{u}^*) - \frac{u_\ell^{2/\ell} - v_\ell^{2/\ell}}{u_\ell - v}G(\mathbf{u}^*)^{\frac{2(\ell-1)}{\ell}}) \geq$$

$$(u_\ell - v)G(\mathbf{u}^*)\frac{2}{\ell}(1 - v^{2/\ell-1}G(\mathbf{u}^*)^{\frac{\ell-2}{\ell}}) =$$

$$u_\ell^{1-\frac{1}{\ell-1}}(1 - \frac{v}{u_\ell})G(\mathbf{u})^{\frac{\ell}{\ell-1}}\frac{2}{\ell}(1 - (\tfrac{G(\mathbf{u}^*)}{v})^{\frac{\ell-2}{\ell}}), \quad (23)$$

in the first inequality we used the arithmetic-geometric inequality twice, while in the second ineqality the fact that $x \mapsto x^{2/\ell}$ is a concave function.

If $u_\ell > 4/3$, then $G(\mathbf{u}^*) < G(\mathbf{u}) = \sqrt{G_2(\mathbf{u})}$, so if we choose $v = 1.2$, we have a lower estimate of $1/(37\ell)$ for the RHS of (23) (minimizing the product term by term).

On the other hand, if $u_\ell < 2/3$, then $G(\mathbf{u}^*) > G(\mathbf{u}) = \sqrt{G_2(\mathbf{u})}$, so if we choose $v = 0.8$, we get a lower estimate of $1/(72\ell)$.

So the statement holds if we choose

$$C(\ell) = 1/(72\ell). \quad (24)$$

□

Let $a_i = u_i\sqrt{m}$, $u_i \leq \log^5 m, 1 \leq i \leq \ell, u_\ell \notin [2/3, 4/3]$. So in this case (16) has the following form:

$$\log^\alpha m - \sum_{i=1}^{\ell}\log^\alpha(u_i\sqrt{m}) \geq \frac{\alpha}{\ln 2}\log^{\alpha-1} m(1 - \frac{\sum_{1\leq j<k\leq\ell} u_j u_k}{\binom{\ell}{2}}). \quad (25)$$

By the generalized binomial theorem we have

$$(\log u_i + \log\sqrt{m})^\alpha = \sum_{j=0}^{\infty}\binom{\alpha}{j}(\log^j u_i)(\log^{\alpha-j}\sqrt{m}) \geq$$

$$\log^\alpha\sqrt{m} + \alpha(\log u_i)\log^{\alpha-1}\sqrt{m} + (\ell - \alpha - 1)(\log^2 u_i)\log^{\alpha-2}\sqrt{m}, \quad (26)$$

using the fact, that $\ell = 2^\alpha = \sum_{j=0}^{\infty}\binom{\alpha}{j}$. If m is large enough (by some counting it can be checked that $\log m > 10^{12}\ell^6$ is sufficient) the following inequality holds,

$$\log^2 u_i < \frac{\alpha}{144(\ln 2)\ell^2}\log\sqrt{m}. \quad (27)$$

Substituting (27) into (26), and (26) into (25) we get the following inequality to prove after simplifications,

$$-\frac{2}{\ell}(\sum_{i=1}^{\ell}\ln u_i) - \frac{\ell - \alpha - 1}{72\ell^2} \geq 1 - \frac{\sum_{1\leq j<k\leq\ell} u_j u_k}{\binom{\ell}{2}}.$$

Let $\mathbf{u} = (u_1, u_2, \ldots, u_\ell)$, then we can write the previous inequality in the following form,

$$A_2(\mathbf{u}) - 1 - \ln G_2(\mathbf{u}) \geq \frac{\ell - \alpha - 1}{72\ell^2},$$

which is true by Proposition 2.4 (more precisely, by (24)).

Case B.2. $\exists 1 \leq i \leq \ell : m_i > \sqrt{m} \log^5 m$.

