EuroComb 2005 DMTCS proc. **AE**, 2005, 293–298

An extremal problem on labelled directed trees and applications to database theory

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

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

We consider an extremal problem on labelled directed trees and applications to database theory. Among others, we will show explicit keysystems on an underlying set of size n, that cannot be represented by a database of less than $2^{n(1-c\cdot\log\log n/\log n)}$ rows.

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

1 An extremal problem on labelled directed trees

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 $\mathcal{F}^{(m)}_{k,\ell}$, for which the vertices of each tree $F \in \mathcal{F}^{(m)}_{k,\ell}$ 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.

$$A(v) \subseteq A(n(v)) \ \ (v\neq v_0), \tag{1}$$

$$|A(v)| \geq k + 1, \tag{2}$$

$$w_1, w_2 \in N_i(v) \Rightarrow A(w_1) \cap A(w_2) = \emptyset \ \ (1 \leq i \leq \ell), \tag{3}$$

$$w_1 \in N_i(v), w_2 \in N_j(v) \Rightarrow |A(w_1) \cap A(w_2)| \leq k \ \ (1 \leq i < j \leq \ell). \tag{4}$$

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

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

†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.

Theorem 1.1

$$T_2(m) \leq \frac{1}{2} m \log m. \tag{5}$$

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

Theorem 1.2

$$T_\ell(m) = \Theta_\ell(m \log^\alpha m) \tag{6}$$

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

In [7], the magnitude of $T_{k,\ell}$ was determined.

Proof of Theorem 1.1: We will use the concept of *entropy* [2] in the proof. Entropy is a measure of a random variable X:

$$H(X) = -\sum_i p_i \log p_i, \tag{7}$$

where Prob$(X = i) = p_i$. It is known, that

$$H((X,Y)) \leq H(X) + H(Y). \tag{8}$$

The proof is by induction on m. (5) 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 $N(v_0) = \{v_1, \dots, v_s, w_1, \dots, w_t\}$, $N_1(v_0) = \{v_1, \dots, v_s\}$, $N_2(v_0) = \{w_1, \dots, w_t\}$. 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

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

holds. So it is enough to prove that

$$\sum_{i=1}^{s} a_i \log a_i + \sum_{i=1}^{t} b_i \log b_i \leq m \log m, \tag{9}$$

since then

$$T(m) = |\ell(F)| = \sum_{i=1}^{s+t} |\ell(F_i)| \leq \sum_{i=1}^{s} T(a_i) + \sum_{i=1}^{t} T(b_i)$$
$$\leq \sum_{i=1}^{s} \frac{1}{2} a_i \log a_i + \sum_{i=1}^{t} \frac{1}{2} b_i \log b_i \leq \frac{1}{2} m \log m.$$

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 $(1 \leq i \leq s')$ and B_j $(1 \leq j \leq t')$ have the following properties:

$$\{A_i, 1 \leq i \leq s'\} \ \text{ is a partition of } \{1, 2, \ldots, m\}, \tag{10}$$

$$\{B_j, 1 \leq j \leq t'\} \ \text{ is a partition of } \{1, 2, \ldots, m\}, \tag{11}$$

$$|A_i \cap B_j| \leq 1, 1 \leq i \leq s', 1 \leq j \leq t'. \tag{12}$$

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, $Y(X \in A_i) = i,\ 1 \leq i \leq s'$ and $Z(X \in B_j) = j,\ 1 \leq j \leq t'$. Then

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

The random variables Y and Z are well defined by (10) and (11). Furthermore, by (12) we get

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

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

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

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

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

Therefore, by (8) 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 (9). □

To prove Theorem 1.2 we need somewhat more counting. We could not find a straightforward way to generalize the concept of entropy for this case, altough the idea of the proof is based on the previous proof. Note, that the constant in Theorem 1.2 for the upper bound can be roughly upperestimated by

$$2^{2 \cdot 10^{12} \ell^6}. \tag{13}$$

2 An application to database theory

A subset K of the set $\Omega(|\Omega| = n)$ of columns of a matrix, is called a *key*, if there are no two rows of the matrix agree in each of the columns of K. If K is a key, then clearly all of its supersets are keys as well. On the other hand, for every nonempty $\mathcal{K} \subseteq 2^{\Omega}$ having this property there always exists a matrix, in which the system of keys is exactly $\mathcal{K}$[1, 3]. Since the system of keys determines the system of minimal keys and vica versa, it is enough to consider Sperner families. Let $\mathcal{K}$ be a Sperner system, then let $s(\mathcal{K})$ denote the minimum number of rows of such a matrix.