Case B.2.1. $\exists i \neq j \quad m_i m_j \geq m \log^2 m$

In this case we have

$$\mathrm{RHS}(16) \leq \frac{\alpha}{\ln 2} \log^{\alpha-1}(1 - \frac{\log^2 m}{\binom{\ell}{2}}) \tag{28}$$

and

$$\mathrm{LHS}(16) \geq -(\ell - 1) \log^\alpha m, \tag{29}$$

so by (28) and (29), (16) holds if m is large enough. (Say, $\log m > (3\ell^3 \ln 2)/(4\alpha)$ is a good choice.)

Case B.2.2. $\forall i \neq j \quad m_i m_j < m \log^2 m$

We would like to minimize the LHS of (16), for such m_is satisfying the conditions

$$\sqrt{m} \log^5 m \leq m_1 \leq m/ \log^{1/2} m \quad \text{and} \quad m_1 m_i \leq m \log^2 m (2 \leq i \leq \ell). \tag{*}$$

The value of the LHS of (16) will not increase if we replace all m_is by $(m \log^2 m)/m_1$. So

$$\min_{(*)} \mathrm{LHS}(16) \geq \min\{\sqrt{m} \log^5 m \leq x \leq m/ \log^{1/2} m |$$
$$\log^\alpha m - \log^\alpha x - (\ell - 1) \log^\alpha \frac{m \log^2 m}{x}\}.$$

Let $f(x) = \log^\alpha m - \log^\alpha x - (\ell - 1) \log^\alpha(m \log^2 m/x)$. This function is monotonically decreasing for $x > (\ell - 1)^{1/(2(\alpha-1))} \sqrt{m} \log m$, consider the derivative of $f(x)$, which is smaller than the lower bound for m_1 if $m > 2$. So the minimum is achieved when $x = m/ \log^{1/2} m$.

Substituting into (16) we get

$$\min_{(*)} \mathrm{LHS}(16) \geq \frac{\alpha}{2} \log^{\alpha-1} m \log\log m - \frac{a_2(\alpha)}{4} \log^{\alpha-2} m \log^2 \log m -$$
$$(\ell - 1) \left(\frac{5}{2}\right)^\alpha \log^\alpha \log m \geq (\alpha/8) \log^{\alpha-1} m \log\log m. \tag{30}$$

In the first inequality we used the generalized binomial theorem. The second inequality holds if m is large enough. Say, $m > \max\{2^{((\ell+2\alpha+2)/\alpha)^2}, \ell 2^{10^4}\}$. On the other hand,

$$\mathrm{RHS}(16) \leq \frac{\alpha}{\ln 2} \log^{\alpha-1} m. \tag{31}$$

If $m > 2^{3100}$ then (16) holds by (30) and (31). □

Remark 2.1. (16) holds for $m > M$, where

$$M = \max\{\frac{\ell \ln 2}{2\alpha}, \ell^5 2^{2200\ell - 5}, 2^{10^{12}\ell^6}, 2^{(3\ell^3 \ln 2)/(4\alpha)}, 3, 2^{((\ell+2\alpha+2)/\alpha)^2}, \ell 2^{10^4}, 2^{3100}\},$$

so the constant in Theorem 2.2 is at most

$$\binom{M}{2} < 2^{2 \cdot 10^{12} \ell^6}.$$

□

3 Construction of a Badly Representable Key System

In [7], it was shown that there exist badly representable Sperner systems, namely of size

$$s(\mathcal{K}) > \frac{1}{n^2} \binom{n}{\lfloor \frac{n}{2} \rfloor}. \tag{32}$$

The proof of this theorem is not constructive. *L. Rónyai*'s observation is that the number of Sperner families that can be represented by a matrix of at most r rows is quite limited, and so r should be at least as big as in (32) to get a representation even for all antichains at the middle level of the Boolean lattice. In the following, we show an explicit badly representable Sperner system close to the middle level, no worse is known up to now.

Remember, that if $\mathcal{K}$ is a Sperner system, $\mathcal{K}^{-1}$ denotes the set of maximal elements, that are not contained in any element of $\mathcal{K}$. $\mathcal{K}_k^n$ denotes the complete k-uniform hypergraph and $K + L$ is the disjoint union of the hypergraphs K and L on the union of the vertices.

Theorem 3.1. *[11] Let $n = n_1 + n_2 + \ldots + n_t$, $n_i \leq N (1 \leq i \leq t)$. Let $\mathcal{K}_n = \mathcal{K}_k^{n_1} + \mathcal{K}_k^{n_2} + \ldots + \mathcal{K}_k^{n_t}$. Then*

$$|\mathcal{K}_n^{-1}| \leq T_\ell(s(\mathcal{K}_n)) \tag{33}$$

holds for $\ell = \binom{N}{k-1}$.

Proof (For details, see [11]). Suppose, that $\mathcal{K}_n$ is represented by a relation I of size $s(\mathcal{K}_n)$. We can recursively construct a labelled directed tree, $F \in \mathcal{F}_\ell^{(s(\mathcal{K}_n))}$ having the property, that there is an injection from $\mathcal{K}_n^{-1}$ to the leaves of F. A maximal antikey contains exactly $k-1$ elements from each clique, so there are $\binom{n_i}{k-1}$ possibilities for the intersection of a maximal antikey and the ith clique. The key observation is that if A_1 and A_2 are two maximal antikeys and $A_j^i = A_j \cap V(\mathcal{K}_k^{n_i})$, $(j = 1, 2)$, then there is no u and v satisfying both $\pi_{A_1^i}(u) = \pi_{A_1^i}(v)$ and $\pi_{A_2^i}(u) = \pi_{A_2^i}(v)$. (By Proposition 1.1 we know, that a representation of a Sperner family $\mathcal{K}$ should contain two rows, for each $A \in \mathcal{K}^{-1}$, that are equal in A, but should not contain two rows that are equal in an element of $\mathcal{K}$.)

The labels of the vertices of the tree are subsets of I. Let the label of the root be I, the whole relation. The out-neighborhood of the root can be devided into $\binom{n_1}{k-1}$ classes, each corresponding to a $(k-1)$-subset of $V(\mathcal{K}_k^{n_1})$. For a $(k-1)$-subset S consider $\pi_S(u)$ for all $u \in I$. By the equality of $\pi_S(u)$'s we get a partition of I. For each element of the partition of size at least two, add a new vertex to the tree, and label it by its elements. By the above remark labels of vertices from the same class are disjoint, while labels of vertices from distinct class can intersect in at most one.

We can continue building up the tree the similar way. The only difference is that instead of considering subsets of I we consider subsets of the label of the actual vertex. It is easy to see, that we get a tree of $\mathcal{F}_\ell^{(s(\mathcal{K}_n))}$, and by Proposition 1.1 the above mentioned correspondence is really an injection.

Corollary 3.1. *There exists a sequence of Sperner systems $\mathcal{K}_n$, such that*

$$s(\mathcal{K}_n) > 2^{n(1-(26/3)\log\log n/\log n)} \tag{34}$$

holds for n large enough.

Proof. We would like to apply Theorem 3.1, let $n_i = s-1$ or $n_i = s$, $1 \leq i \leq \lceil n/s \rceil$. So our task remains to choose s and k. Let $k = g(n)+1$ and $s = 2g(n)+1$, $g(n)$ to be chosen later.

$$\log|\mathcal{K}_n^{-1}| \geq \frac{n}{s}\log\binom{s-1}{k-1} > \frac{n}{2g(n)}\log\frac{2^{2g(n)}}{2\sqrt{2g(n)}} \geq n(1-\frac{1}{3}\frac{\log g(n)}{g(n)}),$$

if $g(n) \geq 2^9$. So by Theorem 2.2 and Theorem 3.1,

$$n(1-\frac{1}{3}\frac{\log g(n)}{g(n)}) < \log C(2^{2g(n)}) + \log s(\mathcal{K}_n) + 2g(n)\log\log s(\mathcal{K}_n). \tag{35}$$

where $C(x) = 2^{2\cdot 10^{12}x^6}$.

$$2g(n)\log\log s(\mathcal{K}_n) < \frac{\log g(n)}{6g(n)}\log s(\mathcal{K}_n) \tag{36}$$

holds if, say, $g(n) < n^{1/4}/7$ (using only, that $s(\mathcal{K}_n) > 2^{n/4}$, n is large enough).