In [6], it was shown that there exist badly representable Sperner systems, namely of size

$$s(\mathcal{K}) > \frac{1}{n^2}\binom{n}{\lfloor \frac{n}{2} \rfloor}. \tag{14}$$

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 (14) 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 (quite far from the middle level), no worse is known up to now.

If $\mathcal{K}$ is a Sperner system, let $\mathcal{K}^{-1}$ denote the set of maximal elements, that are not contained in any superset of $\mathcal{K}$. Let $K_n^{(k)}$ denote the complete k-uniform hypergraph.

Theorem 2.1 *[7] Let $n = n_1 + n_2 + \ldots + n_t$, $n_i \leq s(1 \leq i \leq t)$. Let $\mathcal{K}_n = K_{n_1}^{(k)} + K_{n_2}^{(k)} + \ldots + K_{n_t}^{(k)}$. Then*

$$|\mathcal{K}_n^{-1}| \leq T_\ell(s(\mathcal{K}_n)) \tag{15}$$

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

Proof Sketch: [7] Suppose, that $\mathcal{K}_n$ is represented by a matrix of $s(\mathcal{K}_n)$ rows. We can construct recursively 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. Easy to see [4], 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 construction is based on these properties and labelling is due to the equalities of the entries of the matrix in a certain (set of) column(s). □

Corollary 2.2 *There exists a sequence of Sperner systems $\mathcal{K}_n$, such that*

$$s(\mathcal{K}_n) > 2^{n(1-(26/3)\log\log n/\log n)} \tag{16}$$

holds for n large enough.

Proof Sketch: The elements of $\mathcal{K}_n^{-1}$ are exactly those, that contain precisely $k-1$ elements from each clique, $K_{n_i}^{(k)}$. We apply Theorem 2.1 and Theorem 1.2 with the estimation (13), let $n_i = s-1$ or $n_i = s$, $1 \leq i \leq \lceil n/s \rceil$. We choose $k = g(n)+1$ and $s = 2g(n)+1$, so

$$\binom{2g(n)}{g(n)}^{n/(2g(n)+1)} \leq \prod_{i=1}^{\lceil n/s \rceil} \binom{n_i}{k-1} = |\mathcal{K}_n^{-1}|. \tag{17}$$

Some calculation support to choose $g(n) = \lceil \log n/13 \rceil - 4$. Substituting into (17) we get (16). □

The above theorem on labelled directed trees has been proved surprising widely applicable for proving, that certain Sperner families are badly representable. Recently, another interesting application has been found by the authors in the theory of combinatorial search.

References

[1] W.W. Armstrong, *Dependency structures of database relationship*, Information processing 74 (North Holland, Amsterdam 1974) 580–583.

[2] 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.

[3] J. Demetrovics, *On the equivalence of candidate keys with Sperner systems*, Acta Cybernetica **4** (1979) 247–252.

[4] J. Demetrovics, Z. Füredi, G. O. H. Katona, *Minimum matrix representation of closure operations*, Discrete Appl. Math. **11** (1985), no. 2, 115–128.

[5] J. Demetrovics, Gy. Gyepesi *A note on minimal matrix representation of closure operations*, Combinatorica **3** (1983), no. 2, 177–179.

[6] J. Demetrovics, Gy. Gyepesi, *On the functional dependency and some generalizations of it*, Acta Cybernetica **5** (1980/81), no. 3, 295–305.

[7] K. Tichler, *Extremal theorems for databases*, Annals of Mathematics and Artificial Intelligence **40** (2004) 165–182.