On the other hand,

$$\log C(2^{2g(n)}) < \frac{\log g(n)}{6g(n)}\log s(\mathcal{K}_n), \tag{37}$$

if $g(n) \leq \log n/13 - 4$. By (35)-(37), we get

$$\log s(\mathcal{K}_n) > n(1-\frac{1}{3}\frac{\log g(n)}{g(n)})/(1+\frac{1}{3}\frac{\log g(n)}{g(n)}) > n(1-\frac{2}{3}\frac{\log g(n)}{g(n)}) \tag{38}$$

By (38), the statement holds if we choose $g(n) = \lfloor \log n/13 \rfloor - 4$. □

4 Well Representable Key Systems

4.1 Improving the Upper Bound on Complete k-Uniform Key Systems

In Theorem 1.3 $c_2(k)$ depends on k exponentially. We can replace this by a constant 2.

Theorem 4.1.

$$s(\mathcal{K}_k^n) < 2n^{(k-1)/2} + o(n^{(k-1)/2}).$$

Proof (For details, see [5]). Let p be a prime. The original proof defines polynomials of degree at most k over the finite field GF(p). It's coefficients are chosen from a *difference set* D. D is called a difference set, if $D - D =$GF(p), i.e., each element of GF(p) can be written as a difference of two elements from D. The jth coordinate of the ith vector of the representing relation instance I is $P_i(j)$ over GF(p), where P_i is the ith polynomial. The size of a difference set is about $2\sqrt{p}$. Our first observation is one can take 2 classes of polynomials. One of the classes contains polynomials of degree $k-1$ coefficients from D_1 (except for the coefficient of x^{k-1}, which is 1). The other one contains polynomials of degree $k-2$ and coefficients from D_2.

One can choose D_1 and D_2 so, that each of them have size around $\sqrt{p}$ and each element of GF(p) is of the form $d_1 - d_2$, where $d_i \in D_i (i = 1, 2)$. Let $D_1 = \{0, \lceil\sqrt{p}\rceil, \ldots, (\lceil\sqrt{p}\rceil - 1)\lceil\sqrt{p}\rceil\}$, $D_2 = \{0, 1, \ldots, \lceil\sqrt{p}\rceil\}$.

One can easily check, that the constructed instance represents $\mathcal{K}_k^p$. Two tuples can not have k equal coordinates. That would mean that the difference of the corresponding polynomials have k roots, but it is a polynomial of degree at most $k-1$. On the other hand, the polynomial

$$w(x) = (x - t_1)(x - t_2)\cdots(x - t_{k-1}) = x^{k-1} + a_{k-2}x^{k-2} + \ldots + a_1 x + a_0$$

can be written as a difference of two polynomials that correspond to tuples for arbitrary different $t_1, \ldots, t_{k-1}$. Each a_i can be written as a differnce of $d_{1,i} \in D_1$ and $d_{2,i} \in D_2$. Then $w(x)$ is the difference of the polynomials

$$z_j(x) = (2 - j)x^{k-1} + d_{j,k-2}x^{k-2} + \ldots + d_{j,1}x + d_{j,0} \quad (j = 1, 2).$$

If $n < p$, then $\pi_{\{1,2,\ldots,n\}}(I)$ is an appropriate representation of $\mathcal{K}_k^n$.

Instead of Chebyshev's theorem (Bertrand's postulate) on the density of primes (between n and $2n$ there is a prime) the statement follows from a theorem of Luo and Yao [9], stating that for the nth prime p_n

$$p_{n+1} - p_n \ll p_n^{6/11+\varepsilon} \tag{39}$$

holds for any $\varepsilon > 0$. □

This improvement shows that complete k-uniform key systems are well representable, even if $k = f(n)$ slightly tends to infinity.

4.2 New Well Representable Key Systems

We can use the idea of Theorem 1.3 to prove well representability for key systems, that differ in not too many elements from $\mathcal{K}_k^n$.

First, let us consider the case when the key system $\mathcal{K}$ contains one element $\{(t_1, \ldots, t_{k-1})\}$ of size $k-1$ and all k element subsets but its supersets. One can construct sets $D_1^{(c)}$ and $D_2^{(c)}$, such that $D_1^{(c)} - D_2^{(c)} = \mathrm{GF}(p)\setminus\{c\}$, for each $c \in GF(p)$. Let $P(x) = (x-t_1)(x-t_2)\cdots(x-t_{k-1}) = x^{k-1}+c_{k-2}x^{k-2}+\ldots+c_0$. Then the polynomials will be:

$$\begin{gathered}
x^{k-1} + d_1^{(k-2)}x^{k-2} + d_{1,k-3}x^{k-3} + \ldots + d_{1,0},\\
x^{k-1} + d_{1,k-2}x^{k-2} + d_1^{(k-3)}x^{k-3} + \ldots + d_{1,0}, \ldots,\\
x^{k-1} + d_{1,k-2}x^{k-2} + d_{1,k-3}x^{k-3} + \ldots + d_1^{(0)},\\
d_2^{(k-2)}x^{k-2} + d_{2,k-3}x^{k-3} + \ldots + d_{2,0},\\
d_{2,k-2}x^{k-2} + d_2^{(k-3)}x^{k-3} + \ldots + d_{2,0}, \ldots,\\
d_{2,k-2}x^{k-2} + d_{2,k-3}x^{k-3} + \ldots + d_2^{(0)},
\end{gathered}$$

where $d_i^{(t)} \in D_i^{(c_t)}, (i = 1, 2;\ 0 \le t \le k-2), d_{i,j} \in D_i (i = 1, 2)$ (see Theorem 4.1). So we get each $(k-1)$-degree polynomial, except for $P(x)$. The size of this relation is about $2(k-1)p^{(k-1)/2}$.

Based on the idea above one can construct a good representation for a key system having $o(n^{1/4})$ elements of cardinality $k-1$ and containing all k element sets but their supersets. Let $\nabla(A) = \{B | B \supseteq A, |B| = |A|+1\}$ and $\nabla(\mathcal{A}) = \bigcup_{A\in\mathcal{A}} \nabla(A)$.

Theorem 4.2. *Let $\mathcal{K}_n = \mathcal{A}_n \cup \mathcal{B}_n$ be a Sperner system, such that $\mathcal{A}_n \subseteq \mathcal{K}_{k-1}^n$, $\nabla(\mathcal{A}_n) \cup \mathcal{B}_n = \mathcal{K}_k^n$, and $|\mathcal{A}_n| = o(n^{1/4})$. Then*

$$s(\mathcal{K}_n) \le 2(k-1)n^{(k-1)/2} + o(n^{(k-1)/2}). \tag{40}$$

Proof. It follows from the conditions, that $\mathcal{K} \subseteq \mathcal{K}_{k-1}^n \cup \mathcal{K}_k^n$ and $\mathcal{K}^{-1} \subseteq \mathcal{K}_{k-2}^n \cup \mathcal{K}_{k-1}^n$. Let $\mathcal{A}_n = \{(t_1^{(1)}, \ldots, t_{k-1}^{(1)}), \ldots, (t_1^{(m)}, \ldots, t_{k-1}^{(m)})\}$ and $p > n$. For $1 \le r \le m$ let us consider the polynomial

$$w_r(x) = (x-t_1^{(r)})(x-t_2^{(r)})\cdots(x-t_{k-1}^{(r)}) = x^{k-1} + c_{k-2}^{(r)}x^{k-2} + \ldots + c_1^{(r)}x + c_0^{(r)}.$$

For each $0 \le h \le k-2$ let $J_1^h = [a_1^h, b_1^h], J_2^h = [a_2^h, b_2^h], \ldots, J_{m_h}^h = [a_{m_h}^h, b_{m_h}^h]$ $(m_h \le m)$ be the (possibly 1-length) intervals of (the cyclically ordered) GF(p), such that $\{a_1^h, \ldots, a_{m_h}^h\} = \{c_h^{(1)}, \ldots, c_h^{(m)}\}$ and $\{b_1^h, \ldots, b_{m_h}^h\} = \{c_h^{(1)} - 1, \ldots, c_h^{(m)} - 1\}$. So $a_1^h \le b_1^h = a_2^h - 1, \ldots, a_{m_h}^h \le b_{m_h}^h = a_1^h - 1$. Let $l_i^h = b_i^h - a_i^h + 1$ be the length of J_i^h.

Let $R_t^h = \{1 \le i \le m_h \mid p^{1/2^t} < l_i^h \le p^{1/2^{t-1}}\}$ and $r_t = \lfloor p^{1/2^t} \rfloor$ $(t = 1, 2, \ldots, \lceil \log p \rceil)$. Let $D_t^h = \bigcup_{i\in R_t^h} \{a_i^h + r_t, a_i^h + 2r_t, \ldots, \lfloor (l_i^h - 1)/r_t \rfloor r_t, b_i^h\}$ and $E_t = \{0, 1, 2, \ldots, r_t - 1\}$. Furthermore let $D = \{0, r_1, 2r_1, \ldots, r_1^2,\}$ and $E = E_1$.

We construct a representation I_n of $\mathcal{K}_n$. We give p-tuples of three types. The coordinates of the tuples triples of form $(P_i(j), h, t)$. For each h and t we take all polynomials P_i over GF(p) of the form

$$P_i(x) = x^{k-1} + d_{k-2}x^{k-2} + \ldots + d_1 x + d_0,$$

where $d_h \in D_t^h$ and $d_i \in D$ $(0 \le i \le k-2, i \ne h)$.

Tuples of the second type have coordinates of a triple $(Q_i(j), h, t)$, too. For each h and t we take all polynomials Q_i over GF(p) of the form

$$Q_i(x) = e_{k-2}x^{k-2} + \ldots + e_1 x + e_0,$$

where $e_h \in E_t^h$ and $e_i \in E$ $(0 \le i \le k-2, i \ne h)$.

Tuples of the third type have coordinates of the form $(P(j), i)$, two polynomials for each i, where i corresponds to a $(k-1)$-tuple of coefficients, $(c_{k-2}, c_{k-3}, \ldots c_0)$. All coordinates are a coefficient of some $w_r(x), 1 \le r \le m$, while the polynomial $z(x) = x^{k-1} + c_{k-2}x^{k-2} + \ldots + c_0$ is not among these polynomials. Choose the two polynomials so, that their difference is $z(x)$. (E.g, $z(x)$ and the 0 polynomial.)

We get I_n by deleting the last $p-n$ coordinates of the tuples of the constructed relation. There are no two tuples of I_n having the same value in k coordinates (polynomials of degree $k-1$ can not have k common roots). It is also easy to check that there are no two tuples having the same coordinates in A, $|A| = k-1$ if and only if $A \notin \mathcal{A}_n$.

There are $p^{(k-1)/2} + o(p^{(k-1)/2})$ tuples of type $(P_i(j), h, 1)$ and $(Q_i(j), h, 1)$. The number of $(P_i(j), h, 2)$ and $(Q_i(j), h, 2)$ type tuples are at most $mp^{1/4}p^{(k-2)/2} = o(p^{(k-1)/2})$. Similarly, there is no more than $mp^{1/8}p^{(k-2)/2} = o(p^{(4k-5)/8})$ tuples of type $(P_i(j), h, t)$ and $(Q_i(j), h, t)$ $(t \ge 3)$. Finally, the number of the tuples of type $(P(j), i)$ is at most $m^{k-1} = o(p^{(k-1)/4})$.

The statement follows from (39).

5 Further Research

It remains an open problem to determine the exact value of $T_\ell(m)$. Even in the case of $\ell = 2$ for general m.

The above improvements on the labelled directed tree lemma (Theorem 2.1 and Theorem 2.2) opens a new dimension of the minimum representation problem. Is the log factor needed? If yes/no, what is the exact constant?

Example 5.1. Let $\mathcal{C}_n = \{\{1,2\}, \{2,3\}, \ldots, \{n-1, n\}, \{n, 1\}\}$ be the cycle. We know from [11]

$$|\mathcal{C}_n^{-1}| \le T_2(s(\mathcal{C}_n))$$

So by (1) and we have for $n \ge 5$

$$\frac{2|\mathcal{C}_n^{-1}|}{\log|\mathcal{C}_n^{-1}|} \le s(\mathcal{C}_n) \le |\mathcal{C}_n^{-1}| + 1. \tag{41}$$

Note, that in this case the upper bound cannot be the truth. One can construct an instance in which the rowpairs (see Lemma 1.1) having a non-trivial density, proving say $0.99|\mathcal{C}_n^{-1}|$ for n large enough. □

Improvements similar to (41) can be obtained for the problems considered in [11], maximizing $s(\mathcal{K})$ over $\mathcal{K}$'s from a special class, such as e.g, for all graph.

Theorem 5.1. *Let $\mathfrak{G}_n$ denote the set of all graph on n vertices.*

$$\frac{1}{2^{10^{15}}} \cdot \frac{3^{n/3}}{n^{\log 3}} \leq \max_{\mathcal{G}\in\mathfrak{G}_n} s(\mathcal{G}) \leq 3^{n/3} + 1. \tag{42}$$

It still remains open to show a key system, that is as badly representable as the Demetrovics-Gyepesi probabilistic estimate.

References

1. S. Abiteboul, R. Hull, V. Vianu, Foundations of Databases, Addison-Wesley, Reading, 1995.
2. W.W. Armstrong, *Dependency structures of database relationship,* Information processing 74 (North Holland, Amsterdam 1974) 580–583.
3. I. Csiszár, J. Körner, Information theory. Coding theorems for discrete memoryless systems. Probability and Mathematical Statistics. Academic Press, Inc. (Harcourt Brace Jovanovich, Publishers), New York-London, 1981.
4. J. Demetrovics, *On the equivalence of candidate keys with Sperner systems,* Acta Cybernetica **4** (1979) 247–252.
5. J. Demetrovics, Z. Füredi, G. O. H. Katona, *Minimum matrix representation of closure operations,* Discrete Appl. Math. **11** (1985), no. 2, 115–128.
6. J. Demetrovics, Gy. Gyepesi, *On the functional dependency and some generalizations of it,* Acta Cybernetica **5** (1980/81), no. 3, 295–305.
7. J. Demetrovics, Gy. Gyepesi *A note on minimal matrix representation of closure operations,* Combinatorica **3** (1983), no. 2, 177–179.
8. Z. Füredi, *Perfect error-correcting databases,* Discrete Appl. Math. 28 (1990) 171-176.
9. Lou Shi Tuo, Yao Qi, *A Chebychev's type of prime number theorem in a short interval. II.,* Hardy-Ramanujan J. 15 (1992), 1–33 (1993).
10. Van Lindt, Wilson, A Course in Combinatorics, Cambridge Univ. Press, 1992, Chapter 22.
11. K. Tichler, *Extremal theorems for databases,* Annals of Mathematics and Artificial Intelligence **40** (2004) 165–182.